Identifying Innovative Companies From Their Website

Suzanne van der Doef (Statistics Netherlands) Piet Daas (Statistics Netherlands) - Presenting Author Dick Windmeijer (Statistics Netherlands)

Getting an overview of the innovative companies in a country is a challenging task. One of the ways of doing this is setting up a survey to contact a sample of companies; for instance, by phone or via a questionnaire. The response can be used to derive how many innovative companies there are in a country or area. This approach, however, puts a burden on companies and may result in a considerable non-response. Another downside is the fact that usually the focus of such a survey is on large companies and less on smaller companies. We therefore looked for an alternative approach and came up with the idea of determining if a company is innovative by studying the text on the main page of their website. To enable this the following steps were applied, namely: I) Selecting a set of known innovative and non-innovative companies; II) Making sure that for each company the corresponding URL of their web site is available; III) Scraping the main page of each web site and pre-processing the text displayed; IV) Developing a model to determine if a company is innovative or not based on the pre-processed texts. We started with a sample of 3000 innovative and 3000 non-innovative companies according to the Community Innovation Survey of Statistics Netherlands. The first thing observed was that two-thirds of the URL's of the companies selected were absent in the business register. These URL's were added via the URL finding approach developed in WP2 of the ESSnet Big Data (Deliverable 2.2). Since the companies included in the survey all had 10 or more employed persons, we decided to additionally add a considerable number of smaller innovative companies. Here, Dutch companies listed in the yearly SME-innovation top-100 for the years 2008-2017 were used; any duplicates were removed. Next, the text displayed on the main web page of each company was scraped with Python. After language detection (usually Dutch or English), punctuation marks and stop words were removed and the remaining words were stemmed. This was used as input for the model. Here, it was found that logistic regression with L1-norm performed well. With a 70%-30% training and test set, the trained model was able to determine if a company was innovative or not with 93% accuracy. However, special attention needed to be paid to two character length words. Excluding them resulted in a decrease of the model's accuracy to 63%. Web pages that displayed a lot of email-addresses and URL's were found to produce large amounts of these words. As a result, including two character words would make the approach developed very sensitive to such features. Therefore, it was decided to focus on an approach solely using words of three character lengths or more. Here the combination of unigrams and word embeddings performed well; resulting in an accuracy of 91%. More details will be described in the paper and in the presentation.

Part of the session: Big Data Applications to Enterprise Statistics: Businesses, Employers, and Consumers Room: 40.063, October 27, Barcelona, Spain