Suggested Framework for Big Data Analysis of Enterprise Websites. A Case Study for Web Intelligence Network

Jacek Maślankowski, Dominika Nowak
Statistics Poland, Poland.

Abstract

Big Data gives an opportunity for the researchers and scholars to make surveys in various domains. In this paper we will concentrate on websites as a data source which can be used to provide lots of valuable information for enterprise statistics. In this field, Big Data allows to get various information, including the type of the enterprise (e-commerce etc.), whether the enterprise is present in social media, the frequency of updating the website etc. The main goal of the paper is to present what Big Data methods are the most efficient in acquiring and processing the information from websites. The discussion shows different variants of conducting the work, based on the case studies conducted as experimental statistics at European Union level over the last 6 years.

This paper is based on the experience in processing the data from websites in ESSnet grants on Big Data I (2016-2018), Big Data II (2018-2020) and Web Intelligence Network (2021-2025).

The process of getting enterprise data from websites can be divided into the following steps: (1) Defining the population of enterprise websites; (2) Web scraping; (3) Data processing (extracting); (4) Data validation (de-duplication, quality indicators); (5) Data analysis; (6) Data dissemination.

Each of the steps needs additional validation, especially the first step in this process have an impact on the final results that may not be comparable to the official statistical data. The essential part is also the way the data will be extracted to find the interesting data. In this sense, we need to choose between text mining methods, e.g. machine learning and regular expressions, that gives different results according to the information which should be provided. The paper shows how the use of appropriate methods can increase the overall value of the analysis.