Business Outcomes of Big Data Analysis

project:

WORD COUNT ANALYSIS

for:

www.wp.pl and www.interia.pl

I rok II stopnia informatyka i ekonometria (spec. analiza danych Big Data, S42-11), UG 2021

1. General outline of Big Data project

The purpose of our project is to compare two polish news portals with the most common words. We decided to make analyze between two sites - “Wirtualna Polska” and “Interia”. Our decision is explained by the fact, that both websites are among to the 5 most reading by people news websites in Poland[1]. The most common words may indicate, which type of information has priority to keep interested from readers. By matching the most frequent phrases on a website with the users needs, news portals can gain a larger number of readers, which makes them more popular. It may also illustrate, which type of news is most reading by users. Comparing the results of the analyzes, creates also an opportunity to find out if both sites present similar information on main pages. Properly analyzed results give the authors of news portals the opportunity to specify news on websites to recipients needs, which provide more interested by users.

The data type that is used to create the analysis belongs to human-sourced information. It means, that these informations has been written by people. We used main websites of “Wirtualna Polska” and “Interia” - www.wp.pl for the first one and www.interia.pl for the second one.

In our opinion quality of data on the websites is on good requirements. The titles usually contain some of the information contained in articles. People whose writing these news are journalists, and usually they're checking what they're writing about. Because of it is writing by people means, that data on sites are usually new. On websites occurs also great opportunity to possibility of collecting data in time series. There is also exist risk, that data which we mined can be not available in the future, because news portals can delete some redundant information to get some new space.

[1] https://mansfeld.pl/webdesign/najpopularniejsze-strony-www-w-polsce/ (access to website: 03.01.2021)

2. Project implementation

Method used in our work is web mining and word count analysis. We decided, that these methods are most relatable for our assumptions. To implement the word count analysis project, we used the framework from the lecture written in python language. The BeautifulSoup library was very helpful for pulling data out of website source. The code was launched via Jupyter Notebook.

First, we checked robots.txt for both of websites to find out what we have an access to. Then we downloaded the websites and we removed any extra characters and keywords (e.g. punctuation, html tags). The third step was to do the word count. The last step was to exclude all polish stopwords (via https://www.ranks.nl/stopwords/polish), that add nothing extra to text. With the help of python we also developed a "wordcloud", the result of which can be seen at the bottom. Wordcloud shows the result as an illustration of the most common words on the page. As we can interpret, on both pages the most common word is "pogoda" (eng. weather), because this word is the biggest one on the pictures.

The biggest problem with the dataset is that there are many words (stopwords), html tags etc., that are useless for parsing in the website code. We had to find polish stopwords (which are around of 150) and exclude them from analysis to get better results.

Figure 1. Wordcloud for wp.pl (02.01.2020)

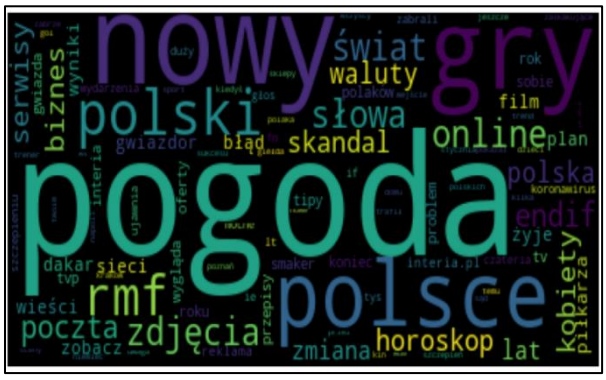


Figure 2. Wordcloud for interia.pl (02.01.2020)

3. Results of analysis

The results of our analysis are presented on figure 3 and figure 4. On both websites "pogoda" (weather) is the most common word. This allows us to conclude that the weather (pogoda) is an important part of both of these websites. This seems natural, because for many people, the information about the weather forecast is one of the most important information, allowing to properly plan the upcoming days. As a rule, news portals such as wp.pl and interia.pl should inform about the weather forecast and they fulfill this function.

The other common words, for example: nowy (new), praca (work), gry (games) and polish names (Kraków, polsce, polski), confirm the belief that interia.pl and wp.pl websites are the Polish portals that inform web users about many areas of life and current events. Thanks to a properly tailored offer, adapted to current trends, both of these websites are among to the 5 most reading by people news websites in Poland[2].

Figure 3. Wordcount analysis for interia.pl (02.01.2020)

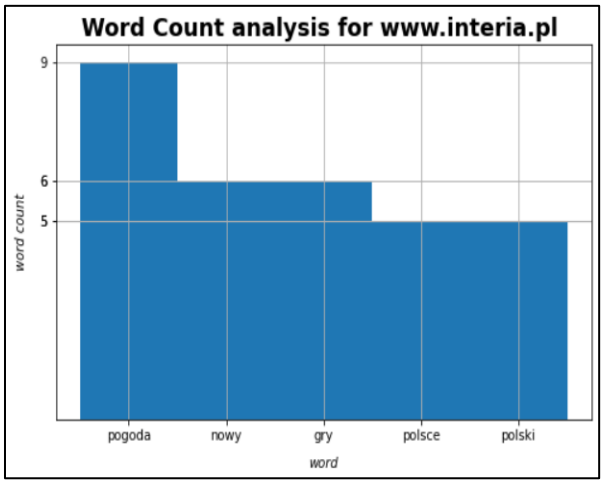
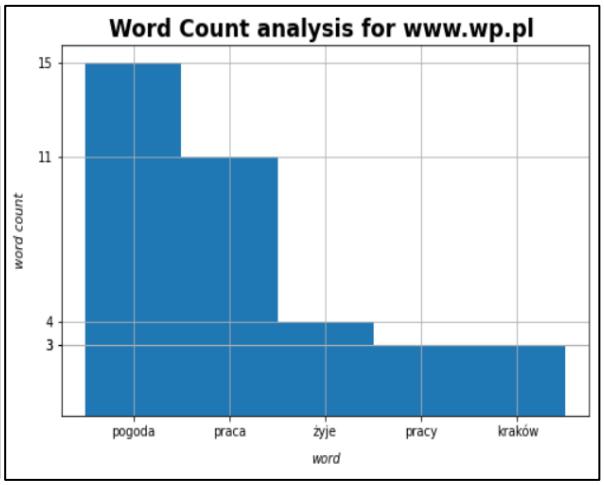


Figure 4. Wordcount analysis for wp.pl (02.01.2020)



[2] https://mansfeld.pl/webdesign/najpopularniejsze-strony-www-w-polsce/ (access to website: 03.01.2021)