IN4252 Web Science & Engineering

Social web Assignment 1

Marios Marinos (M.Marinos@student.tudelft.nl)

November 28, 2020

Part I

Task 1: Retrieving via Twitter API

1. What is the starting and ending time of the data that you have crawled?

For the first two questions of the first part of the assignment, **we do not take into account the deleted tweets**. The starting and ending time are show in table 1.

Starting Time :	Wed Nov 18 16:30:57 +0000 2020
Ending Time :	Wed Nov 18 16:40:55 +0000 2020

Table 1: Table of starting and ending time of the tweets.

2. What is the id of the first tweet you got? And the last one?

Same thing as question 1 applies here. The deleted tweets are not taken into account for the results. The first and last id of the tweets I have received are shown in table 2.

First tweet id :	1329099788256743431
Last tweet id :	1329102296467255296

Table 2: Table of first and last id of the tweets.

3. How many tweets did you get?

The count of tweets are shown in table 3.

Tweets count **included deleted :**	44092
Tweets count**without including deleted :**	33969

Table 3: Table of how many tweets crawled in total.

4. How large is the result file (uncompressed file in JSON format)?

The size of the file **as it was crawled(with the deleted tweets)** is equal to 179.9MB.

5. How many tweets sent from Amsterdam did you get?

To answer the 2 questions remaining from part 1, we need to adjust the code provided a bit. Instead of using the stream.sample() to monitor the incoming tweets, we use the method stream.filter. To filter the tweets sent from Amsterdam we use the location provided : stream.filter(locations=[4.61, 52.27, 5.07, 52.50]). The code for crawling the data is in the file Twitter_API.py whereas the code for calculating how many tweets we received in total is in the file Twitter_manipulation.py. In table 4 is shown how many tweets were sent from Amsterdam.

Tweets sent from Amsterdam :	353

Table 4: Table of how many tweets were sent from Amsterdam in 2 hours.

6. How many tweets are related to COVID-19?

Again, as on question 5. we need to make a minor adaption in the code. Again there is the need of the method stream.filter but now with different filter. So, to find tweets related to COVID-19 with specific keywords provided this statement is used : stream.filter(track=['covid','covid-19','corona','coronavirus','lockdown']). In table 5 is shown how many tweets were related to COVID-19 keywords.

Tweets related to COVID-19 :	286266

Table 5: Table of how many tweets were related to COVID-19 in 2 hours.

Part II

Task 2: Exploratory and Confirmatory Data Analysis

7. Visualization of the features.

The part of visualizing the features is done with the hypothesis testing on question 9.

8. Descriptive statistics for each feature.

In this task we are going to show the descriptive statistics for each feature based on the relevant tweets and non relevant tweets. The 5th feature I chose is if the author of the tweeter is Verified or not of each tweet.

Feature : entities		
Statistics for :	Relevant tweets	Non-Relevant tweets
count	2817.000000	37138.000000
mean	2.367057	1.882304
std	1.606369	1.706187
min	0.000000	0.000000
25%	1.000000	0.000000
50%	2.000000	2.000000
75%	3.000000	3.000000
max	10.000000	11.000000

Feature : entityTypes		
Statistics for :	Relevant tweets	Non-Relevant tweets
count	2817.000000	37138.000000
mean	0.795527	0.597340
std	0.787920	0.754422
min	0.000000	0.000000
25%	0.000000	0.000000
50%	1.000000	0.000000
75%	1.000000	1.000000
max	3.000000	4.000000

Feature : tweetsPosted		
Statistics for :	Relevant tweets	Non-Relevant tweets
count	2817.000000	37138.000000
mean	29862.847710	28888.871641
std	48384.225953	57288.566101
min	0.000000	0.000000
25%	2988.000000	2481.000000
50%	12094.000000	10184.000000
75%	34790.000000	29961.750000
max	545006.000000	1399152.000000

Feature : sentiment		
Statistics for :	Relevant tweets	Non-Relevant tweets
count	2817.000000	37138.000000
mean	-0.024494	0.041925
std	0.268697	0.412782
min	-1.000000	-1.000000
25%	0.000000	0.000000
50%	0.000000	0.000000
75%	0.000000	0.000000
max	1.000000	1.000000

Feature : SemanticOverlap		
Statistics for :	Relevant tweets	Non-Relevant tweets
count	2817.000000	37138.000000
mean	0.253461	0.046529
std	0.435070	0.210631
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	0.000000
75%	1.000000	0.000000
max	1.000000	1.000000

9. Hypothesis testing for each feature.

In this section we are going to perform hypothesis tests for each different feature. For all test we will assume ρ-value threshold $= 0.05$, so if the ρ-value calculated from our data is **less than** the threshold we will reject the main hypothesis.

1. Firstly we need to test how significant is feature entities in . In figure 1 it is shown that in both groups (relevant and non-relevant) there are some outliers but in general it's obvious that the **data follows the normal distribution.** When the data are normally distributed in unpaired groups, usually it is best to test whether the null hypothesis holds or not by using the **unpaired t-test** is used. By intuition the expectation is that **the hypothesis should be rejected as in the table with the descriptive statistics, it is seen that the mean of relevant tweets with entities is slightly bigger (+0.6) than the mean of non-relevant tweets.**

   Hypothesis H_0: The number of entities of relevant and non-relevant tweets are not statistically different.
   The p-value is 3.9186347482289249e-48 $<$ 0.05, therefore with 5% probability of being wrong it is safe to conclude that indeed the two populations of relevant and non-relevant tweets are different so the tweet entities feature is a discriminative feature for relevance judgment.

2. Second feature we need to test if it is discriminative or not is the EnityTypes. The same procedure as in the first feature applies also here. Lookin in figure 2 again it is obvious that the data follows somehow the normal distribution so the unpaired t-test is only applicable here. The descriptive statistics on this feature also tends to be slightly bigger so the expectation is that the null hypothesis should be rejected.

   Hypothesis H_0: The type of Entity of relevant and non-relevant tweets are not statistically different.
   The p-value is 7.442790483859708e-41 $<$ 0.05, therefore with 5% probability of being wrong it is safe

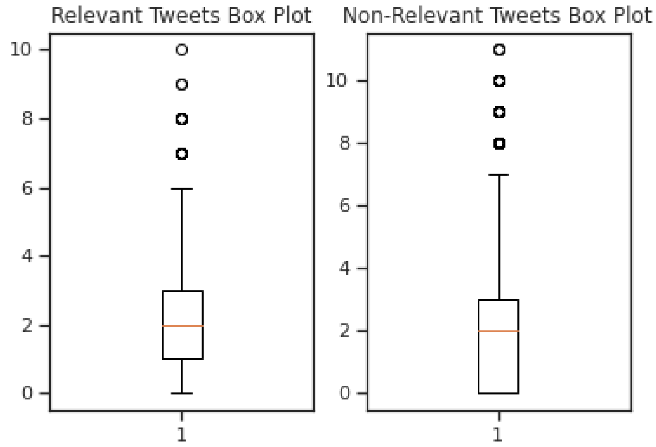

Figure 1: Box plots between relevant and non relevant tweets based on feature Entities.

to conclude that indeed the two populations of relevant and non-relevant tweets are different so the tweet entity types feature is a discriminative feature for relevance judgment.

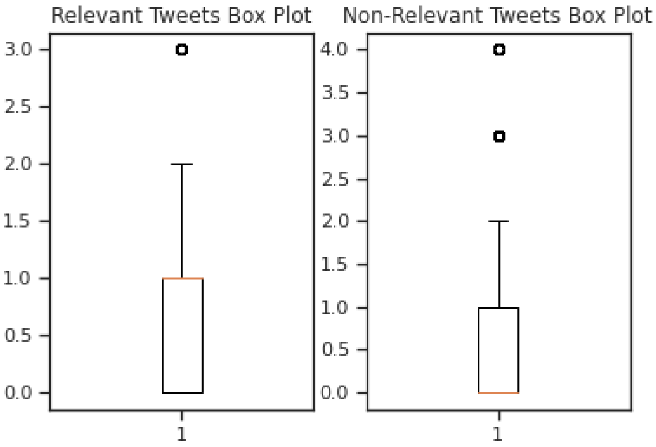

Figure 2: Box plots between relevant and non relevant tweets based on feature EntityTypes.

3. Third feature that needs to be explored is tweetsPosted. It is way obvious from figure 3 that the data **do not follow the normal distribution**. Also, from the same figure it looks like the data are derived from gamma distribution and a good test for non-normal distribution data is the Mann-Whitney U test. Again here, the difference between the means is quite high but there is also a tremendous high variance in both groups.

   Hypothesis H_0: The tweetsPosted by the author of relevant and non-relevant tweets are not statistically different.
   The p-value of Mann-Whitney U test is $0.00000055196679421734 < 0.05$, therefore with 5% probability of being wrong it is safe to conclude that indeed the two populations of relevant and non-relevant tweets

are different so the tweet posts feature is also a discriminative feature for relevance judgment.

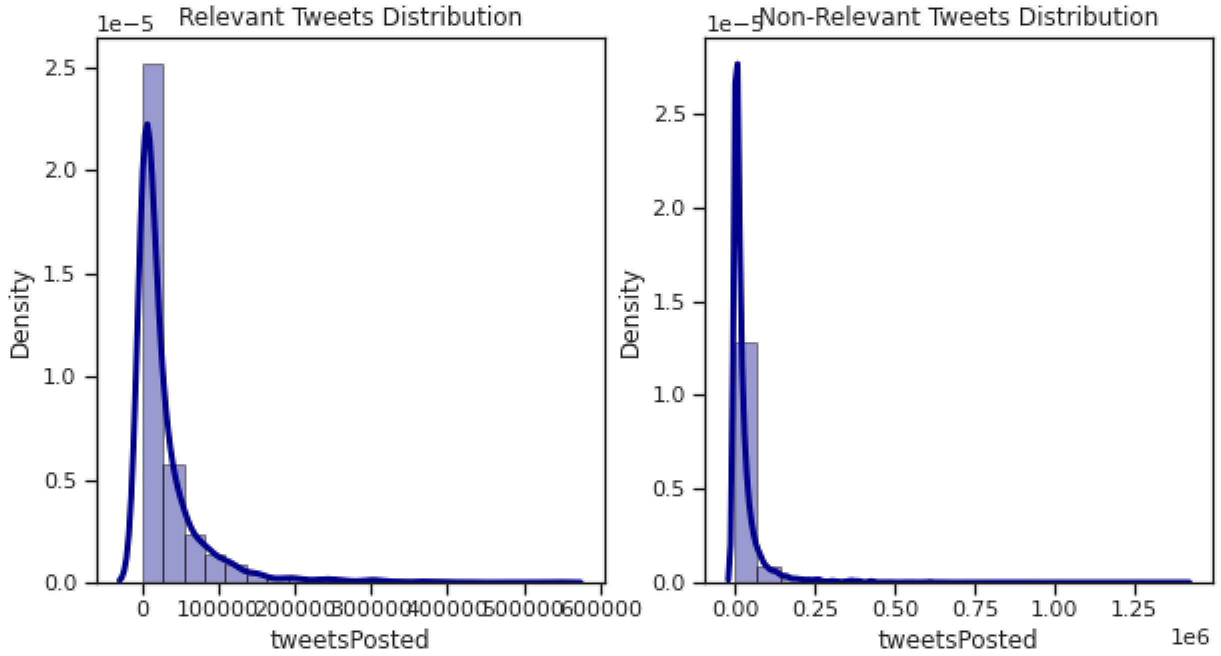

Figure 3: Distribution histogram between relevant and non relevant tweets based on feature tweetsPosted.

4. The fourth feature that is going to be explored is the sentiment of a tweet. In figure 5 it is seen that most relevant tweets are neutral (around 95%) and not positive/negative instead whereas on non relevant tweets the neutral tweets are barely 80%. Again, there is a need to test whether the two populations is statistical significantly different, in order to determine if the sentiment have a role in deciding whether a tweet is relevant or not.

   Hypothesis H_0: The sentiment of relevant and non-relevant tweets are not statistically different.
   The p-value of Mann-Whitney U test is 3.373231842213983e-18 $<$ 0.05, therefore with 5% probability of being wrong it is safe to conclude that indeed the two populations of relevant and non-relevant tweets are different so the sentiment of the tweet is a discriminative feature for relevance judgment.

5. Finally, the feature I found out is interesting to explore is the semantic overlap. In the first sight, it should be definitely related to how easy is to distinguish between relevant and non-relevant tweets. Therefore, we are going to find out if it is discriminative or not based on the data. Firstly, by seeing the table with the descriptive statistics is way obvious that the mean of SemanticOverlap of relevant tweets is almost 5 times as for the non relevant ones, so the expectations is again that the test should lead us to the fact that the semantic overlap is an important feature.

   Hypothesis H_0: The Semantic Overlap of relevant and non-relevant tweets are not statistically different.
   The p-value of Mann-Whitney U test is 0.000000000 $<$ 0.05, therefore with 5% probability of being wrong it is safe to conclude that indeed the two populations of relevant and non-relevant tweets are different so the sentiment of the tweet is a discriminative feature for relevance judgment.

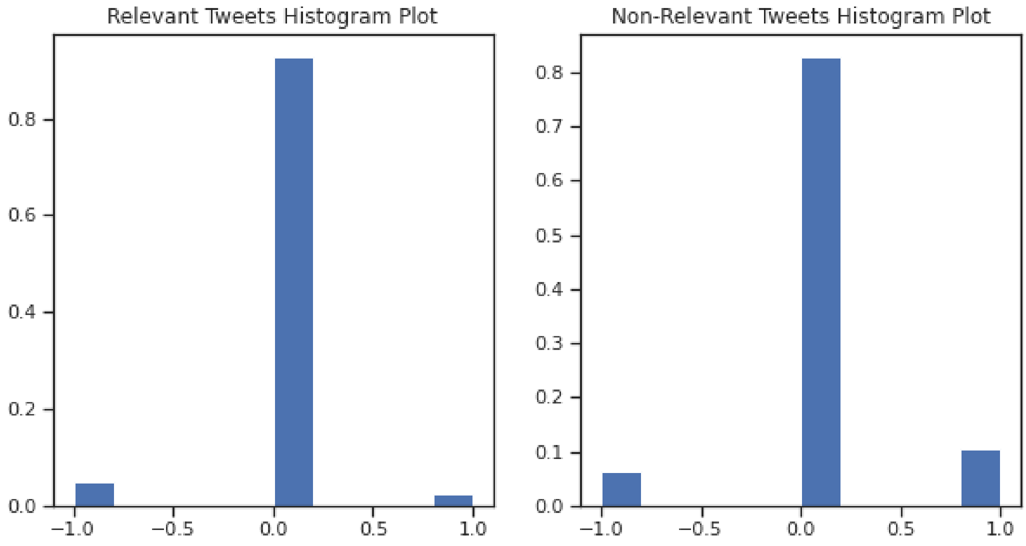

Figure 4: Histograms between relevant and non relevant tweets based on feature tweetsPosted.

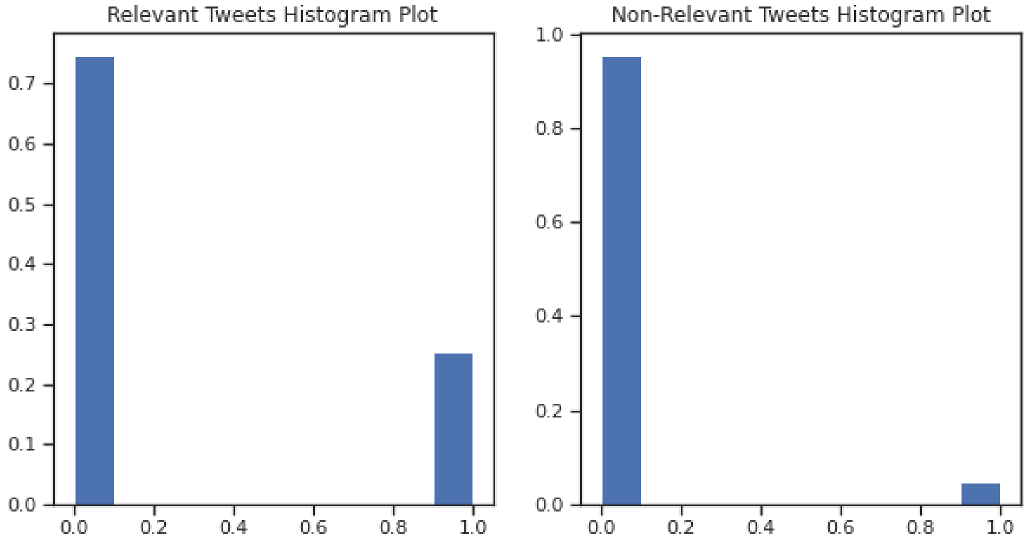

Figure 5: Histograms between relevant and non relevant tweets based on feature semanticOverlap.