Web Science and Engineering 2015

Homework 4

Eric Camellini
4494164
e.camellini@student.tudelft.nl

1. TASK 1

1.1 Accessing Public Streaming API

In this first part of the homework I monitored the Public Stream for 10 minutes using the console and saving the samples in a json file. I used Vim[1] to open the file and easily find the answers to the following questions (Vim automatically shows every json tweet in a different page).

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

The starting time (timestamp) of the crawled data is Mon, 19 Oct 2015 21:50:49 GMT (i.e. Mon, 19 Oct 2015 23:50:49 in the Amsterdam timezone). The ending time is Mon Oct 19 22:00:49 2015 (i.e. Tue, 20 Oct 2015 00:00:49 in the Amsterdam timezone)

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

The first id is 656226018953326592.
The last one is 656228535502290944.

3) How many tweets did you get?

The sampling file contains 29892 tweets.

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

The file size is 110.022.106 bytes (nearly 110,0 MB).

1.2 Filtering Tweets sent from Amsterdam

For this second part of the task, I monitored the Public Stream for 2 hours using the tweepy Python library[2] and saving the samples in a file. To answer the questions I used Pandas[3] inside Ipython Notebook[4].

1) How many tweets did you get?

I obtained 868 tweets.

2) How many tweets did you get that were sent from Schipol?

To count the tweets from Schipol I created a script that extracts the coordinates from the Amsterdam samples and then checks if they are in the Schipol bounding box. The coordinates field was empty in most of the tweets, because the location filtering is often based on other parameters (e.g. on the intersection with predefined places[5]), so I also decided to run again the whole sampling using the Schipol bounding box. I obtained 26 tweets with the first approach and 522 with the second one.

I also made a further check: I checked if all the tweets crawled from Amsterdam were in the Amsterdam bounding box, using the script I made, and I did the same with the Schipol ones. I found that in the Amsterdam samples 176 tweets have coordinates inside the Amsterdam bounding box (over 197 tweets that has the coordinates field not empty) and in the Schipol case only 2 (over 122 with coordinates). From this I can deduce that filtering tweets using a bounding box doesn't give precise results: filtering by 'Place' could be a better solution.

2. TASK 2

In this second task, I conduct exploratory and confirmatory data analysis for 4 features taken from the Tweet Relevance Judgment problem defined in [1]. These features, with the related hypothesis, are:

- #entities: the more entities a tweet mentions, the more likely it is to be relevant and interesting;

[1] http://www.vim.org/

[2] https://github.com/tweepy/tweepy

[3] http://pandas.pydata.org/

[4] http://ipython.org/notebook.html

[5] https://dev.twitter.com/streaming/overview/request-parameters

- #entityTypes: the greater the diversity of concepts mentioned in a tweet, the more likely it is to be interesting and relevant;
- #tweetsPosted: the higher the number of tweets that have been published by the creator of a tweet, the more likely it is that the tweet is relevant.
- sentiment: the likelihood of a tweet's relevance is influenced by its sentiment polarity.

To exctract the feature descriptive attributes, plot the distributions and perform the data analysis I used the Scipy packages[6] Pandas, Numpy and Matplotlib inside IPython Notebook.

The descriptive attributes and the results of hypothesis testing of the features can be seen in Tables 1, 2, 3 and 4 and the plots of the distributions in Figures 1, 2, 3 and 4. For some of the the plots I decided to use a log-scaled y axis to better represent the power-law distribution of the values.

Since the p-values are all less than 0.05, we can say that the hypothesis are confirmed. In Figures 5, 6, 7 and 8 I also plot how the mean value of the four features (on the y-axis) variates w.r.t. to the relevance (on the x-axis). We can see that these four features have different mean value depending on the relevance judge (0 or 1), so this is a further confirmation of the fact that they are discriminative for relevance judgement.

References

[1] K. Tao, F. Abel, C. Hauff, and G.-J. Houben. What makes a tweet relevant for a topic. *Making Sense of Microposts (# MSM2012)*, pages 49–56, 2012.

#entities		
	relevant	non relevant
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
U	42581699.5	
p values	$2.5034102868429897e^{-61}$	

Table 1: #entities descriptive statsitics and results of hypothesis testing

#entityTypes		
	relevant	non relevant
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
U	44719177.5	
p values	$3.794083107935882e^{-38}$	

Table 2: #entityTypes descriptive statsitics and results of hypothesis testing

#tweetsPosted		
	relevant	non relevant
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
U	49433364.0	
p values	$5.519667942173387e^{-07}$	

Table 3: #tweetPosted descriptive statsitics and results of hypothesis testing

sentiment		
	relevant	non relevant
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
U	49024653.0	
p values	$1.314227158467661e^{-08}$	

Table 4: Sentiment descriptive statsitics and results of hypothesis testing

[6] http://www.scipy.org/index.html

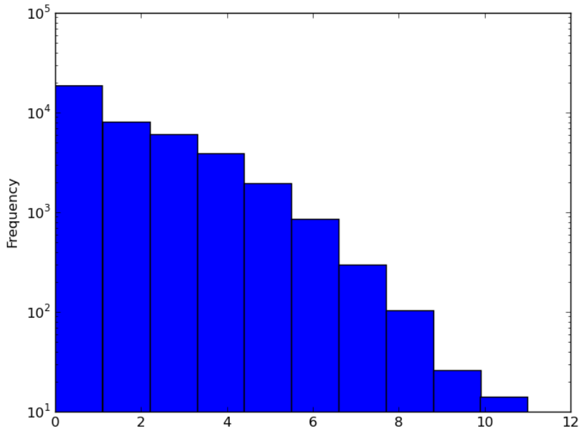

Figure 1: #entities feature distribution

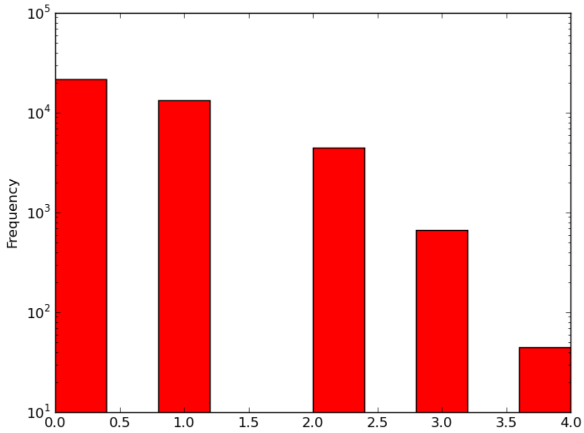

Figure 2: #entitityTypes feature distribution

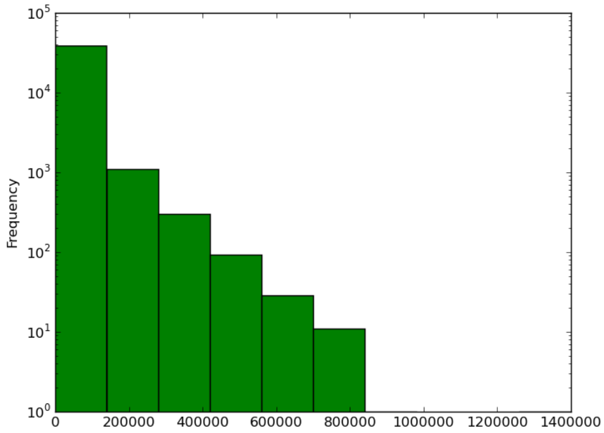

Figure 3: #tweetsPosted feature distribution

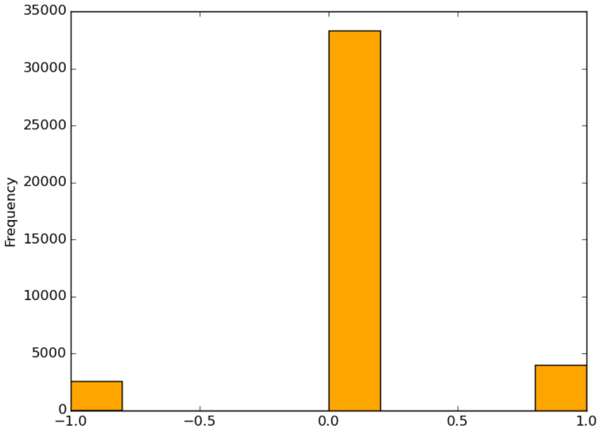

Figure 4: sentiment feature distribution

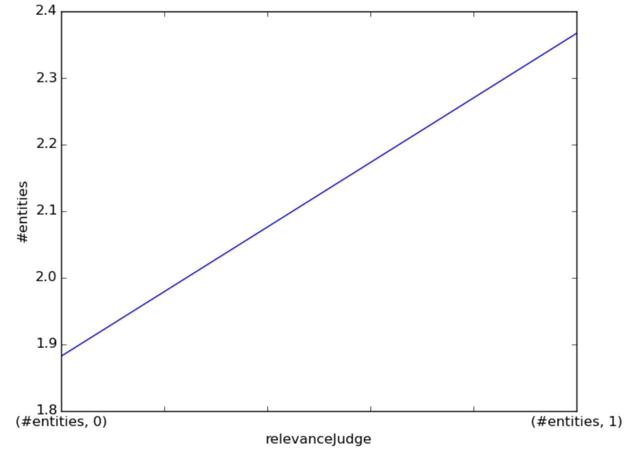

Figure 5: #entities mean value w.r.t. the relevance

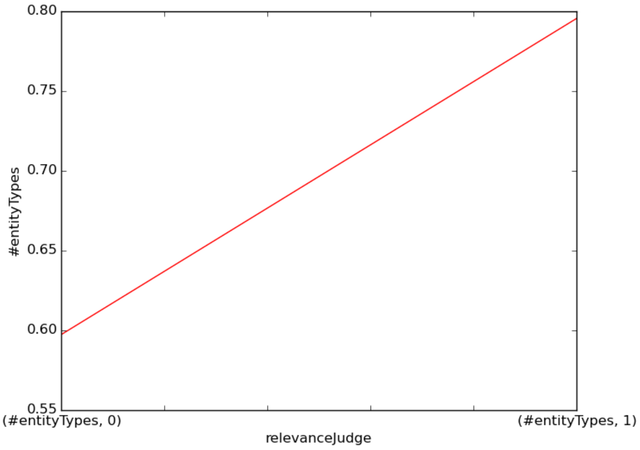

Figure 6: #entitityTypes mean value w.r.t. the relevance

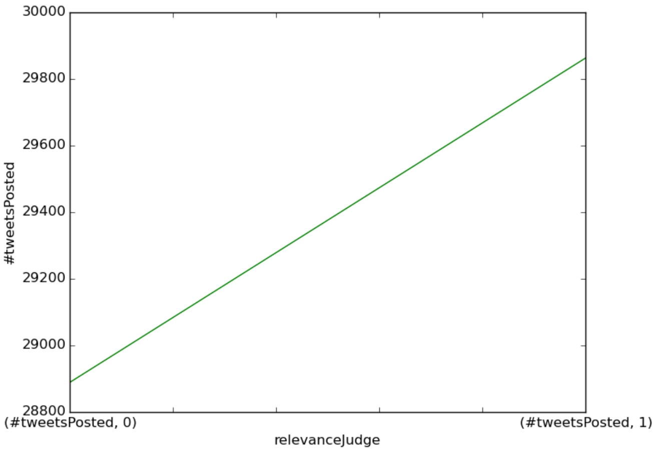

Figure 7: #tweetsPosted mean value w.r.t. the relevance

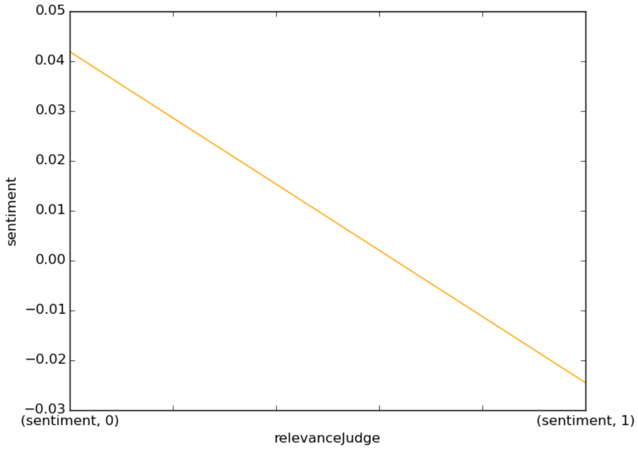

Figure 8: sentiment mean value w.r.t. the relevance