ESSnet Big Data: Dissemination Workshop

WP 2: Webscraping / Enterprise Characteristics
WP2 Session Organization

- Overview of WP2 Objectives and Approach
  - Monica Scannapieco (Istat, Italy)

- Description of Legal aspects related to Web scraping of Enterprise Web Sites
  - Ingered Jansson (SCB, Sweden)

- Pilots
  - URLs Retrieval and Preliminary prediction results for Ecommerce, Social Media and Job Advertisement
    - Giulio Barcaroli, Donato Summa (Istat, Italy)
  - URLs Retrieval and Ecommerce
    - Olav Ten Bosch (CBS, Netherlands)
  - URLs Retrieval, Ecommerce, Social Media
    - Kostadin Georgiev (BNSI, Bulgaria)

- Social Media
  - Jacek Maślankowski (GUS, Poland)
WP2 Session Organization

- Overview of WP2 Objectives and Approach
  - Monica Scannapieco (Istat, Italy)

- Description of Legal aspects related to Web scraping of Enterprise Web Sites
  - Ingered Jansson (SCB, Sweden)

- Pilots
  - URLs Retrieval and Preliminary prediction results for Ecommerce, Social Media and Job Advertisement
    - Giulio Barcaroli, Donato Summa (Istat, Italy)
  - URLs Retrieval and Ecommerce
    - Olav Ten Bosch (CBS, Netherlands)
  - URLs Retrieval, Ecommerce, Social Media
    - Kostadin Georgiev (BNSI, Bulgaria)
  - Social Media
    - Jacek Maślankowski (GUS, Poland)
Web Scraping Enterprise Web Sites

- General Objective: to investigate whether web scraping, text mining and inference techniques can be used to collect, process and improve general information about enterprises.

Enterprises Web sites

National Business Register

Business Statistics Surveys

Crawling  Scaping  Indexing  Searching
Workpackage 2 Participation

- 6 countries:
  - ISTAT, Italy (coordinator)
  - BNSI, Bulgaria
  - CBS, Netherlands
  - GUS, Poland
  - ONS, UK
  - SCB, Sweden

- Same countries in both phases of the ESSnet (SGA-1 and SGA-2)
Main Challenges

- Main challenges:
  1. Massive web scraping
     - Several thousands sites
  2. Information extraction from (almost fully) unstructured data
     - Limited chances to rely on sites’ structure
  3. Define a legal framework

Methodological and IT issues: pilots

Deliverable with current situation and recommendations
Scope: Four Use cases
ESSnet first phase (SGA1)

Use case 1: URLs Inventory
Use case 2: Web sales - ECommerce
Use Case 3: Social Media Presence
Use Case 4: Job Advertisement
Use Case 1: URLs Inventory

- Main Identified Population (ICT Survey):
  - Enterprises with at least 10 Employees
  - Not all of them have a web site, but for those of them that do have it, the URLs of the web sites are not fully available

- The URLs Retrieval problem:
  - Given a set of identifiers (denomination, fiscal code, economic activity, etc.) for the enterprise X, searching the Web for
    - Retrieving a set of associated URLs
    - Estimate (if any) which is the URL corresponding to the web site of X
The goal of this use case is to predict whether an enterprise provides or not web sales facilities on its website.

<table>
<thead>
<tr>
<th>Use of a Website or Home Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>B7.</strong> In January 2013, did your enterprise have a Website or Home Page? (Filter question)</td>
</tr>
<tr>
<td><strong>B8.</strong> In January 2013, did the Website or Home Page have any of the following?</td>
</tr>
<tr>
<td>a) Online ordering or reservation or booking, e.g. shopping cart</td>
</tr>
<tr>
<td>b) A privacy policy statement, a privacy seal or certification related to website safety</td>
</tr>
<tr>
<td>c) Product catalogues or price lists</td>
</tr>
<tr>
<td>d) Order tracking available online</td>
</tr>
<tr>
<td>e) Possibility for visitors to customise or design the products</td>
</tr>
<tr>
<td>f) Personalised content in the website for regular/repeated visitors</td>
</tr>
<tr>
<td>g) Advertisement of open job positions or online job application</td>
</tr>
</tbody>
</table>

*Optional*
Use Case 3: Social Media Presence

- The goal of this use case is to provide information on existence of the particular enterprise in social media (mainly Twitter and Facebook)
Use Case 4: Job Advertisement

- The goal of this use case is to investigate how enterprises use their websites to handle the job advertisements, and in particular if they publish job advertisement or not.
Method of work

- Candidature by participating countries to develop *pilots*
- Shared methodological process
  - Identification of a reference architecture for each use case in terms of «logical building blocks»
  - Mapping of each pilot to such an architecture in terms of specific technological solutions
- With one exception due to legal constraint (SCB-Sweden)
  - Constraints to storage scraped data
Use case 1: Istat’s pilot
**Use case 1:** CBS’s pilot
Legal Framework

- **Deliverable 2.1 **«Legal aspects related to Web scraping of Enterprise Web Sites»
  - Each country worked in close cooperation with internal legal offices (in case of Sweden also external)
  - Country view and a comparative analysis in the report
  - Dedicated presentation on that
WP2 Session Organization

- Overview of WP2 Objectives and Approach
  - Monica Scannapieco (Istat, Italy)

- Description of Legal aspects related to Web scraping of Enterprise Web Sites
  - Ingerged Jansson (SCB, Sweden)

- Pilots
  - URLs Retrieval and Preliminary prediction results for Ecommerce, Social Media and Job Advertisement
    - Giulio Barcaroli, Donato Summa (Istat, Italy)
  - URLs Retrieval and Ecommerce
    - Olav Ten Bosch (CBS, netherlands)
  - URLs Retrieval, Ecommerce, Social Media
    - Kostadin Georgiev (BNSI, Bulgaria)
  - Social Media
    - Jacek Maślankowski (GUS, Poland)
Content of the report

Deliverable 2.1 from work package 2

- Relevant legislation at EU level
- Present situation in the participating countries
  - Relevant regulations
  - Challenges and recommendations
- Case study: access to job portals
- Proposal for a Netiquette
Main legal issues

- Copyright protection, including protection of data bases as property
- Privacy and personal data protection
- Ethical principles for web scraping
Present situation

- Regulations are similar between countries
  - All countries have regulations on copyright
  - All countries have a national statistical law
- Difference in interpretation
  - Italy, Netherlands, Bulgaria
  - UK, Poland, Sweden
- Ethical codes exist in some countries
Recommendations and strategies

- Transparency, clear policy, communication through web sites (avoid consent)
- Minimizing response burden
- New legislation for use of non-commercial data
- Cooperation with web site owners
  - Give aggregates back, informative meetings
- Ethical guidelines
Netiquette

- Developed by CBS and ONS
- Key principles for good practice in web scraping
- Proposal for ESS, but more stringent principles can be adopted by the NSIs
Netiquette

- Respect the ‘robots.txt’ robots exclusion protocol and nofollow links
- Identify yourself in the user-agent string, and provide a means for contacting you
- Be transparent about your web-scraping activities
- Inform website owners if a considerable amount of data is collected on a regular basis.
- Seek to minimise burden on website owners
- Do not crawl sensitive areas of a website
- Only scrape data for the production of official statistics within the scope of your mandate, and do not re-use or distribute the data for any other purpose
- Handle web-scraped data securely according to all relevant protocols and laws
WP2 Session Organization

- Overview of WP2 Objectives and Approach
  - Monica Scannapieco (Istat, Italy)

- Description of Legal aspects related to Web scraping of Enterprise Web Sites
  - Ingered Jansson (SCB, Sweden)

- Pilots
  - URLs Retrieval and Preliminary prediction results for Ecommerce, Social Media and Job Advertisement
    - Giulio Barcaroli, Donato Summa (Istat, Italy)
  - URLs Retrieval and Ecommerce
    - Olav Ten Bosch (CBS, Netherlands)
  - URLs Retrieval, Ecommerce, Social Media
    - Kostadin Georgiev (BNSI, Bulgaria)
  - Social Media
    - Jacek Maślankowski (GUS, Poland)
Use Cases: URLs retrieval

• Enterprise URLs Inventory
  – Generate a URL inventory of enterprises for the Business register, update and maintain the inventory.
  – Participating countries: all (IT, UK, NL, SE, PL, BG)
Use Cases: prediction

• Web sales facility (e-commerce)
  – The goal of this use case is to predict whether an enterprise provides or not web sales facilities on its website
  – Participating countries: IT, UK, BG

• Job advertisements on enterprises’ websites
  – The goal of this use case is to investigate how enterprises use their websites to handle the job advertisements. The goal is to search answers for questions like do they publish the advertisements on their own websites or do they entrust job mediators. The result can be used for examining if the enterprises’ websites can be used as information channels for WP1.
  – Participating Countries: UK, PL, SE, IT

• Social Media Presence on Enterprises webpages
  – The goal of this use case is to provide information on existence of the particular enterprises in social media. It means that all social media accounts will be taken into account. Especially we will search for Facebook and Twitter accounts as well as LinkedIn and less like Instagram.
  – Participating Countries: BG, PL, IT, SE
Internet access  Storage  Data preparation  Analysis

URL searcher  
Retrieved URLs  
Word filters (eg. stopwords)  
Language specific lemmatization  
Feature extraction  
Term document matrix generation  
Build training & test sets  
Train classifier  
Apply classifier  
Information Extraction: NLP  
Information Extraction: Deep learning

Scraper  
Scraped content  
Essnet on Big Data Dissemination Workshop – Sofia 22/24 February 2017
Use Cases: prediction

- Survey on Enterprise ICT usage (2015 round)
- Learners employed: logistic model, classification trees, random forests, boosting, bagging, neural net, Naive Bayes, SVM

- Performance indicators:
  1. **accuracy** (rate of correct predictions on the total of cases)
  2. **sensitivity** (or **recall**: rate of true positives on total number of positives)
  3. **specificity** (rate of true negatives on total number of negatives)
  4. **difference** between (i) proportion of positives on the total calculated on observed and (ii) proportion of positives calculated on predicted
  5. **F1-measure**: harmonic mean of **recall** and **precision** (rate of true positives on the total of predicted positives)
  6. **p-value** related to the test: [Accuracy > Non Informative Rate]
## Learners employed for e-commerce

<table>
<thead>
<tr>
<th>Learner</th>
<th>Indicators (e-commerce)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Sensitivity</td>
<td>Specificity</td>
<td>Est. diff.</td>
<td>F1 measure</td>
<td>p-value</td>
</tr>
<tr>
<td>1. Logistic</td>
<td>0.83</td>
<td>0.53</td>
<td>0.89</td>
<td>0.01</td>
<td>0.53</td>
<td>0.01625</td>
</tr>
<tr>
<td>2. Naïve Bayes</td>
<td>0.80</td>
<td>0.46</td>
<td>0.87</td>
<td>0.00</td>
<td>0.46</td>
<td>0.99490</td>
</tr>
<tr>
<td>3. Rand Forest</td>
<td>0.83</td>
<td>0.53</td>
<td>0.90</td>
<td>0.01</td>
<td>0.55</td>
<td>0.00006</td>
</tr>
<tr>
<td>4. Bagging</td>
<td>0.82</td>
<td>0.44</td>
<td>0.90</td>
<td>0.03</td>
<td>0.48</td>
<td>0.11520</td>
</tr>
<tr>
<td>5. Boosting</td>
<td>0.81</td>
<td>0.50</td>
<td>0.88</td>
<td>0.00</td>
<td>0.50</td>
<td>0.56530</td>
</tr>
<tr>
<td>6. Neural Net</td>
<td>0.82</td>
<td>0.52</td>
<td>0.89</td>
<td>0.01</td>
<td>0.52</td>
<td>0.10180</td>
</tr>
<tr>
<td>7. SVM</td>
<td>0.83</td>
<td>0.64</td>
<td>0.88</td>
<td>0.01</td>
<td>0.59</td>
<td>0.00018</td>
</tr>
<tr>
<td>8. SLAD</td>
<td>0.84</td>
<td>0.62</td>
<td>0.90</td>
<td>0.01</td>
<td>0.60</td>
<td>0.00018</td>
</tr>
</tbody>
</table>
Prediction performance for the three use cases

<table>
<thead>
<tr>
<th>Variable</th>
<th>Indicators (random forests)</th>
<th></th>
<th></th>
<th>Est. (diff.)</th>
<th>F1 measure</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Sensitivity</td>
<td>Specificity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e-commerce</td>
<td>0.83</td>
<td>0.53</td>
<td>0.90</td>
<td>19.1 (0.01)</td>
<td>0.55</td>
<td>&lt; 0.0006</td>
</tr>
<tr>
<td>Online job application</td>
<td>0.80</td>
<td>0.64</td>
<td>0.86</td>
<td>29.1 (0.01)</td>
<td>0.64</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>Presence in social media</td>
<td>0.72</td>
<td>0.65</td>
<td>0.78</td>
<td>41.8 (0.01)</td>
<td>0.67</td>
<td>&lt; 2e-16</td>
</tr>
</tbody>
</table>
Improvements of prediction

Areas of improvement (SGA2):

• Web scraping: use of OCR to capture images content

- Text processing: use of Natural Language Processing to individuate most relevant terms to use as features in learners

- Machine Learning: use of Deep Learning and other advanced learners
WP2 Session Organization

- Overview of WP2 Objectives and Approach
  - Monica Scannapieco (Istat, Italy)

- Description of Legal aspects related to Web scraping of Enterprise Web Sites
  - Ingerged Jansson (SCB, Sweden)

- Pilots
  - URLs Retrieval and Preliminary prediction results for Ecommerce, Social Media and Job Advertisement
    - Giulio Barcaroli, Donato Summa (Istat, Italy)
  - URLs Retrieval and Ecommerce
    - Olav Ten Bosch (CBS, Netherlands)
  - URLs Retrieval, Ecommerce, Social Media
    - Kostadin Georgiev (BNSI, Bulgaria)
  - Social Media
    - Jacek Maślankowski (GUS, Poland)
CBS: URLs identification Enterprises

- Goal: find missing URLs of enterprises using web scraping and machine learning
- 1/3 of 1.5 Million enterprises in Dutch business register have a URL

- Method:
  - Automated Google Search, 5 different search queries per enterprise using combinations of enterprise name and address
  - Calculate scores (features)
  - Train and test model
  - Apply model

- A test on a sample of 1000 enterprises (>10 persons) indicated:
  - The search found in 70% of the cases a valid URL
  - The model predicted URL validity correctly in 73% of the cases
Goal: automatically detect web shops from a list of 1200 foreign companies paying Dutch VAT

URL finding step found in 97% of the cases a domain with indication of correctness (Good, Fair, Mediocre, Poor)

About 1000 sites crawled to determine E-commerce activities automatically via text analysis

Manual check on sample indicated:
- URL finding in class “Good”: accuracy of about 90%
- Webshop detection algorithm: accuracy of about 85%
Conclusions and further work

- Finding URLs using a search engine and machine learning proofed to be valuable
- Can be used as a starting point for retrieving additional characteristics of a population / developing new indicators
- URL finding and text mining successfully applied to detect E-commerce of foreign companies

Further work:
- Improve URL finding accuracy by additional scraping
- Experiment with other search queries and search engines
- Scale up experiments
- Keep methodology aligned with other NSI’s as much as possible: *circumstances may differ but concepts are the same*
WP2 Session Organization

- Overview of WP2 Objectives and Approach
  - Monica Scannapieco (Istat, Italy)

- Description of Legal aspects related to Web scraping of Enterprise Web Sites
  - Ingered Jansson (SCB, Sweden)

- Pilots
  - URLs Retrieval and Preliminary prediction results for Ecommerce, Social Media and Job Advertisement
    - Giulio Barcaroli, Donato Summa (Istat, Italy)
  - URLs Retrieval and Ecommerce
    - Olav Ten Bosch (CBS, Netherlands)
  - URLs Retrieval, Ecommerce, Social Media
    - Kostadin Georgiev (BNSI, Bulgaria)
  - Social Media
    - Jacek Maślankowski (GUS, Poland)
URLs Search

SBR enterprises with 10 and more employees
26836 enterprises
20649 e-mails, 2006 urls
Addresses, Phones, NACE codes...

Search in JABSE, Google – BNSI software

Search in Bing – ISTAT software
UrlSearcher

9809 urls were found

36.6 % of enterprises with 10 and more employees have websites
**E-commerce**

PHP script with 3 logics (4 positive and 1 negative lists of key words)
- e-commerce: 1139
- e-commerce v2: 1048
- e-commerce v3: 662

manual checked: 856 e-commerce

10% sample: 27 e-commerce

856 + 27 * 10 = 1126 e-commerce

11.5% of enterprises (10+ employees) with websites do e-commerce

4.2% of enterprises with 10 and more employees do e-commerce
Social Media

PHP script (crawling only the first page of enterprise website)
facebook: 2356
twitter: 922
linkedin: 560
google: 871
youtube: 527
pinterest: 139
instagram: 127

24.9% of enterprises (10+ employees) with websites have at least one social media profile

9.1% of enterprises with 10 and more employees use at least one of the listed social media
Use of ISTAT software, next steps

- UrlSearcher is Cyrillic friendly
- Run ISTAT software (RootJuice, etc.) with the results of UrlSearcher (also possibly extract the scraped content of Google and JABSE into UrlSearcher’s seed files to run in RootJuice)
- Compare the results of BNSI findings with the results of ISTAT software implementation
WP2 Session Organization

- Overview of WP2 Objectives and Approach
  - Monica Scannapieco (Istat, Italy)

- Description of Legal aspects related to Web scraping of Enterprise Web Sites
  - Ingered Jansson (SCB, Sweden)

- Pilots
  - URLs Retrieval and Preliminary prediction results for Ecommerce, Social Media and Job Advertisement
    - Giulio Barcaroli, Donato Summa (Istat, Italy)
  - URLs Retrieval and Ecommerce
    - Olav Ten Bosch (CBS, netherlands)
  - URLs Retrieval, Ecommerce, Social Media
    - Kostadin Georgiev (BNSI, Bulgaria)

- Social Media
  - Jacek Maślankowski (GUS, Poland)
Social media presence - use case PL: general overview

(1) Web scraping
- HTML File
- Extracting all links to find social media links
- Next iteration if the link is not present

(2) Twitter API
- Scrap the tweet
- Process the post/tweet by Machine Learning algorithm

(3) Machine Learning
- Classify the tweet:
  - based on C11 (ICT 2015)

with open ('wp2_social.csv', 'a') as plikcsv:
  kolumny=['URL', 'Facebook', 'Twitter', 'Youtube', 'LinkedIn', 'Instagram', 'GooglePlus']
  zapis=csv.DictWriter(plikcsv, delimiter=';', dialect=csv.excel, fieldnames=kolumny)

C10. Does the Website have any of the following? (not solely used for paid adverts)
- Description of goods or services, price lists
- Online ordering or reservation or booking, e.g. shopping cart
- Possibility for visitors to customise or design online goods or services
- Tracking or status of orders placed
- Personalised content in the website for regular/recurrent visitors
- Links or references to the enterprise’s social media profiles
- Advertisement of open job positions or online job application
  - Optional

The following question (C11) should only be answered if any of the above social media is used (i.e. C10 has at least one “Yes”).

C11. Does your enterprise use any of the above mentioned social media? (not solely used for paid adverts)
(a) Social networks (e.g. Facebook, LinkedIn, Xing, Viadeo, Yammer, etc.)
(b) Enterprise’s blog or microblogs (e.g. Twitter, Present.ly, etc.)
(c) Multimedia content sharing websites (e.g. YouTube, Flickr, Picasa, SlideShare, etc.)
(d) Wiki based knowledge sharing tools

https://circabc.europa.eu/ie/a/a39ae859-8a16-43d6-8020-a06d3f2c91/Questionnaire%20ENT%202015.pdf
Internet access

Storage
- URL list
  - CSV file

Data preparation
- Feature extraction
  - HTMLParser/BeautifulSoup
- Tokenization
  - HTMLParser/BeautifulSoup
- Data parser
  - HTMLParser/BeautifulSoup

Analysis
- Machine Learning
  - Build training & test sets
    - CSV file
  - Train classifier
    - sci-kit learn
  - Apply classifier
    - sci-kit learn
- Information Extraction: Deep learning
- Saving results
  - CSV file

Scraped website
- In-Memory

Scraper
- Python/Spark (pyspark)
- Tweepy (Twitter API)

Social media URLs and Training Set
- CSV file
Results and result sets

• Social media URL finding

• Machine learning

' LIVE!!! GRAMY NA DRAGON-SURVIVAL.EU REKRUTACJA DO GILDII GRIM! : https://t.co/IkIUZs2zzS przez @YouTube' => inne
'Witamy w nowym roku i dziękujemy za korzystanie z naszego medium' => inne
'Na te pytania startupowiec powinien odpowiadać każdego dnia. Chyba, że nie chce się rozwijać' => inne
Algorithms and technological obstacles

- Difficulties with accessing the data, e.g., from Facebook
- Some social media links does not exist, e.g., Present.ly
- The classification should reflect the real usage of social media
- Providing the reliable training set
- Improving Machine Learning algorithms
  - Multinomial Naive Bayes - MultinomialNB()
    - `Out[131]: 0.7142857142857143`
    - Precision: 0.88, Recall: 0.88, F1-score: 0.88, Support: 16
  - Stochastic Gradient Descent - SGDClassifier()
    - `Out[132]: 0.80952380952380953`
    - Precision: 0.81, Recall: 0.81, F1-score: 0.81, Support: 21
  - ...
Thanks for the attention

By the WP2 Team