ML 2022: Web-scraping Data Theme Group

webinar - 30/11/2022

Michael Reusens - Statistics Flanders
Bilal Kurban – Turkish Statistical Institute
Klaudia Peszat – Statistics Poland
Contact: michael.reusens@vlaanderen.be

Flanders
State of the Art

Overview

- Theme group overview
- Web scraping use cases
  - Turkish Statistical Institute
  - Statistics Poland
  - Statistics Flanders
- Conclusions

Theme group overview

- Lots of attention for web scraped data from statistical organisations
  - Frequency, survey burden, …
- Also lots of questions
  - Methods, quality, …
- Theme group goal: **Facilitate the adoption of web scraped data in official statistics via**
  - **Knowledge sharing**
  - **Parallel implementation**

Knowledge sharing

- Existing resources for web scraping in official statistics
- Use case presentations
- Advanced natural language processing techniques
- Quality issues related to web scraped methods;
- …

Parallel implementation

- 3 organisations with web scraping goals
  - Statistics Poland
  - Turkish Statistical Institute
  - Statistics Flanders
- Each organisation implements their own web scraping project
- Monthly progress report and feedback sessions

Use case – Turkish Statistical Institute

- **Projects**
  1. Scrape an ICT variable

     Obtaining an ICT variable (social media presence) by web scraping and comparing it to the TurkStat ICT Usage in Enterprises survey results.
  2. Gain insights from an open-ended question

     Extracting insights from responses to an open-ended question in biotechnology survey without examining individual responses one by one.
  3. Create a framework for government R&D survey

     Through web scraping, Identifying public organizations that use R&D-related terms on their web pages at least once and creating a government R&D survey framework with them.

Use case – Turkish Statistical Institute

- Project 1 – Scrape an ICT variable: "does the enterprise website have links to its social media profiles"

		Web Scraping		Total
		Yes	No	
Survey	Yes	392 (36%)	325 (30%)	717
	No	**89 (8%)**	281 (26%)	370
Total		606	481	1,087

Use case – Turkish Statistical Institute

- Project 2 – Gain insights from an open-ended question: "Briefly inform about the biotechnology activities carried out by your enterprise and the techniques and applications it uses"

STATISTICS FLANDERS

Flanders State of the Art

Use case – Turkish Statistical Institute

- Project 3 – Create a framework for government R&D survey

STATISTICS FLANDERS

Flanders State of the Art

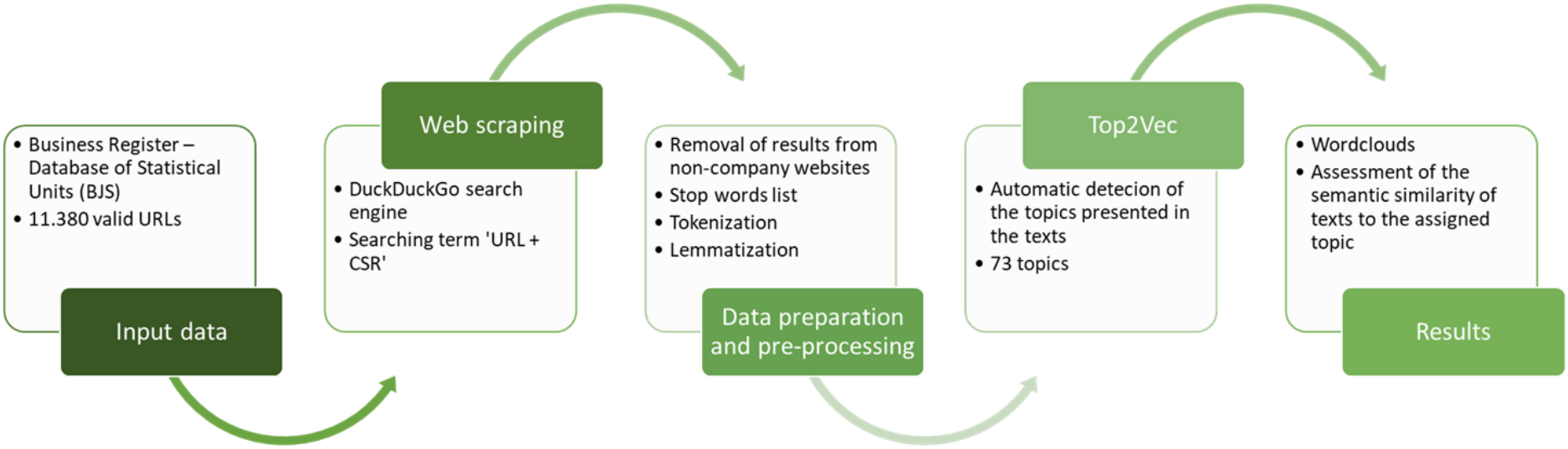
Use case – Statistics Poland

Number of companies undertaking activities in the field of *corporate social responsibility* (CSR).

CSR defined as „a company's sense of responsibility towards the community and environment (both ecological and social) in which it operates”.

Companies express this citizenship:
(1) through their waste and pollution reduction processes,
(2) by contributing educational and social programs, and
(3) by earning adequate returns on the employed resources” (*the Business Dictionary*).

Use case – Statistics Poland



Use case – Statistics Poland

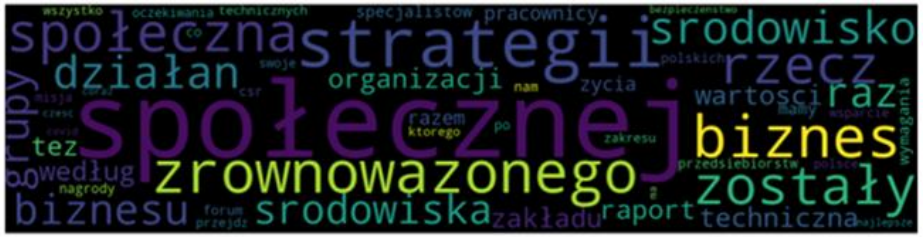


Figure 1. The wordcloud for the topic 0 (in Polish)

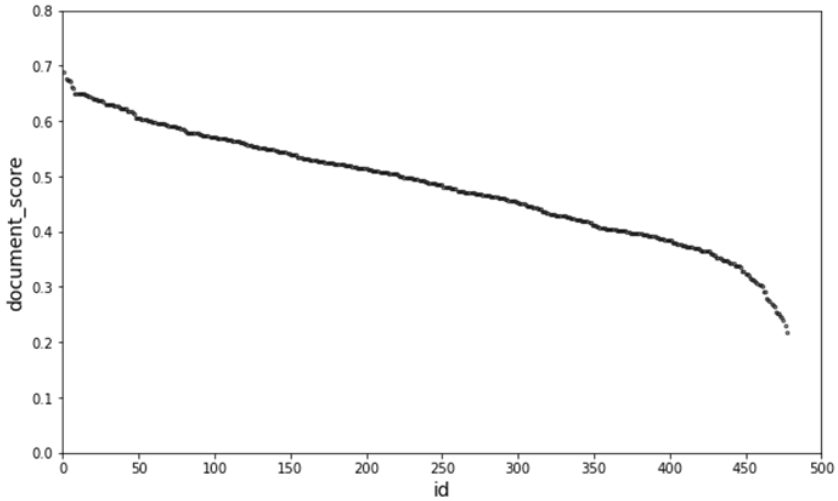


Figure 2. The scatter plot for the topic 0

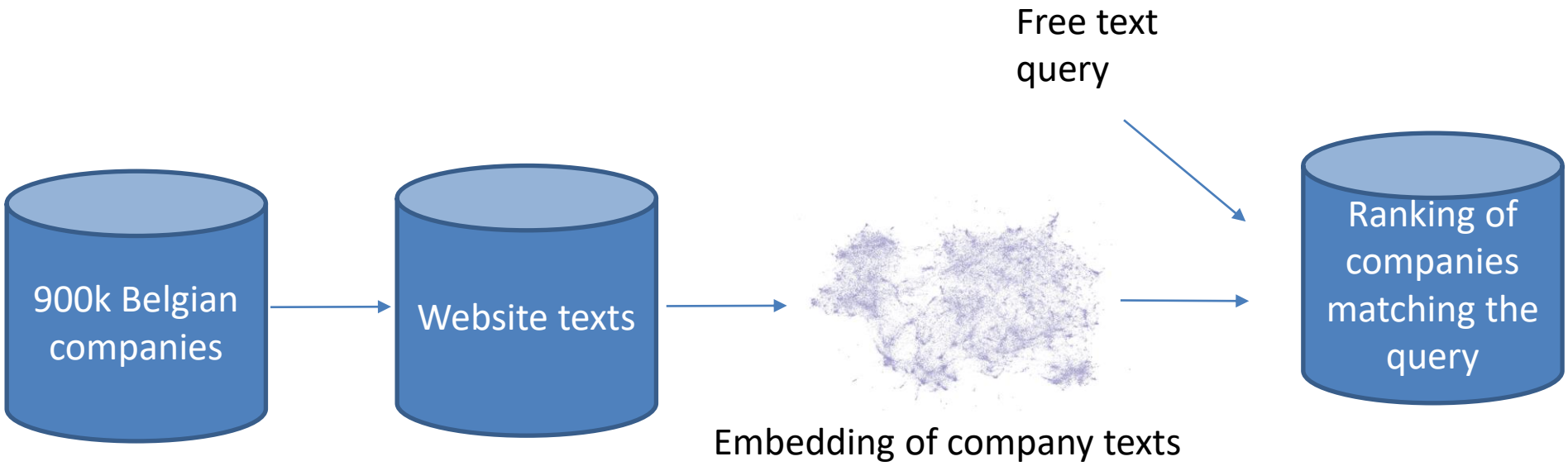
- The most related to the CSR concept topic consists of 477 semantically similar texts.
- The number of unique companies assigned to this topic is 280 (2.5% of all companies).

Use case – Statistics Flanders

- Automatic categorisation of company activity
  - Better statistics (frequency, timeliness, accuracy,...)
  - Lower response burden

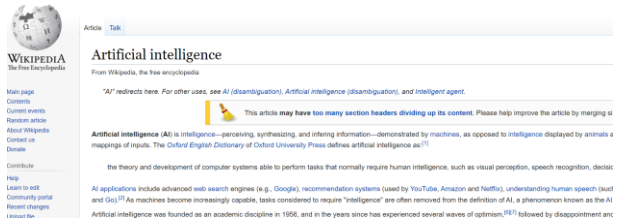
Use case – Statistics Flanders

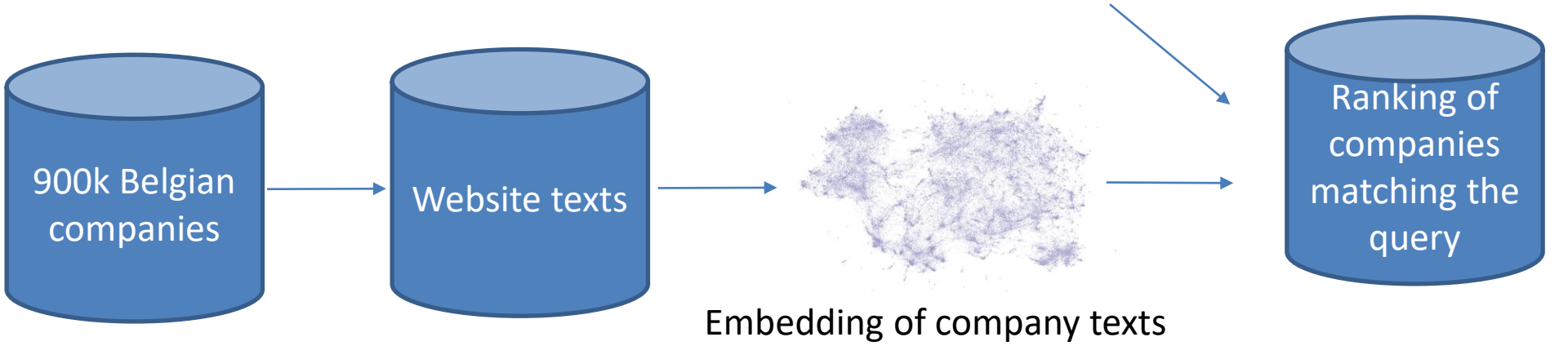
- Automatic categorisation of company activity
  - Better statistics (frequency, timeliness, accuracy,...)
  - Lower response burden



Use case – Statistics Flanders

- Automatic categorisation of companies
  - Better statistics (coverage, frequency, accuracy)
  - Lower response burden





Embedding of company texts

Use case – Statistics Flanders

"**Artificial intelligence** (**AI**) is intelligence—perceiving, synthesizing, and infering information—demonstrated by machines, as opposed to intelligence displayed by animals and humans. Example tasks in which this is done include speech recognition, computer vision, translation between (natural) languages, as well as other mappings of inputs. The *Oxford English Dictionary* of Oxford University Press defines artificial intelligence as:[1]..."

Query the scraped texts

score	kbonr	url
0.433342	764955559	http://minetech.com/
0.420682	886962456	http://www.questra-consulting.be/
0.380442	833198425	http://www.infofarm.be/
0.362002	686858087	http://www.boltzmann.be/
0.349857	741648439	http://www.nerai.io/
0.349857	726888207	http://nerai.io/
0.344207	568854619	http://biztory.com/
0.344207	690558638	http://biztory.com/

STATISTICS FLANDERS

Flanders State of the Art

Use case – Statistics Flanders

- 2 experimental categorisations
  - Artificial Intelligence
  - Bio economy

- First results highly promising
  - Purely unsupervised

- Ongoing data labelling to assess quality
  - Lead generation for government institute for innovation

Conclusions and lessons learned

- Knowledge sharing
  - Inspirational
  - Efficient start for implementors
- Co-implementation
  - Faster solution for problems
  - Peek behind the curtain on very practical approach in different organisations
- --> Huge benefits of collaboration

Thank you for your attention

- Questions?

- michael.reusens@vlaanderen.be