IN4252 Web Science & Engineering
Hands-on Assignment
Social Web Data Analytics

Kevin Nanhekhan
k.r.nanhekhan@student.tudelft.net, 4959094

Task 1: Retrieving via Twitter API

1.2 Accessing Public Streaming API

1. *What is the starting and ending time of the data that you have crawled?*
Starting time: **2022-12-02 17:56:43**
Ending time: **2022-12-02 18:06:44**

2. *What is the id of the first tweet you got? And the last one?*
First tweet id: ***1598722836272644099***
Last tweet id: ***1598725327697711108***

3. *How many tweets did you get?*
34551 tweets

4. How large is the result file (uncompressed file in JSON format)?
Total result file size is 10,6 Mb

1.3 Filtering Tweets sent from Amsterdam

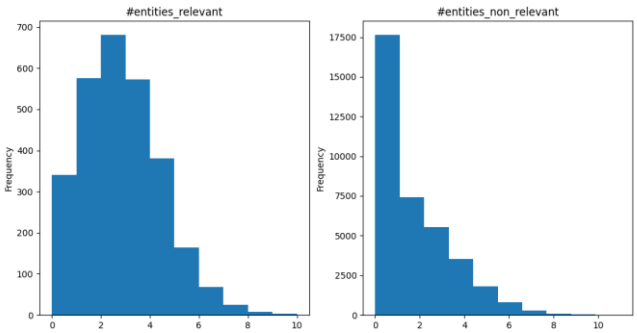
1) How many tweets sent from Amsterdam did you get?
58394 tweets

2) How many tweets are related to COVID-19?
90769 tweets.

Task 2: Exploratory and Confirmatory Data Analysis

Besides the four mandatory features *(#entities, #entityType, #tweetsPosted and sentiment*) also a look has been taken at the feature *nFavorties* as the amount of times a tweet has been favorited could have an effect in its relevance. For the hypothesis testing either the *Mann Whitey U test* as *T-test* has been applied, depending on whether the data is normally distributed or not.

#entities:



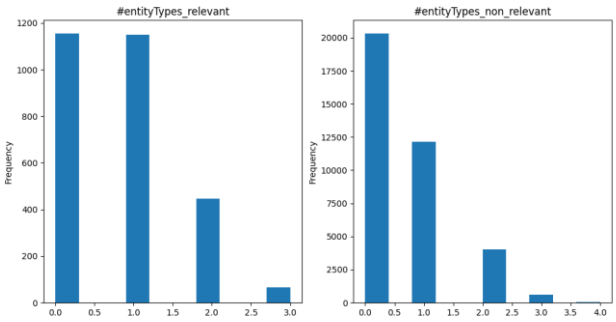
Feature relevant:
count 2817.000000
mean 2.367057
std 1.606369
min 0.000000
25% 1.000000
50% 2.000000
75% 3.000000
max 10.000000
Name: #entities, dtype: float64

Feature non-relevant
count 37138.000000
mean 1.882304
std 1.706187
min 0.000000
25% 0.000000
50% 2.000000
75% 3.000000
max 11.000000
Name: #entities, dtype: float64

Mann Whitney U test – p-value: **1.9277452753941775e-63**

From the plotted histograms we see that the data follows a skewed distribution and not a normal distribution so *Mann Whitney U test* can be applied. The p-values is below 0.05 ($p < 0.05$) which shows that using *#entities* (number of entities) is a useful feature in discerning whether tweets are relevant or not.

#entityType



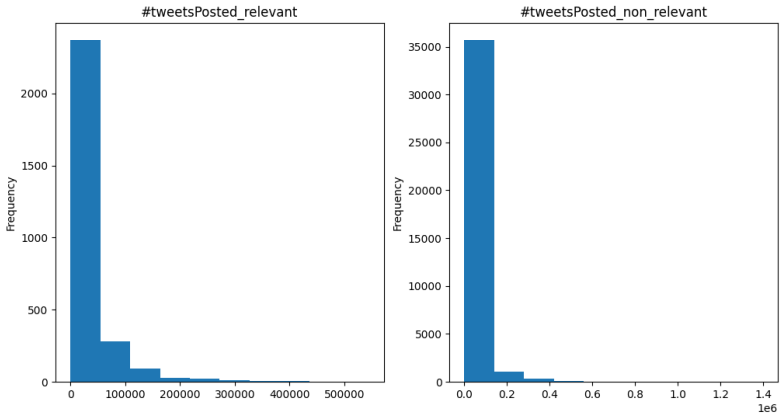
Feature relevant:
count 2817.000000
mean 0.795527
std 0.787920
min 0.000000
25% 0.000000
50% 1.000000
75% 1.000000
max 3.000000
Name: #entityTypes, dtype: float64

Feature non-relevant
count 37138.000000
mean 0.597340
std 0.754422
min 0.000000
25% 0.000000
50% 0.000000
75% 1.000000
max 4.000000
Name: #entityTypes, dtype: float64

Mann Whitney U test – p-value: **1.6603458547298032e-46**

From the plotted histograms we see that the data follows a skewed distribution and not a normal distribution so *Mann Whitney U test* can be applied. The p-values is below 0.05 ($p < 0.05$) which shows that using *# entityTypes* (number of entity types) is a useful feature in discerning whether tweets are relevant or not.

#tweetsPosted:



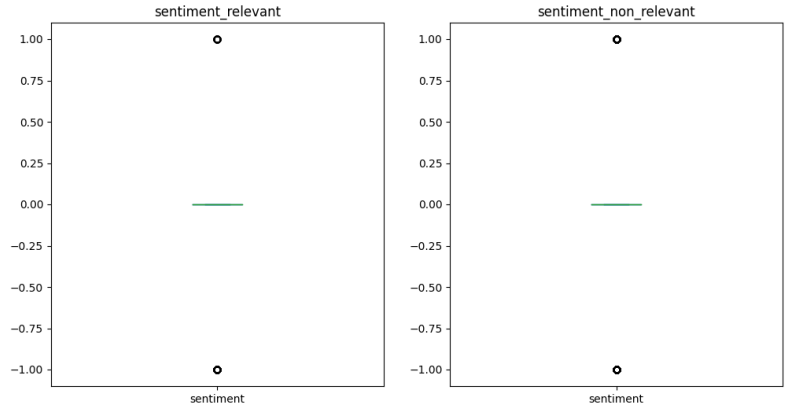
Feature relevant:
count 2817.000000
mean 29862.847710
std 48384.225953
min 0.000000
25% 2988.000000
50% 12094.000000
75% 34790.000000
max 545006.000000
Name: #tweetsPosted, dtype: float64

Feature non-relevant
count 3.713800e+04
mean 2.888887e+04
std 5.728857e+04
min 0.000000e+00
25% 2.481000e+03
50% 1.018400e+04
75% 2.996175e+04
max 1.399152e+06
Name: #tweetsPosted, dtype: float64

P-values:
Mann Whitney U test – p-value: **1.1039335884346776e-06**

From the plotted histograms we see that the data follows a skewed distribution and not a normal distribution so *Mann Whitney U test* can be applied. The p-values is below 0.05 ($p < 0.05$) which shows that using *#tweetsPosted* (number of tweets posted) is a useful feature in discerning whether tweets are relevant or not.

sentiment:



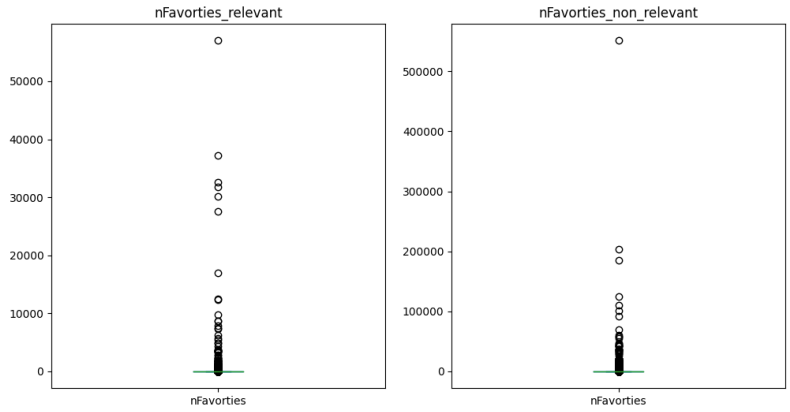
Feature relevant:
count 2817.000000
mean -0.024494
std 0.268697
min -1.000000
25% 0.000000
50% 0.000000
75% 0.000000
max 1.000000
Name: sentiment, dtype: float64

Feature non-relevant
count 37138.000000
mean 0.041925
std 0.412782
min -1.000000
25% 0.000000
50% 0.000000
75% 0.000000
max 1.000000
Name: sentiment, dtype: float64

T-test – p-value: **4.378557294479076e-17**

From the plotted histograms, we see that the data follows a normal distribution so *T-test* can be applied. The p-values is below 0.05 ($p < 0.05$) which shows that using *sentiment* is a useful feature in discerning whether tweets are relevant or not.

nFavorites:



Feature relevant:
count 2817.000000
mean 184.261981
std 1856.418620
min 0.000000
25% 0.000000
50% 1.000000
75% 11.000000
max 57064.000000
Name: nFavorties, dtype: float64

Feature non-relevant
count 37138.000000
mean 185.978755
std 3648.010529
min 0.000000
25% 0.000000
50% 2.000000
75% 25.000000
max 551473.000000
Name: nFavorties, dtype: float64

Mann Whitney U test – p-value: **2.6749752490079864e-21**

From the boxplots we see that the data follows a skewed distribution and not a normal distribution so *Mann Whitney U test* can be applied. The p-values is below 0.05 ($p < 0.05$) which shows that using *nFavorties* (number of times a tweet has been as favorite by others) is a useful feature in discerning whether tweets are relevant or not.