OTTAWA 2023
64TH WORLD STATISTICS CONGRESS

SIPS Abstract

Exploring the Use of Web Pages Texts, in Brazilian Portuguese, for Classifying Main Economic Activity of Companies

Author: Mrs Ana Gabriela Faria da Silva

Submission ID: 1741

Reference Number: 1741

Brief Description

The role of statistics is to produce information that aims to portray reality.

To make this possible, it is necessary to establish standards.

Economic statistics in Brazil, following international guidelines, adopt the National Classification of Economic Activities (CNAE).

The CNAE has a hierarchical structure where the greater the number of digits the more specific the activity described.

The purpose of the present study is to evaluate the use of supervised learning, in the context of text mining, to achieve the CNAE corresponding to the companies' main economic activity.

Therefore, it was used texts as predictor variables, obtained via web scraping, from business websites and URLs.

Both URLs and the response variable, the CNAE, were derived from the Annual Business Surveys, from the Brazilian Institute of Geography and Statistics (IBGE).

Due to the hierarchical structure of the classification, two approaches were tested to fit the models.

The first, called flat classification, aims to directly obtain the most specific class.

The second approach, which is framed in the category of hierarchical classification, consists of training several independent local classifiers for each level of the class hierarchy.

In both cases, among the tested algorithms, the Logistic Regression classifier presented the best performance, being able to extract patterns fit to identify the classification.

The two approaches provided different results by class, having the flat classifier apparently exhibited a more adequate behavior in categories that tended to be more difficult to characterize in the higher levels, that is, in those that represent less specific activities.

Despite this, both approaches’ results were similar when considering all classes.

Abstract

The role of statistics is to produce information that aims to portray reality. To make this possible, it is necessary to establish standards. Economic statistics in Brazil, following international guidelines, adopt the National Classification of Economic Activities (CNAE). The CNAE has a hierarchical structure where the greater the number of digits the more specific the activity described. The purpose of the present study is to evaluate the use of supervised learning, in the context of text mining, to achieve the CNAE corresponding to the companies' main economic activity. Therefore, it was used texts as predictor variables, obtained via web scraping, from business websites and URLs. Both URLs and the response variable, the CNAE, were derived from the Annual Business Surveys, from the Brazilian Institute of Geography and Statistics (IBGE). Due to the hierarchical structure of the classification, two approaches were tested to fit the models. The first, called flat classification, aims to directly obtain the most specific class. The second approach, which is framed in the category of hierarchical classification, consists of training several independent local classifiers for each level of the class hierarchy. In both cases, among the tested algorithms, the Logistic Regression classifier presented the best performance, being able to extract patterns fit to identify the classification. The two approaches provided different results by class, having the flat classifier apparently exhibited a more adequate behavior in categories that tended to be more difficult to characterize in the higher levels, that is, in those that represent less specific activities. Despite this, both approaches’ results were similar when considering all classes.