Individual report: Statistics Flanders

The individual report from Statistics Flanders is written in paper-form, and was submitted for presentation at NTTS 2023.

Unsupervised ranking and categorisation of companies using web scraping and machine learning

Michael Reusens, michael.reusens@vlaanderen.be

Keywords: web scraping, natural language processing, official statistics, machine learning

INTRODUCTION

This paper presents a method to automatically categorise companies based on the text scraped from their website. The method is demonstrated by applying it to categorising companies as being active in the domain of artificial intelligence (AI) or not.

This study has been set up to evaluate if the text scraped from company websites can be used as a complement to existing data sources to produce official statistics. Today, most official statistics are produced using survey data (such as the Community Innovation Survey [1]) and administrative data sources. The web scraping methodology proposed in this paper has the following advantages compared to these traditional data sources. First, scraping instead of surveying alleviates the response burden of companies. Second, using the web scraping method, companies can be surveyed at any desired frequency. This leads to more up-to-date data, resulting in increased quality and timeliness of the company statistics. Finally, our approach can be generalised to any categorisation of interest, resulting in the ability of statistical organisations to create new company statistics relatively quickly.

In recent years, there have been other studies that show the opportunities of web-scraped company information for use in statistics production [2,3]. Our study contributes to these existing studies in the following ways. First, we apply and evaluate a method that to the best of our knowledge has not yet been applied for use in official statistics. Next, our approach allows for the ranking of companies without any labelled data, and for the categorisation of companies with only a small amount of labelled data. Existing methods of categorising companies based on their website texts require a relatively large set of labelled data. Finally, the method evaluated in this paper is generically applicable to a large amount of different company categorisations.

METHODS

As a specific use case to demonstrate the method, we discuss our experiments categorising companies as being active in AI or not. Keep in mind that the same method can be applied to other company categorisations (e.g. active in bioeconomy or not, being a transportation company, etc.).

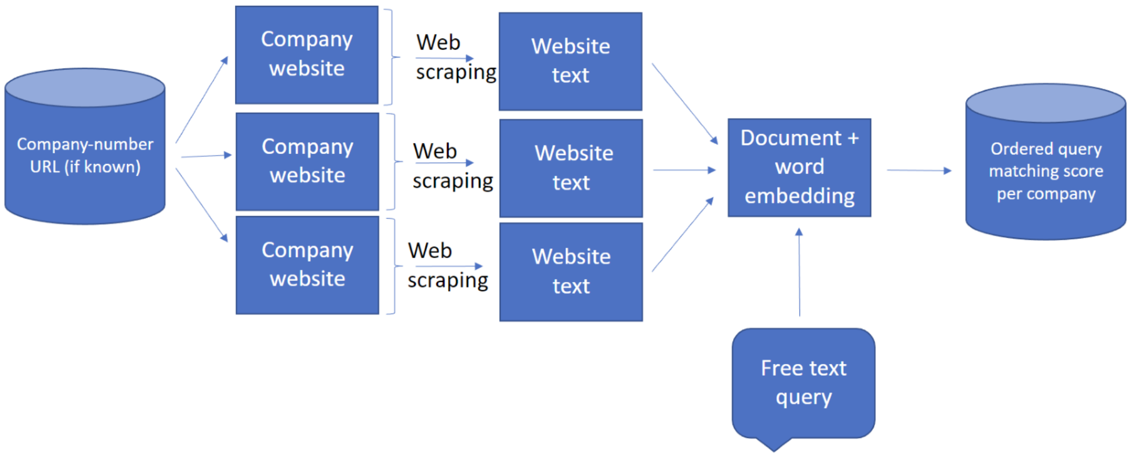

Figure 1. Overview of the method

Figure 1. shows an overview of the method. In the following subsections, we will discuss each step in the figure.

2.1 Input data

The input data for the method is a list of companies described by their company number and URL if known. For our experimental dataset we used the list of all companies with a legal entity in Belgium, excluding one-person businesses. The one-person businesses are excluded for privacy reasons. This results in a dataset of 914,000 Belgian companies, of which 320,000 had a known URL. The URLs in this dataset were purchased from a business partner. The challenge of automatically finding and validating company URLs is out of the scope of this study.

2.2 Web scraping and data cleaning

For each of the URLs, the visible text from the homepages was scraped using Python in combination with the requests [4] and beautifulsoup [5] libraries. Scraping text from deeper webpages (such as the 'about us' page) is an improvement we will tackle in future work. The following cleaning steps are performed on the resulting texts. Only texts in English and Dutch are retained. Language detection are done using the langdetect [6] library in Python. Only texts with more than 50 characters are retained and stop words and a custom list of web-technology words are removed from the texts. The goal of the cleaning is to obtain texts that are dense in information on company activities. After scraping and cleaning, the dataset is reduced to 200,000 clean texts.

2.3 Document and word embedding

The cleaned texts and individual words are jointly embedded using a fine-tuned version of a pretrained multilingual transformer model [7]. For the implementation of this joint embedding the Top2Vec [8] library is used. The resulting embeddings of company texts and words lie close together if they are semantically similar and far apart if they are dissimilar. The embedding model used in our demonstration allows for the combination of 16 different languages for which the model is pretrained. This makes it trivial to deal with different companies using a different language on their website (as long as the languages are included in the pretrained set of languages).

2.4 Query selection and embedding

Next, a free text query is defined that describes the categorisation of interest. For our example use case, AI, we concatenated the Wikipedia introductions of ‘artificial intelligence’ and of 'machine learning' with a description of ‘data science’ found on an IBM webpage. There are infinitely many options to define a query. Designing a method to find the optimal query for a given categorisation will be tackled in future work.
Once it is defined, the query is embedded using the same model as the company texts.

2.5 Company ordering and categorisation

The distance between the embedded query and each embedded company text is calculated. This allows for a ranking of companies with the first company having a website text that has the shortest distance to the query and the last company having the website text with the highest distance to the query. If a sorted list of companies given a specific activity is the desired output, the method can stop here and is completely unsupervised. For business-facing government agencies this is already a valuable outcome. If a binary categorisation is needed, such as for statistics production, some labelled data is necessary, making the approach semi-supervised. To go from ordering to categorisation, a cut-off score must be decided. Companies with a shorter distance to the query than the cut-off are considered part of the category, the others as not part of the category. The choice of cut-off can be made by optimising recall@N and precision@N, with N the number of companies being included by the cut-off.

RESULTS

In order to validate the approach, a dataset of 50 known AI companies was created.

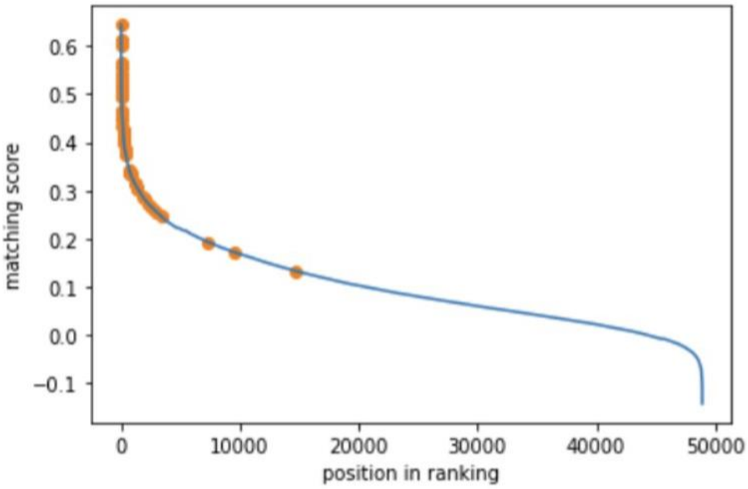

Figure 2. Matching score per company (blue) and placement of known AI companies (orange)

Figure 2. shows the sorted matching scores per company given its position in the ordering. This figure also shows that the known AI companies are placed highly in the ordering, which is desirable. The quantitative performance of the method can be seen in Table 1. The median ranking of the known AI companies is 199, with median score of 0.41.

Table 1. Quality of ranking known AI companies

	Mean rank	**Median rank**	**Mean score**	**Median score**
Known AI companies	1290	199	0.40	0.41

Inspection of 50 randomly selected unknown companies that were ranked higher than the median rank showed only 1 company not active in AI. An inspection of 50 random companies with matching score lower than the lowest-scoring known AI company showed no companies active in AI. This indicates desirable false-positive - and false-negative rates.

CONCLUSIONS

The method presented in this paper demonstrates ta new way of complementing traditional company information with website texts. Our first experiments show that the method is successful in ordering companies active in AI. Following these promising results, we identify the following gaps in this paper for future research. First, a more elaborate validation approach needs to be set-up to assess the quality of the ordering and give guidance to the choice of cut-off point for categorisation. In order to do so, the ordered list of AI companies is currently being used by business-facing government consultants, who provide further feedback on the quality of the results of this method. Next, the performance of the method for other categorisations should be verified. Besides AI, we are investigating categorising companies as being active in bioeconomy and circular economy. Finally, further research should be done on the general properties of website texts for the production of company statistics. For example, bias in the type of companies having a website could be of concern.

References

[1] https://ec.europa.eu/eurostat/web/microdata/community-innovation-survey, (2022)

[2] P. Daas and S. van der Doef, "Using Website texts to detect Innovative Companies.", (2021).

[3] https://ec.europa.eu/eurostat/cros/content/WPC_Enterprise_characteristics_en, (2022)

[4] Python Software Foundation, https://requests.readthedocs.io/en/latest/, (2022)

[5] L. Richardson, https://www.crummy.com/software/BeautifulSoup/, (2022)

[6] https://github.com/Mimino666/langdetect , (2022)

[7] https://tfhub.dev/google/universal-sentence-encoder-multilingual/3 , (2022)

[8]D. Angelov, "Top2vec: Distributed representations of topics.", arXiv preprint arXiv:2008.09470, (2020).