WP2 Meeting in Gdansk: SGA2 Kick OFF

Agenda of the Meeting: 5th October

- 1. Welcome / Introduction / Overview 9:00
- 2. Country updates – WP2 related activities and software demonstrations (format 15 minutes presentation followed by 10 minutes discussion) 9:30
- Coffee break 11:00
- 3. Discussion on deliverables for experimental statistics (tables, indicators – who can deliver the data) 11:30
- 4. Review of SGA-2 plans 12:30
- Lunch 13:00
- 5. Discussion on SGA-2 activities and deliverables 14:00
- Close 17:00

Agenda of the Meeting: 6th October

- 1. Discussion on indicators and tables 9:00
- Coffee break 10:30
- 2. Discussion on countries responsibilities for SGA-2 pilots 11:00
- 3. AOB 12:00
- Coffee break with sandwiches 12:30
- Close 13:00

Expected Outputs from the Meeting

- Revision/confirmation of use cases planned for SGA2
- Definition of new approaches for SGA2
- What we release as result of SGA1 and SGA2 (output)
- Overall workplan and country responsibilities

SGA1 Deliverables

- Deliverable on Legal Aspects of Web Scraping of Enterprises Web sites
  - https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/a/a0/WP2_Deliverable_2_1_15_02_2017.pdf
- Deliverable on methodological and IT issues and solutions
  - https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/b/b9/WP2_Del2.2_20170731.pdf

Deliverable on Methodological and IT Issues and Solutions

- 16 different pilots were implemented by participanting countries

Use Cases\Countries	IT	SE	UK	NL	BG	PL
1: URLs retrieval	**x**		x	x	x	x
2: Ecommerce	**x**		x	x	x	
3: Job advertisement	x	**x**	x			
4: Social Media	x				x	**x**

Bulgaria implemented two pilots for use case 1 (one using Istat's sw)

Both Deterministic and Machine learning methods used

- Both deterministic as well as machine learning approaches have been applied
- Still room to evaluate quality of results

	IT	**SE**	**UK**	**NL**	**BG**	**PL**
1 URLs retrieval	ML	-	D	ML	ML, D	D
2 Ecommerce	ML	-	ML	D	D	-
3 Job Advertisements	ML	ML	ML	-	-	-
4 Social Media	ML	-	-	-	D	D

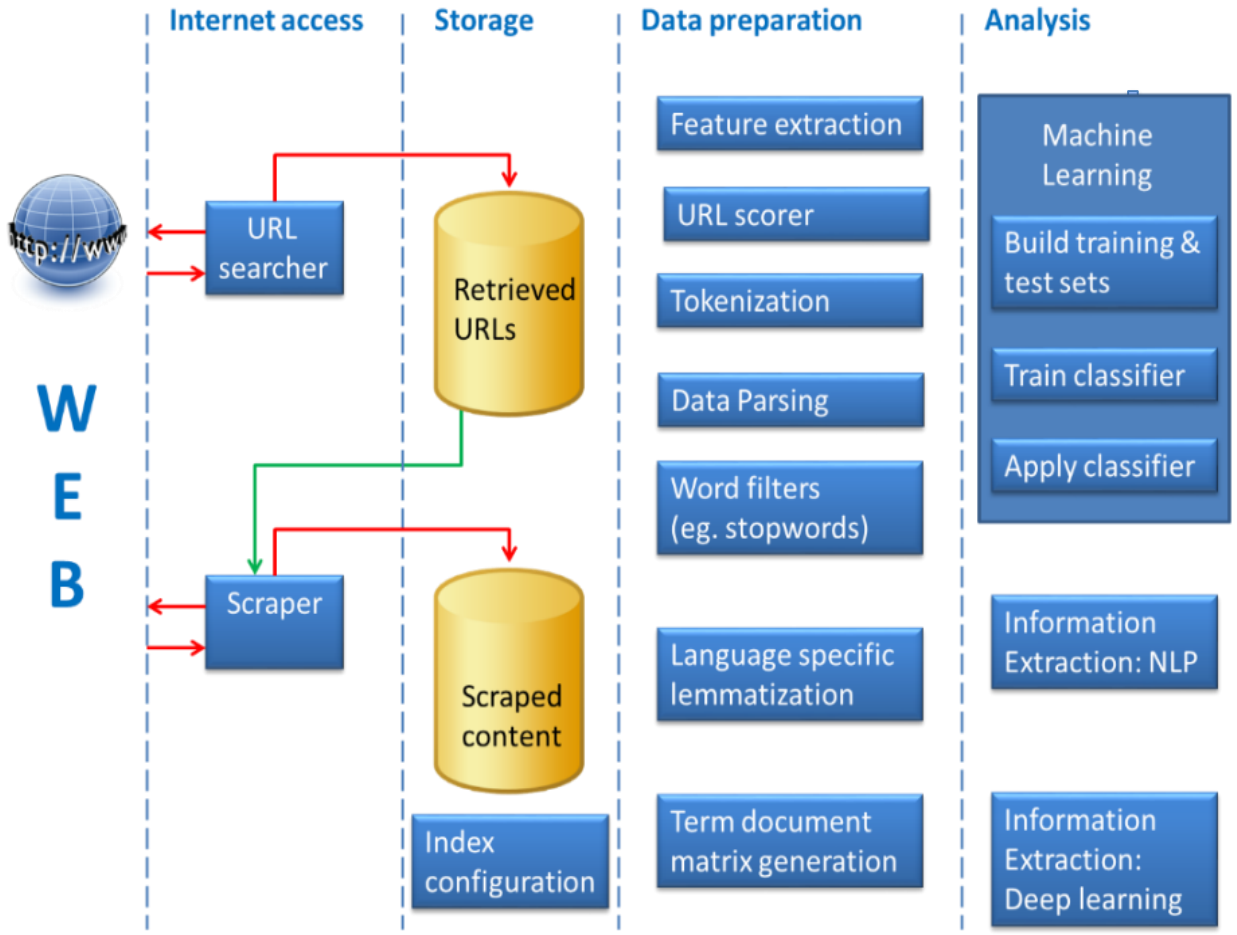
IT solutions

Table 1. Overview of programming languages, libraries and tools used in pilots

Use Case	BG	IT	NL	PL	SE	UK
URL Retrieval	(1) PHP language Java URL searcher Jabse Search API Google Custom Search API Bing Search API (2) ISTAT URL Retrieval Java language	Java language R language	Python language JavaScript language NodeJS ElasticSearch engine Natural library for NodeJS Scikit-learn library for Python	Java language (ISTAT URLSearcher) Python language		Python language Py-whois API Bing API
E-commerce	PHP language	Java language R language TreeTagger library for lemmatization SnowballStemmer for stemming Crawler4J	Python language R language Scrapy for webscraping			Python language Scrapy for webscraping *NLTK python Library* *ML (Naïve Bayes)*
Soci				execute scripts		

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 PHP programming language.

IT solutions



Reviews

3. Discuss the effects of variation

There is a lot of variation in methodology and technology that you use in the pilots e.g., tables on page 12 and page 15. Would we expect to get the same results/outputs if we used another set of “tools” e.g. table 1 if you would apply the same set of tools used in NL in IT would that yield the same results or would we expect different results?

4. What are the future challenges?

Obviously, a lot of work remains to be done when it comes to web scraping and there are things such as ethics, selection biases beside development of technology that affects strategy in this domain. Add a discussion on future challenges.

Outputs as Experimental Statistics

- URL Retrieval - Rate(s) of retrieved URLs from an enterprises' list
- Ecommerce - Rate(s) of enterprises engaged in ecommerce from enterprises websites
- Job advertisements - Rate(s) of enterprises that have job advertisements on their websites
- Social media presence:
  - Rate(s) of enterprises that are present on social media from their websites
  - Percentage of enterprises using Twitter for a specific purpose, i.e.
    - a) Develop the enterprise's image or market products (e.g. advertising or launching products, etc);
    - b) Obtain or respond to customer opinions, reviews, questions;
    - c) Involve customers in development or innovation of goods or services
    - d) Collaborate with business partners (e.g. suppliers, etc.) or other organisations (e.g. public authorities, non governmental organisations, etc.)
    - e) Recruit employees
    - f) Exchange views, opinions or knowledge within the enterprise

Outputs as Experimental Statistics

- Predicted values can be used for a twofold purpose:
  - at unit level, to enrich the information contained in the register of the population of interest;
  - at population level, to produce estimates.

SGA 2

- Official start date: August 2017
- Planned end date: May 2018
- New Use cases
  - sustainability reporting on enterprises' websites
  - identifying categories relevant to Enterprises' types of activity (NACE)
- To be evaluated
  - web sites accessibility
  - support to Euro group register

SGA2 Tasks according to the proposal

Task 1 –Data access: URLs Retrieval

- 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.
- Evaluation of the quality of the results, also with respect to a tradeoff between automated and manual tasks.

Task 2 – Data Handling: Application of the Web scraping task

- For all the identified use cases, the scraping task will be carried out on the URLs of the respective reference populations (either available or retrieved according to Task 1).
- Construction of the database of the scraped data and reporting of its characteristics and metadata (reference population metadata, statistics on dimensionality, etc.).

SGA2 Tasks according to the proposal

Task 3 – Methodology and Technology and Statistical Output: Application of the text and data mining techniques

- For all identified use cases, apply text and data mining techniques (learners) to predict characteristics of the enterprises. Evaluate quality indicators for some of them (e.g. accuracy, sensitivity, specificity). On the basis of the quality indicators, choose the best predictor.
- 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.
- 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 (evaluated on the basis of survey benchmark data like the ones available from “Survey on ICT usage and e-Commerce in Enterprises”) .

- **Task 4** – Future Perspectives: Testing information extraction techniques and Applicability of Findings
  - This task has the purpose of testing whether some of the results obtained in task 3 can be improved by using different approaches to text processing. Indeed, the techniques used in task 3 adopt a “bag of words” approach to model text data resulting from the scraping activity. In this task, we will test techniques for information extraction world in order to go beyond the bag of words approach. These techniques include Natural Language Processing techniques (like e.g. the ones offered by the toolkit NLTK). In particular we will:
    - Identify some use cases (or part of them) for which it can be supposed an improvement by means of information extraction techniques.
    - Set up of an information extraction procedure to be applied to the selected cases.
    - Application of the procedure and evaluation of the results. In this step a detailed analysis will be carried out on the quality improvement that can be obtained by complementing text and data mining procedures with information extraction techniques.
    - In addition, these task will assess the applicability of the WP ouputs, in particular considering production scenarios making use of the proposed approaches.

SGA 2 Deliverable

Del		Due Date	Review Board
2.1	Final report describing final procedures set up for accessing Enterprises web sites and use them for the different uses cases	May 2018	15 April 2018

SGA2 Meeting

- SGA2 Meetings
  - Internal workshop on enterprise web scraping (Oct 2017)
  - Joint internal workshop of WP 1 and WP 2, with report (March 2018)
    - in accordance with Nigel not to be held as a joint workshop

New Methods

- Access to images on web sites:
  - Refinement of some use cases (e.g. Ecommerce)
  - Self-contained use case (e.g. Social Media)
- Testing new techniques
  - Word embeddings
  - Deep learning

Desiderata for wp2 in sga2

1. More joint work
   - In SGA1 some software sharing
   - Less management meetings and more technical sharing
2. Set of guidelines for using enterprises web sites for enriching business registers (joint paper?)
3. Demo session at the final workshop