2021

POLITÉCNICA
"Ingeniamos el futuro"

CAMPUS
DE EXCELENCIA
INTERNACIONAL

TOPIC MODELLING ON SPOOKY AUTHORS DATASET

IS NLP DELIVERABLE

FERNANDO CASABÁN BLASCO

Problem to solve

The problem to be solved in this natural language processing project is to discover possible abstract themes that occur in a collection of documents. Topic modeling is a very useful tool when trying to discover hidden semantic structures in texts.

The dataset to be used contains excerpts from horror novels by authors Edgar Allan Poe (EAP), Mary Shelley (MWS), and HP Lovecraft (HPL). This dataset comes from a competition in 2017 on the popular Kaggle web community, it can be found at the following web URL:

https://www.kaggle.com/c/spooky-author-identification

In this competition, participants were asked to build a model that was capable of predicting the author of a specific document, also to share new knowledge about the data.

The project has been developed with R in notebook format although it is also available in script format. The notebook contains comments along with the code so this document will be very similar. Both the script and the notebook can be found in my github:

https://github.com/Fcabla/Topic-modeling-spooky-authors

Experiments done and Analysis of results

The first thing to do in almost any NLP project is to load the packages and the data. In this case only the train.csv file has been used.

After that, the dataset has been transformed in to a corpus format, since is easier to work with this type of data.

When exploring the corpus, it was noticed that the data had to be processed before since they contained characters that would worsen the result of the project. The transformations that were made to the corpus were: remove punctuation characters, put all characters in lowercase, remove numbers and remove extra white spaces.

In addition to those transformations, stemming was also applied. This technique consists in removing the prefixes or suffixes of a word to his root. In this way you can reduced all the variants of a word in to a single term, for example the terms: consultant, consulting and consultants would be transformed in to consult.

However, after exploring the resultant corpus it was detected some strange terms like "howev". Since this would make it more difficult to interpret the results, it was decided to use lemmatization.

Lemmatization consists in determine the lemma of a term based on its meaning, considers the morphological analysis of the words. Unlike stemming, which it was tested on the corpus, Lemmatization.

The next step in processing the data was to remove the stop words. This are those terms which appear the most in any language, they usually are words without any meaning by themselves.

Apart from the stopwords of the tdm package we can also identify the terms that are repeated the most and delete them from the corpus since they probably do not provide any new knowledge.

The common words that where deleted were: little, even, find, know, come, will, make, good, may, now, say, see, upon, one, can, much.

Most of the terms are verbs, which makes sense since they are widely used words in general and they probably will not be very helpful.

Once the preprocessing was done it was time to explore and understand the data. The first thing to plot was the distribution of the classes.

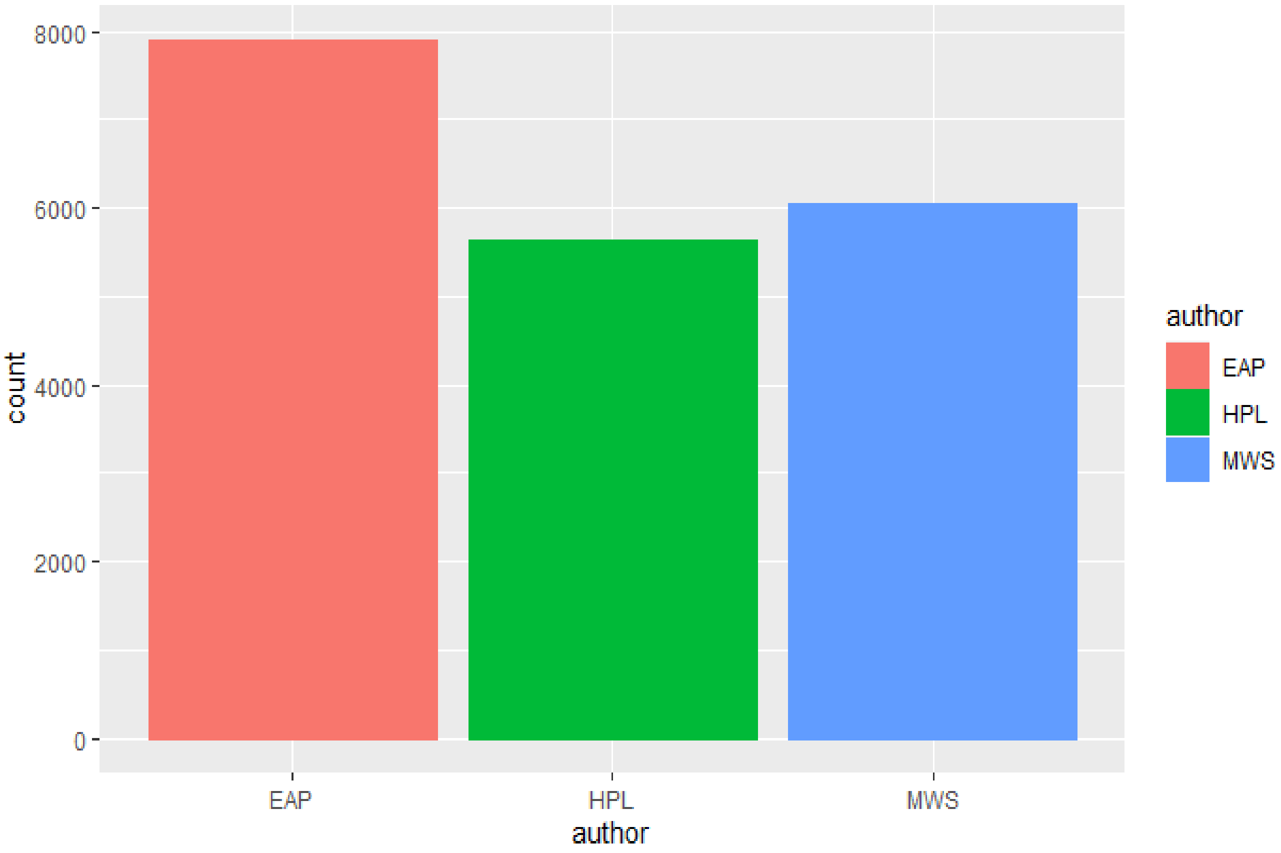

Ilustración 1 Instances per autor

It was also explored which are the most frequent words in the dataset with a barplot.

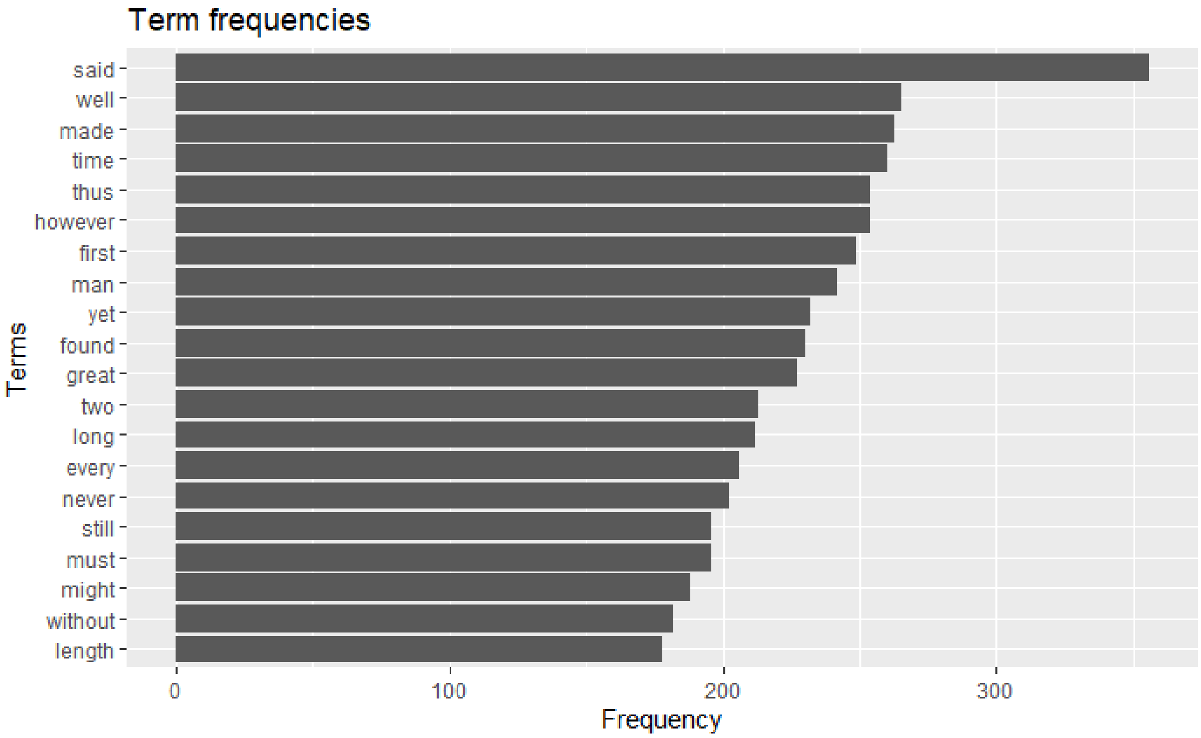

Ilustración 2 Barplot most common words

Then the most common words were explored by each author using a barplot and a wordcloud. In addition, bigram barplots and wordclouds were plotted to check which were the most common bigrams.

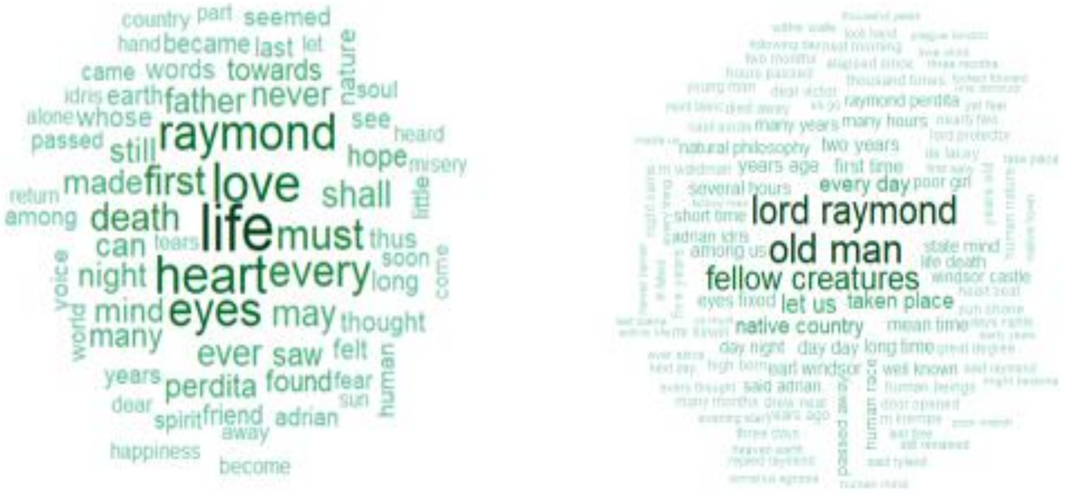

Ilustración 3 wordcloud and bigram wordcloud of MWS

The last plot of the exploration part was to check the total number of words and the number of distinct words written by author.

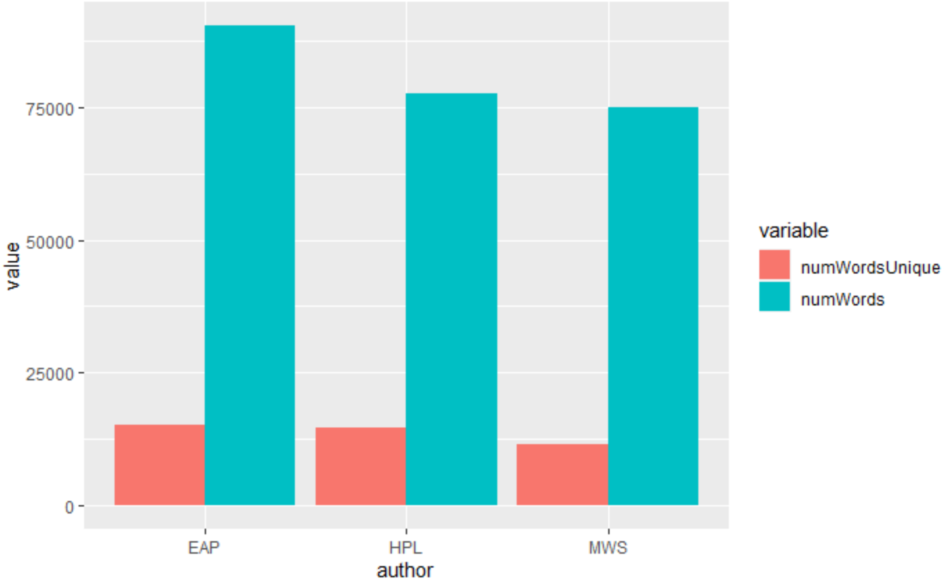

Ilustración 4 total num of words vs distinct words

The last step was to perform topic modelling. Topic modeling is a technique for unsupervised classification of documents. This is similar to perform clustering on numerical data, finding natural groups of items with no previous knowledge of the data.

In this project we are going to use Latent Dirichlet allocation (LDA).

After exploring some of the models, with different number of topics, a model with 3 topics was built, hopping each topic belongs to each author.

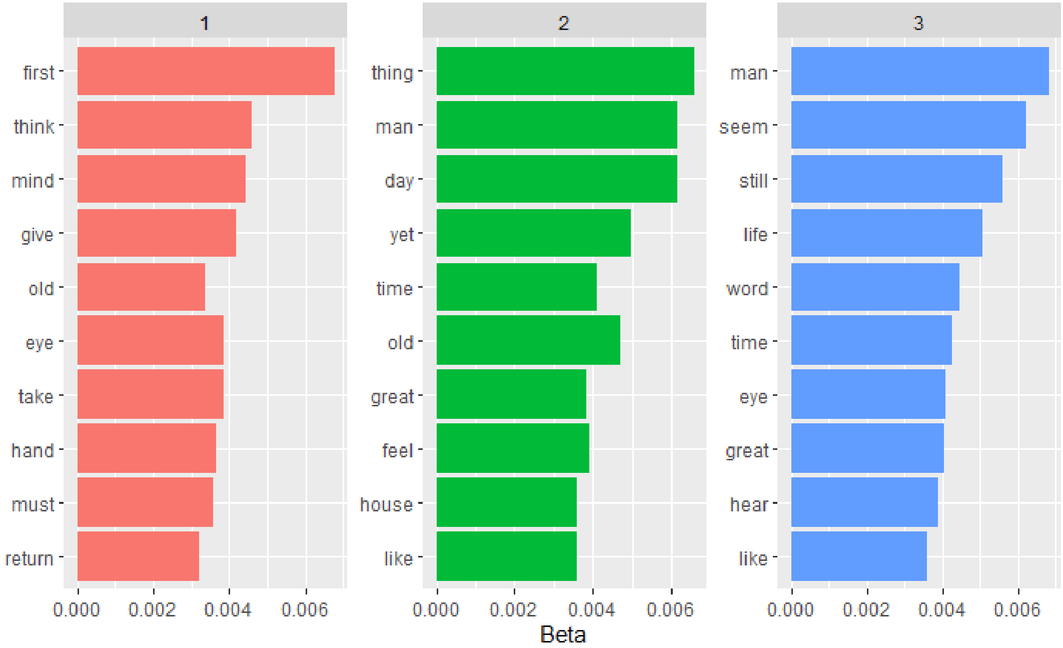

Ilustración 5 Top terms by topic

Only topic 3 seems to be representing the author HPL since some of the top terms of the topic are also the top terms of the author (seem, old, night).

In the notebook was also explore the distribution of authors by topic and also some models were tested with data about one specific author instead of all 3. This is not shown in this document because it is too much and no good results appear.

As it can be seen in the notebook and in this document the results are not very interpretable,it may be because of the fact that I do not know these authors and their works thoroughly. One posible solution for the future would be to try these algorithms on a dataset containing only nouns or nouns + adjectives.