You are always welcome to ask us any questions in Piazza! With the gained knowledge during the lecture and lab session you should be able to solve the tasks below.

Please proceed to the course web page where Homework 7 can be found and download “apple1.csv” and “apple2.csv” that contains apple tweets.

In apple1.csv and apple2.csv you have comments about apple before (apple1.csv) and after (apple2.csv) announcing the quarterly profits.

1. (1 point) Draw one bar plot to compare existing sentiments from two datasets (see Figure 3 in the lab session).
2. (3 points) Clean both datasets:
   - change all characters to lowercase;
   - remove URLs;
   - remove words starting with ‘@’;
   - remove words starting with ‘$’ (use ‘\$’ for ‘$’ in regular expression);
   - remove punctuation;
   - remove stopwords;
   - remove numbers;
   - remove whitespaces;

   **Don’t forget to inspect the text after each cleaning to see if the outputs are desirable.**

   Tip: Think about the best order in which you would like to do the cleaning for the above cleanings.

   After cleaning the datasets, perform following tasks:

3. (2 point) Create word cloud with words that occur at least 8 times for each dataset. Which is the most frequent word in each of the word clouds

4. (1 point) Build one bar plot with the total scores of the all sentiments and feelings from both datasets. Which tweets (before or after) are more positive?

5. (1.5 points) Use KNN and random forest to predict sentiments of the tweets after combining the two datasets. Divide the data into 80 to 20 % ratio for training and testing. Build confusion matrix and calculate accuracy of the predicting model. For which model you achieved better accuracy?

Please note: Use the same training and testing data for KNN and random forest.