Introduction to Data Science and Analytics

Group Number 1 / Delivery 4

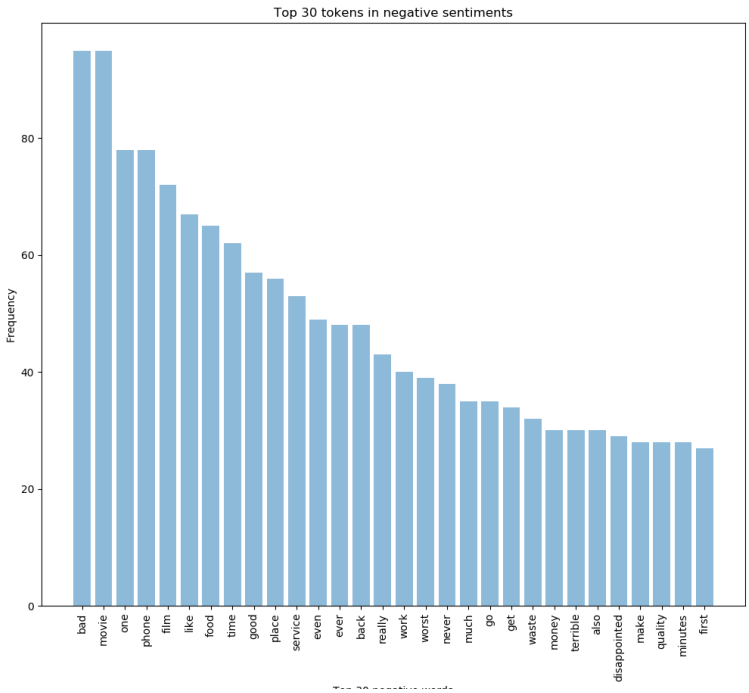
5 May 2019

1 Explanation Before Finding Top Tokens

We used *python* language in our project. We use *pandas* library for reading, representing data as a data frame. We use *numpy* library for array operations. Also for drawing graphics of our final results we use matplotlib library.
First of all, we converted all uppercase to lowercase. Before finding tokens in sentiments we made recovery for each sentences. For example : *i'm to i am*
After making recovery we defined stop words in English by using *nltk* library and we deleted them from sentences.
Negative sentences represented with 0 and positive sentences with 1 in our data set. We calculated term frequencies of words as positive or negative after deleting stop words.
We stored that words in arrays and plotted graphics with the help of this arrays.

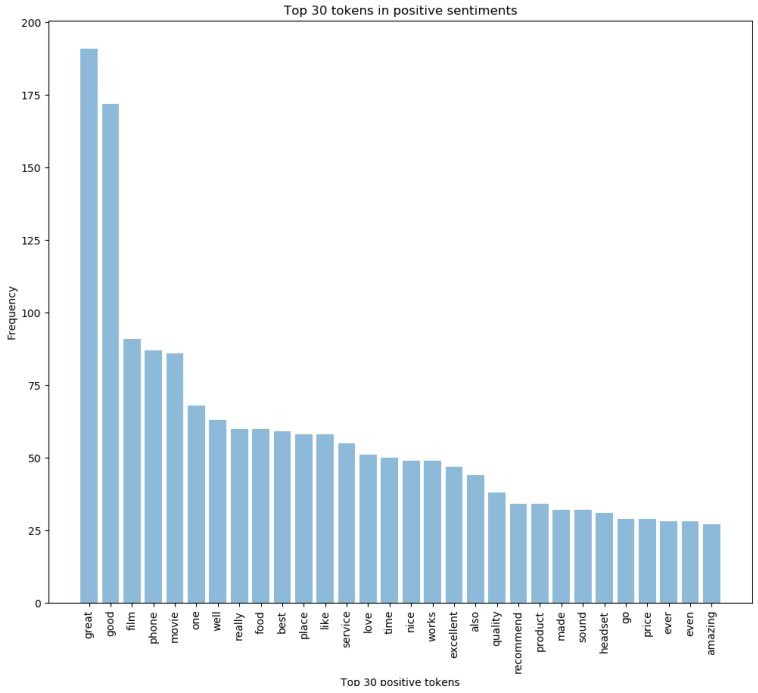
2 Top 30 Tokens in Negative Sentiments

For this part of project we found top 30 words in negative sentiments as frequency.
We found this tokens in negative sentences as we mentioned in the first section of this document. We can see from the graphic the most used word is *bad* in negative sentences.



3 Top 30 Tokens in Positive Sentiments

For this part of project we found top 30 words in positive sentiments as frequency We found this tokens in positive sentences as we mentioned in the first section of this document. We can see from the graphic the most used word is *great* in negative sentences.



4 Negative Frequency vs Positive Frequency

For this part we compare negative and positive word's term frequencies that we found before in the graph.
We made this comparison with using scatter plot.
We can see that most of the distribution is in the region near zero. The reason for this is that most words have been used for 0-20 times. Each word may not be as distinctive as *bad* or *good*.