Topic modeling on social media

Progress report II.

Bálint Hantos, Supervisor: István Márkusz

In the last progress report I have written about the change of topics and the approach of gathering the data. In this report I am writing about the data preprocessing pipeline which I have assembled to clean the data and transform it to an appropriate format to use the topic detection algorithm on.

As a first step I have cleaned the data from a technical perspective. I have removed web links, user mentions, punctuations, redundant whitespaces and other string escape characters. Then, processed the text in linguistic approach: removed stopwords and utilized stemming and lemmatizing separately. Stemming and lemmatizing is different in the sense that stemming chops down some characters from the end of a word, which may result in a loss of meaning. On the other hand, lemmatizing is a bit more sophisticated as with this method the output is closer to the root word. I later compared the two approach because it lead to somewhat different results.

The second step was to encode the features. I used the bag of words method to encode the text. Later I transformed this matrix using the Term Frequency – Inverse Document Frequency (TF-IDF) method. The TF-IDF matrix is a more sophisticated representation of the text when doing topic modeling. This is because the IDF gives more weight to less frequent words and decreases the weight of more common words.

The last step was the dimension reduction. It was done by Singular Value Decomposition (SVD). The way it was done that I took some of the most significant vectors from the SVD: these are the most representative and distinct topics of the corpus.

The results were shown at the first presentation, but two of the topics is worth the mention. One topic was a falsely discovered, as it was formed by the messages of the moderating bot – which should've been filtered out. The other topic was about the scandal surrounding one of the constituency candidate's, which was followed by the candidate's withdrawal from the primary election. The detection of this other topic was quite remarkable.

Other than that, I tried to explore the data looking at other subreddit mentions, but it didn't turn out to be as meaningful as the topics found through LSA.

The next steps are going to be using the same preprocessing pipeline while utilizing other topic detection algorithms, namely pLSA, LDA and hSBM (perhaps).