ESSnet Big Data

Specific Grant Agreement No 1 (SGA-2)

https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata

Framework Partnership Agreement Number 11104.2015.006-2015.720
Specific Grant Agreement Number 11104.2015.007-2016.085

Work Package 2
Web scraping Enterprise Characteristics

Deliverable 2.4
Final report
Version 2018-05-30

Prepared by:

Galia Stateva (BNSI, Bulgaria)
Olav ten Bosch, Dick Windmeijer (CBS, Netherlands)
Jacek Maślankowski (GUS, Poland)
Giulio Barcaroli, Monica Scannapieco, Donato Summa (ISTAT, Italy)
Matthew Greenaway (ONS, United Kingdom)
Ingegerd Jansson, Dan Wu (SCB, Sweden)

ESSnet co-ordinator:

Peter Struijs (CBS, Netherlands)
p.struijs@cbs.nl
telephone : +31 45 570 7441
mobile phone : +31 6 5248 7775
# Table of contents

Table of contents

1 Executive Summary ............................................................................................................. 2

2 General Motivations for Web Scraping of Enterprises’ Web Sites ...................................... 5

3 Description of a Reference Framework for Web Scraping of Enterprises’ Websites ............... 9

4 Methodological Issues and Solutions ..................................................................................... 12

  4.1 Specific vs generic scraping: when/what ........................................................................ 12

  4.2 Analysis techniques: machine learning vs deterministic approaches ............................. 13

  4.3 Review of methods used in pilots .................................................................................. 15

5 Technological Issues and Solutions ....................................................................................... 21

  5.1 Review of the technological environments used in pilots ............................................. 21

  5.2 Issues and Solutions ..................................................................................................... 25

6 Outputs Provided by Use Cases ............................................................................................. 27

7 Conclusions, Recommendations and Future Challenges ..................................................... 29

  7.1 Conclusions and recommendations on quality ............................................................ 29

  7.2 Conclusions and recommendations on IT techniques .................................................. 31

  7.3 Future challenges .......................................................................................................... 32

8 Appendix: Methodological Notes by Use Case
1 Executive Summary
Monica Scannapieco (Istat)

This document describes the technical work performed within the workpackage “Web Scraping Enterprise Characteristics” (WP2) of the ESSnet Big Data in both SGA1 and SGA2 phases.

The general objective of this workpackage was to infer some enterprises characteristics by accessing their websites through Web scraping techniques. In order to reach this objective, the work performed within WP2 had to face methodological, technological and legal challenges. In the deliverable 2.1, legal issues and possible solutions are discussed in detail. In the deliverable 2.2, methodological and technological aspects of the first implementation phase (SGA1) are described. This document starts from what described in deliverable 2.2, and extends it in the following directions:

- SGA1 work had in scope 4 use cases (URLs retrieval, e-commerce/web sales, social media detection, job advertisement detection). SGA2 work considered two additional use cases, namely: NACE detection and Sustainable Development Goals (SDGs) detection.
- Pilots enhancements: all the pilots implemented in the first phase have been enhanced by (i) consolidating adopted techniques and (ii) extending the number of enterprises to which the scraping activity was targeted.
- New pilot development: new pilots were implemented not only for new use cases but also for SGA1 use cases for which additional countries committed to develop pilots.
- Final methodological and technological conclusions based on all the work done within the ESSnet.
- Production of output indicators and comparison with related survey statistics.

Some figures on the performed work are:

- Six different use cases were identified, and for all of them pilot implementations where realized.
- A total number of 24 different pilots were implemented, namely: 6 for use case 1 (with Bulgaria implementing two pilots with two different technologies), 6 for use case 2 (with Bulgaria implementing two pilots with two different technologies), 3 for use case 3, 5 for use case 4, 3 for use case 5 and 1 for use case 6.
- The pilots resulted in a set of output indicators that will be published as experimental statistics on the project wiki. Each published figure will be accompanied by a methodological note describing in details the choices done for computing.
- Generalized software development: Italy, Bulgaria and Poland developed software tools and made them available for reuse. Italian and Polish software was already reused by other countries within the work of WP2.

1 Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/a/a0/WP2_Deliverable_2_1_15_02_2017.pdf
2 Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/66/WP2_Deliverable_2.2_2017_07_31.pdf
3 Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP2_Links
The main findings can be summarized as follows:

- The complex pipeline for processing data scraped from enterprises’ websites has been defined in detail and shared among the participants. This pipeline can be considered as a reference one to which mapping specific technological and methodological choices. A set of logical building blocks have been identified for each phase of the pipeline.

- From a methodological perspective, both deterministic and machine learning methods were used in the pilots. On one side, we learned that even with different methods good results can be achieved. On the other side, however, we saw that in some cases there can be a convergence of methods (e.g. the URL retrieval pilot where Italy, Bulgaria and the Netherlands applied the same methodology). Predicted values can be used for a twofold purpose: (i) at unit level, to enrich the information contained in the register of the population of interest; (ii) at population level, to produce estimates. Both unit level and population level quality have been evaluated in the piloting phase. With respect to quality evaluation at unit level, when employing supervised machine learning methods the quality can be measured by comparing against labelled data in a hold-out ‘test’ set. If this test set is representative of the whole population and not used for training the model, the performance measures (like accuracy and F1-score) calculated for the test set can be considered a good estimate for the overall performance. The unit level quality evaluation has been performed by all pilots. The issue of measuring the quality of population estimates making use of predicted values has also been addressed, for instance by Italian pilots, but more work is needed for a drawing conclusions on this issue.

- From an IT perspective, performance is a key issue especially when downloading and processing the whole websites. Processing unstructured information is very CPU and memory consuming, especially with machine learning algorithms, and as a result not very efficient. A sustainability issue is also very relevant, due to the fact that big data tools are changing very frequently as well as the website technology, there is a need to provide an agile-like development of tools. For the storage, the possible choices are between filesystems (CSV, JSON etc.), NoSQL database (Solr, Cassandra, Hbase etc.) or relational database (MySQL, PostgreSQL, SQL Server etc.). The decision of the use of particular data storage should be taken according to the volume of the data and the type of data to be stored. Although some software are developed in particular countries, it is possible to apply them in other countries as well without any major changes. For instance, URLSearcher developed by Istat was tested on Bulgarian and Polish websites as well.

- Some output indicators will be published as experimental statistics. These include:
  - Rate of retrieved URLs from an enterprise list
  - Rate of enterprises engaged in web sales on their website
  - Rate of enterprises with job advertisements on their website
  - Rate of enterprises that are present on social media
  - Percentage of enterprises using Twitter for a specific purpose, estimated from web data

The overall quality of the results at unit level was good, in some cases (e.g. Rate of enterprises that are present on social media) even excellent. The quality of the results at population level has been only partially evaluated but comparisons with survey estimates are promising.
For those NSIs that would like to implement similar use cases our main recommendations can be summarized as follows:

1. It is proved that a whole pipeline from data collection to analysis can be put in place and produce quite good results. The pipeline implemented so far is an experimental one but it is complete, ranging from the data collection phase to the analysis one. Some relevant issues, from the methodological and technological perspective, have already available solutions.

2. Moving to production the solutions that have been proposed in this work requires still to face two main categories of issues, namely:
   a. Methodology: quality evaluation in the new framework of estimates from Web data still deserves some investigation.
   b. Technology: web scraping and text processing tools need to be evolved for an enterprise-level usage in terms of volume, access, processing speed and technical organisation. Massive scraping might require improved bandwidth and parallel scraping.
   c. Organizational: re-organization of production processes and capability building actions should be designed and implemented.

The document is organized as follows:

- Section 2 describes use cases and candidatures by involved countries to implement them.
- Section 3 deals with generic building blocks identified as reference ones to guide the implementation of the different pilots.
- Section 4 details methodological challenges and solutions, including:
  - Generic and specific web scraping approaches and their possible usages.
  - Deterministic vs. machine learning approaches for the analysis phase.
- Section 5 details IT challenges and solutions, including: performance, sustainability, storage, de-duplication, national languages and diacritic characters dependencies, software licensing issues.
- Section 6 reports the value of output indicators compared to survey estimates.
- Section 7 details recommendations for the work done within the workpackage.
- Finally, Section 8 reports the methodological notes associated to the output figures of the pilots.
2 General Motivations for Web Scraping of Enterprises Web Sites

Galia Stateva (BNSI, Bulgaria)

The purpose of workpackage 2 - “Web Scraping of Enterprise Web Sites” - is to investigate whether web scraping, text mining and inference techniques could be used to collect, process and improve general information about enterprises. In particular, the aim is twofold:

1. to demonstrate whether business registers can be improved by using web scraping techniques and by applying model-based approaches in order to predict for each enterprise the values of some key variables;
2. to verify the possibility to produce statistical outputs with more predictive power combined or not with other sources of data (survey or administrative data). The benchmark of big data output could be the data produced by the “ICT use by enterprises” survey, which is carried out in all EU Member States.

Companies tend to be very active in the internet. Their internet activities range from selling products, looking for employees, sharing info about their new products and investments in the future to tweeting. By collecting this information and presenting it in a structured way we can gain an insight what is going on in the economy. This information is much timelier than national accounts and standard business surveys. National accounts in particular may lag economic processes and have no predictive power.

The whole set of identified use cases were:

1. Enterprise URLs Inventory. This use case is about the generation of a URL inventory of enterprises for the Business register.
2. E-Commerce on enterprises’ websites. This use case is about predicting whether an enterprise provides or not web sales facilities on its website.
3. Job advertisements on enterprises’ websites. This use case is about investigating how enterprises use their websites to handle the job ads.
4. Social Media Presence on enterprises’ webpages, aimed at providing information on existence of enterprises in social media.
5. Sustainability reporting on enterprises’ websites. One of the Sustainability Development Goals set up by the UN is to encourage enterprises to produce regular sustainability reports highlighting sustainability actions which the business has taken. In order to measure companies’ response to this, this use case will look at what companies publish on their official website and track changes over time.
6. Identifying categories relevant to enterprises’ types of activity (NACE). Aimed at identifying relevant categories of Enterprises’ activity sector from enterprises’ web sites to check or complete Business registers.

Use case 3 is particularly useful for WP1 in understanding whether the enterprises’ websites contain useful information on enterprise job vacancies. However, in general, WP2 faces some challenges not faced in WP1, including application of much larger-scale scraping of websites and collecting and analysing more unstructured data.

Work package 2 has six participating countries, namely IT (leader), BG, NL, PL, SE and UK.

The main activities carried out within the work package are:
- Task 1 – Data access
- Task 2 – Data handling
- Task 3 – Application of the text and data mining techniques
- Task 4 – Testing information extraction techniques

Within Task 1, the main activities were: (i) identification of a set of methods to retrieve URLs for enterprises for which they were not available; (ii) study of the legal aspects to access data on enterprises web sites; (iii) for all the identified use cases, retrieval of the URLs of the reference population by applying the URLs retrieval procedure set-up in SGA-1 and (iv) evaluation of the quality of the results, also with respect to a tradeoff between automated and manual tasks.

URL identification is the very first step toward getting access to company information and presenting it to decision makers in a structured way.

The study of the legal aspects was a particulary important initial step and it was carried out in SGA-1. Each participating country had to involve internal legal offices and there were much iteration before getting the desired answers. The result of this activity is the Deliverable 2.1 “Legal aspects related to Web scraping of Enterprise Web Sites”.

Within Task 2, the main activities were: (i) detailed definition of use cases; (ii) carrying out scraping activities; (iii) for all the identified use cases, the scraping task was carried out on the URLs of the respective reference populations (either avialble or retrieved according to Task 1) and (iv) construction of the database of the scraped data and reporting of its characteristics and metadata (reference population metadata, statistics on dimensionality, etc.).

In order to carry out the scraping activities, some software tools needed to be developed and shared. All the countries worked on performing scraping activities according to the defined use cases. Some countries, namely Sweden and the UK, were at the beginning of SGA1 not able to perform large-scale scraping of many websites due to internal organizational constraints. For both Sweden and the UK, the work done within this ESSnet helped demonstrate the benefits of web-scraping, such that by the end of SGA2 both countries have authorization to do large-scale scraping.

Within Task 3 the main activities were: (i) build of a proof of concept for each of the selected use cases to predict characteristics of the enterprises by applying text and data mining techniques; (ii) for all identified use cases, apply text and data mining techniques (learners) to predict characteristics of the enterprises. Evaluate quality indicators for all of them (e.g. accuracy, sensitivity, specificity). On the basis of the quality indicators, choose the best predictor; (iii) application of the best predictor to the whole set of scraped data in order to predict characteristics of enterprises. Compare and possibly integrate the Business Registers with the obtained information and (iv) on the basis of predicted values, for the different use cases production of estimates (means and totals) of population parameters (for instance, percentage of enterprises offering e-commerce, present on social media, etc.), and evaluation of related Mean Square Error. Steps iii and iv were not done for all the use case, as explained in the remainder of the document.

Within Task 4 the main activities were: (i) identify some use cases (or part of them) for which it can be supposed an improvement by means of information extraction techniques; (ii) set up of an information extraction procedure to be applied to the selected cases and (iii) application of the
procedure and evaluation of the results. In this step a detailed analysis was carried out on the quality improvement that can be obtained by complementing text and data mining procedures with information extraction techniques.

The selected use cases for SGA 1 are:

1. Enterprise URLs Inventory.
2. E-Commerce in Enterprises.
3. Job vacancies ads on enterprises’ websites.

Use-case 5 – on NACE identification – and 6 – on SDGs - were implemented in SGA2.

The distribution of countries participating to each selected use case in SGA 1 are shown in (Figure 1 bold X identifies the country responsible for the use case).

<table>
<thead>
<tr>
<th>Use Cases Countries</th>
<th>IT</th>
<th>SE</th>
<th>UK</th>
<th>NL</th>
<th>BG</th>
<th>PL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: URLs retrieval</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>2: Ecommerce</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3: Job advertisement</td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4: Social Media</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 1: Use cases’ assignment in SGA 1**

In the pilots implementation in SGA1, UK and SE were not able to large-scale massive web scraping due to legal reasons as detailed on the WP 2 deliverable 2.1 “Legal aspects related to Web scraping of enterprise web sites”.

The distribution of countries participating to each selected use case in SGA 2 are shown in Figure 2, where the participation involvement in SGA1 use cases has increased and UK was able to carry out large-scale web-scraping.

<table>
<thead>
<tr>
<th>Use Case</th>
<th>Country</th>
<th>IT</th>
<th>SE</th>
<th>UK</th>
<th>NL</th>
<th>BG</th>
<th>PL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: URLs Retrieval</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>2: Ecommerce</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>3: Job advertisement</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>4: Social media</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>5: NACE</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6: SDG</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 2 Use cases’ assignment in SGA2**
Use case analysis results

As remarked a detailed use case definition was carried out by first sharing a use case template. The use case template involves the following fields:

- ID
- Name
- Description
- Actors
- Preconditions
- Postconditions
- Frequency of use
- Scenarios
- Special Requirements
- Issues

Participant countries filled the template concerning the use cases they are involved in.

The detailed definition of use cases was carried out according to a shared template by all the participating countries which have been published on the project wiki platform (https://webgate.ec.europa.eu/ffpis/fwikis/essnetbigdata/index.php/WP2_Working_area).
3 Description of a Reference Framework for Web Scraping of Enterprises’ Websites
Monica Scannapieco (Istat), Donato Summa (Istat)

From a conceptual point of view we designed a generic reference logical architecture made of several building blocks organized into four main layers (see Figure 3), namely:

- « Internet access »,
- « Storage »,
- « Data preparation » and
- « Analysis ».

For each layer we reported the logical functionalities to be implemented by specific software products; so, for instance, the « Internet access layer » has two logical functionalities, i.e. « URL searcher » and « «Scraper » . In the following we will provide some details on each block.

Figure 3: Logical Reference Architecture

The objective of the URL searcher block is to retrieve a list of websites related to a given enterprise. Usually this list is obtained by querying a search engine on the web using the name of the enterprise as a search term. The underlying assumption is that, if an enterprise has an official website, this should be found within the results provided by the search engine.
Two specific software were implemented within the ESSnet for this block by Italy and Bulgaria. These are available at the link:

One other software implementation was done by the Netherlands which uses the Google search engine for this goal. It can be found here:
https://github.com/SNStatComp/S4SGoogleSearch

The Retrieved URLs block is basically a container of URLs obtained in the previous step, it can be implemented in different ways, ranging from a file system to a relational DBMS.

The Scraper block is responsible for acquiring the content available on each URL in the list of URLs provided as input. It can have additional features such as URL filtering (if a list of URL to filter out is provided) and is usually configurable by setting different parameters such as the level of scraping (e.g. just the homepage or the homepage plus a first level of links from that, etc.).

Three software were implemented for the Scraper block: two general purpose software by Italy and Bulgaria, and a specific scraping software for social media detection by Poland. The reference URL is:

The Scraped content block is a container of the content scraped by the Scraper block. Usually it is necessary to implement this block using Big Data technological solutions due to the fact that the amount of information could be huge and mainly consisting of unstructured data.

The Index configuration block represents a strategy of indexing the scraped data stored into the Scraped content block. In a Big Data context, the huge amount of data that can be stored may make data indexing a mandatory step, in order to easily retrieve information in subsequent phases. This block is normally included in the storage platform.

The URL scorer block is used to assign a score to an URL on the basis of some predefined parameters such as the presence of some features of interest inside the URL's content. Given a list of URLs related to an enterprise, this block can be used alone or in conjunction with other blocks in order to prepare the decision strategy to identify the most probable official URL for a given enterprise.

A specific software was proposed by Italy to implement this block. This is available on the cited website.

The Tokenization block processes the textual content of the scraped resources by transforming it in a text that becomes input for further processing such as parsing and text mining or for analyses blocks. Normally, in lexical analysis, tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens.

The Data parsing block focuses on the analysis of the tokens produced by the Tokenization block by searching for a specific regular expressions, matching sentences etc.

The Word filters block is used to filter out some words/tokens (if a list of words to be filtered out is provided) from the scraped textual content or to enrich it with a list of go words.
The **Language specific lemmatization** block lemmatizes the tokens found in the scraped textual content in order to reduce the number of textual elements to be analyzed. Lemmatization in linguistics is the process of grouping together the inflected forms of a word so they can be analysed as a single item, identified by the word’s lemma, or dictionary form. In this case (computational linguistics), lemmatization is the algorithmic process of determining the lemma of a word based on its intended meaning. When it is not possible to determine the intended meaning usually the base form of a token is obtained by using a **stemmer** that compute the base form of a token by operating on a single word without knowledge of the context.

The **text representation** block includes the main paradigms according to which text could be represented for subsequent processing, namely:

- **Bag of words**: this very well-known and used to prepare document classification tasks where the (frequency of) occurrence of each word is used as a feature for training a classifier.
- **Engineered Features**: it is based on the selection of keywords intended to represent the text for the subsequent analysis task.
- **Language encoding**: this is the approach adopted by Word Embeddings. Word Embeddings have been successfully exploited to automatically summarize huge text corpora, and to subsequently encode the summarized texts in order to feed a deep learning algorithms for downstream analysis (e.g. to predict whether a given enterprise performs e-commerce or not)

The **Machine learning** block (and its sub-blocks) produces the final output statistics by using one (or more) learner(s).

To implement the Machine Learning block some scripts were implemented and shared by Italy and Poland. This can be found at: [https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP2_Links](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP2_Links)

The **Deterministic rules** block is designed from a set of rules with known characteristics of the sites and data in mind.

To implement the Deterministic rules block some software was made available at the cited website by Bulgaria.

The reference logical architecture has been adopted by all the developed pilots. In the Deliverable “Methodological and IT Issues and Solutions”, available at [https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/66/WP2_Deliverable_2.2_2017_07_31.pdf](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/66/WP2_Deliverable_2.2_2017_07_31.pdf), it is described how each pilot has been developed as an “instantiation” of this architecture.
In this chapter we focus on the methodological side of web scraping for statistics. We try to answer questions such as:

- “What can we learn methodologically from the six pilots executed in the field of web scraping of enterprises web sites in six countries?”
- “How easy is it to compare the approaches taken in different circumstances and what are the general underlying principles?”
- “Can we identify some methodological best practices, common problems and solutions from the pilots that were executed?”

We will compare and review the pilots according to three dimensions: (i) specific versus generic, (ii) the use of machine learning or deterministic approaches and (iii) the methods used in the pilots. We will use the pilots executed in both SGA1 and SGA2 as a framework of reference to describe these issues.

### 4.1 Specific vs generic scraping: when/what

The crucial factor when web-scraping is whether you know the structure of the content where the information is to be found or whether you cannot make any assumptions beforehand about the structure of the data. We will refer to the first case as **Specific scraping**, in the second case as **Generic scraping**.

We define the two different concepts more specifically:

1. **Specific web scraping** as when both the structure and the (type of) content of websites to be scraped are perfectly known. In this case, crawlers just have to replicate the behaviour of a human being visiting the website to collect the information of interest. Typical examples of specific web scraping can be found in the area of price statistics, where most of the items in a web store have the same product listing or product page and scraping software can retrieve specific elements, such as the original price, the sales price, the label, the description, the quantity, colour, material etc. from many web pages for many products.

2. **Generic web scraping** as when no a priori knowledge on the structure and content is available and the whole website must be scraped and processed in order to infer some information of interest. A clear example is the discovery of web pages of enterprises to derive some general characteristics of the population of enterprises. In this case more general scraping methodologies are needed using scraping and processing software.

Of course there are examples where the object of interest has to be attacked by a mix of specific and generic scraping. An example is the retrieval of job vacancies from a job portal, where the main elements of each job vacancy, such as the title, the domain, the branch and maybe closing date are usually well-structured and specifically scrapable. However, the job description tends to be an unstructured piece of text that has to be interpreted using techniques for generic scraping, such as text mining and machine learning.
So why do we make this distinction? We do this because it makes a difference technically and methodologically. The less we know of the subject to be scraped, the more generic our scraping techniques and methodologies should be. Technologically speaking with specific scraping it is possible beforehand to study the object of interest (website or sites) and design a specific scraper that uses this knowledge to navigate through the site using page structure identifiers such as html id’s, xpaths and css selectors. With generic scraping we usually strip out all html markup and apply text and document mining software on the remaining content, although in some cases html and the structure of the site may be useful in itself (for example, in determining e-commerce). From a methodological viewpoint with specific scraping we have reasonably well defined variables (keep in mind that in scraping there is always a certain degree of uncertainty about your data) to use in our processing, whereas in generic scraping we usually apply (machine) learning methods on the results of the data processing steps.

This idea is depicted graphically below:

<table>
<thead>
<tr>
<th>Know structure and content</th>
<th>know nothing</th>
</tr>
</thead>
<tbody>
<tr>
<td>specific scraping</td>
<td>generic scraping</td>
</tr>
<tr>
<td>use page structure (html id’s, css selectors)</td>
<td>text mining, vocabularies</td>
</tr>
<tr>
<td>specific methods on known variables</td>
<td>generic (learning) methods</td>
</tr>
<tr>
<td></td>
<td>pattern recognition</td>
</tr>
</tbody>
</table>

Although advanced text mining and machine learning methods might be applicable to specific scraping contexts as well, we see in the pilots that these methods are particularly applicable to generic scraping. In the next section we will dive deeper into this subject based on the analysis techniques used in the various pilot studies.

### 4.2 Analysis techniques: machine learning vs deterministic approaches

In the analysis phase we see a distinction between methodological approaches based on (machine) learning techniques and approaches based on deterministic techniques. Before we roughly classify our pilots into either one of these two approaches we give an informal definition:

1. In machine learning we have 2 flavour: supervised learning and unsupervised learning. We speak about *supervised machine learning* approaches in scraping for official statistics when algorithms or models/classifiers are derived from a set of training data which is supposed to be reasonably representative for the problem at hand. The parameters of the machine learning model are usually tuned with a validation set before measuring its performance on a so called test set with known characteristics. Finally the model is then applied to other sets of data of which we do not know anything but for which we suppose the model performs well, in order to produce statistics. We speak about *unsupervised machine learning* if we do not have a training data and try to train the model from the data at hand. In this ESSnet project we mainly applied supervised machine learning methods, except for the SDG case.

2. We speak about *deterministic approaches* in scraping for official statistics when algorithms are designed from a set of rules with known characteristics of the sites and data in mind. Put
in a different way, the knowledge of an expert is used to design an algorithm to process and interpret input data from web and other sources into statistical target variables. We call this method deterministic, because the algorithm applied to the same data will always have the same (deterministic) result, where the result of a machine learning approach heavily depend on the training set being used.

In general turning web data into statistics usually takes many steps, and in each step a different approach may be taken. One of the factors that influence this choice is the complexity of the relationship between the input variables or the features derived from the input data and the statistical target variables. If this relationship is fairly straightforward a deterministic approach might be obvious. But if this relationship is complex, unknown or difficult to model in a deterministic algorithm, which might well be the case when working with web data, a machine learning approach might perform better.

One thing to be noted here is that in machine learning approaches the availability of training data of sufficient quality is essential. This is an important challenge in many cases. In some of the pilots this training data is available or can be derived from earlier surveys. This might be in the short term – at the point at which a machine learning approach is introduced in official statistics to (partly) replace a traditional statistical process – however, in the long run, survey-based training data might become a rarity and other means have to be found to (re)train machine learning models. Other pilots do not have any training data at all available and rely on manually-labelled data. Obviously, deterministic approaches do not rely on training data, but have other pitfalls, and still rely on the availability of correctly labeled data in order to evaluate the performance of the deterministic algorithm (although not as much data is needed).

The following table (Table 1) roughly classifies the methods used in the pilots into the use of deterministic (D) and Machine learning (ML), although that some of them are actually a mix of both – for example, a deterministic approach may be taken to identify businesses with a particular attribute such as having a Social Media presence, and Machine Learning used to categorise the nature of the use of social media.

---

4 One could argue that a machine learning approach, once the model has been trained, is in fact also deterministic. That is true, however, these terms seem to be used widely to indicate the difference, so we will continue to use them in this document.
### Table 1: Methods used in pilots

This table shows that both deterministic as well as machine learning approaches are used. In the next section we take a closer look at the methods used in the pilots.

#### 4.3 Review of methods used in pilots

In this section we take a look at methods used within the pilots, bearing in mind our classification of specific vs generic scraping and deterministic vs machine learning approaches.

It’s useful to understand something about the machine learning (‘ML’) flow in order to have a better understanding of the work done within ML-based pilots.

Under the machine learning approach, the approach taken for turning the text collected by generic scraping into ‘features’ – representations of the text which can be used by the model (also known as independent variables in a regression context) - is crucial. HTML, and other information contained in tags and images, is not structured, and contain a lot of noise which, if not filtered, would make signals unintelligible. For this reason, text and data mining techniques must be applied in order to (i) structure and standardize data and (ii) turn this data into features.

In most of the experiences carried out in use cases, collected texts were processed through a preparation phase, which, as mentioned in Section 3, included:

- A set of text processing functions, like tokenization, lemmatization, data parsing, etc.;
- A text representation step, carried out in one (or more) of the following ways:
  - Keywords/’engineered’ features
  - Bag of words;
  - Language encoding (word embeddings)

Using keywords/’engineered’ features means turning the text into features by simply searching for keywords related to the use-case – for example, ‘basket’ and ‘shop’ in the e-commerce use-case. The
features in the machine learning model are then simply the normalized counts of these keywords. This is potentially a powerful and simple approach, but requires subject-matter knowledge to specify the features. In contrast, the ‘bag of words’ and language encoding (word embedding) both automatically generate features from the extracted text, which does not require subject-matter knowledge.

There is no consistent finding from the pilots as to which text-representation method is superior – it may be worth trying a number of approaches in combination with different classifiers. However, it’s worth noting that keywords/‘engineered’ features performed well in some cases, and it’s certainly worth trying this approach in cases where one has some subject matter knowledge.

It’s also worth noting that it was often the case that a feature selection step had to be applied to reduce dimensionality. In particular, for each dependent variable (e.g. e-commerce, online job-application, presence in social media) terms were processed treated in order to detect the most relevant for the prediction: a number of techniques have been employed, such as the correspondence analysis, regularization techniques (LASSO, Ridge, Elastic Net), importance in Random Forest generation.

The subsequent analysis phase mainly involved to solve a classification task either via a deterministic or via a ML approach.

In the case of a ML approach, a number of classifiers were fitted by different countries – approaches include Random Forests, Support Vector Machines, and Naïve Bayes. The most suitable classifier will vary according to the text representation and amount of training data – again, it’s worth trying a number of classifiers. All pilots utilising ML followed the standard supervised machine learning practice of splitting the training data into the training set on which the model was trained and a test set which was used to evaluate the model by comparing observed and predicted values. Training data were either prepared manually (e.g. Poland Use case on Social Media presence) or resulted from survey data (e.g. Italy and the Netherlands use case on Ecommerce).

Some important lessons learnt on training sets and in general on the quality evaluation of the classification task were:

- Check possible label noise (that is, incorrectly-labelled cases) present in training data and (i) either remove it (e.g. manually) or (ii) make your models robust with respect to it. We found that the quality of the results was deeply affected by the presence of label noise (see e.g. Italy Use Case Ecommerce).
- Keep in mind that training sets should be representative of the population – if a training set has been derived from a survey, it’s probably necessary to use the survey weights (as, for example, a survey will typically over-represent larger businesses).
- If the goal is to produce an estimate using the results of the classifier, it’s important to balance False Positives with False Negatives by adjusting the classification threshold. Simply choosing the highest-accuracy classifier is insufficient – since (for example) if this classifier produces more False Positives than False Negatives, the estimate will be biased upwards.

We will now discuss the six pilots in detail.
The **URLs retrieval pilot** has usually three steps: 1) getting starting points from a search engine, 2) do some scraping on the urls found, or otherwise extract information about them and 3) determining which of the results belongs to the enterprise. This first step is either performed via a structured API or by scraping the results page of a search engine. Two different approaches were used: (i) more information in the query to the engine\API (e.g. by Netherlands) or (ii) less info in the query and features retrieved on the results (e.g. by Italy). In either case the first step is a clear example of specific scraping with deterministic analysis techniques - the search results can be viewed as structured content.

The second step is an example of generic scraping as nothing is known from the site to be scraped beforehand. In the analysis phase, all countries used an ML approach to determine the correctness of a found url, but BG and PL applied a deterministic approach and a manual validation of results.

The **Ecommerce pilots** are mostly examples of generic scraping. Starting from a list of enterprise urls, nothing is known about the structure of the website beforehand. However, there are certain common characteristics on the way that web shops advertise their commercial activities. These characteristics have been used by NL and BG to design a simple deterministic approach based on a vocabulary of keywords that are commonly used on the home page of a webshop. NL and BG also applied a ML approach, like the other countries, for the identification of Ecommerce activities, where the ICT survey was mainly used as a training set. In the case of Italy, the best models have been used to predict the values of the target variables for all the units in the population of interest for which it was possible to identify the websites and scrape them.

Notably, UK compared a number of different approaches for text representation: bag-of-words, engineered features and language embeddings (doc2vec). The performances were overall comparable.

We conclude for now that Ecommerce detection seems to be an area where multiple approaches can be successful.

The **Job advertisements pilot** is an example where a mix of specific and generic scraping is applied. Sweden did some generic scraping on public sector websites. In addition they used data from job portals not retrieved by scraping. In the analysis phase Sweden used a machine learning approach. Bulgaria used a generic web scraping followed by a deterministic approach. Italy used a generic web scraping followed by a ML approach as in the case of the Ecommerce pilot. No clear conclusion emerged about which approach is superior – a number of different approaches appear useful.

In the **Social media presence pilot**, Italy used a generic web scraping followed by a ML approach. Poland developed a script to detect some of the most common social media buttons on webs pages of enterprises. As it heavily relies on a known structure that this is an example of specific scraping (although the structure can be anywhere on the website) and utilises a deterministic rule. The work on social media in SGA2 has gone a step deeper, not only detecting social media activity but also the kind of activity – which involves the use of machine learning against unstructured data, so for this kind of use-case the scraping has evolved into a mix of specific and generic scraping. After all, Twitter messages, which indeed were the focus of these pilots, are always the same, but nothing can be said beforehand on their content and must be interpreted.
In the **NACE pilot**, UK used bag-of-words, engineered features and language embeddings (doc2vec). For each text representation three classifiers were compared, namely: Support Vector Classifier (SVC) with polynomial kernel, Naive Bayes and Random Forests. The combination Doc2Vec with an SVC with a polynomial kernel performed best, but this finding is probably not generalisable to different contexts. For the Italian pilot the best results were obtained by the Random Forest classifier.

The **SDG pilot** carried out by the UK utilised a deterministic approach to identify websites containing sustainability information, and then unsupervised machine learning - Latent Dirichlet allocation (LDA) - for topic modelling.

The following table (Table 2) gives a quick view of the Machine learning classifiers used by ML-based pilots:

<table>
<thead>
<tr>
<th>1 URLs retrieval</th>
<th>IT</th>
<th>SE</th>
<th>UK</th>
<th>NL</th>
<th>BG</th>
</tr>
</thead>
<tbody>
<tr>
<td>URLs retrieval</td>
<td>Neural Networks</td>
<td>Random Forest</td>
<td>Support Vector Machines</td>
<td>Logistic model was chosen</td>
<td>Neural Networks</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td></td>
<td>Random Forest</td>
<td></td>
<td>Random Forest</td>
</tr>
<tr>
<td></td>
<td>Logistic Model</td>
<td>Random forest</td>
<td>Naive Bayes</td>
<td></td>
<td>Logistic model</td>
</tr>
<tr>
<td></td>
<td>Logistic model</td>
<td></td>
<td>Decision trees</td>
<td></td>
<td>was chosen</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2 Ecommerce</th>
<th>Logistic Model</th>
<th>Support Vector Classifier</th>
<th>Random Forest</th>
<th>Random Forest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Neural Networks</td>
<td>Naive Bayes classifier</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Naive Bayes</td>
<td>Random Forest classifier</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Support Vector Machines</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>Decision tree</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bagging</td>
<td>Keras sequential neural networks</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Boosting</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3 Job Advertisements</th>
<th>Logistic Model</th>
<th>Support Vector Classifier</th>
<th>Support Vector Classifier</th>
<th>Random Forest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Neural Networks</td>
<td>Naive Bayes</td>
<td>Naive Bayes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Naive Bayes</td>
<td>Decision Tree</td>
<td>Random Forest classifier</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Support Vector Machines</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bagging</td>
<td>Keras sequential neural networks</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Boosting</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The following table (Table 2) gives a quick view of the Machine learning classifiers used by ML-based pilots:
Looking back to the questions raised in the introduction of this chapter and briefly present the most relevant answers.

- “What can we learn methodologically from the four pilots executed in the field of web scraping of enterprises web sites in six countries?”
  One thing we learned is that it is useful and feasible to apply web scraping techniques in the field of official statistics to compute experimental indicators. However, there is not one preferred way of doing these very different pilots in different countries. Even per pilot the methods being used differ, which may have been caused by different data landscapes per country or other circumstantial differences. However some common machine learning methods have been applied in some of the pilots, especially the URL finding pilot where all participant country applied (almost) the same methodology. In addition, we learned some practical lessons that have been discussed in this section, e.g. on building training sets and on evaluating the quality of the results.

- “How easy is it to compare the approaches taken in different circumstances and what are the general underlying principles?”
  Using the same terminology and describing the work being done in general building blocks has been a big advantage to compare the different approaches, not only from an IT viewpoint, but from a methodological viewpoint as well. The concept of generic versus specific scraping and deterministic versus machine learning approaches (and within machine learning a wide spectrum of different classifiers and text representation methods) form a useful general underlying basis for scraping for official statistics.

- “Can we identify some methodological best practices, common problems and solutions from the pilots that were executed?”

<table>
<thead>
<tr>
<th>4 Social Media</th>
<th>-</th>
<th>Naïve Bayes</th>
<th>Naïve Bayes</th>
<th>-</th>
<th>Random Forest</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 NACE</td>
<td>-</td>
<td>Logistic Model</td>
<td>Keras sequential neural network</td>
<td>-</td>
<td>Support Vector Classifier</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Neural Networks</td>
<td></td>
<td></td>
<td>Naive Bayes classifier</td>
<td>Random forest</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Naive Bayes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Support Vector Machines</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bagging</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Boosting</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Random Forest</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Random Forest was chosen</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 SDG</td>
<td>-</td>
<td>-</td>
<td>Latent Dirichlet Allocation (LDA) topic model</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Machine learning classifiers used in pilots
All of the approaches used in the pilots in the different countries produced results that will be detailed in section 6. It is very difficult to generalize the approaches being taken. Both deterministic as well as machine learning approaches have been successfully applied. In the latter case some practical lesson learnt have been described.
5 Technological Issues and Solutions

Jacek Maślankowski (Statistics Poland)

Web scraping is not a new technique - it has been used since the beginning of the Internet and has evolved in recent years. For instance, proxy servers scraped the content of websites and serve them for local computers more than 20 years ago. This long history of web scraping has resulted in the variety of different tools and methods that can be used to gather information from websites.

Web scraping tools can be divided into generic and dedicated. Typical generic web scraping tools include the following: import.io, Scrapy, imacros, Apache Nutch or similar. The second group includes libraries for a specific purposes, such as Tweepy for Python programmers – to scrap the data from Twitter. The extended list of web scraping tools and libraries can be found on various repositories, including a Github.

The aim of this section is to give a general overview of the programming languages, tools and libraries used in WP2 pilots. Section 5.1 shows the current software used for pilots implementation. In section 5.2 there is information on the technological issues we have to tackle with during implementation of use cases.

5.1 Review of the technological environments used in pilots

Due to the variety of tools and methods for web scraping, each pilot can be implemented in various way. As mentioned in the previous point, there are repositories with hundreds of different tools and libraries that support web scraping. The goal for the WP2 members was to select popular open source or free web scraping software, that NSI’s employees are familiar with. It was the reason to include traditional as well as big data dedicated software that can be downloaded and implemented in NSI without extra costs.

In Table 3 we included the current programming languages, libraries and tools used in pilots.

---

5 https://github.com/lorien/awesome-web-scraping, accessed 30th May 2017
Table 3: Overview of programming languages, libraries and tools used in pilots

<table>
<thead>
<tr>
<th>Use Case</th>
<th>BG</th>
<th>IT</th>
<th>NL</th>
<th>PL</th>
<th>SE</th>
<th>UK</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-commerce</td>
<td>(1) PHP language</td>
<td>Java language</td>
<td>Python language</td>
<td>Python-SKLearn (Naive Bayes, Random Forests, Support vector classifiers)</td>
<td>Python-Scrapy NLTK Library ML (Naive Bayes)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2) ISTAT E-commerce:</td>
<td>R language</td>
<td>Scrapy for webscraping, scikit-learn</td>
<td>Python language</td>
<td>Python-Scrapy NLTK Library ML (Naive Bayes)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Java language</td>
<td>TreeTagger library for lemmatization SnowballStemmer for stemming</td>
<td>Crawler4J</td>
<td>Scrapy</td>
<td>Crawler4J</td>
<td></td>
</tr>
<tr>
<td>Job vacancies</td>
<td>PHP language</td>
<td>Java language</td>
<td>Python, the libraries are Pandas, re, urllib3, urllib, BeautifulSoup, numpy, glob, nltk, skleearn, matplotlib</td>
<td>Python language</td>
<td>Python language</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R language</td>
<td>TreeTagger library for lemmatization SnowballStemmer for stemming</td>
<td>Scrapy</td>
<td>Scrapy</td>
<td>Scrapy</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Crawler4J</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social media</td>
<td>(1) PHP language</td>
<td>Java language</td>
<td>Python language</td>
<td>Polish Socialmedia program: Python language (pandas, bs4, os, requests, csv, tweepy)</td>
<td>Python language</td>
<td>Python language</td>
</tr>
<tr>
<td></td>
<td>(2) Polish Socialmedia:</td>
<td>R language</td>
<td>Sci-kit learn library BeautifulSoup4 library Tweepy library Apache Spark/Hadoop to execute scripts</td>
<td>Python language</td>
<td>Python language</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Python language (pandas, bs4, os, requests, csv, tweepy)</td>
<td>TreeTagger library for lemmatization SnowballStemmer for stemming</td>
<td>Scikit-learn</td>
<td>Python-SKLearn (Naive Bayes, Random Forests, Support vector classifiers)</td>
<td>Python-Scrapy NLTK Library ML (Naive Bayes)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>tweepy)</td>
<td>Crawler4J</td>
<td>Tweepy library</td>
<td>Python language</td>
<td>Python-Scrapy NLTK Library ML (Naive Bayes)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Based on the Table 3, we can say that four different programming languages are used: Python, Java, PHP and R. For Netherlands, Sweden, Poland and United Kingdom the main programming language is Python. For Italy most of the work was done in Java and R. Bulgarian statisticians prefer to use open source PHP programming language.

There are several libraries used for pilots. Regarding Python examples, Poland and Sweden decided to develop a solution based on BeautifulSoup4, while Netherlands and UK decided to use Scrapy for these purposes. Based on their experience and lesson learnt we can say that Scrapy is a useful, scalable tool for web-scraping. On the other hand BeautifulSoup4 is a HTML parser that can process various forms of websites. For WP2 purposes there is a need to have any HTML parser that is able extract HTML tags from the file and exclude any CSS or JavaScript code in the analysis. It is also suggested to choose tools and libraries that are able to scrap the data from datafiles such as XML/JSON/CSV/TXT. Before scraping the webpage, it is recommended to inspect it by using an inspect tool, such as page inspector for Firefox web browser. It allows to identify which HTML tags are responsible on the web page to select a specific library.

For text processing and machine learning purposes in Python we have used NLTK library and scikit-learn. The NLTK library in Python is useful for investigating more advanced natural language processing. In text processing, the use of regular expressions is particularly powerful - especially in Polish Social Media Presence pilot, it allowed to identify all hyperlinks, even they are not included in the anchor HTML tag. Sci-kit learn allows access to various machine learning algorithms, including Naive Bayes, SVM, Decision trees or Random forest. R provides a subetute for Python libraries - ISTAT developed several scripts regarding machine learning in R, as presented in Table 3.

---

6 https://pypi.python.org/pypi/beautifulsoup4, accessed 30th May 2017
7 https://scrapy.org, accessed 30th May 2017
9 http://www.nltk.org, accessed 30th May 2017
10 http://scikit-learn.org, accessed 30th May 2017
The purpose of the choice of the tools for storage layer was to have an environment easy to maintain, and CSV files where therefore a common option. The comparison of the storage use for different pilots was included in Table 4.

<table>
<thead>
<tr>
<th>Use Case</th>
<th>BG</th>
<th>IT</th>
<th>NL</th>
<th>PL</th>
<th>SE</th>
<th>UK</th>
</tr>
</thead>
<tbody>
<tr>
<td>URL Retrieval</td>
<td>(1) MySQL database, CSV</td>
<td>(2) ISTAT software:</td>
<td>Apache Solr</td>
<td>CSV</td>
<td>CSV</td>
<td>CSV, json, MongoDB</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Apache Solr</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-commerce</td>
<td>(1) MySQL database</td>
<td>(2) ISTAT software:</td>
<td>Apache Solr</td>
<td>CSV</td>
<td>CSV</td>
<td>CSV, MongoDB</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Apache Solr</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Job vacancies</td>
<td>MySQL database</td>
<td>Apache Solr</td>
<td></td>
<td>CSV</td>
<td></td>
<td>CSV</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social media</td>
<td>(1) MySQL database</td>
<td>(2) Polish software:</td>
<td>Apache Solr</td>
<td>CSV</td>
<td>CSV</td>
<td>CSV, MongoDB</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CSV</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 4: Data storage used in pilots**

The most common way of data storing is a filesystem with CSV file type. Selected pilots use the Apache Solr database (NoSQL) and MySQL database (relational). The decision of using the filesystem as a primary data storage is a result of the fact that most of the tools used in pilots have embedded libraries to access CSV files. On the other hand, CSV files are used to store the results of analysis. The use of this filetype helps to load it into popular applications, such as R or MS Excel. To increase the performance of accessing such files, there is a possibility to store them in HDFS filesystem (Hadoop Distributed File System) to perform automatic and very efficient parallel data access.

The purpose of the use of Apache Solr is to provide a scalable environment that is able to store different types of data. However the main purpose of the use of Apache Solr in WP2 pilots is to store websites in NoSQL database. This type of database allows dynamic searching through its storage including full-text search, hit highlighting, faceted search, dynamic clustering, database integration, rich document handling, distributed search, index replication and high scalability.

For the WP2 purposes several different tools has been developed (see Table 5). For example, ISTAT developed a software URLSearcher in Java that allows to retrieve URLs associated with different enterprises based on their attributes, such as name, city, contact address. This tool can be applied in Java environment in every operating system. The output of this tool is a CSV file or a set of files including links identified. Further analysis of URLs can be performed with other tools developed by ISTAT for the URL Retrieval use case.
<table>
<thead>
<tr>
<th>Use Case</th>
<th>BG</th>
<th>IT</th>
<th>NL</th>
<th>PL</th>
<th>SE</th>
<th>UK</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>URL Retrieval</strong></td>
<td>(1) Custom PHP scripts: Conf, geturl, checkoldurl, google_search, jabse_search, list, info Custom Java application: URLSearcher URLSearcher (custom Java application) RootJuice (custom Java application) UrlScraper (custom Java application) UrlMatchTableGenerator</td>
<td>URLRetriever</td>
<td>NodeJS package: GoogleSearch Custom Javascript: feature extraction Custom Python: feature selection, model selection, ML model training and prediction and creation of reports</td>
<td>URLSearcher (custom Java application) RootJuice (custom Java application) UrlScraper (custom Java application)</td>
<td></td>
<td>Python scripts created for performing API querying and breadth-first scraping crawls.</td>
</tr>
<tr>
<td></td>
<td>(2) Custom Java application: URLSearcher, RootJuice, UrlScraper, UrlMatchTableGenerator Custom R scripts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 5: Tools used or developed for the pilots**

Most of the software was developed as a short scripts used for a specific actions regarding data collection and processing. The most popular environment for the pilots testing was to use virtual machine (e.g., Linux Mint) or dedicated machine (e.g., Linux Ubuntu Server, MS Windows).
Our experience of pilots sharing shows that they WP2 scripts and applications are platform independent. For instance, it is possible to run Python scripts or Java applications in Linux as well as Windows environment without any changes in the source code.

During our pilots implementation we have learnt that Apache Solr that is not only a NoSQL database but also an enterprise search platform. It is possible to store any type of data, including web pages, which represents the main input files of the WP2 cases.

5.2 Issues and Solutions

Performance
The performance is a key issue especially when downloading and processing the whole websites. Processing unstructured information is very CPU and memory consuming, especially with machine learning algorithms, and as a result not very efficient. Because of the fact that most of the environment used for WP2 pilots have limited CPU and memory resources it was difficult to estimate how efficient the algorithm will be in a production environment. Based on the Bulgarian example we can say that conventional IT tools are sufficient for the URL inventory creation with tens of thousands enterprises. Based on the Italian experience, Apache Solr has technical problems in dealing with Solr Connection pool in the loading phase right after the scraping phase.

Sustainability
Due to the fact that big data tools are changing very frequently as well as the website technology, there is a need to provide an agile-like development of tools. The technologies used for the pilots are likely to change in the next few years, and for that reason we do not recommend a specific language. One example of this is the Polish implementation of the social media use case, which was started in Python 2 with HTML Parser as the main library and was eventually migrated to Python 3 with BeautifulSoup library.

It is possible to switch easily to other platforms. ISTAT example allows to use another similar solution, e.g., elastic Search instead of Apache Solr. The use of CSV files implemented in most of the use cases can also be replaced by any filesystem more efficient, such as HDFS.

Storage
Based on the Bulgarian experience, the requirements for the storage of a database about 27,000 enterprises takes around 1 GB of HDD, including BR data, scraped Search APIs data and enterprises web sites and e-stores first pages titles, key words, descriptions and URLs data. It allows to create a conclusion that for this specific use case it is possible to use traditional technology.

As mentioned in the previous section of this document, the possible choices are between filesystem (CSV, JSON etc.), NoSQL database (Solr, Cassandra, Hbase etc.) or relational database (MySQL, PostgreSQL, SQL Server etc.). The decision of the use of particular data storage should be taken according to the volume of the data and the type of data to be stored.

De-duplication issues
There is a need for de-duplication framework that will automatically exclude all duplicates of websites and particular information taken from them, e.g., job offers. Based on the Swedish experience, de-duplicating the pages is needed since the same pages are retrieved from different links.
National languages and diacritic characters

Although the frameworks are developed in particular countries, it is possible to apply them in other country without any major changes. For example, URLSearcher, a tool developed by ISTAT was tested with Polish websites that contains specific diacritic characters, such as “ą”, “ę” or “ś”. Since the results are stored in UTF-8 coding page (65001), it allows to be read by other tools, e.g., MS Excel or MS SQL Server via import/export wizard.

However the web scraper must be able to recognize different coding pages when scraping the web page. Polish web pages are usually published in UTF-8, WINDOWS-1250 (CP-1252) or ISO-8859-2. Therefore when storing the data from web pages, it is recommended to unified the coding page. Because of possible international comparability the suggested option is to use UTF-8.

Licensing

All software used for the WP2 pilots implementation is free and open-source. This means that everyone can easily test and improve the tools. On the other hand, using such software means and not always we could rely on a good documentation or on a detailed guide to make everything work. A variety of different tools and methods used in pilots may lead to the question whether it is possible to achieve a high quality output using different tools. The simple answer is yes, as programming languages and data storage technologies are independent from the methodology. It means that we can choose what software is the most convenient for us to solve a specific problem, such as collecting a particular information from unstructured data published on websites.

In fact the decision of using a specific programming language may be dependent on the skills of IT staff employed in NSI. Different criteria are used when selecting types of data storages to be used. The reason for the use of a specific storage is based on the type of the data that will be stored. When it is necessary to store a whole webpage, the most expected is a database with NoSQL features, such as Apache Solr (see ISTAT Use Case 1). When there is a necessity of gathering a structured data, any structured or quasi structured database or file can be used (such as MySQL or CSV file).

However results of analysis may vary depending on the methodology used. It means that the tool prepared with the specific Machine Learning algorithm, e.g., Naïve Bayes, may provide different results, depending on the training dataset used or the type of an algorithm. It was briefly explained in the appendix by comparing results of the use of different machine learning algorithms (see Appendix, Example 1 IT 1).
6 Outputs Provided by Use Cases

This section presents the statistics produced by several of the pilots alongside comparisons against survey estimates (except for the ‘rate of retrieved URLs’ use-case, where there is no relevant estimate available). Methodological notes accompanying these estimates can be found in Section 8.

We remark that survey estimate has not to be considered as a gold standard, though however providing reference figures that are currently published to which compare the predicted ones.

Table 6 Rate of retrieved URLs from an enterprise list

<table>
<thead>
<tr>
<th></th>
<th>Estimate from web data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italy</td>
<td>63%</td>
</tr>
<tr>
<td>UK</td>
<td>24%</td>
</tr>
<tr>
<td>Netherlands</td>
<td>78%</td>
</tr>
<tr>
<td>Bulgaria (BNSI software)</td>
<td>82%</td>
</tr>
<tr>
<td>Bulgaria (ISTAT software)</td>
<td>74%</td>
</tr>
<tr>
<td>Poland</td>
<td>74% (Pomeranian voivodship only)</td>
</tr>
</tbody>
</table>

Table 6 reports the results obtained by the different countries for use case 1. The lowest result is provided by UK. This is due to sites not scraped (28%) either for technical or for ethical reasons, and to potential web sites not found by the search engine API (around 15%).

Table 7 Rate of enterprises engaged in websales on their website

<table>
<thead>
<tr>
<th></th>
<th>Estimate from web data</th>
<th>Survey Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italy</td>
<td>16%</td>
<td>15%</td>
</tr>
<tr>
<td>UK</td>
<td>30%</td>
<td>21% (unweighted ICT survey 2015)</td>
</tr>
<tr>
<td>Netherlands</td>
<td>14%</td>
<td>33%</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>6,2 %</td>
<td>8,6% (unweighted ICT survey 2017)</td>
</tr>
</tbody>
</table>

Table 7 reports the results obtained by the different countries for use case 2. Please notice that the NL result is based on a keyword/feature engineering approach for text representation that could be affected by errors. This could explain at least part of the difference that is reported.
Table 8 Rate of enterprises with job advertisements on their website

<table>
<thead>
<tr>
<th></th>
<th>Estimate from web data</th>
<th>Survey Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italy</td>
<td>14%</td>
<td>11%</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>19%</td>
<td>9% (ICT survey 2016)</td>
</tr>
</tbody>
</table>

Table 8 reports results for Italy and Bulgaria with Bulgaria output having some more relevant difference with respect to survey estimates.

Table 9 Rate of enterprises that are present on social media

<table>
<thead>
<tr>
<th></th>
<th>Estimate from web data</th>
<th>Survey Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italy</td>
<td>37%</td>
<td>31%</td>
</tr>
<tr>
<td>UK</td>
<td>80%</td>
<td>66% (unweighted ICT survey 2015)</td>
</tr>
<tr>
<td>Netherlands</td>
<td>65%</td>
<td>69%</td>
</tr>
<tr>
<td>Bulgaria (BNSI software)</td>
<td>31%</td>
<td>34%</td>
</tr>
<tr>
<td>Bulgaria (Polish software)</td>
<td>37%</td>
<td>34%</td>
</tr>
<tr>
<td>Poland</td>
<td>26% (Pomeranian voivodship only)</td>
<td>25% (Pomeranian voivodship only)</td>
</tr>
</tbody>
</table>

Table 9 reports results for all the countries with quite aligned figures between estimateds from web data and survey estimates.

Table 10: Percentage of enterprises using Twitter for a specific purpose, estimated from web data

<table>
<thead>
<tr>
<th></th>
<th>Recruit employees</th>
<th>Develop the enterprise's image or market products (e.g. advertising or launching products, etc)</th>
<th>Others – any tweet that does not fit into the other two categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Netherlands</td>
<td>90%</td>
<td>93%</td>
<td>Not estimated</td>
</tr>
<tr>
<td>Bulgaria (Polish software)</td>
<td>0%</td>
<td>61%</td>
<td>39%</td>
</tr>
<tr>
<td>Poland</td>
<td>90% (Pomeranian voivodship only)</td>
<td>46% (Pomeranian voivodship only)</td>
<td>100% (Pomeranian voivodship only)</td>
</tr>
</tbody>
</table>

Table 10 reports results on the use of Twitter data by specific purposes as estimated from web data. Due to the fact that we decided to use different classification than ICT in Enterprises it is not possible to use the accurate comparison of the results from Table 11 with the data from the survey. We also calculated the indicators using the target population of the enterprises using social media. The target population for these indicators in the survey is the total number of enterprises, not only using social media. For example, for Poland in Pomeranian voivodship, 25% of enterprises are using social media (26% based on Big Data), of them ca. 41% are using social media (90% based on Big Data) to recruit employees and ca. 95% to build enterprise image (46% based on Big Data).
7 Conclusions, Recommendations and Future Challenges
Matthew Greenaway (ONS), Ingegerd Jansson (SCB), Dan Wu (SCB)

We learned that it is useful and feasible to apply web scraping techniques in the field of official statistics to compute experimental indicators. The framework in section 3 describes the common approach. Using the same terminology and describing the work being done in general building blocks has been a big advantage to compare the different approaches, not only from an IT viewpoint, but from a methodological viewpoint as well. The concept of generic versus specific scraping and deterministic versus machine learning approaches (and within machine learning a wide spectrum of different classifiers and text representation methods) form a useful general underlying basis for scraping for official statistics.

Some common machine learning methods have been applied in some of the pilots, especially the URL finding pilot where all participant country applied (almost) the same methodology.

However, there is not one preferred way of doing these very different pilots in different countries. Even per pilot the methods used differ, which may have been caused by different data landscapes per country or other circumstantial differences. We learned some practical lessons that have been discussed in section 4, e.g. on building training sets and on evaluating the quality of the results. In section 6, the detailed practical lessons in pilots are described.

We conclude the package work from two perspectives, the quality and the IT techniques.

7.1 Conclusions and recommendations on quality
We observe that predicted values can be used for a twofold purpose:

1. **At unit level**, to enrich the information contained in the register of the population of interest; The quality of data pertaining the unit level can be measured by considering the same indicators used to evaluate the trained model. If the test set is representative of the whole population and not used for training the model, the performance measures (like accuracy and F1-score) calculated for the test set can be considered a good estimate for the overall performance. Another way to train a model and measure the performance is to use n-fold cross validation. Assuming that the train/test set is representative for the whole population, the average performance of the model against the train/test set is a good estimate for the overall performance of the model. Table 11 reports the qualitative judgement on the quality at unit level. For the detailed information, please refer to the methodological notes.
Table 11 General quality evaluation per use case

2. **At population level**, to produce estimates. The quality estimation is much more complex comparing with the unit level. Italy proposed a strategy applied in the pilots and is summarized in Figure 4.

<table>
<thead>
<tr>
<th>Estimator Type</th>
<th>Formula</th>
<th>Weighting</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design-based/model-assisted</td>
<td>$\hat{P} = \sum_i \gamma_i \bar{w}_i$</td>
<td>$\sum_{i=1}^N w_i = N_U$</td>
<td>$w_i$ weights are obtained by calibration procedure of base weights (inverse of inclusion probabilities) making use of known totals in the population in order to reduce bias due to non-response and the variability due to sampling errors.</td>
</tr>
<tr>
<td>Model-based</td>
<td>$\hat{P} = \sum_i \bar{y}_i w_i$</td>
<td>$\sum_{i=1}^N w_i = N_{P(1)}$</td>
<td>The estimate of the total number of enterprises offering web ordering facilities on those websites is given by the count of the predicted values $\bar{y}_i$ for all units for which it was possible to retrieve websites (population $U^1$), calibrated in order to make them representative of all the population having websites ($U^2$).</td>
</tr>
<tr>
<td>Combined</td>
<td>$\hat{P} = \sum_{i \in U^2 \setminus U^1} \bar{y}<em>i + \sum</em>{i \in U^1} (\bar{y}<em>i - y_i) w_i^0 + \sum</em>{i \in U^2 \setminus U^1} \bar{y}_i w_i^{0'}$</td>
<td>$\sum_{i=1}^N w_i^0 = N_{U^0}$ and $\sum_{i=1}^N w_i^{0'} = N_{U^0 U^2}$</td>
<td>Estimates are produced by summing three components: 1. the counting of predicted values in the subpopulation $U^2 \setminus U^1$ of units for which it was impossible to scrape and process corresponding websites; 2. an adjustment based on the consideration of the differences between the $y_i$ reported values and the predicted values (expanded to the same subpopulation $U^2$); 3. the counting of observed values for the $y_i$ respondents that described a website, that was not found nor scraped, expanded to the whole subpopulation $U^0 \setminus U^1$.</td>
</tr>
</tbody>
</table>

**Figure 4: Estimators**

Three Estimators are proposed, namely:

(i) Design-based/model-assisted;

(ii) Model based, the main flaws of the model based estimator are in the presence of

- prediction errors;
- undercoverage of the population of enterprises owning websites, part of which has not been reached by web scraping.

(iii) Combined. Looking at coherence as one important dimension of quality, both
combined estimates and full model based estimates can be considered as equally acceptable. But two considerations can be made.

- The second component of the combined estimator is based on an assumption of perfect correctness of reported values, and considers predicted values as errors when they do not coincide with the reported ones. But controls have been carried out when fitting models, and in half of the cases in which predicted values were contradictory with reported ones, this was not due to model fault, but to response errors. So, this assumption does not always hold. In any case it would be advisable to deepen this phase also by returning to the respondents to verify if it is an error in response or if, for example, the model has evaluated the content of a site different from that one considered by the respondent.
- If a medium-term aim is to make multi-annual frequency of the questions in the survey related to the websites characteristics, then the combined estimator cannot be applied, as it relies on the current availability of reported values from the survey, and the full model based estimators remains the only alternative. In this case, there would be an issue in time series analysis due to problems in comparability between survey estimates and model based ones.

The work done so far could be extended in multiple ways. In particular, if we consider the different use cases:

1. the e-commerce detection algorithms could be refined to distinguish between different levels of e-commerce maturity (for instance, determined by the presence of only an ordering facility, or also payment and deliver tracking ones).
2. The job advertisement spiders could be trained to additionally take the job details and the enterprise characteristics into consideration. The identification of the characteristics of each single job (economic activity, profession) and even the skills required, is a much more ambitious task that implies a different approach, more oriented to “information retrieval” than to “machine learning”.
3. The social media presence detection could be extended to not only observe the enterprise website, but also inspect the social media itself in order to investigate what kind of use of social media is being done in a more detailed way. This is of course subject to ethical and legal considerations.

7.2 Conclusions and recommendations on IT techniques

Python, Java and PHP are the main web scraping languages. R and Python are used for data analysis. Both NoSQL (Apache Solr and MongoDB) and relational databases (MySQL, MS SQL) are used for data storage; in some cases, the data are stored in CSV files. We can choose the most convenient programming language depending on the in-house competence.

However, libraries for different programming languages provide possibility for parameters setting. Such as scikit-learn of Python, have options for algorithm “tuning”. The varieties of parameters setting can generate different results on the same dataset. This should not be interpreted as the different results. More efforts need to be investigated on the parameters’ setting and the result comparison.

It is the differences in methods cause the main differences in results. For example, the varieties on the method of preparing the training data sets and the method of text representation produce different results.

Data storage must be appropriate to the goal of the specific use case. The selection criteria can be performance, possibility of data processing and analysis and the data amount and type.
### 7.3 Future challenges

More and more data pertaining to individuals and enterprises are being placed on Internet. This is a significant opportunity; however, challenges exist in the several areas, i.e., ethics and legal framework, Internet as the data source, methodology for integrating big data for OS and the complete IT platform of handling the Internet data life cycle.

- **Ethics and legal framework**

Extracting knowledge from online data draws attention and public concerns how governments are utilising online data, including relatively uncontroversial cases such as NSIs utilising textual data on company websites. GDPR also adds further requirements to the web scraping scenario. As a response to this, NSIs and Eurostat need to develop transparent web scraping policies in order to allay public concerns about the data collected and the usage of them. The ‘netiquette’ developed as a part of this work-package is an important first step. It is remarkable that NSIs adopt different interpretations on the question. Sharing good experiences among NSIs need to continue.

- **Challenge of adopting Internet data**

Websites are created with various techniques and different standards. Some are even behind firewalls. Some are built with Javascript, which make the content unseen to an ordinary scraper. To develop efficient scrapers, in the case of generic web scraping, intense web technique is necessary in order to handle the webpages’ varieties. As the internet evolves, data in forms of audio or video files are increasing in volume; and the use of interactive or user-specific content increases. The new data formats set even higher requirements on the scraper and the storage. To make the web scraping approach lasting in the longer term, it is necessary to build up new competence, both for web scraping and for data storage at NSIs. Sharing software and knowledge is very important for competence building. The participating countries of this package have been able to share software for web scraping and social media presence checking, which prove to be an efficient way.

- **Methodological challenge**

Not only the internet technologies evolving rapidly. The stability of Internet as a data source for OS need also consideration. Internet policies for enterprises vary depending on their economic activities, size and many other issues. To handle the data source stability bring the challenges to the methodology.

The enterprises’ URL inventory contains millions of URLs and millions of webpages are scraped for analysis; it is not predictable how websites are changing, and hence how they can be used for various purposes. Without knowing the web data fully, it is impossible to estimate the coverage and the bias in the estimates. A source of bias, for some statistics, will simply be that some businesses are less likely to have a website than others. For example, we may be more likely to identify websites for businesses that conduct e-commerce, and e-commerce statistics based purely on data scraped from the websites will therefore be biased without adjustment.

A key challenge will be to understand the web data and the bias for different use-cases. This is likely to involve methodological work on combining web-scraped data and survey or administrative data. The questions need answers are, e.g., what is the quality framework of the statistics generated from web data? How compare the statistics from the web data with statistics based on the survey data?
• Data management platform

The web scraping system is separated from the ordinary systems at some NSIs for security reasons. The gap is inconvenient when transferring the scraped data from the system on one network to the system on another network, especially for big data. The data handling process is different from other types of data. The data are stored, not in the traditional relational database, but in a NoSql data storage; then they are analysed and extracted for loading into databases. It is a challenge to build up the pipeline to manage the entire data life circle i.e., the stage area for storing data scraped, data analysis, and extracting, transferring and loading process. In the data analysis, machine-learning methods are often applied. For developing good classifiers, challenges are to build up good quality training datasets and robust data processing procedures.
Appendix with Methodological Notes by Use Case

Use Case 1: URLs Retrieval

ESSnet Big Data WP2: Webscraping Enterprise Characteristics

Methodological note

The ESSnet BD WP2 performs joint web scraping experiments in multiple countries, using as much as possible the same methodological concepts. The aim is to derive experimental statistics on enterprises from information found on the web, especially the websites of enterprises. It should be noted that these statistics have not reached maturity in terms of harmonisation, coverage or methodology. At this point they are to be treated as the output of research experiments and they do not necessarily align with the official statistics published on this subject.

Use case: URL retrieval

Country: Italy
Date: 2018-03-05
Authors:

Data sources

- The Italian Business Register (ASIA)
- Administrative data (CONSODATA)
- Indications of website address from 2015, 2016 and 2017 rounds of Italian ICT survey (responses from about 14,000 enterprises)
- Websites from enterprises (2017)

Population
The 184,000 enterprises as defined by the ICT survey (>10 persons, limited NACE)
Methodology

The overall procedure is based on the following steps:

1. **Building the input training dataset**: We consider the units for which the URLs are known as well as the specific features of such units that could be useful to identify them on the Web. We built such a dataset by integrating different sources, namely: third party information (Consodata) and “ICT in enterprises” survey editions of 2014, 2015, 2016. The resulting dataset was composed by 81912 enterprises with URLs.

2. **URLs searching**: The objective is to retrieve for each unit in the input training dataset, one or more URLs to scrape based on identifying information that is present in such a dataset. We decided to set up an automated procedure that used “enterprises’ denominations” (in Italian “Ragioni Sociali”) for searching the enterprise on the Internet via a search engine. In particular, we used the denomination of the enterprise as a search string, then we queried the search engine (Bing) and collected the first ten links returned as the result of the search query. Each of the link was visited.

3. **URLs crawling**: The objective is to retrieve, for each potential URL collected at the previous step, the textual content of the corresponding page. We used ad-hoc developed software that is available at the links: [https://github.com/SummaIstat/UrlSearcher](https://github.com/SummaIstat/UrlSearcher), [https://github.com/SummaIstat/RootJuice](https://github.com/SummaIstat/RootJuice)

4. **URLs scoring**: For each text collected in the previous step, a score vector is computed and a score is assigned. For our case study, the resulting dataset, i.e. the output scoring dataset, contains the following fields: `enterpriseId`, `linkPosition`, `URL`, `scoreVector`, `score`. The score vector contains the following elements:

   - Simple URL (is the URL in the form www.name.com or not ?)
   - VAT (is it present in the page or not ?)
   - city (is it present in the page or not ?)
For each element/characteristic we computed the confusion matrix obtained by using just that element for classification purposes. Based on that confusion matrix we computed the standard performance measures, i.e. precision, recall and f-measure. Then, we assigned the f-measures of the corresponding confusion matrixes as raw weights to elements; lastly we normalized the raw weights so that the sum of the final (adjusted) weights is 1000. In order to assign a final score to a link, we summed up the normalized weights of those elements that were actually present in the vector. In a nutshell, we multiplied each number of the vector with a specific coefficient and summed up all the results in order to obtain a score for that link.

In step 5 (machine learning) taking into account that a subset of enterprises for which the correct link is already indicated is available, it is possible to adopt a machine learning approach under which a model is fitted in this “training” set, and then applied to the set represented by all other enterprises. Our input training dataset consisted 57,000 records that had at least one page fetched. On the basis of the output scoring dataset we first associated to each enterprise the link with the highest score. As we know if the link is correct or not, a dichotomous variable correct_Yes_No says if the URL is the right one or not: this variable plays the role of the Y variable, to be predicted by the model. Together with this information, variables indicating success or failure of the search of telephone, VAT code, municipality, province and zip code play the role of the X variables (predictors), together with the link position and coincidence of the central part of the URL with the name of the enterprise (simple URL). This initial set is split into two equal size subsets, the first acting as the proper training set to fit the model, the second as the test set used to evaluate the performance of the model. Three different models were fitted: logistic, neural network and random forest. Their performance was almost equivalent, and logistic was chosen for the interpretability of its parameters.

In step 6 and 7 we applied the logistic model to the set of enterprises for which the website URL was not known.

Results

Accordingly to the predicted score, we decided if the found link was acceptable or not in terms of reliability. By so doing, we were able to find about 36,000 new URLs that, added to the already available, allowed a total number of about 101,000 URLs. Considering that the estimate (from the ICT survey) of enterprises with a website is around 133,000, the obtained coverage is 76%.
Conclusions

Our results show the feasibility of addressing the URL retrieval problem with solution that gets good results both in terms of quality and efficiency. This solution can be in place to retrieve URLs in all cases where they are missing or a quality control on collected ones is appropriate.

<table>
<thead>
<tr>
<th>score_class true false classification_error group</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 [0.101,0.209] 641 2221 0.77603075 1</td>
</tr>
<tr>
<td>2 (0.209,0.269] 703 2294 0.76543210 2</td>
</tr>
<tr>
<td>3 (0.269,0.32] 630 2142 0.77050360 3</td>
</tr>
<tr>
<td>4 (0.32,0.414] 853 2055 0.70667125 4</td>
</tr>
<tr>
<td>5 (0.414,0.48] 1178 1556 0.56912948 5</td>
</tr>
<tr>
<td>6 (0.48,0.66] 1759 1080 0.38041564 6</td>
</tr>
<tr>
<td>7 (0.66,0.812] 2497 564 0.18425351 7</td>
</tr>
<tr>
<td>8 (0.812,0.875] 2459 361 0.12801418 8</td>
</tr>
<tr>
<td>9 (0.875,0.917] 2424 267 0.09921962 9</td>
</tr>
<tr>
<td>10 (0.917,0.943] 2536 287 0.10166490 10</td>
</tr>
</tbody>
</table>

Recall: 0.7441994
Precision: 0.8202192
F1 measure: 0.7803623
ESSnet Big Data WP2: Webscraping Enterprise Characteristics

Methodological note – Test Statistic Produced: Yes

The ESSnet BD WP2 performs joint web scraping experiments in multiple countries, using as much as possible the same methodological concepts. The aim is to derive experimental statistics on enterprises from information found on the web, especially the websites of enterprises. It should be noted that these statistics have not reached maturity in terms of harmonisation, coverage or methodology. At this point they are to be treated as the output of research experiments and they do not necessarily align with the official statistics published on this subject.

Use case: URLs retrieval

Country: UK
Date: 2018-04-04

Data sources
Online Registers, Business Websites, ICT survey (called the ‘E-commerce survey’ in the UK)

Population
The enterprises as defined by the ICT survey (called the ‘E-commerce survey’ in the UK) (>10 persons, limited NACE)

Methodology

1. Extraction of a set of enterprises with known websites from an online website register as a ‘training’ set

2. For each enterprise in this training set, query a web search API with the enterprise name and collect the first 10 responses – call these ‘candidate’ websites

3. Scrape each ‘candidate’ website and store the scraped text (note – all scraping carried out according to ONS web-scraping policy)

4. Extract features from the scraped text, including: whether the enterprise address is present on the website, search rank of website, whether the enterprise name is in the URL

5. Train a machine learning model (random forests) to predict whether each website is a genuine match to the enterprise

6. Apply this pipeline and model to population of interest to identify websites for enterprises currently without a known website
Results

The confusion matrix below summarises the performance of the machine learning classifier (from step ‘5’ in the ‘methodology’ section). Precision is very high – where we do identify a website, it is probably the correct one – but recall is somewhat lower – we are still missing a fairly large number of genuine websites.

<table>
<thead>
<tr>
<th></th>
<th>Feature found</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct website</td>
<td></td>
</tr>
<tr>
<td>True</td>
<td>25</td>
</tr>
<tr>
<td>False</td>
<td>88</td>
</tr>
<tr>
<td>False</td>
<td>943</td>
</tr>
<tr>
<td>True</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.96</td>
<td>0.78</td>
<td>0.86</td>
</tr>
</tbody>
</table>

However, when applying this methodology to the reference population, the quality of our results also depends on steps in the pipeline other than the machine learning. For an estimated 28% of websites, we were unable to scrape the website for technical or ethical reasons. There are also an estimated 15% of potential websites not found by the search engine API. For this reason, we identify websites for only 24% of the reference population, compared to approximately 81% of the population that have a website.

Limitations and future work

We are only able to identify a website for a relatively small proportion of companies using this method, but we are reasonably certain that the websites we have identified are correct.

An additional issue is that the websites we have identified are biased towards large companies, possibly because of technical capabilities & web presence are large in companies with more resources.

Future work can include combining these data with website registry information and improving our scraping methods to increase coverage. To improve classifier performance, we can utilise a more explicit, curated training set tailored towards UK websites. The initial search engine API queries can be expanded and optimised, especially for retrieving small company urls.
ESSnet Big Data WP2: Webscraping Enterprise Characteristics

Methodological note

The ESSnet BD WP2 performs joint web scraping experiments in multiple countries, using as much as possible the same methodological concepts. The aim is to derive experimental statistics on enterprises from information found on the web, especially the websites of enterprises. It should be noted that these statistics have not reached maturity in terms of harmonisation, coverage or methodology. At this point they are to be treated as the output of research experiments and they do not necessarily align with the official statistics published on this subject.

Use case: URL Retrieval
Country: NL
Date: 2018-03-28
Authors: Olav ten Bosch and Dick Windmeijer

Data sources
- The General Business Register (GBR) (about 1.5 M enterprises of which 1/3 have a known URL)
- Responses from Dutch ICT use survey (responses from about 9 000 enterprises)
- Websites from enterprises via Google search engine

Population
The enterprises as defined by the ICT use survey (>10 persons, limited NACE)

Methodology
1. Extraction of a set of 1 000 enterprises with URL from the GBR, > 10 persons, any NACE code.
2. For all enterprises in this set, we executed 6 search queries on Google, using the Google API. 5 queries use a combination of enterprise name and address info. One query is designed to find the contact page of an enterprise.
3. We split this set into a training set and a validation set and created a model to predict if a URL from the search results belongs to the enterprise being searched for. Features are created by checking whether the name or postal code is contained in the URL, title or snippet. Also the rank in the search results list and the type of query is used as a feature.
4. For the 40 000 enterprises without URL from the GBR, > 10 persons, any NACE code, we performed the 6 search requests and used the model derived in step 3 to select the URL that is most likely to belong to the enterprise.
5. We selected the set of enterprises that responded to the ICT use survey and for which we have a URL either from the GBR or as a result from step 4.
5. We compared the answers from the ICT use survey on the question “Do you have a website” with our prediction based on the web.
**Results**

We describe the confusion matrix and some common performance metrics\(^1\).

**Question:** *Does your enterprise have a website (Yes, No)?*

<table>
<thead>
<tr>
<th>Predict</th>
<th>True</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>6846</td>
<td>1158</td>
<td>8004</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>113</td>
<td>792</td>
<td>905</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>6959</td>
<td>1950</td>
<td>8909</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.98</td>
<td>0.86</td>
<td>0.92</td>
</tr>
<tr>
<td>No</td>
<td>0.41</td>
<td>0.88</td>
<td>0.55</td>
</tr>
<tr>
<td>avg / total</td>
<td>0.93</td>
<td>0.86</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Accuracy: 0.86

Balanced accuracy: 0.87

Prediction having a website: 78 %

---

Comparison with published figures
Enterprises in ICT population having a website:
Published: 86%
Predicted: 78%

Comparison per number of employees:

Comparison per NACE class:

C Manufacturing
F Construction
G Wholesale and retail trade; repair of motor vehicles and motorcycles
H Transportation and storage
Conclusion
- Detecting the existence of a website for an enterprise can be done successful via this method.
- The predicted values are consistently smaller compared to the published figures for all size classes and NACE classes (except “M Professional, scientific and technical activities”). The relative distribution of the percentage of businesses with a website over size class and over NACE classes is comparable for true versus predicted.

Limitations and future work
- The datasets were created at different points in time; the ICT use survey was executed beginning 2017, the URL retrieval was carried out January 2018. This could explain some differences.
- One has to keep in mind that not only the web scraper could be wrong, an answer on the ICT use survey could be wrong as well.
- Probably the model could be further improved by increasing the size of the training set, which would also add more negative examples (enterprises without a website). Also manual checks would further improve the quality of the training set.
- The method described here does only involve search via a search engine, not any additional scraping of the website itself. Adding this scraping step after the search step could further improve the results.
ESSnet Big Data WP2: Webscraping Enterprise Characteristics

Methodological note

The ESSnet BD WP2 performs joint web scraping experiments in multiple countries, using as much as possible the same methodological concepts. The aim is to derive experimental statistics on enterprises from information found on the web, especially the websites of enterprises. It should be noted that these statistics have not reached maturity in terms of harmonisation, coverage or methodology. At this point they are to be treated as the output of research experiments and they do not necessarily align with the official statistics published on this subject.

Use case: URL Inventory of enterprises

Country: BG
Date: 2018-04-10
Authors: Kostadin Georgiev and Galya Stateva

Data sources
- The Statistical Business Register (SBR)
- Responses from Bulgarian ICT survey
- URL Inventory of enterprises from the SGA-I
- Search results from Jabse Search API, Google Custom Search API and Bing Search API on the base of the enterprise’s name

Population
The enterprises as defined by the ICT survey (>10 persons, limited NACE category)

Methodology

1. Methodological procedures with BNSI software:
   1. Get enterprises from the ICT population 2017 with the following data: IDs, Names, URLs, E-mails and other characteristics form the SBR in CSV file.
   2. Create MySQL database table with IDs, Names, URLs, E-mails and other characteristics fields from the CSV file of the enterprises.
   3. Upload the data from CSV file to the database table.
   4. Get Google Search API Key and Jabse Search API Key.
   5. Configure the information in the conf.php file.
   6. Add necessary database table fields according to the information in the conf.php file where the output information from the project’s software will be saved.
   7. Upload the known and verified URLs of the enterprises from the previous project SGA-I to the database table for sustainability results.
   8. Check if the known URLs from SBR have working web sites, generate URLs from the e-mails domains of the enterprises (excluding the popular e-mails domains: Yahoo, Gmail, etc. form the list in the conf.php file.) and check the generated URLs for working web sites with geturl.php script.
   9. Check the known and verified URLs of the enterprises from the previous project SGA-I for working web sites with checkoldurl.php script since the changes have been likely occurred from the previous period.
10. Run google_search.php script to get up to 10 suggested URLs of the enterprises from the results of running Google Search API with the enterprises names.

11. Run jabse_search.php script to get up to 20 suggested URLs of the enterprises from the results of running Jabse Search API with the enterprises names in Bulgarian and transliterated in Latin.

12. Make a list with enterprises IDs and a corresponding list with enterprises names from MySQL database containing enterprises’ data for the ICT survey 2017 target population.

13. Run ISTAT software UrlSearcher.jar either stands alone or through UrlSearcher.php script to get to up to 10 suggested URLs of the enterprises from the results of running Bing Search API with the enterprises names.

14. Run list.php script to make manual verification of the correct URLs of the enterprises from the checked URLs of the SBR and previous project SGA-I and the suggested URLs from the Google, Jabse and Bing Search APIs. After running this step, you have a List of enterprises with known URLs in the DB table.

15. Run the info.php script to see statistics from the above executed steps for URLs Retrieval at regional level and NACE categories.

The BNSI URLs retrieval scripts are available at https://github.com/kostadingeorgiev/bnsi_bigdata

The ISTAT software UrlSearcher is available at https://github.com/SummaIstat/UrlSearcher

II. Methodological procedures with ISTAT software:

1. Make a list with enterprises IDs and a corresponding list with enterprises names from MySQL database containing enterprises’ data for the ICT survey 2017 target population.

2. Run ISTAT software UrlSearcher.jar either stands alone or through UrlSearcher.php script to get to up to 10 suggested URLs of the enterprises from the results of running Bing Search API with the enterprises names (URLs Searching).

3. Make a negative list with domains of yellow pages sites.

4. Run RootJuice.jar with the negative domain list and the result seed.txt file from the execution of the UrlSeracher.jar program.

5. Get running Solr 4.10.4 and create a data collection.

6. Configure the data collection and run the SolrTSVImporter.jar with the result file from the RootJuice.jar program, to populate the Solr data collection (URLs crawling).

7. Configure the UrlScorer.jar parameters, crate a list with territorial units, create a list with enterprises IDs, Names and Address information, and run UrlScorer.jar program (URLs scoring).

8. Create a list with enterprises IDs, Names and known URLs, and run the UrlMatchTableGenerator.jar program using the result file from the execution of the UrlScorer.jar program (Machine learning with Custom R script) taking into account that a subset of enterprises for which the correct link is already indicated is available. Our input training dataset consists 24 268 records that had at least one page fetched. On the basis of the output scoring dataset we first associated to each enterprise the link with the highest score. As we know if the link is correct or not, a dichotomous variable correct_Yes_No says
if the URL is the right one or not: this variable plays the role of the Y variable, to be predicted by the model. Together with this information, variables indicating success or failure of the search of telephone, VAT code, municipality, province and zip code play the role of the X variables (predictors), together with the link position and coincidence of the central part of the URL with the name of the enterprise (simple URL). This initial set is split into two equal size subsets, the first acting as the proper training set to fit the model, the second as the test set used to evaluate the performance of the model. Three different models were fitted: logistic, neural network and random forest. Their performance was almost equivalent, and logistic was chosen for the interpretability of its parameters (exactly the same as ISTAT procedure, which is well described in the Methodological note of the use case: URL retrieval).

9. Apply the logistic model to the set of enterprises for which the website URL was not known.

The ISTAT software is available at https://github.com/Summalstat/.

Related figures

I. The results from URLs Retrieval (with BNSI software) procedure is:

The total size of population (enterprises with 10+ employees) of the ICT survey 2017: 27489
Number of e-mails of the enterprises in the population: 19888
Initial number of URLs of enterprises in the SBR: 1994
Verified URLs from the initial ones: 1822
Verified URLs from the e-mails: 6844
Number of searches in www.jabse.com: 18245
Number of searches in www.jabse.com Latin: 19043
Number of searches in www.google.com: 27356
Number of searches in www.bing.com (with ISTAT software): 27489
URLs of enterprises found with previous project SGA-I for the target population of the ICT survey 2017: 9024
URLs of enterprises found with the project SGA-II: 11442

The benchmark analysis was carried out between the ICT survey 2017 data and URL Retrieval data (ICT survey question: Does your enterprise have a website (Yes, No)?). The total 27489 enterprises were object to URL Retrieval procedure and 4776 of them were in the scope of the ICT survey 2017. The number of full matches (Yes/Yes and No/No results in the both sources) are 3907 or 81.80% success of the URLs retrieval procedure.

II. The results from URLs Retrieval (with ISTAT software) procedure is:

The URLMatchTableGenerator takes the results from the URLScorer and compares them with known list of enterprise’s URLs (11442 URLs). The result shows that software predicts the right URLs of 74 % of the enterprises. The results are slightly better than those from SGA-I, but there is room for improvement by adopting a better list of yellow pages and internet catalogues. Also, there were differences between the expected data fields from the software and the provided fields from the Bulgarian SBR, for example the area code and phone number of the enterprise are concatenated in Bulgarian SBR, compare to their separate use of in the ISTAT software.
Matching | Count
--- | ---
URLs that match | 8067
URLs that don’t match | 2810

The results from Analyse phase with Machine learning method are the following:

Values “0.281” and “0.488” are such that all first seven classes are defined as “non-links”, while the last two classes are defined as “links”; the eighth class is destined to manual inspection.

```plaintext
using threshold 8
URL to be taken
<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8022087</td>
<td>0.1977913</td>
</tr>
</tbody>
</table>

URL to be excluded
<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4310615</td>
<td>0.5689385</td>
</tr>
<tr>
<td>0.7667299</td>
<td>0.2332701</td>
</tr>
</tbody>
</table>
```

Accordingly to the predicted score, we decided if the found link was acceptable or not in terms of reliability.

We were able to find 898 new URLs, which were added to the already available URLs from deterministic approach and SBR, meaning that a total number of URLs is 12 340 or 44.9% of enterprises (with 10+ employees) have web-sites. Considering that the estimate (from the ICT survey) of enterprises with a website is 50.8% the obtained coverage is 88%.

<table>
<thead>
<tr>
<th>score_class</th>
<th>true</th>
<th>false</th>
<th>classification</th>
<th>error</th>
<th>group</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[0.0025, 0.0176]</td>
<td>49</td>
<td>1171</td>
<td>0.9598361</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>(0.0176, 0.0411]</td>
<td>58</td>
<td>1157</td>
<td>0.9522634</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>(0.0411, 0.0994]</td>
<td>84</td>
<td>1320</td>
<td>0.9401709</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>(0.094, 0.139]</td>
<td>108</td>
<td>1076</td>
<td>0.9087938</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>(0.139, 0.201]</td>
<td>273</td>
<td>1568</td>
<td>0.8517110</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>(0.201, 0.255]</td>
<td>127</td>
<td>361</td>
<td>0.7397541</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>(0.255, 0.281]</td>
<td>367</td>
<td>915</td>
<td>0.7137255</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>(0.281, 0.488]</td>
<td>470</td>
<td>624</td>
<td>0.5703839</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>(0.488, 0.802]</td>
<td>538</td>
<td>332</td>
<td>0.3848160</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>(0.802, 0.949]</td>
<td>1263</td>
<td>278</td>
<td>0.1804023</td>
<td>10</td>
</tr>
</tbody>
</table>

Recall: 0.8283313
Precision: 0.5237192
F1 measure: 0.6417112
The first results from usage of the machine learning techniques with ISTAT software are promising in terms of quality and efficiency.

**Limitations and future work**

- The ICT survey was carried out in the beginning of 2017, where the URLs retrieval procedures were performed in January 2018. We intend to repeat the comparison when the ICT survey 2018 results are available. It’s probably will lead to the more accurate results;
- We used SBR data, SGA-I results and three search engines for improving of the accuracy and to facilitate the manually verification of the URLs enterprises;
- The output results are better than SGA-I results and we may conclude that URLs Retrieval procedure could be used in the real statistical production and improving the quality of the SBR data.
- The ISTAT score vector is not completely relevant to the Bulgarian SBR data. The fine-tuning of the scores should be applied and the vector could be modified more precisely to the Bulgarian SBR, e.g. municipality name to be replacing with street name; telephone number format to be precise and etc.
- We could use the known URLs from SGA-I or SGA-II results in the future to retrieve the URLs only for the rest of the enterprises from the target population. It’ll decrease the staff burden and the time spent.
ESSnet Big Data WP2: Webscraping Enterprise Characteristics
Methodological note, draft

The ESSnet BD WP2 performs joint web scraping experiments in multiple countries, using as much as possible the same methodological concepts. The aim is to derive experimental statistics on enterprises from information found on the web, especially the websites of enterprises. It should be noted that these statistics have not reached maturity in terms of harmonisation, coverage or methodology. At this point they are to be treated as the output of research experiments and they do not necessarily align with the official statistics published on this subject.

Use case: URL Retrieval
Country: PL
Date: 2018-03-13
Authors: Jacek Maślankowski and Joanna Wyzina

Data sources
- The Business Register (BJS) by Statistics Poland (about 3.3 M enterprises), of which 189 thous. has a known URL; for enterprises having 10 and more employees – total number is 165 thous., of which 47 thous. has URLs (28.4%),
- Websites from enterprises via Istat URL Retrieval software,
- Education Information System (System Informacji Oświatowej) by the Ministry of Education – having population of education institutions.

Population
The enterprises in Pomeranian voivodship (Poland) with 10 and more employees having URL in business register – 2961

Methodology
1. Extraction of a set of 2961 enterprises with URL from the GBR >= 10 persons, any NACE code.
2. Istat URL Retrieval software was used to find first 10 links based on Bing search engine.
3. Comparison of URLs in business register and the ones received by executing Istat URL Retrieval software – we selected educational institutions for this purposes.

Results
Our analysis to compare the results of analysis of the use of Istat URL Retrieval software and data from business register (BJS) shows that ca. 74% of hyperlinks stored in BJS were the same as identified in the first 10 results from Bing search engine.

We decided to make deep analysis of the links from 199 enterprises providing educational services. The results from BJS and Istat URL Retrieval software were stored in Excel file, as shown in the figure below.
The figure above shows that the same link as in the BJS usually were not on the first position on the Bing search engine result list. The link that redirected to the school was on the 7th, 2nd and 2nd position, according to the list above. In two presented cases, the link was different than the one stored in the BJS. It means that the Istat URL Retrieval software may be used to check if the URL was changed.

The use of detailed information on enterprise (e.g., full address + business id number) may increase the accuracy but it also resulted in decreasing the number of enterprises found via Bing API. For instance, 704 from 2961 enterprises were found by Istat URL Retrieval software. It is the result of the fact that full information about enterprise is usually not present on the website.

**Conclusion**

Comparison of URLs based on Istat URL Retrieval software and stored in business register provides the following conclusions:

1) The most frequent links refer to:
   - Social media – ca. 40% units were successfully found on Facebook
   - Maps – 49% were found on Targeo, Zumi, Wikimapia
   - Governance registers – 37% units were identified in KRS (Krajowy Rejestr Sądowy – National Court Register)
   - News in press
   - Regional portals
   - Trade portals
   - Other portals

2) Our suggestions to increase the accuracy in Polish conditions (regarding business register and website URLs) are as follow:
   - Add additional attributes, such as:
     - REGON (ID)
     - KRS number
     - Postal code
Limitations and future work

- The limitation of this work is that lots of microenterprises do not have a website. Some of them have the same name (e.g., self-employed persons) what decrease the accuracy.
- We would like to use URL retrieval to monitor the sustainability of the URLs – whether the URL identified by this software is the same as the one in business register.
Use Case 2: E-Commerce (Web Ordering)

ESSnet Big Data WP2: Webscraping Enterprise Characteristics

Methodological note

The ESSnet BD WP2 performs joint web scraping experiments in multiple countries, using as much as possible the same methodological concepts. The aim is to derive experimental statistics on enterprises from information found on the web, especially the websites of enterprises. It should be noted that these statistics have not reached maturity in terms of harmonisation, coverage or methodology. At this point they are to be treated as the output of research experiments and they do not necessarily align with the official statistics published on this subject.

Use case: E-commerce (Web-ordering)

Country: Italy
Date: 2018-03-05
Authors:

Data sources

- The Italian Business Register (ASIA)
- Responses from 2017 Italian ICT survey (responses from about 21,000 enterprises) to the question: “Online ordering or reservation or booking, e.g. shopping cart (Yes/No)”
- Websites from enterprises (2017)

Population

The 184,000 enterprises as defined by the ICT survey (>10 persons, limited NACE)

Methodology

The overall procedure is based on the following steps:
1. get the websites address (Uniform Resource Locator) potentially for all enterprises included in the population of reference (URL retrieval);
2. access websites with available URL and scrape their content (web scraping);
3. process the content of the scraped websites in order to identify the best predictors for the target variables (text mining);
4. fit models (machine learning) in the subset of enterprises where both Internet data and survey data were available (considering survey data as the true values) and predict the values of target variables for all the enterprises for which the retrieval and scraping of their websites was successful;
5. apply the best predictor to all units for which steps 1 and 2 were successful;
6. produce estimates applying different estimators (full based model approach and combined approach);
7. compare to current ICT survey design based estimates.

Results

Results of steps from 1 to 3 are reported in the following figure:
In step 4 a number of different machine learning models have been applied: logistic model, neural networks, classification trees, naïve bayes, support vector machine, bagging, boosting, random forest. In order to handle non response in training set, it has been expanded by using calibrated weights obtained by using known totals in U1 population. The best resulted to be random forest, with the following performance

<table>
<thead>
<tr>
<th>Confusion Matrix and Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>observTest</td>
</tr>
<tr>
<td>predictedTest</td>
</tr>
<tr>
<td>0  4027  493</td>
</tr>
<tr>
<td>1  491  887</td>
</tr>
<tr>
<td>Accuracy : 0.7320</td>
</tr>
<tr>
<td>95% CI : (0.6225, 0.8418)</td>
</tr>
<tr>
<td>No Information Rate : 0.7699</td>
</tr>
<tr>
<td>F-Value [Acc &gt; NIR] : &lt;2e-18</td>
</tr>
<tr>
<td>Kappa : 0.5264</td>
</tr>
<tr>
<td>McNemar’s Test F-Value : 0.8746</td>
</tr>
<tr>
<td>Sensitivity : 0.6398</td>
</tr>
<tr>
<td>Specificity : 0.8913</td>
</tr>
<tr>
<td>Pos Pred Value : 0.6358</td>
</tr>
<tr>
<td>Neg Pred Value : 0.8909</td>
</tr>
<tr>
<td>Prevalence : 0.2301</td>
</tr>
<tr>
<td>Detection Rate : 0.1460</td>
</tr>
<tr>
<td>Detection Prevalence : 0.2297</td>
</tr>
<tr>
<td>Balanced Accuracy : 0.7691</td>
</tr>
<tr>
<td>‘Positive’ Class : 1</td>
</tr>
<tr>
<td>F1 measure : 0.6352854</td>
</tr>
</tbody>
</table>
It is important to remark that in 2016 round of the ICT survey for each case in which a conflict between reported value from the survey and predicted value from random forest was observed, a manual inspection of the corresponding website was carried out in order to decide the correct value. In about 50% of cases the right value was the predicted one. Once eliminated response errors in training set the performance of the random forest predictor resulted to be:

![Confusion Matrix and Statistics](image)

that is, much more than when not considering response errors.

In step 6 the following estimators have been applied:

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Formula</th>
<th>Weights</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design based/</td>
<td>$\hat{P} = \sum_{k} y_k w_k$</td>
<td>$\sum_{k=1}^{K} w_k = N_0$</td>
<td>$w_k$ weights are obtained by calibration procedure of basic weights (inverse of inclusion probabilities) making use of known totals in the population in order to reduce the bias due to non-response and the variability due to sampling errors</td>
</tr>
<tr>
<td>model assisted</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model based</td>
<td>$\hat{P} = \sum_{k} \hat{y}_k w_k$</td>
<td>$\sum_{k=1}^{K} w_k = N_0$</td>
<td>The estimate of the total number of enterprises offering web ordering facilities on their websites is given by the count of the predicted values $\hat{y}_k$ for all units for which it was possible to reach their websites (population $U_1$), calibrated in order to make them representative of all the population having websites ($U$).</td>
</tr>
</tbody>
</table>
| Combined             | $\hat{P} = \sum_{(k'-k) \in\pi} \hat{y}_k + \sum_{k} (\hat{y}_k - y_k) w_k = \sum_{(k'-k) \in\pi} \hat{y}_k w_k''$ | $\sum_{k=1}^{K} w_k = N_{0''}$ | Estimates are produced by summing these components:
3. the counting of observed values for the $\pi$ respondents that declared a website, that was not found nor stayed, expanded to the whole subpopulation $U'' - U''$. |
|                      |                                                                         |         |                                                                             |
As for step 7, estimates have been produced for different domains. In the following plot we report a comparison of estimates of web-ordering rate for enterprises cross-classified by macro-sector of economic activity (1-4) and class of employees (1-4):

The dashed lines define the area delimitated by the lower and upper limits of the confidence intervals calculated in correspondence of each design based estimate.

**Conclusions**

The analysis of the estimates related to web-ordering, obtained with the two alternative estimators, compared to the estimates produced by the official survey, allows some preliminary conclusions.

The three different sets are not incoherent. For instance, considering web-ordering the estimates for the total are well inside the confidence interval of the survey estimate, and this is the same for many values in the different domains.

Looking at coherence as one important dimension of quality, both combined estimates and full model based estimates can be considered as equally acceptable.

However, they cannot be considered as equivalent, as each has some specificities that is important to remark.

The main flaws of the model based estimator are in the presence of prediction errors and in the undercoverage of the population of enterprises owning websites, part of which has not been reached by
web scraping. As for the first, taking into consideration the presence of response errors in the test set, once eliminating them by manual inspection, the accuracy of the model predictions increases to more than acceptable levels (around 90% for web ordering), in any case comparable with the accuracy of survey data. As for the second, pseudo-calibration allow to limit the bias, especially when the difference in the values of the parameters in the two sub-populations is not high, as it is the case.

As for the combined estimator, its second component (aiming at compensate for prediction errors) is based on an assumption of perfect correctness of reported values, and considers predicted values as errors when they do not coincide with the reported ones. But controls have been carried out when fitting models, and in half of the cases in which predicted values were contradictory with reported ones, this was not due to model fault, but to response errors. So, this assumption does not always hold.
ESSnet Big Data WP2: Webscraping Enterprise Characteristics
Methodological note - Test Statistic Produced: Yes

The ESSnet BD WP2 performs joint web scraping experiments in multiple countries, using as much as possible the same methodological concepts. The aim is to derive experimental statistics on enterprises from information found on the web, especially the websites of enterprises. It should be noted that these statistics have not reached maturity in terms of harmonisation, coverage or methodology. At this point they are to be treated as the output of research experiments and they do not necessarily align with the official statistics published on this subject.

Use case: E-commerce

Country: UK
Date: 2018-04-04

Data sources
Online Registers, Enterprise Websites, ICT survey (called the ‘E-commerce Survey’ in the UK)

Population
The enterprises as defined by the ICT survey (called the ‘E-commerce Survey’ in the UK) (>10 persons, limited NACE)

Methodology
7. Extraction of a set of companies in the reference population with known websites and ecommerce status

8. Scrape website for each company and store the scraped text (note – all scraping carried out according to ONS web-scraping policy)

9. Extract features from the scraped text based on different text representation models:
   a. TF-IDF (term frequency – inverse document frequency)
   b. Doc2Vec
   c. ‘Engineered features’ – the presence/non-presence of certain e-commerce related terms

10. For each text representation model, train three different machine learning classifiers to predict whether the website is engaged in e-commerce:
    a. Support Vector Classifier (SVC) with polynomial kernel
    b. Naïve Bayes
    c. Random Forests

11. Evaluate the performance of different text representation models and classifiers

12. Compare to estimates from deterministic methods; searching for e-commerce keywords/terms or technologies on a website

Results
Estimating e-commerce engagement using classifiers
The table below summarises the performance of each text representation model/classifier combination:
The confusion matrix below can be used to evaluate the performance of the best performing model/classifier combination (engineered features & random forests) -

<table>
<thead>
<tr>
<th>Text Representation Model</th>
<th>Classifier</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF</td>
<td>Naive Bayes</td>
<td>65 %</td>
</tr>
<tr>
<td>TF-IDF</td>
<td>SVC (polynomial)</td>
<td>71 % ALL FALSE</td>
</tr>
<tr>
<td>TF-IDF</td>
<td>Random Forest</td>
<td>75 % (Out-Of-Bag)</td>
</tr>
<tr>
<td>Engineered Features</td>
<td>Naive Bayes</td>
<td>72 %</td>
</tr>
<tr>
<td>Engineered Features</td>
<td>SVC (polynomial)</td>
<td>70 % ALL FALSE</td>
</tr>
<tr>
<td>Engineered Features</td>
<td>Random Forest</td>
<td>79 % (Out-Of-Bag)</td>
</tr>
<tr>
<td>Doc2Vec</td>
<td>Naive Bayes</td>
<td>70 %</td>
</tr>
<tr>
<td>Doc2Vec</td>
<td>SVC (polynomial)</td>
<td>74 %</td>
</tr>
<tr>
<td>Doc2Vec</td>
<td>Random Forest</td>
<td>72 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>N. Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No web sales</strong></td>
<td>0.80</td>
<td>0.94</td>
<td>0.86</td>
<td>777</td>
</tr>
<tr>
<td><strong>Does web sales</strong></td>
<td>0.74</td>
<td>0.43</td>
<td>0.54</td>
<td>323</td>
</tr>
<tr>
<td><strong>Avg/total</strong></td>
<td>0.78</td>
<td>0.79</td>
<td>0.77</td>
<td>1100</td>
</tr>
</tbody>
</table>

The classifier has relatively low precision, many websites identified as doing e-commerce websites are in fact non-e-commerce, and very low recall, many e-commerce enterprises are not identified as such by the classifier. The classifier would determine that 21 % of enterprises are engaged in e-commerce.

Examining the website data itself, it appears there is a representativeness problem. With large (especially international) enterprises that own multiple domains, the website retrieved by the URL retrieval process may not be the one through which an enterprise conducts e-commerce. On the other hand, a number of smaller enterprises appear to have simply answered the original survey incorrectly. The classifiers also suffer for being trained on a relatively small dataset. 1100 is not a large sample size for NLP-type problems. This probably accounts for the higher performance of classifiers using engineered features, which fundamentally include words that are also included in the language (BOW-type) text representations.
Estimating e-commerce engagement using website features

An estimate of websites that engage in ecommerce is created by searching human readable text and underlying html code for keywords related to web-sales or supporting technologies (e.g. “visa” and “mastercard”, or “woocommerce” – a company that makes shopping websites for other enterprises). Using a comprehensive list of words that are only present if web-sales are being made produces an estimate of 30% of enterprise websites making web sales. Note this figure is not adjusted for bias in retrievals or enterprise category/size. The survey yields a naïve, unweighted estimate of 21% of enterprises making sales through e-commerce. The figure below compares the survey answers to whether they engage in web sales with the prediction based upon whether diagnostic words are found on their website (for sample with linked enterprise and website only).

Limitations and future work

Our best-performing classifier has good accuracy but poor predictive performance in practice, and produces estimates which are quite far away from the survey estimate. The performance would be significantly improved by increasing the number and quality of samples for training, manually constructing a set through web research is feasible if laborious for a required sample size of ~2000. Such a dataset would also allow for expanding the deterministic estimation method, and gauging its performance metrics as a classifier.

If these methods were to be scaled up to production, the largest issue would be that e-commerce identification is biased with respect to company size, as with the UK URL identification use-case. These systematic biases would need to be accounted for by stratification and weighting of the data, possibly using the e-Commerce survey answers.
ESSnet Big Data WP2: Webscraping Enterprise Characteristics

Methodological note

The ESSnet BD WP2 performs joint web scraping experiments in multiple countries, using as much as possible the same methodological concepts. The aim is to derive experimental statistics on enterprises from information found on the web, especially the websites of enterprises. It should be noted that these statistics have not reached maturity in terms of harmonisation, coverage or methodology. At this point they are to be treated as the output of research experiments and they do not necessarily align with the official statistics published on this subject.

Use case: Ecommerce

Country: NL
Date: 2018-03-28
Authors: Olav ten Bosch and Dick Windmeijer

Data sources
- The General Business Register (GBR) (about 1.5 M enterprises), of which 1/3 have a known URL
- Responses from Dutch ICT survey (responses from about 9 000 enterprises)
- Websites from enterprises via dedicated scraping software

Population
The enterprises as defined by the ICT survey (>10 persons, limited NACE)

Methodology

1. Determination of a set of enterprises from the GBR that responded to the 2017 ICT survey (8909).
2. URL retrieval (see other use case) on the subset of enterprises in this set not having a known URL. Creation of a smaller set for which we have a URL either from our GBR or from URL retrieval (7222).
3. Ecommerce detection using the deterministic web shop detection script based on keywords developed earlier at Statistics Netherlands.
4. Ecommerce detection using the keywords from the script in step 3 but now applying a ML approach to predict Ecommerce based on the answers in the ICT use survey instead of counting words.
5. For the methods in step 3 and 4 we compared the results from the most relevant answers from the ICT survey with the prediction on Ecommerce based on the web. The questions from the ICT survey taken into consideration are:
   - Does your website contain:
     - a possibility to order online, to make a booking or to make a reservation? (ORDER)
     - a mechanism that a client can use to design / adapt its product (goods or services)? (PRODUCT)
     - information about the status of its order? (TRACK)

The frequencies of these answers in the ICT survey are:

<table>
<thead>
<tr>
<th></th>
<th>ORDER</th>
<th>PRODUCT</th>
<th>TRACK</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>6000</td>
<td>8071</td>
<td>7789</td>
</tr>
<tr>
<td>Yes</td>
<td>2909</td>
<td>838</td>
<td>1120</td>
</tr>
</tbody>
</table>
Results
We describe the confusion matrix and some common performance metrics per case\(^2\).

Prediction using deterministic web shop detection script
In an earlier project we developed a script to determine Ecommerce activities from websites of foreign companies paying tax in the Netherlands. This script use the following deterministic approach for detecting webshops:
- Count the occurrences of the following words on the homepage of a website: *winkel, shop, cart, wagen, bag, mand, basket, warenkorb, klant*\(^3\)
- a site has a web shop if the sum of these counts is equal to or greater than 10.

<table>
<thead>
<tr>
<th>Ecommerce Deterministic</th>
<th>Predict</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>Yes</td>
</tr>
<tr>
<td>Yes</td>
<td>655</td>
</tr>
<tr>
<td>No</td>
<td>383</td>
</tr>
<tr>
<td>Total</td>
<td>1038</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.63</td>
<td>0.21</td>
<td>0.32</td>
<td>3116</td>
</tr>
<tr>
<td>No</td>
<td>0.69</td>
<td>0.93</td>
<td>0.79</td>
<td>5793</td>
</tr>
<tr>
<td>avg / total</td>
<td>0.67</td>
<td>0.68</td>
<td>0.63</td>
<td>8909</td>
</tr>
</tbody>
</table>

Accuracy: 0.68
Balanced accuracy: 0.57
Prediction Ecommerce Deterministic: 12 %

---


\(^3\) Since this vocabulary was designed to detect Ecommerce activities from foreign enterprises it contains a mixture of languages.
Prediction using Random Forest

<table>
<thead>
<tr>
<th>Set</th>
<th>number of records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>5612</td>
</tr>
<tr>
<td>Test</td>
<td>2406</td>
</tr>
<tr>
<td>Validation</td>
<td>891</td>
</tr>
<tr>
<td>Total</td>
<td>8909</td>
</tr>
</tbody>
</table>

Features: scaled wordcount on homepage of the words **winkel, shop, cart, wagen, bag, mand, basket, warenkorb, klant**

We report on Random Forest only as this appeared to yield the best results. The results for the test and validation set are very comparable. We report on the validation set only.

**Validation set**

Ecommerce Random Forest (validation set)

<table>
<thead>
<tr>
<th>Predict</th>
<th>True</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>76</td>
<td>235</td>
<td>311</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>49</td>
<td>531</td>
<td>580</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>125</td>
<td>766</td>
<td>891</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.61</td>
<td>0.24</td>
<td>0.35</td>
</tr>
<tr>
<td>No</td>
<td>0.69</td>
<td>0.92</td>
<td>0.79</td>
</tr>
<tr>
<td>avg / total</td>
<td>0.66</td>
<td>0.68</td>
<td>0.64</td>
</tr>
</tbody>
</table>

Accuracy: 0.68

Balanced accuracy: 0.58

Prediction Ecommerce Random Forest (validation set): 14 %
Comparison with published figures
The figures were published on the level of individual questions, where the predictions were done on a combination of questions. The table below summarizes the results.

<table>
<thead>
<tr>
<th></th>
<th>ORDER</th>
<th>PRODUCT</th>
<th>TRACK</th>
<th>ORDER+PRODUCT+TRACK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Published</td>
<td>33 %</td>
<td>7 %</td>
<td>10 %</td>
<td></td>
</tr>
<tr>
<td>Prediction</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>deterministic</td>
<td></td>
<td></td>
<td></td>
<td>12 %</td>
</tr>
<tr>
<td>Prediction</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random Forest</td>
<td></td>
<td></td>
<td></td>
<td>14 %</td>
</tr>
</tbody>
</table>

Conclusions
- According to the performance metrics, detection of Ecommerce activities in general can be done reasonably effective via the methods described in this methodological note.
- It makes not much of a difference to use the deterministic method, designed earlier, or the random forest on the same vocabulary.

Limitations and future work
- The datasets were created at different points in time; the ICT survey was executed beginning 2017, the URL retrieval was carried out January 2018. This could explain some differences.
- One has to keep in mind that not only the web scraper could be wrong, an answer on the ICT survey could be wrong as well.
- Instead of using the vocabulary of multi-language words that was designed earlier and applied in this project, it was suggested to add another Ecommerce detection features, based on the use of pictures of shopping baskets. We leave this to future work.
- It is difficult to compare de prediction of Ecommerce in general with the more detailed results being published on the level of individual questions. We leave it to future work to use one question in the machine learning, most probable ORDER would be the best option.
ESSnet Big Data WP2: Webscraping Enterprise Characteristics

Methodological note

The ESSnet BD WP2 performs joint web scraping experiments in multiple countries, using as much as possible the same methodological concepts. The aim is to derive experimental statistics on enterprises from information found on the web, especially the websites of enterprises. It should be noted that these statistics have not reached maturity in terms of harmonisation, coverage or methodology. At this point they are to be treated as the output of research experiments and they do not necessarily align with the official statistics published on this subject.

Use case: E-commerce

Country: BG
Date: 2018-02-06
Authors: Kostadin Georgiev and Galya Stateva

Data sources
- The Statistical Business Register (SBR)
- Responses from Bulgarian ICT survey
- URL Inventory of enterprises
- Search results from Bing Search API on the enterprises’ names (as defined by the ICT survey, >10 persons, limited NACE category)

Population
- The List of enterprises with known websites found with URLs Retrieval use-case (11442)
- Number of searches in www.bing.com (with ISTAT software): 27489

Methodology

III. Methodological procedures with BNSI software:


17. Add necessary database table fields according to the information in the conf.php file.

18. Run scrap.v3.php script to get up to 3 prediction for e-commerce URLs of the enterprises based on scraped data from the enterprises web sites.

19. Run list_estore.php script to choose manually the correct e-commerce URLs of the enterprises.

20. Run list_estore_lost.php script to find missed e-commerce URLs with 10% sample extracted from enterprises we don’t predict e-commerce websites with scrap.v3.php script.

21. Run the info.php script to see statistics from the above executed steps for E-commerce URLs at regional level and NACE categories.

The BNSI E-commerce URLs scripts are available at https://github.com/kostadingeorgiev/bnsi_bigdata

IV. Methodological procedures with ISTAT software:

1. Make a list with enterprises IDs and URLs from MySQL database containing enterprises with known websites found with URLs Retrieval use-case (11442).

2. Run RootJuice.jar with the list with enterprises IDs and URLs.

3. Get running Solr 4.10.4 and create a data collection.
4. Configure the data collection and run the SolrTSVImporter.jar with the result file from the RootJuice.jar program, to populate the Solr data collection.

5. Create a positive and negative list for e-commerce key words.

6. Run FirmDocTermMatrixGenerator.jar with created positive and negative lists for the Solr data collection.

7. Analyze the results obtained from the execution of the UrlMatchTableGenerator.jar with Custom R script.

The ISTAT software is available at https://github.com/SummaIstat/.

Related figures

III. The results from URLs E-commerce (with BNSI software) procedure is:

Predicted e-stores of enterprises with script 1: 2148
Predicted e-stores of enterprises with script 2: 1809
Predicted e-stores of enterprises with script 3: 1279
E-stores of enterprises found with the project: 1106
Missed e-stores of enterprises in 10% sample: 59

**Expected number of e-stores of enterprises: 1106+59*10^-5=1696**

The benchmark analysis was carried out between the ICT survey 2017 data and URLs E-commerce (ICT survey question: Enterprises which sold via a website or apps - via their own website or apps (Yes, No)?). The total 11442 enterprises with known URLs were object of the URLs E-commerce procedure and 4776 of them were in the scope of the ICT survey 2017. The number of full matches (Yes/Yes and No/No results in the both sources) are 4386 or 91.83% success of the URLs E-commerce procedure.

IV. The results from URLs E-commerce (with ISTAT software) procedure is:

The obtained result after execution of the UrlMatchTableGenerator.jar is a matrix with 10022 rows of enterprises and 65 columns of words for e-commerce. This matrix is a starting point for R script analysis.

Limitations and future work

- The ICT survey was carried out in the beginning of 2017, where the URLs E-commerce procedure was performed in February 2018. We intend to repeat the comparison when the ICT survey 2018 results for e-commerce are available. It’s probably will lead to the more accurate results;

- The output results are better than SGA-I results and we may conclude that URLs E-commerce procedure could be used in the real statistical production and improving the quality of the ICT survey, including could be integrated to the ICT methodology;

- We are just in the beginning of the analysis with R script and BNSI staff isn’t so experienced with R. Also, the analysis phase with R script isn’t well documented and hasn’t clear methodological guidelines yet;

- Compared to SGA-I, we use more extended keyword lists. Script predicts more e-commerces, but at the same time, number of missed e-commerces in a 10% sample are twice as many as the previous project. There is a room for improving the prediction script and refining the key words lists;
There were some problems with running the FirmDocTermMatrixGenerator.jar caused by lack of support for Bulgarian language in some of the components used. Particularly, words which are lemmatized with TreeTagger and stemmed with SnowballStemmer.
Use Case 3: Job Advertisement

ESSnet Big Data WP2: Webscraping Enterprise Characteristics

Methodological note

The ESSnet BD WP2 performs joint web scraping experiments in multiple countries, using as much as possible the same methodological concepts. The aim is to derive experimental statistics on enterprises from information found on the web, especially the websites of enterprises. It should be noted that these statistics have not reached maturity in terms of harmonisation, coverage or methodology. At this point they are to be treated as the output of research experiments and they do not necessarily align with the official statistics published on this subject.

Use case: Job advertisements

Country: Italy
Date: 2018-03-05
Authors:

Data sources

- The Italian Business Register (ASIA)
- Responses from 2017 Italian ICT survey (responses from about 21,000 enterprises) to the question: “Online job advertisements (Yes/No)”
- Websites from enterprises (2017)

Population

The 184,000 enterprises as defined by the ICT survey (>10 persons, limited NACE)

Methodology

The overall procedure is based on the following steps:

8. get the websites address (Uniform Resource Locator) potentially for all enterprises included in the population of reference (URL retrieval);
9. access websites with available URL and scrape their content (web scraping);
10. process the content of the scraped websites in order to identify the best predictors for the target variables (text mining);
11. fit models (machine learning) in the subset of enterprises where both Internet data and survey data were available (considering survey data as the true values) and predict the values of target variables for all the enterprises for which the retrieval and scraping of their websites was successful;
12. apply the best predictor to all units for which steps 1 and 2 were successful;
13. produce estimates applying different estimators (full based model approach and combined approach);
14. compare to current ICT survey design based estimates.

Results

Results of steps from 1 to 3 are reported in the following figure:
In step 4 a number of different machine learning models have been applied: logistic model, neural networks, classification trees, naïve bayes, support vector machine, bagging, boosting, random forest. In order to handle non response in training set, it has been expanded by using calibrated weights obtained by using known totals in U1 population. The best resulted to be random forest, with the following performance:

<table>
<thead>
<tr>
<th>Confusion Matrix and Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>observTest</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>Accuracy: 0.8179</td>
</tr>
<tr>
<td>95% CI: (0.8078, 0.8277)</td>
</tr>
<tr>
<td>No Information Rate: 0.6938</td>
</tr>
<tr>
<td>P-Value [ACC &gt; NIR]: &lt;2e-16</td>
</tr>
<tr>
<td>Kappa: 0.5729</td>
</tr>
<tr>
<td>Mcnemar's Test P-Value: 0.5205</td>
</tr>
<tr>
<td>Sensitivity: 0.7088</td>
</tr>
<tr>
<td>Specificity: 0.8661</td>
</tr>
<tr>
<td>Pos Pred Value: 0.7002</td>
</tr>
<tr>
<td>Neg Pred Value: 0.8708</td>
</tr>
<tr>
<td>Prevalence: 0.3062</td>
</tr>
<tr>
<td>Detection Rate: 0.2170</td>
</tr>
<tr>
<td>Detection Prevalence: 0.3099</td>
</tr>
<tr>
<td>Balanced Accuracy: 0.7874</td>
</tr>
<tr>
<td>'Positive' Class: 1</td>
</tr>
</tbody>
</table>

F1 measure 0.7044826
As for the real performance when considering response errors in the survey, the same considerations reported for the use case related to e-commerce are valid.

In step 6 the following estimators have been applied:

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Formula</th>
<th>Weighting</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design based/ model assisted</td>
<td>( y = \sum_{k=1}^{n} y_k w_k )</td>
<td>( \sum_{k=1}^{n} w_k = N_0 )</td>
<td>( w_k ) weights are obtained by calibration procedures of basic weights (inverse of inclusion probabilities) making sure of known totals in the population in order to reduce the bias due to non-response and other variability due to sampling errors.</td>
</tr>
<tr>
<td>Model based</td>
<td>( \hat{y} = \sum_{k=1}^{n} \hat{y}_k w'_k )</td>
<td>( \sum_{k=1}^{n} w'<em>k = N</em>{0}^{(2)} )</td>
<td>The estimate of the total number of enterprises offering web ordering facilities on their websites is given by the count of the predicted values ( \hat{y}_k ) for all units for which it was possible to visit their websites (population ( U_2 )), calibrated in order to make them representative of all the population having websites (( U_1 )).</td>
</tr>
</tbody>
</table>
| Combined         | \( \hat{y} = \sum_{k=1}^{n} \hat{y}_k w''_k \) | \( \sum_{k=1}^{n} w''_k = N_{0}^{(3)} \) and \( \sum_{k=1}^{n} w'''_k = N_{0}^{(4) - (3)} \) | Estimates are produced by summing these components:
  1. the counting of predicted values in the subpopulation \( U_2 \) of units for which it was possible to scrape and process corresponding websites;
  2. an adjustment based on the consideration of the difference between the \( n'' \) reported values and the predicted values (expanded to the same subpopulation \( U_2 \));
  3. the counting of observed values for the \( n''' \) respondents that declared a website that was not found is scrapped, expanded to the whole subpopulation \( U_2 - U_3 \). |

As for step 7, estimates have been produced for different domains. In the following plot we report a comparison of estimates of job advertisements rate for enterprises cross-classified by macro-sector of economic activity (1-4) and class of employees (1-4):

[Image of a chart comparing job advertisements by 4 groups NACE and 4 classes of employees]
The dashed lines define the area delimitated by the lower and upper limits of the confidence intervals calculated in correspondence of each design based estimate.

**Conclusions**

Same as for e-commerce use case.
ESSnet Big Data WP2: Webscraping Enterprise Characteristics

Methodological note

The ESSne
St BD WP2 performs joint web scraping experiments in multiple countries, using as much as possible the same methodological concepts. The aim is to derive experimental statistics on enterprises from information found on the web, especially the websites of enterprises. It should be noted that these statistics have not reached maturity in terms of harmonisation, coverage or methodology. At this point they are to be treated as the output of research experiments and they do not necessarily align with the official statistics published on this subject.

The result in this note is not published in the experimental statistics output, because of the data set used is not comparable to the population in the official statistics.

Use case: Job Advertisements
Country: SE
Date: <2018-03-28>

Data sources
- URLs extracted from other datasets, 700 unique URLs of the public sector.
- A test dataset i.e., about 800 pages scraped from Internet; mixed with 800 randomly chosen Job advertisements from a dataset that contain ~4 million advertisements.
- A training dataset including text of 6 508 advertisements and 8 415 non-advertisements

Population
The population is the test dataset of roughly 1600 pages, only from the public sector.

Method
1. Scraping websites according to URLs: a) starting from each URL, the links on each page are retrieved and tested by a pattern; if the address or the text of the link contain the pattern, in this case, “arbetta” or “jobb” (“work” or “job” in English), the scraper will go deeper along the qualified links and check if the link on the following pages contain the correct pattern; b) the qualified links are saved; c) the scraper scrapes the websites according to the qualified links and saves the page content.
2. Data cleaning: a) the websites’ content extraction, by python module BeautifulSoup; b) pre-processing the text: remove stop-words in Swedish and English, delete punctuations, remove numbers, Swedish word stemming, by python package nltk.
3. Feature selection: a) apply tfidf(term frequency inverse document frequency) on the text corpus cleaned, one variation is choosing 1 000 of the most common words from the entire corpus, i.e. of advertisements and non-advertisements together; the other variation is choosing 1 000 of the most common words from the advertisements corpus only, excluding the common words of non-advertisements; b) use ngram_range = (1,1), i.e. only single words are considered; and ngram_range=(1, 2), which considers both single words and two-word phrases.
4. Model testing (deterministic/ ML): SVC, DecisionTree, NavieBayes and Keras are tested first on the training data. The trained model is applied on the testing data to evaluate the results, how good the models are when applied on an entirely new dataset. The accuracy score is used as the metric for choosing the best parameters for models.
5. Use cross evaluation method to generate the relative stable metrics scores i.e. precision, recall and accuracy. We also set the testing data totally unseen for the classifier to test how well the classifier perform on the unseen data.

Results
1. On the training dataset, the 1000 most common words of total corpus. On the training dataset, cross validation method is applied to generate a relative stable result. When the classifier is applied on the testing dataset, the accuracy dropped heavily, which indicates the generalization problem of the classifier.

   ngram_range=(1,2)

<table>
<thead>
<tr>
<th></th>
<th>Precision(2*std)</th>
<th>Recall(2*std)</th>
<th>Accuracy(2*std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVC</td>
<td>0.98(+/-0.008)</td>
<td>0.98(+/-0.01)</td>
<td>0.97(+/-0.03)</td>
</tr>
<tr>
<td>NB</td>
<td>0.89 (+/-0.03)</td>
<td>0.96(+/-0.01)</td>
<td>0.93(+/-0.01)</td>
</tr>
<tr>
<td>DecisionTree</td>
<td>0.93(+/-0.04)</td>
<td>0.92(+/-0.02)</td>
<td>0.93(+/-0.07)</td>
</tr>
<tr>
<td>KerasSequencialModel</td>
<td>0.98(+/-0.007)</td>
<td>0.98(+/-0.003)</td>
<td>0.98(+/-0.03)</td>
</tr>
</tbody>
</table>

On the testing dataset

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVC</td>
<td>1.00</td>
<td>0.34</td>
<td>0.61</td>
</tr>
<tr>
<td>NB</td>
<td>0.96</td>
<td>0.45</td>
<td>0.73</td>
</tr>
<tr>
<td>DecisionTree</td>
<td>0.98</td>
<td>0.35</td>
<td>0.62</td>
</tr>
<tr>
<td>KerasSequencialModel</td>
<td>1.00</td>
<td>0.33</td>
<td>0.60</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Precision(2*std)</th>
<th>Recall(2*std)</th>
<th>Accuracy(2*std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVC</td>
<td>0.98(+/-0.001)</td>
<td>0.98(+/-0.001)</td>
<td>0.97(+/-0.03)</td>
</tr>
<tr>
<td>NB</td>
<td>0.88(+/-0.04)</td>
<td>0.96(+/-0.005)</td>
<td>0.89(+/-0.08)</td>
</tr>
<tr>
<td>DecisionTree</td>
<td>0.93(+/-0.004)</td>
<td>0.94(+/-0.01)</td>
<td>0.94(+/-0.02)</td>
</tr>
<tr>
<td>KerasSequencialModel</td>
<td>0.98(+/-0.004)</td>
<td>0.98(+/-0.003)</td>
<td>0.98(+/-0.003)</td>
</tr>
</tbody>
</table>

On the testing dataset

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVC</td>
<td>0.98</td>
<td>0.20</td>
<td>0.43</td>
</tr>
<tr>
<td>NB</td>
<td>0.93</td>
<td>0.35</td>
<td>0.59</td>
</tr>
<tr>
<td>DecisionTree</td>
<td>0.93</td>
<td>0.48</td>
<td>0.76</td>
</tr>
<tr>
<td>KerasSequencialModel</td>
<td>0.98</td>
<td>0.22</td>
<td>0.45</td>
</tr>
</tbody>
</table>

2. On the training dataset, using the 700 words from the diff-set of the advertisements and non-adsvertisements common words

   ngram_range=(1,2)

<table>
<thead>
<tr>
<th></th>
<th>Precision(2*std)</th>
<th>Recall(2*std)</th>
<th>Accuracy(2*std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVC</td>
<td>0.92(+/-0.02)</td>
<td>0.89(+/-0.02)</td>
<td>0.91(+/-0.07)</td>
</tr>
<tr>
<td>NB</td>
<td>0.84(+/-0.06)</td>
<td>0.82(+/-0.07)</td>
<td>0.84(+/-0.07)</td>
</tr>
<tr>
<td>DecisionTree</td>
<td>0.87(+/-0.01)</td>
<td>0.88(+/-0.02)</td>
<td>0.89(+/-0.03)</td>
</tr>
<tr>
<td>KerasSequencialModel</td>
<td>0.93(+/-0.01)</td>
<td>0.89(+/-0.01)</td>
<td>0.92(+/-0.002)</td>
</tr>
</tbody>
</table>

On the testing dataset

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVC</td>
<td>0.84</td>
<td>0.32</td>
<td>0.37</td>
</tr>
<tr>
<td>NB</td>
<td>0.86</td>
<td>0.53</td>
<td>0.53</td>
</tr>
<tr>
<td>Model</td>
<td>Precision</td>
<td>Recall</td>
<td>Accuracy</td>
</tr>
<tr>
<td>------------------------</td>
<td>-----------</td>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td>SVC</td>
<td>0.92(+/-0.03)</td>
<td>0.89(+/-0.03)</td>
<td>0.91(+/-0.07)</td>
</tr>
<tr>
<td>NB</td>
<td>0.84(+/-0.06)</td>
<td>0.82(+/-0.06)</td>
<td>0.84(+/-0.07)</td>
</tr>
<tr>
<td>DecisionTree</td>
<td>0.87(+/-0.01)</td>
<td>0.88(+/-0.02)</td>
<td>0.89(+/-0.04)</td>
</tr>
<tr>
<td>KerasSequentialModel</td>
<td>0.93(+/-0.02)</td>
<td>0.89(+/-0.001)</td>
<td>0.92(+/-0.005)</td>
</tr>
</tbody>
</table>

On the testing dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVC</td>
<td>0.84</td>
<td>0.46</td>
<td>0.47</td>
</tr>
<tr>
<td>NB</td>
<td>0.84</td>
<td>0.66</td>
<td>0.60</td>
</tr>
<tr>
<td>DecisionTree</td>
<td>0.84</td>
<td>0.41</td>
<td>0.44</td>
</tr>
<tr>
<td>KerasSequentialModel</td>
<td>0.84</td>
<td>0.41</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Summary of the result:

Of the four testing sets, the n-gram of (1, 2) and the 1 000 common word features return the best total result, the n-gram of (1, 1) on the 700 different word set of advertisements and non-advertisements give the best result on recall, i.e. relatively more true positives comparing with the total true values are reported. Naïve Bayes classifier gives the best result in all cases.

Limitations and future work

1. The scraper was able to extract the related links according to the pattern defined and go deeper along the links, but improvements are needed. The improvements concern error handling during the scraping process. Most of the webpages scraped so far are non-advertisements, we need to examine the reason. More true advertisements from the websites need to be detected and scraped. One guess is that the advertisements need user interaction to reach, e.g. click on search button.

2. The results on the training dataset are much better than the test dataset, showing the generalization problem of applying the classifier on new data. Generalization refers to how well a classifier performs applied on unseen data with the learned concepts from specific examples. Overfitting is suspected to be the problem and should be checked in the future. To increase the amount of training data is the first step.

3 Applying pipe-line and GridSearchCV to automate the testing of parameters in feature selection and classifier evaluation. More parameters need to be tested to improve the stability of the classifier.
ESSnet Big Data WP2: Webscraping Enterprise Characteristics

Methodological note

The ESSnet BD WP2 performs joint web scraping experiments in multiple countries, using as much as possible the same methodological concepts. The aim is to derive experimental statistics on enterprises from information found on the web, especially the websites of enterprises. It should be noted that these statistics have not reached maturity in terms of harmonisation, coverage or methodology. At this point they are to be treated as the output of research experiments and they do not necessarily align with the official statistics published on this subject.

Use case: Job advertisements on enterprises’ websites

Country: BG
Date: 2018-03-20
Authors: Kostadin Georgiev and Galya Stateva

Data sources
- The Statistical Business Register (SBR)
- Responses from Bulgarian ICT survey
- URL Inventory of enterprises

Population
- The List of enterprises with known websites found with URLs Retrieval use-case (11442)

Methodology

22. Configure the information in the conf.php file.

23. Add necessary database table fields according to the information in the conf.php file.

24. Run scrap.jobs.v3.php script to get up to 3 prediction for Job advertisements URLs of the enterprises based on scraped data from the enterprises web sites.

25. Run list_jobs.php script to choose manually the correct Job advertisements URLs of the enterprises.

26. Run list_jobs_lost.php script to find missed Job advertisements URLs with 10% sample extracted from enterprises we don’t predict Job advertisements websites with scrap.v3.php script.

27. Run the info.php script to see statistics from the above executed steps for Job advertisements URLs at regional level and NACE categories.

The BNSI Job advertisements URLs scripts are available at https://github.com/kostadingeorgiev/bnsi_bigdata

Related figures
Predicted Job advertisements URLs of enterprises with script 1: 2222
Predicted Job advertisements URLs of enterprises with script 2: 1085
Predicted Job advertisements URLs of enterprises with script 3: 1143
Job advertisements URLs of enterprises found with the project: 1672
Missed Job advertisements URLs of enterprises in 10% sample: 26

Expected number of Job advertisements URLs of enterprises: 1672+26×10~≈1932
The benchmark analysis wasn’t done between the ICT survey 2017 data and URLs Job advertisements, because the question for job advertisement isn’t included in the ICT 2017 survey.

**Limitations and future work**

- The used software for current use-case is the same as the E-commerce software, only four positive lists of keywords were changed for job advertisements procedure and we didn’t use the negative list of keywords. More detailed lists of keywords gives respectively better results.
Use Case 4: Social Media Presence

ESSnet Big Data WP2: Webscraping Enterprise Characteristics

Methodological note

The ESSnet BD WP2 performs joint web scraping experiments in multiple countries, using as much as possible the same methodological concepts. The aim is to derive experimental statistics on enterprises from information found on the web, especially the websites of enterprises. It should be noted that these statistics have not reached maturity in terms of harmonisation, coverage or methodology. At this point they are to be treated as the output of research experiments and they do not necessarily align with the official statistics published on this subject.

Use case: Presence in social media

Country: Italy
Date: 2018-03-05
Authors:

Data sources

- The Italian Business Register (ASIA)
- Responses from 2017 Italian ICT survey (responses from about 21,000 enterprises) to the question: “Links or references to the enterprise’s social media profiles (Yes/No)”
- Websites from enterprises (2017)

Population

The 184,000 enterprises as defined by the ICT survey (>10 persons, limited NACE)

Methodology

The overall procedure is based on the following steps:
15. get the websites address (Uniform Resource Locator) potentially for all enterprises included in the population of reference (URL retrieval);
16. access websites with available URL and scrape their content (web scraping);
17. process the content of the scraped websites in order to identify the best predictors for the target variables (text mining);
18. apply techniques of Information Retrieval in the subset of enterprises where both Internet data and survey data were available (considering survey data as the true values) and predict the values of target variable for all the enterprises for which the retrieval and scraping of their websites was successful, and evaluate the performance of the prediction;
19. apply the Information Retrieval application to all the units for which steps 2 and 3 were successful;
20. produce estimates applying different estimators (full based model approach and combined approach);
21. compare to current ICT survey design based estimates.

Results

Results of steps from 1 to 3 are reported in the following figure:
In step 4 the application of the Information Retrieval technique produced the following results:

Cases not on the principal diagonal were subject to manual inspection by accessing corresponding website in order to decide what was the right value: 1,289 cases were changed from “No” to “Yes” and 1,064 from “Yes” to “No”.

The performance was calculated on clean data and gave these results:

<table>
<thead>
<tr>
<th>Confusion Matrix and Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 1</td>
</tr>
<tr>
<td>0 9939 1012</td>
</tr>
<tr>
<td>1 1966 4564</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>0.7157</th>
</tr>
</thead>
<tbody>
<tr>
<td>95% CI</td>
<td>(0.7075, 0.7238)</td>
</tr>
<tr>
<td>No Information Rate</td>
<td>0.5114</td>
</tr>
<tr>
<td>P-Value (Acc &gt; NIR)</td>
<td>&lt; 2.2e-16</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.4303</td>
</tr>
<tr>
<td>McNemar's Test P-Value</td>
<td>1.251e-09</td>
</tr>
</tbody>
</table>

Sensitivity = 0.7512
Specificity = 0.4786
Pos Pred Value = 0.7098
Neg Pred Value = 0.7226
Prevalence = 0.8114
Detection Rate = 0.8841
Detection Prevalence = 0.5412
Balanced Accuracy = 0.7149
‘Positive’ Class = 1

F1 measure = 0.7290897
In step 6 the following estimators have been applied:

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Formula</th>
<th>Weights</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design based/</td>
<td>[ \hat{y} = \sum y_k w_k ]</td>
<td>[ \sum w_k = N_0 ]</td>
<td>The estimates of known totals in the population are obtained by calibration procedures of basic weights (inverse of inclusion probabilities)</td>
</tr>
<tr>
<td>model assisted</td>
<td></td>
<td></td>
<td>making sure that the variability due to sampling error is reduced.</td>
</tr>
<tr>
<td>Model-based</td>
<td>[ \hat{\varphi} = \sum w_k' \hat{\varphi}_k ]</td>
<td>[ \sum w_k' = N_0' ]</td>
<td>The estimate of the total number of enterprises offering web ordering facilities on their websites is given by the count of the predicted values ( \hat{\varphi}_k ) for all units for which it was possible to reach their websites (population ( U' )).</td>
</tr>
<tr>
<td>Combined</td>
<td>[ \hat{\varphi} = \sum_{k \in U^{ref}} \hat{\varphi}<em>k + \sum</em>{k \not\in U^{ref}} (\hat{\varphi}<em>k - y_k) w_k' + \sum</em>{k \not\in U^{ref}} y_k w_k'' ]</td>
<td>[ \sum w_k' = N_0' ] and [ \sum w_k'' = N_{ref} - N_0' ]</td>
<td>Estimates are produced by summing these components: 1. the counting of predicted values in the subpopulation ( U' ) of units for which it was possible to scrape and process corresponding websites. 2. an adjustment based on the consideration of the difference between the ( r^2 ) reported values and the predicted values (expanded to the same subpopulation ( U^{ref} )); 3. the counting of observed values for the ( r^2 ) respondents that declared a website, that was not found or scraped, expanded to the whole subpopulation ( U^{ref} - U' ).</td>
</tr>
</tbody>
</table>

As for step 7, estimates have been produced for different domains. In the following plot we report a comparison of estimates of presence in social media rate for enterprises cross-classified by macro-sector of economic activity (1-4) and class of employees (1-4):
The dashed lines define the area delimited by the lower and upper limits of the confidence intervals calculated in correspondence of each design based estimate.

**Conclusions**

Same as reported for e-commerce use case.
ESSnet Big Data WP2: Webscraping Enterprise Characteristics

Methodological note

The ESSnet BD WP2 performs joint web scraping experiments in multiple countries, using as much as possible the same methodological concepts. The aim is to derive experimental statistics on enterprises from information found on the web, especially the websites of enterprises. It should be noted that these statistics have not reached maturity in terms of harmonisation, coverage or methodology. At this point they are to be treated as the output of research experiments and they do not necessarily align with the official statistics published on this subject.

The result in this note is not published in the experimental statistics output, because of the data set used is not comparable to the population in the official statistics.

Use case: Social media
Country: SE
Date: <2018-02-17>

Data sources
- URLs extracted from other datasets, 1466 unique urls of the public sector
- Code source developed by Statistics Poland on https://github.com/jmaslankowski/WP2-Social-Media-Presence

Population
Because of the web scraping restriction at SCB, we scraped the 1466 urls of the public sector. The public sector is not covered in the ICT population. The result is not comparable to other results from the use case.

Methodology
1. Preparing the urls of the public sector
2. Run the software on the urls
3. Manually check the social media presence of 105 units and estimate the accuracy of the software.
   We use the Fixfox plugin “Inspector” searching the keywords, “facebook”, “twitter”, “Instagram”, “linkedin”, “google” and “youtube”. For certain unsure cases, we google to search name of the organization and keywords to ensure presence on the channel.

Results
The program returned result is presented below:

<table>
<thead>
<tr>
<th></th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>SocialMedia</td>
<td>963</td>
<td>503</td>
</tr>
<tr>
<td>Blog</td>
<td>581</td>
<td>885</td>
</tr>
<tr>
<td>Multimedia</td>
<td>653</td>
<td>813</td>
</tr>
</tbody>
</table>

105 units are drawn from the total, the manually checking result is presented in the table.

<table>
<thead>
<tr>
<th></th>
<th>blog</th>
<th>multimedia</th>
<th>socialmedia</th>
</tr>
</thead>
<tbody>
<tr>
<td>False positive</td>
<td>9 (9%)</td>
<td>8 (8%)</td>
<td>5 (5%)</td>
</tr>
<tr>
<td>True positive</td>
<td>49 (47%)</td>
<td>45 (43%)</td>
<td>67 (64%)</td>
</tr>
<tr>
<td>False negative</td>
<td>20 (19%)</td>
<td>17 (16%)</td>
<td>12 (11%)</td>
</tr>
<tr>
<td>True negative</td>
<td>27 (26%)</td>
<td>35 (33%)</td>
<td>21 (20%)</td>
</tr>
</tbody>
</table>
Blog identified correctly 76%, Multimedia identified correctly 76%, and social media correctly identified 88%.

**Limitations and future work**

1. The scraper checks on the first and the second level from the given urls. This may lose some accounts if they are present on an even deeper level. This may explain why the false negative are relatively high. The program so far checks only links for detecting the accounts, during the manually checking, we find that other html tags, e.g. *meta*, *scripts*, *img* contain information about the accounts. The improvement on the scraper can decrease the percentage false negative. Searching on names and the account keywords like “facebook” can also complete the finding of the result.

2. At SCB, the next step is to web scrape scale urls from the private sector and compare the result with ICT usage.

3 Statistics Sweden tested data scraped by Vainu⁴ to examine the possibility of using alternative data for enterprise IT usage statistics. The preliminary result shows that it is possible to use web scraped data, however, it is too expensive to buy Vainu dataset for the quality provided. With the software developed in this work package, we expect that the cost and the benefit can have a better balance.

⁴https://company.vainu.io/
ESNnet Big Data WP2: Webscraping Enterprise Characteristics

Methodological note – Test Statistic Produced: Yes

The ESSnet BD WP2 performs joint web scraping experiments in multiple countries, using as much as possible the same methodological concepts. The aim is to derive experimental statistics on enterprises from information found on the web, especially the websites of enterprises. It should be noted that these statistics have not reached maturity in terms of harmonisation, coverage or methodology. At this point they are to be treated as the output of research experiments and they do not necessarily align with the official statistics published on this subject.

Use case: Social Media

Country: UK
Date: 2018-04-04

Data sources
Business Websites, ICT survey (called the ‘Ecommerce survey’ in the UK)

Population
The enterprises as defined by the ICT survey (called the ‘Ecommerce survey’ in the UK) (>10 persons, limited NACE)

Methodology
13. Extraction of a set of companies in the reference population with known websites and ecommerce status
14. Scrape website for each company and store the scraped text (note – all scraping carried out according to ONS web-scraping policy) as well as all links found on the pages.
15. Search the text of all links for the names of multiple common social media sites, under the assumption that the links lead to the enterprises’ own pages or groups on each site

Results
The method, under the stated assumption of linkage equalling usage, estimates that 80 % of enterprises are using social media sites based on a sample of 1100 linked enterprises and websites (unweighted). This is compared with 66 % that claim to use social media within that 1100-strong sample (unweighted, from e-Commerce/ICT survey 2015). The discrepancy may be due to the rapid increase in the use of social media over the last few years, given the website data was scraped late 2017. The official, weighted estimate in 2015 for the population was 35 %.

Limitations and future work
The method allows assessment of usage of different social media platforms by enterprises, but does not allow one to assess what the social media platforms are being used for, or how intensively. The method allows for directly answering one of the core e-Commerce/ICT survey questions in a reliable manner.
ESSnet Big Data WP2: Webscraping Enterprise Characteristics

Methodological note

The ESSnet BD WP2 performs joint web scraping experiments in multiple countries, using as much as possible the same methodological concepts. The aim is to derive experimental statistics on enterprises from information found on the web, especially the websites of enterprises. It should be noted that these statistics have not reached maturity in terms of harmonisation, coverage or methodology. At this point they are to be treated as the output of research experiments and they do not necessarily align with the official statistics published on this subject.

Use case: Social Media
Country: NL
Date: 2018-03-28
Authors: Olav ten Bosch and Dick Windmeijer

Data sources
- The General Business Register (GBR) (about 1.5 M enterprises), of which 1/3 has a known URL
- Responses from Dutch ICT survey (responses from about 9 000 enterprises)
- Websites from enterprises via dedicated scraping software from Poland
- Twitter messages via Twitter API

Population
The enterprises as defined by the ICT survey (>10 persons, limited NACE)

Methodology
1. Determination of a set of enterprises from the GBR that responded to the 2017 ICT survey (8909).
2. URL retrieval (see other use case) on the subset of enterprises in this set not having a known URL. Creation of a smaller set for which we have a URL either from our GBR or from URL retrieval (7222).
3. Social media detection on these enterprises using software developed by Statistics Poland, customized for local use: https://github.com/jmaslankowski/WP2-Social-Media-Presence
4. Comparison of four different questions on social media presence from the ICT survey with the prediction based on the web.
5. For detecting the goal of social media use, we collected about 4 M. tweets from the enterprises for which we found a Twitter account in step 3. We created a collection of 1-, 2- and 3-grams of words in all tweets per enterprise. After splitting it in a training set and a test set (70/30) we trained a model using random forest.

---

5 One of the refinements was to add a step to determine the Twitter user id from the Twitter-URL being found.
6 Mentioned in the awesome official statistics software list: https://github.com/SNStatComp/awesome-official-statistics-software
## Results
We describe the confusion matrix and some common performance metrics per case⁷.

### Social Media Type

**Question:** Does your enterprise use the following social media?

**Answer:** *Social networks such as Facebook, LinkedIn, GooglePlus? (Yes, No)*

<table>
<thead>
<tr>
<th>Predict</th>
<th>True</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>3815</td>
<td>1885</td>
<td></td>
<td>5700</td>
</tr>
<tr>
<td>No</td>
<td>553</td>
<td>969</td>
<td></td>
<td>1522</td>
</tr>
<tr>
<td>Total</td>
<td>4368</td>
<td>2854</td>
<td></td>
<td>7222</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.87</td>
<td>0.67</td>
<td>0.76</td>
</tr>
<tr>
<td>No</td>
<td>0.34</td>
<td>0.64</td>
<td>0.44</td>
</tr>
<tr>
<td>avg / total</td>
<td>0.76</td>
<td>0.66</td>
<td>0.69</td>
</tr>
</tbody>
</table>

**Accuracy:** 0.66

**Balanced accuracy:** 0.65

Prediction social network use: 60 %

**Answer:** *Blogs or microblogs such as Twitter? (Yes, No)*

<table>
<thead>
<tr>
<th>True</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>2058</td>
<td>1021</td>
<td>3079</td>
</tr>
<tr>
<td>No</td>
<td>1290</td>
<td>2853</td>
<td>4143</td>
</tr>
<tr>
<td>Total</td>
<td>3348</td>
<td>3874</td>
<td>7222</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.61</td>
<td>0.67</td>
<td>0.64</td>
</tr>
<tr>
<td>No</td>
<td>0.74</td>
<td>0.69</td>
<td>0.71</td>
</tr>
<tr>
<td>avg / total</td>
<td>0.68</td>
<td>0.68</td>
<td>0.68</td>
</tr>
</tbody>
</table>

---

Accuracy: 0.68
Balanced accuracy: 0.68
Prediction use of Blog or microblogs such as Twitter: 46 %

**Answer:** *Websites that share multimedia (movies, photos) like Youtube, Instagram? (Yes, No)*

<table>
<thead>
<tr>
<th>Predict</th>
<th>True</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>1439</td>
<td>1592</td>
<td>3031</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>892</td>
<td>3299</td>
<td>4191</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2331</td>
<td>4891</td>
<td>7222</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.62</td>
<td>0.47</td>
<td>0.54</td>
</tr>
<tr>
<td>No</td>
<td>0.67</td>
<td>0.79</td>
<td>0.73</td>
</tr>
<tr>
<td>avg / total</td>
<td>0.65</td>
<td>0.66</td>
<td>0.65</td>
</tr>
</tbody>
</table>

Accuracy: 0.66
Balanced accuracy: 0.63
Prediction use of websites that share multimedia: 32 %

**All answers put together**, which could be taken as an indication of the social media presence in general:

<table>
<thead>
<tr>
<th>Predict</th>
<th>True</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>4126</td>
<td>1720</td>
<td>5846</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>561</td>
<td>815</td>
<td>1376</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4687</td>
<td>2535</td>
<td>7222</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.88</td>
<td>0.71</td>
<td>0.78</td>
</tr>
<tr>
<td>No</td>
<td>0.32</td>
<td>0.59</td>
<td>0.42</td>
</tr>
<tr>
<td>avg / total</td>
<td>0.77</td>
<td>0.68</td>
<td>0.71</td>
</tr>
</tbody>
</table>
Accuracy: 0.68
Balanced accuracy: 0.65
Prediction Social Media use any type (All): 65 %

Social Media Use (goal):
Question: Does your enterprise use social media to:

Answer: **develop the image of your enterprise or for marketing? (Yes, No)**

<table>
<thead>
<tr>
<th>Predict</th>
<th>True</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>715</td>
<td>34</td>
<td>749</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>140</td>
<td>33</td>
<td>173</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>855</td>
<td>67</td>
<td>922</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.84</td>
<td>0.95</td>
<td>0.89</td>
</tr>
<tr>
<td>No</td>
<td>0.49</td>
<td>0.19</td>
<td>0.28</td>
</tr>
<tr>
<td>avg / total</td>
<td>0.77</td>
<td>0.81</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Accuracy: 0.81
Balanced accuracy: 0.57
Prediction use for marketing: 93 %

Answer: **to get opinions/reviews from customers or to answer questions? (Yes, No)**

<table>
<thead>
<tr>
<th>Predict</th>
<th>True</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>202</td>
<td>253</td>
<td>455</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>135</td>
<td>332</td>
<td>467</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>337</td>
<td>585</td>
<td>922</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.60</td>
<td>0.44</td>
<td>0.51</td>
</tr>
<tr>
<td>No</td>
<td>0.57</td>
<td>0.71</td>
<td>0.63</td>
</tr>
<tr>
<td>avg / total</td>
<td>0.58</td>
<td>0.58</td>
<td>0.57</td>
</tr>
</tbody>
</table>
Accuracy: 0.58
Balanced accuracy: 0.58
Predicted use for opinions/reviews/questions: 37%

Answer: for recruitment? (Yes, No)

<table>
<thead>
<tr>
<th>Predict</th>
<th>True</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>661</td>
<td>59</td>
<td>720</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>167</td>
<td>35</td>
<td>202</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>828</td>
<td>94</td>
<td>922</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.80</td>
<td>0.92</td>
<td>0.85</td>
<td>720</td>
</tr>
<tr>
<td>No</td>
<td>0.37</td>
<td>0.17</td>
<td>0.24</td>
<td>202</td>
</tr>
<tr>
<td>avg / total</td>
<td>0.70</td>
<td>0.75</td>
<td>0.72</td>
<td>922</td>
</tr>
</tbody>
</table>

Accuracy: 0.75
Balanced accuracy: 0.55
Predicted use for recruitment: 90%
Comparison with published figures

The figures were published on the level of individual questions, where the predictions were done on a combination of questions. We therefore roughly compare the mean of the statistics on the individual results with the predictions:

<table>
<thead>
<tr>
<th></th>
<th>Social networks</th>
<th>Blogs</th>
<th>Multimedia</th>
<th>ALL</th>
<th>marketing</th>
<th>Opinions</th>
<th>Recruitment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Published</td>
<td>67 %</td>
<td>29 %</td>
<td>29 %</td>
<td>69 %</td>
<td>83 %</td>
<td>46 %</td>
<td>72 %</td>
</tr>
<tr>
<td>Predicted</td>
<td>60 %</td>
<td>46 %</td>
<td>32 %</td>
<td>65 %</td>
<td>93 %</td>
<td>37 %</td>
<td>90 %</td>
</tr>
</tbody>
</table>

Conclusion

- Detecting the social media use for enterprises can be done reasonably effective via this method, with the exception of detecting the use of blogs.
- Detecting the goal of social media use of enterprises is also reasonably successful, comparing the published and predicted figures for the three variables taken into consideration.

Limitations and future work

- The datasets were created at different points in time; the ICT survey was executed beginning 2017, the URL retrieval and social media scraping was carried out January 2018. This could explain some differences.
- Some early manual inspection showed that some enterprises do have a facebook account, however they do not have a link to their facebook account on their webpage. This might explain some of the results. A further refinement could be to also consult facebook directly if no activity is found on the enterprises web site. The same could hold for other social media.
- One has to keep in mind that not only the web scraper could be wrong, an answer on the ICT survey could be wrong as well.
- There was no checking on the URLs found between the URL retrieval step and the social media detection step. Starting with a wrong URL could explain some of the differences.
**ESSnet Big Data WP2: Webscraping Enterprise Characteristics**

**Methodological note**

The ESSnet BD WP2 performs joint web scraping experiments in multiple countries, using as much as possible the same methodological concepts. The aim is to derive experimental statistics on enterprises from information found on the web, especially the websites of enterprises. It should be noted that these statistics have not reached maturity in terms of harmonisation, coverage or methodology. At this point they are to be treated as the output of research experiments and they do not necessarily align with the official statistics published on this subject.

**Use case**: Social Media Presence

**Country**: BG

**Date**: 2018-03-06

**Authors**: Kostadin Georgiev and Galya Stateva

**Data sources**
- The Statistical Business Register (SBR)
- Responses from Bulgarian ICT survey
- URL Inventory of enterprises

**Population**
- The List of enterprises with known websites found with URLs Retrieval use-case (11442)

**Methodology**

**V. Methodological procedures with BNSI software:**

28. Configure the information in the conf.php file.

29. Add necessary database table fields according to the information in the conf.php file.

30. Run scrap.v3.php script to get social media URLs of the enterprises on the first page of their websites.

31. Run the info.php script to see statistics from the above executed steps for Social Media presence URLs at regional level and NACE categories.

The BNSI E-commerce URLs scripts are available at [https://github.com/kostadingeorgiev/bnsi_bigdata](https://github.com/kostadingeorgiev/bnsi_bigdata)

**VI. Methodological procedures with Polish software:**

8. Make a list with enterprises IDs and URLs from MySQL database containing enterprises with known websites found with URLs Retrieval use-case (11442).

9. Run the Social media detection software developed by Statistics Poland, refined by Statistics Nederland and Bulgarian National Statistical Institute using the List with enterprises IDs and URLs. Available at: [https://github.com/jmaslankowski/WP2-Social-Media-Presence](https://github.com/jmaslankowski/WP2-Social-Media-Presence).

10. Import the result of previous step into MySQL database table for a comparison with BNSI software social media results.

11. Run software for gathering enterprises’ Tweets over the output file from step 2 (software developed by Statistics Poland, refined by Statistics Nederland and Bulgarian National Statistical
Institute using the List with enterprises IDs and URLs. Available at: https://github.com/jmaslankowski/WP2-Social-Media-Presence.

12. Run software for detection of the goals of social media use (as defined in the ICT survey questionnaire) with appropriate configuration of stop and go words.

Related figures

The execution of Bulgarian and Polish social media software gave the following results:

<table>
<thead>
<tr>
<th>Software</th>
<th>Facebook</th>
<th>Twitter</th>
<th>Google Plus</th>
<th>Linkedin</th>
<th>Youtube</th>
<th>Instagram</th>
<th>Pinterest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulgarian</td>
<td>3228</td>
<td>1111</td>
<td>1103</td>
<td>858</td>
<td>751</td>
<td>308</td>
<td>202</td>
</tr>
<tr>
<td>Polish</td>
<td>3776</td>
<td>1462</td>
<td>1276</td>
<td>1009</td>
<td>1247</td>
<td>353</td>
<td></td>
</tr>
</tbody>
</table>

V. The results from Social media presence (with BNSI software) procedure is:

The benchmark analysis was carried out between the ICT survey 2017 data and Social media presence use-case (ICT survey question: Usage of social media (Yes, No)?). The total 11442 enterprises with known URLs were object of the Social media presence procedure and 4776 of them were in the scope of the ICT survey 2017. The number of full matches (Yes/Yes and No/No results in the both sources) are 3370 or 70, 56% success of the Social media presence procedure. From the target population of the ICT survey 2017, 12,74 % of enterprises have at least one social media profile.

VI. The results from Social media presence (with Polish software) procedure is:

We found 21 813 web-pages of enterprises with social media links. We collected about 1 M. tweets (1 000 273) from the enterprises for which we found Twitter accounts on the fly in step 4.

The results after execution of the step 5 (with Polish Tweets classification training set) for detection of goals of social media use are:

- recruitment – 685 392 tweets;
- enterprise image – 207 725 tweets;
- commercials – 79 514 tweets;
- marketing – 4132 tweets;
- others – 23 510 tweets;

The results after execution of the step 5 (with Bulgarian Tweets classification training set – 1001 Tweets) for detection of goals of social media use are:

- commercials – 510 644 tweets;
- others – 388 804 tweets;
- enterprise image – 77 570 tweets;
- marketing – 23 137 tweets;
- recruitment – 118 tweets;
Limitations and future work

- The ICT survey was carried out in the beginning of 2017, where the Social media presence procedure was performed in March 2018. We intend to repeat the comparison when the ICT survey 2018 results for social media presence are available. It’s probably will lead to the more accurate results;

- The results on social media detection from the 1st page of the enterprises web-sites and from all web-site pages are almost the same (15% more on all web-site pages). Therefore if you don’t want to make heavy web scrapping on the whole web-site you may go only with the 1st page to get the comparable results.

- The difference between results for detection of goals of social media use with Polish and Bulgarian Tweets classification training sets could be explain with different size of both training set and the presence of Bulgarian Cyrillic key words. Most likely with another different training set will get a different result.

- There was need for some software changes over the Polish software, even after the Netherlands’ refinements to be able to run the methodological steps smoothly. In general, the Polish software works well in the BNSI environment.
ESSnet Big Data WP2: Webscraping Enterprise Characteristics
Methodological note, Draft

The ESSnet BD WP2 performs joint web scraping experiments in multiple countries, using as much as possible the same methodological concepts. The aim is to derive experimental statistics on enterprises from information found on the web, especially the websites of enterprises. It should be noted that these statistics have not reached maturity in terms of harmonisation, coverage or methodology. At this point they are to be treated as the output of research experiments and they do not necessarily align with the official statistics published on this subject.

**Use case**: Social Media Presence
**Country**: PL
**Date**: 2018-03-13
**Authors**: Jacek Maślankowski and Joanna Wyzina

**Data sources**
- The Business Register (BJS) by Statistics Poland (about 3.3 M enterprises), of which 189 thous. has a known URL; for enterprises having 10 and more employees – total number is 165 thous, of which 47 thous. has URLs (28.4%),
- Websites from enterprises via dedicated scraping software from Poland,
- Twitter messages via Twitter API.

**Population**
The selected enterprises in Poland in Pomeranian voivodship (2961 enterprises).

**Methodology**
1. Preparing the list of enterprises according to the presence of their URL in BJS.

2. Social media detection on these enterprises using software developed by Statistics Poland: https://github.com/jmaslankowski/WP2-Social-Media-Presence to provide answer to the following question from ICT in Enterprises 2016 questionnaire:

<table>
<thead>
<tr>
<th>C11. Does your enterprise use any of the following social media?</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) Social networks (e.g. Facebook, LinkedIn, Xing, Viadeo, Yammer, etc.)</td>
</tr>
<tr>
<td>b) Enterprise's blog or microblogs (e.g. Twitter, Present.ly, etc.)</td>
</tr>
<tr>
<td>c) Multimedia content sharing websites (e.g. YouTube, Flickr, Picasa, SlideShare, etc.)</td>
</tr>
<tr>
<td>d) Wiki based knowledge sharing tools</td>
</tr>
</tbody>
</table>

3. Scraping tweets from Twitter by enterprise' account (Twitter accounts).

4. Preparation of the training dataset: manual classification of the tweets according to the classification from ICT in Enterprises Survey 2017 |

---

9 Mentioned in the awesome official statistics software list: https://github.com/SNStatComp/awesome-official-statistics-software
10 https://circabc.europa.eu/sd/a/a39ae859-8a16-4306-8020-ae06d3df3c91/Questionnaire%20ENT%202016.pdf, accessed 13th March 2018
C11. Does your enterprise use any of the above mentioned social media to:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a)</td>
<td>Develop the enterprise’s image or market products (e.g. advertising or launching products, etc.)</td>
</tr>
<tr>
<td>b)</td>
<td>Obtain or respond to customer opinions, reviews, questions</td>
</tr>
<tr>
<td>c)</td>
<td>Involve customers in development or innovation of goods or services</td>
</tr>
<tr>
<td>d)</td>
<td>Collaborate with business partners (e.g. suppliers, etc.) or other organisations (e.g. public authorities, non-governmental organisations, etc.)</td>
</tr>
<tr>
<td>e)</td>
<td>Recruit employees</td>
</tr>
<tr>
<td>f)</td>
<td>Exchange views, opinions or knowledge within the enterprise</td>
</tr>
</tbody>
</table>

5. Detection of the social media use was analysed with Naïve Bayes algorithm – we split the dataset prepared in step 4, in the proportion of 85% for training purposes and 15% for testing purposes. The detection was made by Polish Social Media Presence software [https://github.com/jmaslankowski/WP2-Social-Media-Presence](https://github.com/jmaslankowski/WP2-Social-Media-Presence)

**Results**

The pilot was conducted in two steps. The first step was to identify the number of enterprises present in social media. The population was 2961 enterprises in Pomeranian voivodship in Poland. According to the figure below, all websites are scraped to find the links of different social media links.

The results for the Pomeranian voivodship in Poland is presented below.

![Number of enterprises present in social media in Pomeranian voivodship](chart.png)

The second step is to find the current type of activity in social media. To accomplish that task a training dataset has been prepared with the number of 2 thous. tweets.

---


The accuracy is strictly related to the number of occurrences in the training dataset. Therefore, disproportion in tweets categories (e.g., commercial, enterprise image, recruitment). Table below shows the accuracy of selected categories.

<table>
<thead>
<tr>
<th></th>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.78</td>
<td>0.84</td>
<td>0.81</td>
</tr>
<tr>
<td>2</td>
<td>0.45</td>
<td>0.37</td>
<td>0.40</td>
</tr>
<tr>
<td>5</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>avg / total</td>
<td>0.69</td>
<td>0.71</td>
<td>0.70</td>
</tr>
</tbody>
</table>

It indicates that the training dataset for selected categories must be reinforced. Therefore, the recommended solution for this step is to use machine learning with reinforcement.

**Conclusion**

- It is a reliable solution to provide information on social media presence.
- Detection of the type of social media presence is very prone to the training dataset. We have to be aware of issues related to lemmatization, stop words and stemming to provide better results with higher accuracy.
- Manual preparation of the training dataset is very time consuming activity in this use case and has a big impact on the successful results.
- There is a risk of finding the wrong social media link on the website (e.g., reference to another company fan-page) but the scale of this issue is very small.
- The link to social media presence should be searched in subpages if not found on the main page.

**Limitations and future work**

- The social media presence software must have a well prepared list of URLs for enterprises to be tested. Therefore, it is recommended to execute Istat URL Retrieval software to provide reliable URLs list.
- The future work will be to extend the training dataset in the social media presence software to increase the accuracy of the results regarding detection of the type of social media presence.
Use Case 5: NACE

ESSnet Big Data WP2: Webscraping Enterprise Characteristics

Methodological note

The ESSnet BD WP2 performs joint web scraping experiments in multiple countries, using as much as possible the same methodological concepts. The aim is to derive experimental statistics on enterprises from information found on the web, especially the websites of enterprises. It should be noted that these statistics have not reached maturity in terms of harmonisation, coverage or methodology. At this point they are to be treated as the output of research experiments and they do not necessarily align with the official statistics published on this subject.

Use case: NACE (economic activity classification)

Country: Italy
Date: 2018-03-05
Authors:

Data sources

- The Italian Business Register (ASIA)
- Websites from enterprises (2017)

Population

The 184,000 enterprises as defined by the ICT survey (>10 persons, limited NACE)

Methodology

The overall procedure is based on the following steps:
22. get the websites address (Uniform Resource Locator) potentially for all enterprises included in the population of reference (URL retrieval);
23. access websites with available URL and scrape their content (web scraping);
24. consider half of the cases with scraped websites and known value of economic activity (from the Business Register) as training set, and the remaining half as test set;
25. fit models in the training set and apply to the test set;
26. compute performance indicators for the most aggregate classification of economic activity (4 macro-sectors);
27. repeat steps 4 and 5 with more detailed classification of economic activity (12 and 27 values).

Results

Steps 1 and 2 were already been performed for use cases related to URLs retrieval and the set of enterprises characteristics linked to the ICT survey (e-commerce, job advertisements and presence in social media). Texts from scraped websites (85,360) were already available, and they have been linked to the classification of economic activity available in the Business Register.

In step 3 the training set includes 42,677 enterprises, while the test set includes 42,683 enterprises.

In step 4 the best learner resulted to be random forest.
For the 4 macro-sectors considered, i.e.:

- 0: Manufacturing
- 1: Energy
- 2: Construction
- 3: Non financial services

this was the obtained prediction performance:

<table>
<thead>
<tr>
<th>Confusion Matrix and Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>predictedTest</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
</tbody>
</table>

Overall Statistics

- Accuracy : 0.8169
- 95% CI : (0.814, 0.8193)
- No Information Rate : 0.4734
- P-Value [Acc > NIR] : < 2.2e-16
- Kappa : 0.6589
- Monemar's Test P-Value : < 2.2e-16

Statistics by Class:

<table>
<thead>
<tr>
<th>Class: 0</th>
<th>Class: 1</th>
<th>Class: 2</th>
<th>Class: 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.8583</td>
<td>0.301587</td>
<td>0.64919</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.8529</td>
<td>0.998534</td>
<td>0.98852</td>
</tr>
<tr>
<td>Pos Pred Value</td>
<td>0.8047</td>
<td>0.823105</td>
<td>0.84927</td>
</tr>
<tr>
<td>Neg Pred Value</td>
<td>0.8951</td>
<td>0.984429</td>
<td>0.96584</td>
</tr>
<tr>
<td>Prevalence</td>
<td>0.4138</td>
<td>0.022114</td>
<td>0.09064</td>
</tr>
<tr>
<td>Detection Rate</td>
<td>0.3552</td>
<td>0.006669</td>
<td>0.05884</td>
</tr>
<tr>
<td>Detection Prevalence</td>
<td>0.4414</td>
<td>0.008103</td>
<td>0.06928</td>
</tr>
<tr>
<td>Balanced Accuracy</td>
<td>0.8556</td>
<td>0.650061</td>
<td>0.81885</td>
</tr>
</tbody>
</table>

For macro-sectors with high number of cases (manufacturing and non financial services) the performance is quite high, while for the other two macro-sectors accuracy is not satisfactory.
The second experiment has been on a subset of sections of economic activity, those regarding the reference population of the ICT survey, namely:

0. Manufacturing;
1. Electricity, gas and steam,
2. water supply, sewerage and waste management;
3. Construction;
4. Wholesale and retail trade
5. repair of motor vehicles and motorcycles;
6. Transportation and storage;
7. Accommodation and food service activities;
8. Information and communication;
9. Real estate activities;
10. Professional, scientific and technical activities;
11. Administrative and support activities.

The last section has been dropped because of its too limited number of cases.

This was the obtained performance:

With an overall accuracy of 74%, there are values with very high accuracy rates (up to 92%) and other with low levels of it.
Finally, considering a more detailed level of classification:

0. activities not included in ICT Sector
1. activities included in ICT Sector
2. manufacture of food products, beverages and tobacco products
3. manufacture of textiles, apparel, leather and related products
4. manufacture of wood and paper products, and printing
5. manufacture of coke and refined petroleum products, of chemicals and chemical products, of basic pharmaceutical products and preparations, of rubber, plastic and of other non-metallic mineral products
6. manufacture of basic metals and fabricated metal products, except machinery and equipment
7. manufacture of computer, electronic and optical products
8. manufacture of electrical equipment and of machinery and equipment n.e.c.
9. manufacture of transport equipment
10. manufacture of furniture, other manufacturing, and repair and installation of machinery and equipment
11. electricity, gas steam, air conditioning supply, water supply, sewerage, waste management and remediation activities (d-e)
12. construction
13. wholesale and retail trade and repair of motor vehicles and motorcycles
14. transport and storage, except warehousing and support activities for transportation (h except 53)
15. postal and courier activities
16. accommodation
17. food service activities
18. publishing activities
19. motion picture, video and television programme production, sound recording
20. telecommunications
21. IT and other information services
22. real estate activities
23. professional, scientific and technical activities except veterinary activities
24. administrative and support service activities except travel agency, tour operator and other reservation service and related activities (N except 79)
25. travel agency, tour operator and other reservation service and related activities
26. repair of computers and communication equipment

the obtained performance was the following:
As it can be seen, the variability of results is much higher. For some values we high accuracy levels (up to 96%), while in other cases it is completely unacceptable.

### Future work and conclusions

When increasing the detail of the classification, the overall accuracy tends to decrease. Nevertheless, there are specific values for which the prediction can be quite accurate. It could be worthwhile to test the prediction “one vs all” for all values and analyse the results that could be significantly different.
ESSnet Big Data WP2: Webscraping Enterprise Characteristics

Methodological note

The ESSnet BD WP2 performs joint web scraping experiments in multiple countries, using as much as possible the same methodological concepts. The aim is to derive experimental statistics on enterprises from information found on the web, especially the websites of enterprises. It should be noted that these statistics have not reached maturity in terms of harmonisation, coverage or methodology. At this point they are to be treated as the output of research experiments and they do not necessarily align with the official statistics published on this subject.

The result in this note is not published in the experimental statistics output, because of the data set used is not comparable to the population in the official statistics.

Use case: NACE categorizing
Country: SE
Date: <2018-03-27>

Data sources
- Swedish Companies Registration Office provides NACE description by enterprises
- Swedish Tax Agency provide data since 1999
- Business Register maintained at Statistics Sweden

Population

The training data set has 40,219 enterprises and their business activity descriptions from several years; evaluation data contain 11,613 enterprises and activity descriptions, see table below. NACE code is the first digit, representing the highest category level of NACE code.

Figure 1 Description of the data sets

<table>
<thead>
<tr>
<th>NACE code</th>
<th>Training data</th>
<th>Evaluation data</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (Agriculture, forestry and fishing; Mining and quarrying)</td>
<td>105</td>
<td>31</td>
</tr>
<tr>
<td>1 (Manufacturing e.g. food, textiles and paper)</td>
<td>1,304</td>
<td>396</td>
</tr>
<tr>
<td>2 (Manufacturing, e.g. chemicals and computers)</td>
<td>918</td>
<td>274</td>
</tr>
<tr>
<td>3 (Manufacturing, e.g. machinery and furniture; Gas, electricity supply; water supply and waste management)</td>
<td>795</td>
<td>218</td>
</tr>
<tr>
<td>4 (Construction; Wholesale and retail trade etc.; Land transport etc.)</td>
<td>12,071</td>
<td>3,408</td>
</tr>
<tr>
<td>5 (Transportation and storage; Accommodation and food service activities; Publishing activities)</td>
<td>2,810</td>
<td>826</td>
</tr>
<tr>
<td>6 (Information and communication; Financial and insurance activities; Real estate; Legal and accounting)</td>
<td>5,733</td>
<td>1,705</td>
</tr>
<tr>
<td>7 (Professional, scientific and technical activities; Administrative and support service activities, partially)</td>
<td>8,194</td>
<td>2,337</td>
</tr>
<tr>
<td>8 (Administrative and support service activities, partially: Public administration etc.; Education; Human health and social work activities)</td>
<td>4,674</td>
<td>1,366</td>
</tr>
<tr>
<td>9 (Arts, entertainment and recreation; Other service; Activities of households as employers etc.; Activities of extraterritorial organisations and bodies)</td>
<td>3,615</td>
<td>1,052</td>
</tr>
</tbody>
</table>
Methodology
6. Integrate the description text from both Swedish Companies Registration Office and Swedish Tax Agency, match them to the Business Register by the enterprise ID. Hence the NACE code, the description text are linked by the enterprise ID.
7. Pre-processing the text before classifier fitting, i.e. delete stop words and numbers in the text, tokenizing and generating tfidf matrix on the most common 300 words.
8. Using Keras\textsuperscript{12} module, a simple three layers \textit{sequential} classifier is built on the training data. Sequential is a linear stack of layers. The three layers are input-, hidden- and output- layer. 300 features are chosen from the text corpus, they are the neurons in the input layer. The hidden layer apply \textit{relu} activation function with the default Keras weights \( w \) and bias \( b \). The out is presented in the function.
\[
output = relu\left( \sum w_t * x_t + b \right)
\]
\[
relu(x) = \max(0, x)
\]
The last layer apply the \textit{softmax} activation function that calculate the probabilities distribution of the neuron over the total neurons, as the function shown below
\[
P(x) = \frac{\exp(x_t)}{\sum_{t=0}^{n} \exp(x_t)} \quad t = (0, 1, 2, \ldots, n)
\]
9. The loss function can be given as a parameter. \textit{categorical_crossentropy} is given in our classifier, which is the common loss function for multiple classifications.
10. Generalization tells how a classifier learns the concepts from specific data. Good classifier should have learned the general concepts from the given training data and be able to predict on the unseen data. The evaluation data are used for this purpose.

Results
1. With the training data set, the classifier is generated with the metrics as: loss: 26.7%, accuracy: 91%.
2. The performance on the evaluation data is shown in the table. The relative good accuracy score depends on the unbalanced data, i.e. large numbers of true negatives in the categories increase the accuracy.

\textit{Figure 2 Evaluation scores}

<table>
<thead>
<tr>
<th>NACE code</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy %</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>0.12</td>
<td>0.13</td>
<td>84.6</td>
<td>0.13</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>99</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0.05</td>
<td>0.06</td>
<td>94</td>
<td>0.05</td>
</tr>
<tr>
<td>2</td>
<td>0.02</td>
<td>0.03</td>
<td>96</td>
<td>0.01</td>
</tr>
<tr>
<td>3</td>
<td>0.03</td>
<td>0.01</td>
<td>97</td>
<td>0.03</td>
</tr>
<tr>
<td>4</td>
<td>0.34</td>
<td>0.34</td>
<td>59</td>
<td>0.37</td>
</tr>
<tr>
<td>5</td>
<td>0.11</td>
<td>0.07</td>
<td>88</td>
<td>0.10</td>
</tr>
<tr>
<td>6</td>
<td>0.25</td>
<td>0.23</td>
<td>79</td>
<td>0.23</td>
</tr>
<tr>
<td>7</td>
<td>0.25</td>
<td>0.30</td>
<td>69</td>
<td>0.25</td>
</tr>
<tr>
<td>8</td>
<td>0.16</td>
<td>0.15</td>
<td>81</td>
<td>0.15</td>
</tr>
<tr>
<td>9</td>
<td>0.15</td>
<td>0.17</td>
<td>85</td>
<td>0.14</td>
</tr>
</tbody>
</table>

Related figures
\textsuperscript{12} https://keras.io/layers/core/
Figure 3 shows the accuracy history during the training epochs. We can see that after 30 iterations, the accuracy is increasing very little. The iteration in this classifier is good enough.

The confusion matrix on the evaluation dataset is presented above. The Y-axis is the true label, the X-axis is the predicted label. The figures on the diagonal are the true positive predictions for each group. Except group 4, they are unfortunately not the highest in each row. The false positive for each group is the sum of the column value minus the diagonal value. The false negative for each group is the sum of the row value minus the diagonal value. The total minus the sum of the true labelled (row) and the sum of the predicted label (column) is the true negative.

**Limitations and future work**

1. The classifier did not give good result on recall, precision and F1 score in this study. The grouping of data should be improved. At SCB the coding help is useful at two digits level, which has much better division of the activities than the first digit. The section division carried out in this study contain many mixed activities, which brings problems for the classifier.

2. Even more data will help the neural network performs better. The data used here are quite small, especially in several small sections. If the datasets have a balanced data of each group and positive and negative examples, the result will be much better. We expect the method will work even better on big amounts of web pages.

3. It is useful to learn that accuracy, although it is a good metric for quality indication, it is not for unbalanced data. All the metrics should be examined for classifier evaluation.

4. Apply cross-validation and GridSearch for turning parameters in the classifier is the next step. It is even important to improve the word features, e.g. apply word2doc method.

5. The demand for categorizing NACE code at SCB is on the detailed level, 4 or 5 digits. To develop classifiers for more categories than 10 is another challenge.
The ESSnet BD WP2 performs joint web scraping experiments in multiple countries, using as much as possible the same methodological concepts. The aim is to derive experimental statistics on enterprises from information found on the web, especially the websites of enterprises. It should be noted that these statistics have not reached maturity in terms of harmonisation, coverage or methodology. At this point they are to be treated as the output of research experiments and they do not necessarily align with the official statistics published on this subject.

Use case: NACE

Country: UK
Date: <YYYY-MM-DD>

Data sources
Online Registers, Business Websites, ICT survey (called the ‘Ecommerce survey’ in the UK)

Population
The enterprises as defined by the ICT survey (called the ‘Ecommerce survey’ in the UK) (>10 persons, limited NACE)

Methodology
Rather than attempt to recover the full NACE classification, we only attempted to classify businesses according to whether or not they are in the ‘Infrastructure’ SIC group.

16. Extraction of a set of companies in the reference population with known websites and NACE status

17. Scrape website for each company and store the scraped text (note – all scraping carried out according to ONS web-scraping policy, which is available on request)

18. Extract features from the scraped text based on different text representation models:
   a. TF-IDF (term frequency – inverse document frequency)
   b. Doc2Vec
   c. ‘Engineered features’ – the presence/non-presence of certain e-commerce related terms

19. For each text representation model, train three different machine learning classifiers to predict whether the website is in the ‘Infrastructure’ NACE group:
   a. Support Vector Classifier (SVC) with polynomial kernel
   b. Naïve Bayes
   c. Random Forests

20. Evaluate the performance of different text representation models and classifiers
Results

Representing the text using Doc2Vec and using an SVC with a polynomial kernel can predict whether or not a business is in the Infrastructure NACE with 75% accuracy. However, the data cannot support a more detailed classification, potentially as we do not have sufficient training data.

Limitations and future work

Classifier performance would almost certainly be improved by increasing the amount of training data, and further work could focus on this. However, it is unlikely that we’ll be able to predict at a much more detailed level.
Use Case 6: SDGs

ESSnet Big Data WP2: Webscraping Enterprise Characteristics
Methodological note – Test Statistic Produced: No

The ESSnet BD WP2 performs joint web scraping experiments in multiple countries, using as much as possible the same methodological concepts. The aim is to derive experimental statistics on enterprises from information found on the web, especially the websites of enterprises. It should be noted that these statistics have not reached maturity in terms of harmonisation, coverage or methodology. At this point they are to be treated as the output of research experiments and they do not necessarily align with the official statistics published on this subject.

Use case: Sustainable Development Goals (SDG)

Country: UK
Date: 2018-04-04

Data sources
Business Websites, ICT survey (called the ‘E-commerce survey’ in the UK)

Population
The enterprises as defined by the ICT survey (called the ‘E-commerce survey’ in the UK) (>10 persons, limited NACE)

Methodology

21. Based upon previous methodological research by the Office for National Statistics - https://github.com/AlessandraSozzi/MSc-dissertation

22. Extraction of a set of enterprises in the reference population with known websites and ecommerce status

23. Scrape website for each enterprise and store the scraped text (note – all scraping carried out according to ONS web-scraping policy)

24. Extract the human readable text from any individual web page with a URL containing one of more of a set of keywords related to sustainability topics

25. Use the cleaned and tokenised text from all qualifying pages to train a Latent Dirichlet Allocation (LDA) topic model
Results
The topic model attempts to explain the text of observed documents in terms of a smaller number of unobserved groupings of words or “topics”, out of which each document is composed. The method finds 350 relevant pages total, from the websites successfully linked to e-Commerce survey respondents.

The first five words in each topic of a fifteen-topic model, and their weightings, are displayed in the table below. They appear to relate to coherent subjects (eg; the fourth topic may relate to the environment and sustainability). These topics are both potential features for further classification of business types and activities, and results in and of themselves that can be used for analysis of website content.

Limitations and future work
The number of topics must be chosen manually, this step can be automated in future.

A lot of analysis remains to be done before topic models can be used to usefully inform interested parties on the engagement of enterprises with sustainability development goals.

It’s unclear without further research how accurate the keyword list used to identify relevant pages is, this is something that would need thorough checking/validation.