Seasonality of the newspaper data at the week level
Machine Learning for Natural Language Processing 2022

Julia NICOLAS
ENSAE
julia.nicolas@ensae.fr

Agathe ROSENZWEIG
ENSAE
agathe.rosenzweig@ensae.fr

Abstract

Newspaper articles can inform on many different topics and are often classified into different categories to facilitate the information reader's search. It seems rather logical a priori that depending on the category to which the article belongs, the keywords and the most frequent words will not be the same. More precisely, one can probably say that the nature of the words used in the articles can in many cases allow a naive reader to guess himself to which category the article he is reading belongs. From articles published on the website of the British newspaper The Guardian the first six months of 2018, we are interested here in establishing a list of the most frequent words in all types of categories and asking ourselves if it is possible to compare them. We also look for seasonality in the use of words in each category by trying to predict the day of the week the article is published based on the vocabulary used.

1 Problem Framing

In this project, we are working on articles from The Guardian published in 2018. We are interested in the different categories of articles, which are classified on the site according to their content and the subject they deal with. Our goal is multiple : we first look at the terms used in the different categories, asking ourselves to what extent there is a specific vocabulary for each category (we have the following five categories : News, Lifestyle, Sport, Arts, Opinion). In this first part, we seek to establish a possible list of the most frequent terms in each category and make a comparison of these. Following **dzogang**, who assess the existence of regularities in the topics used in newspapers, we were interested in the seasonality of the words used by journalists in the paper. Indeed, we try to look at the potential seasonality that exists in the use of terms in newspapers for each topic. As we have 6 months of article data, we adopted a much smaller scale than that of the article and looked for potential weekly seasonality. Our goal was then to see if it was possible, from the vocabulary used in an article, to predict the day of the week during which it was published.

2 Experiments Protocol

The first step in our work was obtaining the data. We used data scraped ourselves from the Guardian website, choosing articles published during the first 6 months of the year 2018. The resulting table was thus composed of about 40,000 rows each corresponding to an article that was published during this period of 2018. We started our work with a first part of descriptive statistics, allowing us to characterize the words used in each category of articles. For tokenization, we were limited by the RAM of Google Colab and we first chose to use TokTok tokenizer. We put all words in lower case and excluded punctuation marks and stopwords as performed in one of the lab sessions. Once the tokenization was finished, we were able to refine our descriptive statistics by looking at the frequencies of the most used words in the articles, and used a LDA model to do Topic Modelling. Then, we used TF-IDF to get more precise information on words such as how often a term appears in a document and its relative rarity in the collection of document. We also used TF-IDF as a keyword extraction method. We also seeked for weekly seasonality of article publication by category, and reused TF-IDF to describe the evolution of vocabulary use over the week. In order to predict the day of the week according to the vocabulary used, we first established a baseline model based on TF-IDF techniques and words frequency, and then we used a basic multinomial naive logistic classifier as a baseline model. Subsequently, we use Word2Vec for word embedding which allows our model to learn word associations. In this model, we limit ourselves to the first 100 words of the article and we select 4000 of them on our sample. For the classification, we use a random forest classifier.

3 Results

Thanks to Topic Modelling, we were able to highlight differences in vocabulary use in the articles. We were able to extract three distinct categories between which the most frequently used words were quite different. We concluded that some categories were more similar than others in terms of the vocabulary used, but that the vocabulary was far from homogeneous between the articles as we had thought when we expressed the research question of our work. The use of TF-IDF allowed us to build a search engine function, and we concluded that based on descriptive analysis of the frequency of words within each category of the paper, we can not observe meaningful results (for example the categories Arts and Lifestyle have almost the same most common words, which can be related to the results obtained with Topic Modelling). However, thanks to the TF-IDF method, we have been able to extract meaningful important words for each category. Regarding the study of seasonality, we observe that all the series present a seasonality at the week scale. Indeed, the number of articles published for each category is very stable for a given day of the week. We did not notice tremendous differences between each day of the week in terms of vocabulary. Yet, we noticed that word ”Tuesday” is extracted as keyword only in papers published on Wednesday, and that words related to sport are more present in the weekend. Our two prediction models (baseline model and embedding thanks to Word2Vec) allowed us to classify the articles according to the day of the week of their publication and to see if the model was adapted. The days of the week that we were best able to predict were Monday, Sunday and Tuesday for our first model. The prediction results are worse in our second model, the classification is significantly less accurate.

4 Discussion and Conclusion

Throughout this project, we were able to work on the use of vocabulary and its evolution in newspaper articles. Our data showed that the words used and their frequency in the articles varied according to the category to which they were affiliated. Thanks to several different methods, we were able to extract the keywords of each category, to model stronger or weaker associations between the words according to the category of the article in which they are found. Our data allowed us to observe some seasonality at the weekly level in the publication of articles in each category, which encouraged us to model a prediction of article publication day based on the vocabulary used.

The results on the various metrics to evaluate the quality of the prediction of the day of the week seems to show that the simplest method (indeed tf-idf representation and the multinomial Bayes classifier) were more relevant than the Word2Vec embedding and the random forest classifier. However, these puzzled results seem to be the corollary of the fact that we were only able to use 100 words by article for the embedding of each article, and to limit our sample of articles to 4000.

During this work, we were still limited in several ways. For example, the Google Collaboratory AMR forced us to choose a particular tokenizer and prevented us from trying more advanced classification models. We were also limited to one type of embedding. To run our model, we limited ourselves to the first 100 words. To go further and be more precise, we could extend the work by considering extracting instead the 100 most important words thanks to TF-IDF and keeping only those to run the model.

. biblio