Queen’s Day 2013 on Twitter

Analyzing large scale social media

Hester Bartelsman, Eddie Stok, Jana Wagemaker,
Leonard Wein

Manos Tsagkias
Text Mining and Collective Intelligence
Amsterdam University College
May 2013

Introduction

Social media networks such as Twitter offer a wealth of real-time social data that promises insights into current events and sentiments in the population. In this project we analyze data relating to Queensday 2013 and the inauguration of the new Dutch King on the 30th of April 2013. The event was interesting for us, because it was the first abdication and coronation in The Netherlands that we witnessed and in general a big holiday in Amsterdam. Goal of the project was to obtain insights into social media perceptions of the event, to discover patterns, analyze sentiments and Retweets, relevant keywords, spatial distributions among others. The project was conducted in an exploratory fashion without a clear research question, but rather exploring what insights were obtainable from the vast amount of data. This project report presents the research approach, analyses conducted, results obtained and discussions of the results.

Approach

The data was provided by Manos Tsagkias and obtained by crawling the Twitter-Feed for a set of pre-defined keywords (See Annex A) between April 29th, 10 am and May 4th, 10 am. In total 1.788.070 tweets were crawled (excluding empty/deleted tweets). The data was stored in a MongoDB and analysis was conducted using Python. The database was accessed using Pymongo.

Results

Several questions were addressed during the research and exploration of the dataset that are presented in continuation grouped by individual analyses.

Descriptive Statistics

Script: stats.py

We conducted several analyses to determine general descriptive statistics of our dataset using operations run directly on the dataset (count and aggregate operations, see Stats.py). The findings are summarized Table 1.

Statistic	Figure
Total Tweet Count	1.788.070
Unique Users	766.836
Retweeted Tweets Count	757.950
Percentage of retweeted Tweets	42,4%
Valid Tweets (i.e. containing text, language indication, place information)	1.056.981
Same as above, but requiring non-empty content for these attributes	36.443
Number of Tweets with Geolocation/ coordinates	31.999

It is remarkable that 42,3% of tweets in the database are in fact retweets of previous tweets. How these retweets influence the social structure was analysed and is presented below under Retweet-Analysis.

Furthermore the number of tweets with geolocation is very small (about 1,8%). For the attributes check (text, language, place) the place attribute was the significant factor reducing the total tweet number. Apparently most users do not share their place location (<2% of posts).

Languages

Script: Stats.py

Using pymongo/MongoDB's aggregate function, we assessed the language attribute contained in the tweets as a means to assess spatial/geographical distribution of the tweeters in our dataset. The results are presented in the table below.

Language	**Count**
Undefined	*731089*
Dutch	595254
English	242405
Turkish	68386
Spanish	63091
Indonesia	19652
German	18139
French	11437
Portuguese	7717
Danish	3962
Norwegian	3088
Italian	2832
Vietnamese	2586
Polish	2368
Estonian	2301
Slovenian	1901
Swedish	1810
Japanese	954

Languages and their frequency counts based on language attribute contained in tweets

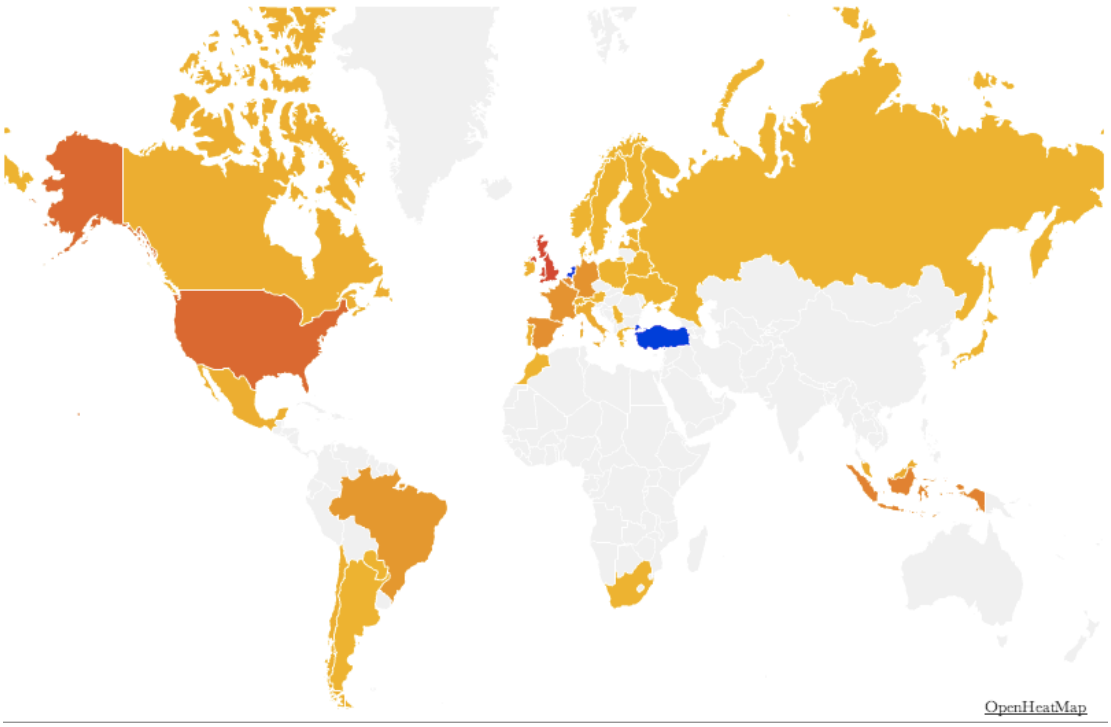
We see a majority of tweets in Dutch, which is not surprising since most of the keywords used for the crawl were Dutch, followed by English. This can be explained considering that a few prominent keywords (Queensday, etc.) were contained in the keyword set, that English is the most prominent language on Twitter and that many Dutch people are twittering in a mix of English and Dutch. The relatively large number of Spanish and Turkish tweets however is surprising, although the number of Turkish tweets can likely be linked to the large Turkish community in the Netherlands. Similarly, the high number of Indonensian tweets can be explained as a colonial heritage and the continuing bonds between the countries. It is also

remarkable that both neighboring languages (Dutch and German), while being prominent, are not very strongly represented.

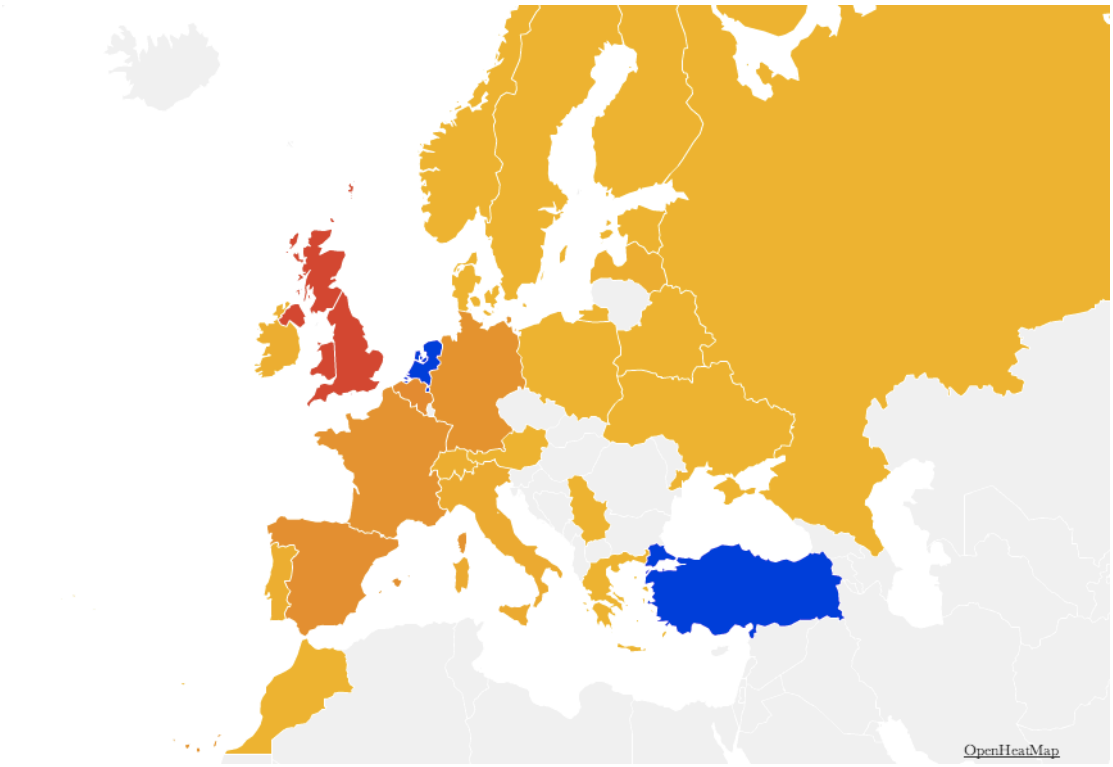
Country Distribution

Script: Stats.py

Using the same approach as above, MongoDB aggregate was used to obtain frequency counts for the countries represented in the dataset. While the country data is analysed more in detail in a later section, we present an overview of the geographical distribution in form of heatmaps as a first orientation that also serves as a comparison and confirmation of language info obtained above. The heatmaps represent Tweet volume per country, where the intensity scale increases with grey-yellow-orange-red-blue.



Heatmap of the world



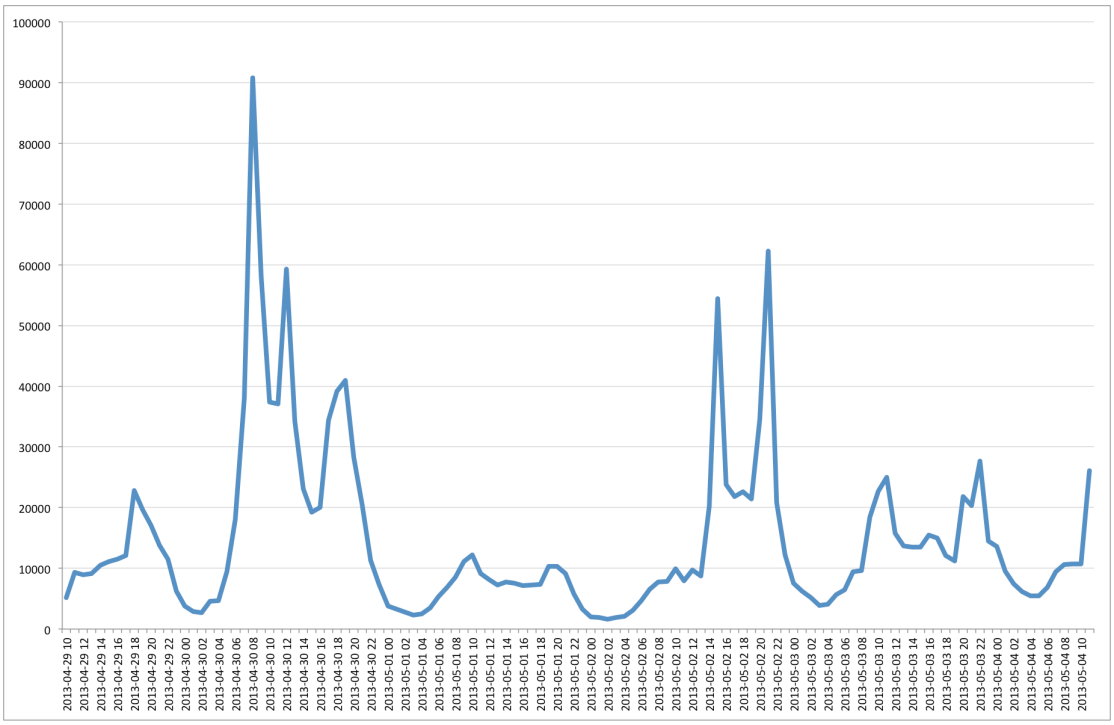
Heatmap of Europe

The heatmpas indicate congruency with the language indications. Although we would expect increased tweets from Turkey, the large amount is still surprising. The best explanations could be that the Turkish community in the Netherlands and their families in Turkey are simply very interested in Queensday and the new King. Alternatively, we might suspect that there are problems with the quality of the dataset and the tweets are in fact unrelated. Through exemplary assessment of individual tweets we got the impressions that quite some tweets were in fact unrelated to our topic of interest.

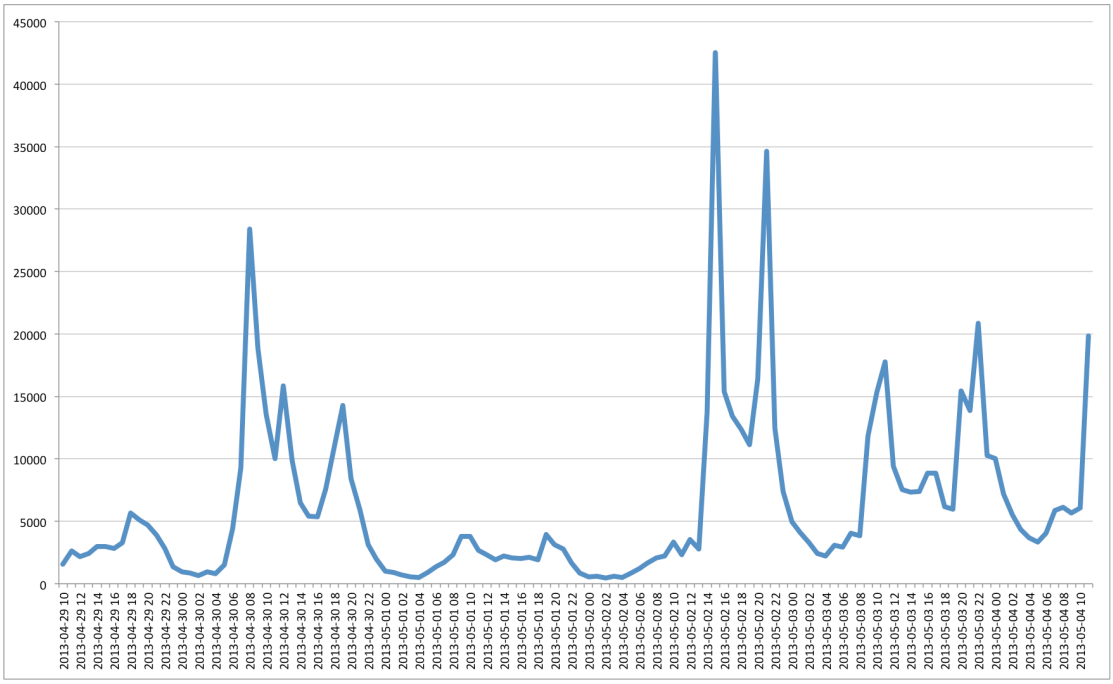
Tweet Volume over Time

Script: TweetVolume.py

In order to assess the quality of our dataset, we aggregated the tweets on an hourly basis and compared the tweet volume over time. We would expect a spike on April 30th, 2013 during the actual Queensday/Coronation.



Plot of Tweet Volume per hour between April 29th, 10 am and May 4th, 10am.



Plot of Volume of Retweeted tweets (tweets containing a 'RT') per hour between the same dates as above.

As expected the plot shows a significant spike on April 30th. However, to our surprise we found a second significant spike on May 2nd. Looking at the individual tweets, it seems that these tweets were not related to our subject, but instead

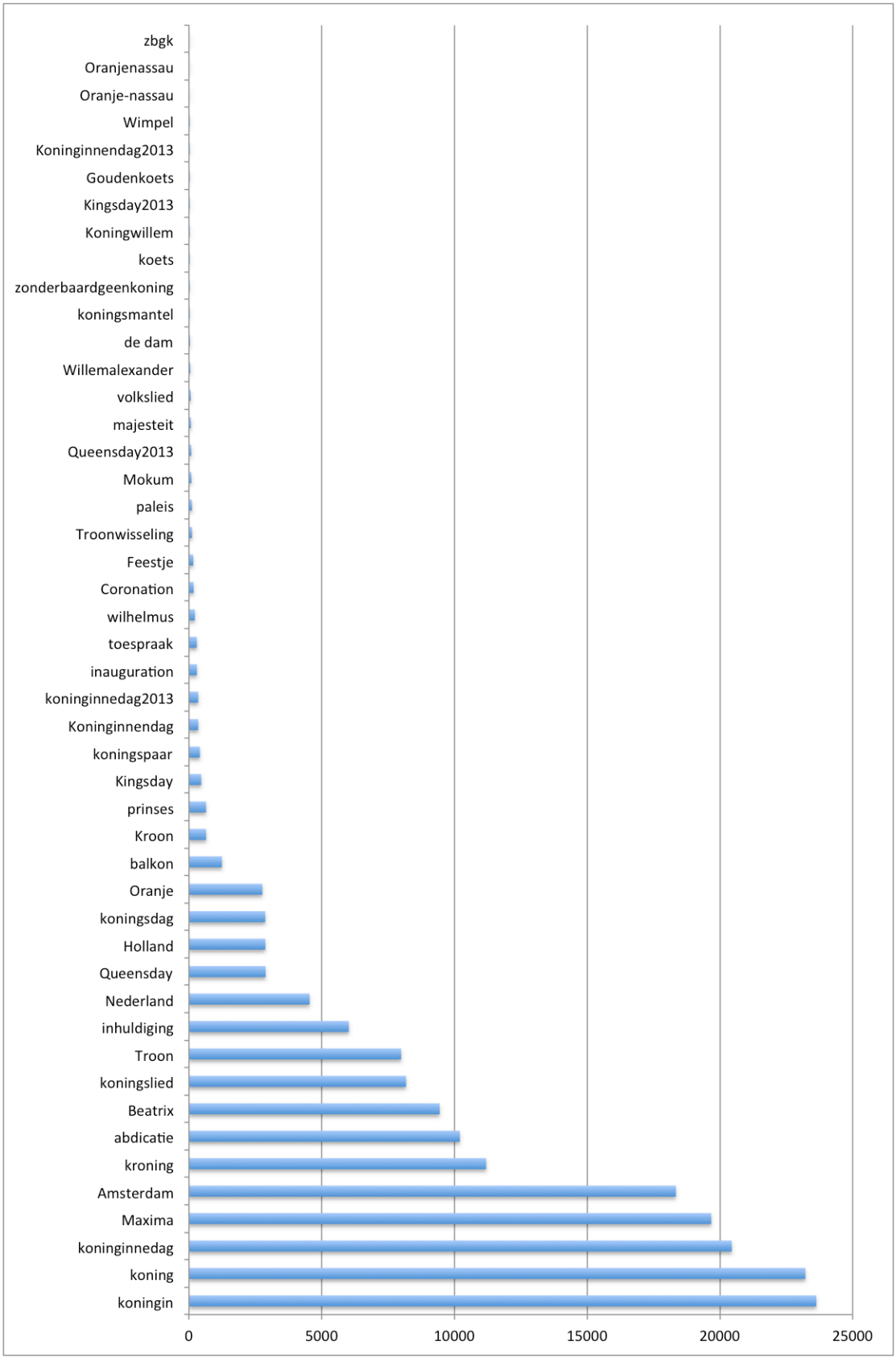
referred to a concert by a popular pop boy-band called One Direction.
In fact the same plot but filtered for Retweets (RT) shows that users were actually more actively following and retweeting about that concert than about Queensday.
We found this factor only at a later stage in our analysis and were hence not able to consider this finding. We expect that this mixing of events might have significantly reduced the quality of our dataset and hence the validity of our insights. It might have helped to consider only tweets dated previous to the second spike for our analysis.

Keyword Frequencies

In order to assess the quality of our dataset, we also conducted a frequency analysis for our keywords. Using a Regular Expression filter on the text attribute we counted the occurences of our keywords in the dataset, both for regular occurences as well as only in hastags (i.e. with a # as a prefix):

```
dbnp.tweets.find({'text':{'$regex':u'inauguration'}}).count()
```

A complete list of frequencies is presented in Appendix A.



Frequency chart for occurences of our keywords in Hashtags

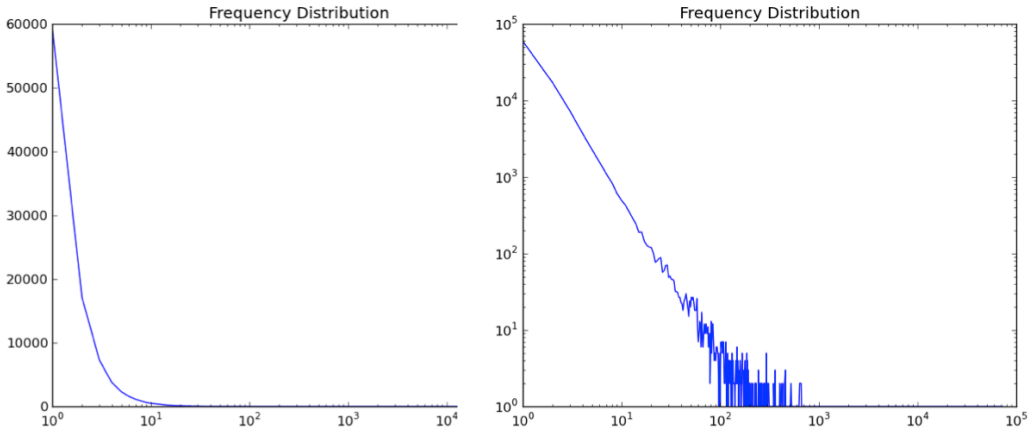
Our analysis showed that we, in fact, filtered all occurences of the predefined keywords – not only hashtags. It also showed

that only about one third of the keywords actually captured significant tweets during that time period. Most keywords were irrelevant considering the overall size of the dataset. A limitation is that ‘Willem’ the name of the new King was not included in the Keywords. Furthermore, while ‘Amsterdam’ appears ‘within range’ in the hashtag-count, it is exceptionally prevalent in the overall count, which was used to gather the dataset (3.5x more frequent than the second most frequent word, see Appendix A). This might also deteriorate the quality of the dataset, since ‘Amsterdam’ is a very general term that can not directly or necessarily be linked to the subject of our investigation. Finally, the hashtags only cover a dataset of about 180.203 tweets, whereas the total occurences of our keyword lead to 1.516.535 tweets. While this is significantly more, it is noteworthy that it is actually still 272.000 tweets short of the total count of our dataset – a finding for which we do not have a good explanation.

Retweet Analysis

Script: RetweetFast.py and Freq_Dist_Plot.py
(and RetweetSlow.py)

We have previously found that more than 42% of the tweets are retweeted tweets. We had the hypothesis that retweeted tweets might be of higher quality and thus higher relevance, since they have been likely through a sort of peer review process, i.e. at least one person would have had to approve of the content and retweet manually (disregarding the possibility of automated retweets).
First we plotted the frequency distribution of retweeted users (see below).



Retweet Frequency blogs. Left on a log-lin scale and right on a log-log scale.

The straight line for a good portion of the log-log plot indicates a Power Law distribution with fat-tail, indicating scale invariance (for the straight part). Essentially, this means that the structure of retweets is very nested around a few very influential hubs that get retweeted more by orders of magnitude than other users. This reflects a general social media theory of multipliers/hubs that are much more relevant for information propagation than other users. Therefore one idea would be to focus on retweeted tweets for further analysis as a means to exclude completely random or unrelated posts. Looking at the tweet volume graphs however, it could be that a focus on retweeted tweets might give even more weight to the concert of One Direction over our actual study subject.

Consequently, we retrieved information for the most influential users in terms of RTs in our dataset. We developed a sequential version where every user is looped over a list and then individually retrieved from the database and a parallel version for which the two loops are separated instead of nested. That is first a list of usernames is created and then we retrieve information for all these users in a separate loop that retrieves all user information at once using the $in operator (see scripts). This lead to significant performance gains.

Usernames	**Retweets**
Harry_Styles	72.666
NiallOfficial	69.718

Real_Liam_Payne	64.468
Louis_Tomlinson	43.884
1DSuperHumans	10.181
NiallSpanish	8.374
JoshDevineDrums	5.797
GaryLineker	5.783
fullmoonlouis	4.513
Queen_UK	4.220

Table with Top10 retweeted users and the retweet counts.

The table above shows the most important users in terms of retweets. In addition, I collected relevant information for all these users, but further analysis of that data is out of scope of this project. Interestingly though, most of the users in that list do not provide geoinformation. The retrieved data can be found on GitHub in serialized format (pickle) for further analysis.

Sentiment Analysis

Script: SentimentAnalysis.py and sentiment_lexicon_nl.txt

Another analysis that we conducted was a sentiment analysis. The goal of which was to identify, whether perception about the event was overall positive or negative. Initially it was planned to compare sentiments across locations, however given the pre-analysis of the data, we concluded that such finer granularity would not yield any meaningful results.

Based on an annotated sentiment/subjectivity lexicon that we obtained from the UvA, we implemented a function to compute the sentiments present in tweets to obtain an idea on whether the overall perception about Queen's Day/the coronation is perceived positively or negatively.

We conducted a sentiment Analysis for all Dutch tweets with a text attribute (not equal to None) and an existing geo-location (N=19.267). Overall, the results show approval and positive attitudes.

Sentiment	**Frequency**
Positive	5.648

Negative	1.575
Neutral	178
NA (no matching words between lexicon and tweet)	11.866

Sentiment Analysis for all Dutch tweets with geo-information

The large number of tweets for which no sentiment could be identified is explicable by the nature of tweets. Due to the character-limit, tweets have developed their own (abbreviated) vocabulary and in addition messages often contain typos, etc. Hence, the word tokenz cannot be matched with the annotated sentiment corpus containing only properly formed and spelled words with variations.
Nevertheless, the large number of positive tweets (78%) show a significant positive attitude relating to Queen's Day and the coronation.
We conducted the same analysis adding retweet ('RT') as a further condition. The Results (Positive: 44, Negative: 12, Neutral: 0, NA: 64), show that adding the RT-filter might be sensible as it preserves the positive/negative ratio (79% positive), while drastically reducing the NA-fraction. This might be an indication that indeed, retweeting introduces a quality control for tweet data and produces more correct word tokens.
However, it should be kept in mind that the quality of our dataset (in terms of being accurately focused on the research subject) might be questionable in parts, especially regarding RT, as we have analyzed before.

In Addition, we extracted the top ten keywords by frequency for both negative and positive tweets in order to 1) check whether there were apparen contradictions in the sentiment analysis and 2) whether we could find characteristic words describing negative/positive feelings about the event. The results show that no clear descriptive words can be found to distinguish between the two. The majority of words appear in both lists.

Positive Tweets	**Negative Tweets**
Troon (1535)	Troon (429)
Oranje (1157)	Amsterdam (202)

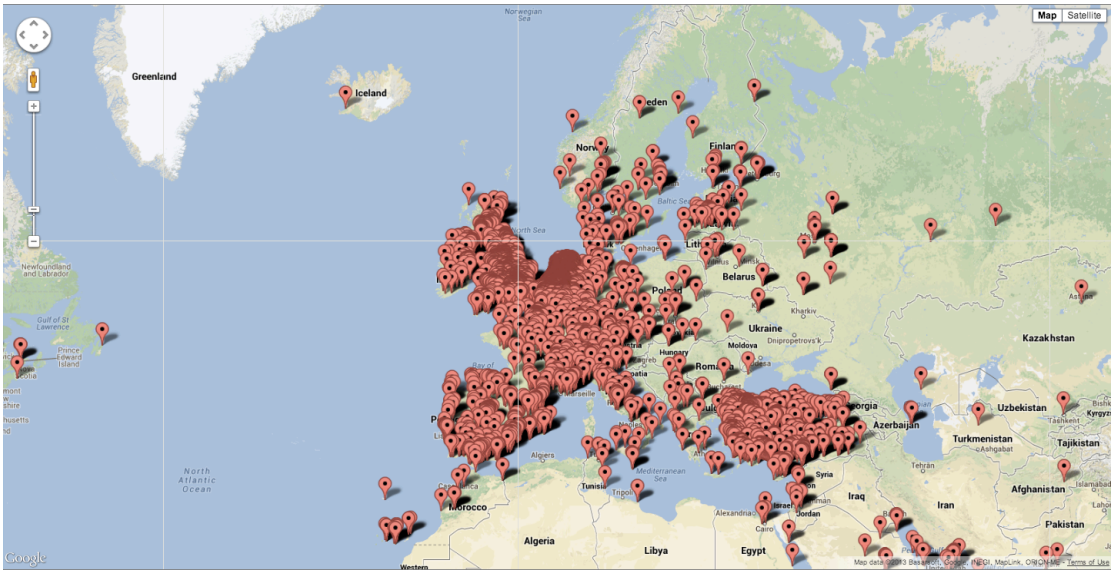
Koning (656)	Koninginnedag (174)
Mooi (580)	Wel (169)
Koninginnedag (566)	Koning (160)
Amsterdam (457)	Nederland (115)
Wel (420)	Oranje (97)
Nederland (415)	We (87)
Weer (382)	Alleen (77)
Lekker (338)	Jammer (75)

Keywords by frequency found in positive and negative tweets based on sentiment analysis.

Geographic Mapping on Google Maps

File: TweetMap.html

Using javascript and the Google Maps API, we mapped all tweets containing geolocation/coordinates (N=31.999) on a Google Map. Each individual tweet can be accessed, by clicking on the respective pin. The file is attached.



Screenshot of the TweetMap.html file that pins each tweet to its geolocation.

Word Clouds

In order to get a clear and quick overview of the words that occurred most in our crawled tweets, we needed to visualize the data somehow. Word clouds seemed a good option: they are

easy to make, they look nice, and they instantly give you an idea of the most occurring words in the tweets. Word clouds are visual representations of text data. The importance of each word is based on the number of times it occurred in all the tweets crawled on Queens day and is shown in the word clouds by font size. The colour and position is random.

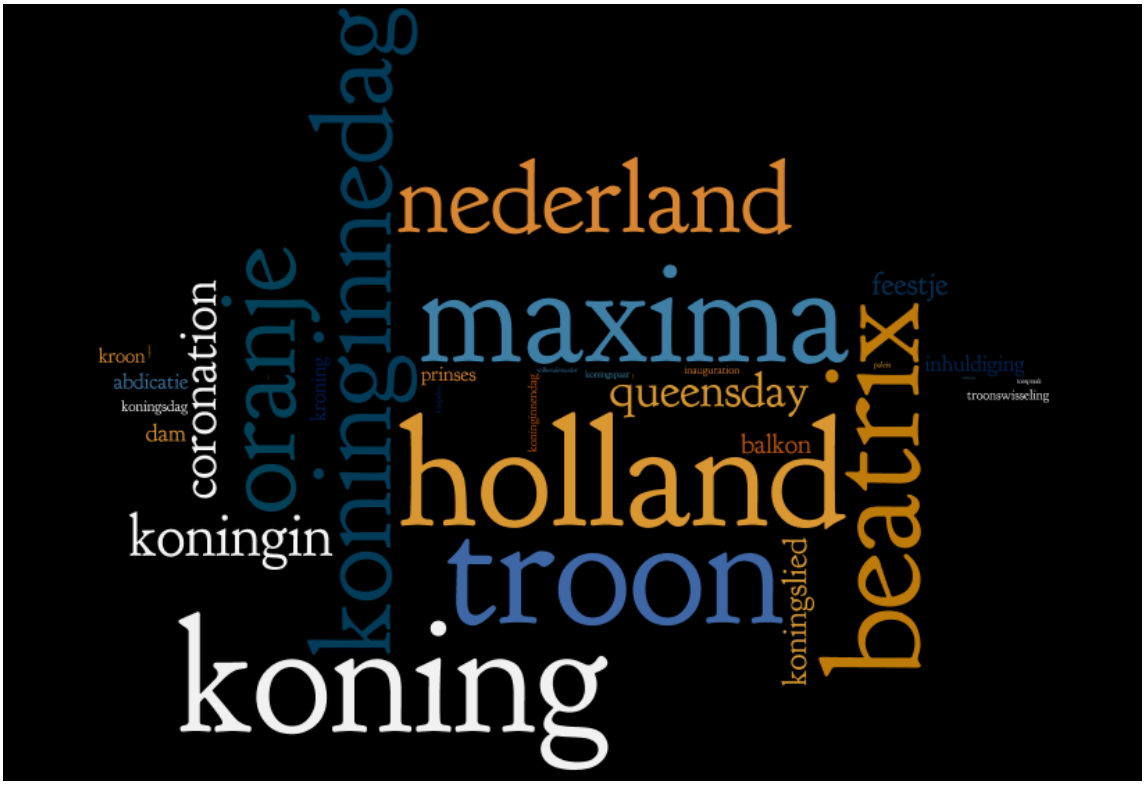
In order for Wordl.net to generate word clouds for us, we needed to put in words and their frequencies. Using nltk, we wrote a code that filters out all separate words from the text-part of the tweets, takes out all Dutch and English stop words (where stop words are the words in the nltk stop words corpus), deletes all non-words and outputs a list with all words and their frequencies. In order to erase the words with an insignificant number of occurrences, only the words that occurred fifty times or more were taken into account. The exact code is in the appendix. Once the list was created, it seemed reasonable to only take the top part of it and put that into the word cloud generator, which resulted in all words with a frequency above 5700. Deleted are ‘rt’ (retweet) and ‘http:’, because those are not interesting ‘words’ for our word clouds.

This first word cloud is one that contains all words with a frequency above 5700. As you can see ‘Amsterdam’ and ‘orange’ are both very prominent: their frequencies are 642855 and 322007 respectively. The other words are therefore so small that many are not readable anymore. To give those other words a chance, we also made a word cloud where ‘Amsterdam’ and ‘orange’ are excluded:

As you can see, this one reveals quite a bit more details than the previous one. We can see that the words are mainly English, which is interesting, since most of the key words we selected the tweets on were in Dutch. ‘Orange’ occurred 322007 times, where ‘oranje’ occurred only 69124 times. This could suggest that Dutch people prefer tweeting in English. It could also suggest that the few English hash tags we used for selecting tweets brought in a lot of English tweets not necessarily related to the coronation.

This next word cloud is one of all key terms we used for crawling all the tweets. Since ‘Amsterdam’ occurs 642855 times and the word next in line (‘koning’) occurs 102501 times, this

word cloud has pretty odd proportions. For this purpose, there is another word cloud below it with all key terms except ‘Amsterdam’. As you can see, the results are quite different.



This reveals the number of occurrences of the keywords we used with respect to one another. It is quite funny that ‘holland’ occurs more often than ‘nederland’. We can also see that ‘koning’ is more frequent than ‘koningin’, which might have been expected since it was the coronation of a king. Unfortunately ‘Willem’ was not one of our key words, so in this word cloud you cannot easily compare ‘Willem’ to ‘Maxima’. ‘Willemalexander’ is in the word cloud, but it occurs really few times compared to the others, so you can hardly see it. However, ‘Maxima’ occurred about 4 times more than ‘Willem’ in the text of all the tweets. Perhaps that would have been different if ‘Willem’ was one of the key words as well.

Since we selected tweets on a specific list of keywords, it is quite obvious those words will be frequent in our tweets. To see what other words are important, we thought it might be interesting to also make word clouds excluding the key words. Here are another two word clouds of all the words with a frequency above 5700 except for the key terms we used for selecting the tweets in the first place. The first word cloud contains the word ‘orange’, the second one does not.


orange
wear
coming
tonight
niallofficial
tomorrow
show
gig
code
la
dress
everyone
advised
see
please
thanks
wearing
everybody
people
today
thank
queen
new
happy
turned
crowd
strict
netherlands
willem
dag
nieuwe
morgen
must
toooooo
day
vandaag
wel
dutch
may
make
troonrt
louis
never
sure
ziggo
unbelieveable
weturnedziggoorange
que
noise
street
memo
king
boys
hi
got
final
amp
el
much
zayn
arena
us
like
wait
heard
mooi
alexander
harry
reina
chelsea
niall
liam
onze
last
one
ca
y
weer
via
madnessrt
also
echt
ready
benfica
love
hotel


tomorrow
wear
niallofficial
show
coming
gig
tonight
wearing
thanks
see
dress
la
everybody
please
advised
strict
code
queen
everyone
happy
new
day
crowd
netherlands
willem
thank
que
hi
like
people
much
never
memo
boys
liam
el
zayn
heard
noise
today
weer
street
amp
wel
ziggo
must
weturnedziggoorange
unbelieveable
wait
final
got
harry
mooi
madnessrt
louis
may
dutch
direction
morgen
turned
king
dag
vandaag
chelsea
also
make
toooooo
nieuwe
arena
troonrt
ca
via
sure
concert
onze
going
reina
last
fans
even
one
time
ready
y
alexander
benfica
hotel
roache
come
dress

As we can see ‘niallofficial’ is quite a big word. Searching the Internet, we found that niallofficial is the twitter account of Niall Horan, a band member of One Direction. It could be that people twittered about Niall Horan because of One Direction’s concert in The Hague on the third of May. Perhaps people spotted him and tweeted something like ‘#best Queensday ever’? That could explain why Niall appeared in the tweets we selected with our keywords. Another big word is ‘tomorrow’. We are not sure why that occurs so many times. It occurs way more often than its Dutch equivalent.

Detailed Geographic Analysis

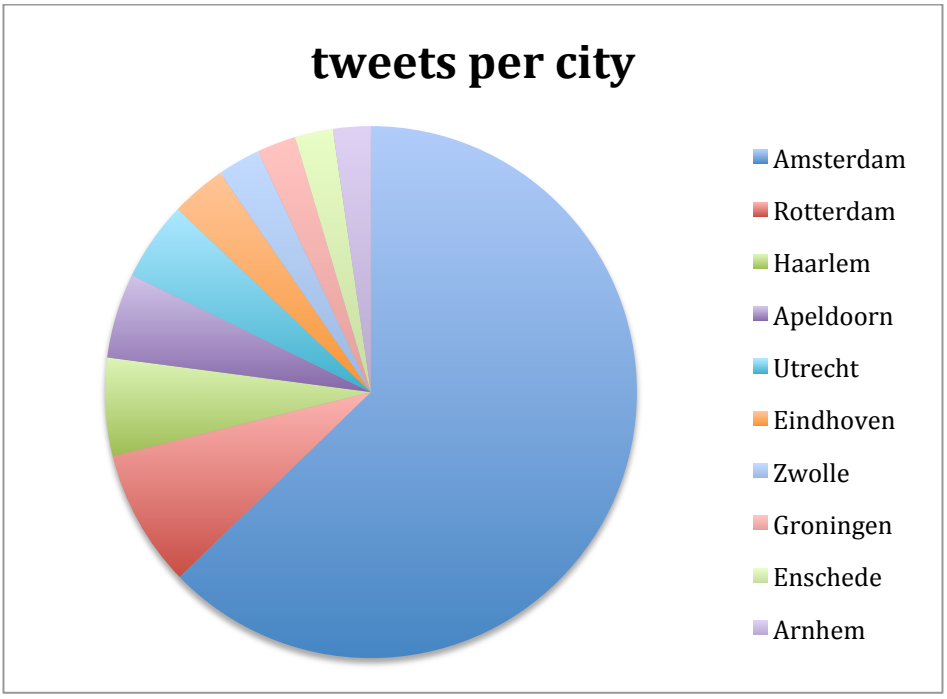
Top tweeting Dutch cities

We decided it would be more informative to make a bar chart of the tweets per city in the Netherlands than a histogram. To do this, we created a dictionary with a city as key, and the number of tweets coming from that city as value. We didn’t want to include the numbers of tweets of every municipality, so we selected the 50 largest municipalities in the Netherlands to include in the chart. To do this, we created a list of these 50 municipalities, and wrote the code to only include the cities in the dictionary that occurred in the list. For some reason many of the locations gave a UnicodeEncode Error, which interrupted the loop in our code. To fix this problem, we included a ‘Try-Except” debugging system.

In Table 1 (in the appendix), you can find all 50 largest municipalities in the Netherlands corresponding to the number of tweets they generated. Keep in mind that these are (most probably) not all the tweets generated, merely the tweets with a geo-location in the code.

To visualize the result we made a pie chart of the 10 most ‘tweeting’ municipalities (pie chart1). It doesn’t come as a surprise that Amsterdam generated the most tweets, as Amsterdam is the capital, and the location of the inauguration. Of course this pie chart gives the impression that over half of the tweets were from people located in Amsterdam at the

moment of the tweet, but this is not necessarily the case as the pie chart is only a rough estimate as it only reflects on the 10 most tweeting municipalities. The pie chart is still an interesting graph, however, because it does give a good idea of the amount tweets of the ten top municipalities with respect to one another. To see the exact number of tweets per city, see Table 1 in the appendix.



Pie Chart 1

In order of largest municipality	tweets per city
Amsterdam	6360
Rotterdam	847
Den Haag	4
Utrecht	488
Eindhoven	339
Tilburg	264
Groningen	241
Almere	184
Breda	210
Nijmegen	168

Top 10 Dutch cities alphabetically	tweets per city
Amsterdam	6360
Rotterdam	847
Haarlem	603
Apeldoorn	524
Utrecht	488
Eindhoven	339
Zwolle	264
Groningen	241
Enschede	234
Arnhem	232

Table 2

Table 2 allows for a comparison between the 10 largest municipalities and the 10 most tweeting municipalities. On might expect these to be very similar, as it would be a natural assumption to think the largest cities would attract more people during Queensday and therefor generate the largest number of tweets. Overall, it seems that this is the case. However, there are some interesting deviations.

The most obvious deviation is that Den Haag (The Hague) doesn't appear on the list of the 20 most tweeting municipalities. The direct reason is obvious: Den Haag only shows to have generated 4 tweets. The reason for this is that Den Haag is probably spelled 's Gravenhage in most tweet geo locations. Were we to run the code again we should include 's Gravenhage in the list of cities instead of Den Haag. It is really a shame we are missing the actual number of tweets for the Hague, as it is one of the major cities in the Netherlands, definitely with respect to the royal family.

As you see there is a lot of overlap in these two tables, which makes sense: The larger municipalities will attract more people for Queensday, and will therefore generate more tweets. Interesting to see though, is that Den Haag, Almere, Breda and Nijmegen did not make it to the top tweet list, and had to make room for Haarlem, Apeldoorn, Enschede and Arnhem. That Den Haag didn't make it to the top tweet list is probably a mistake, which I explained above. The differences in tweet count between Almere, Breda, Nijmegen, Eschede and Arnhem is not very large, and therefor I do not think it is a very significant finding that the list shifted slightly. What I do find very interesting is how high up on the list Haarlem and Apeldoorn are. The reason for Haarlem being so high on the list is probably that it is the capital of Noord-Holland, even though it may not be very large. The reason for Apeldoorn being so high on the list is probably that there aren't many major cities in that area of the Netherlands. Consequently, the whole population surrounding Apeldoorn goes to Apeldoorn to celebrate Queensday, whereas people living in the West of the country have more major cities to choose from. Also, Apeldoorn is the 12th largest municipality in the Netherlands, so it is not that extreme of a surprise.

For future, similar, projects, it might be an interesting idea to make a sentiment analysis between Dutch cities. Can you notice a difference in the protestant regions and the catholic regions (keep in mind that the royal family is protestant)? For this particular research a sentiment analysis would not be very trustworthy, as, relatively speaking, we did not collect many tweets per city.

The code for collecting the data is included in the appendix in DATA1.

Top tweeting countries around the world

The collection of the top tweeting countries is very similar to the collection of the top tweeting Dutch municipalities, so I will not further discuss the technique within python. The code can be found in the appendix in DATA2. However, there were some issues with this data we did not encounter with the collection of tweets per city, as we could not use the same trick of creating a list of countries. This list would have had to be too elaborate, and would have wound up not saving time, but creating extra work.

Once we created a dictionary of all countries selected by geo-location, we deleted the items with a value under 10. Some of these keys were obscure countries, but others were countries with where the geo-location had different spelling, for example: the e-umlaut translated to '\xeb'. We did some manual work with this code as the countries were stated in different languages, and we couldn't get python to recognize the different spellings as the same country. An option might have been to write a code so if the word were similar enough (e.g. a certain amount of letters would be in the right place) the word would be equal to the other, counting Italie and Italia as the same country. However, this probably wouldn't work, as Pays-Bas and 'Nederland' both are the same country, but look completely different.

Looking at table2 (the top 10 tweeting countries) the number of tweets coming from Turkey is astounding. Looking at the geo

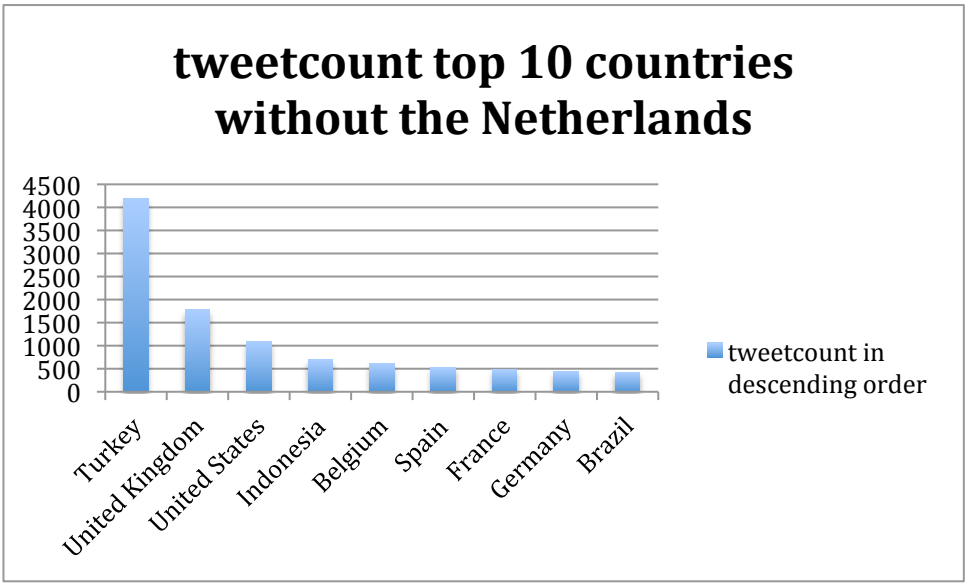
location google maps you can already tell that tweet, but the number of tweets coming from Turkey far exceeds the tweets coming from other countries (excluding the Netherlands), even countries closer to the Netherlands as Belgium and Germany. We have to keep in mind though, that we crawled tweets with keywords that were not necessarily specific to Queensday, like 'inauguration'. Other puzzling observations that can be made in Table 2 are the number of tweets coming from the United States. Countries in Europe, close to the Netherlands are not surprising to have generated many tweets, but why the United States would have such an interest is odd. This might be something to look into with future research. The tweeting results in Indonesia can be explained by the country being a former colony of the Netherlands. Brazil again is an interesting deviation from reasonable expectation.

Country	tweetcount in descending order
The Netherlands	24762
Turkey	4195
United Kingdom	1778
United States	1088
Indonesia	700
Belgium	610
Spain	525
France	483
Germany	442
Brazil	416

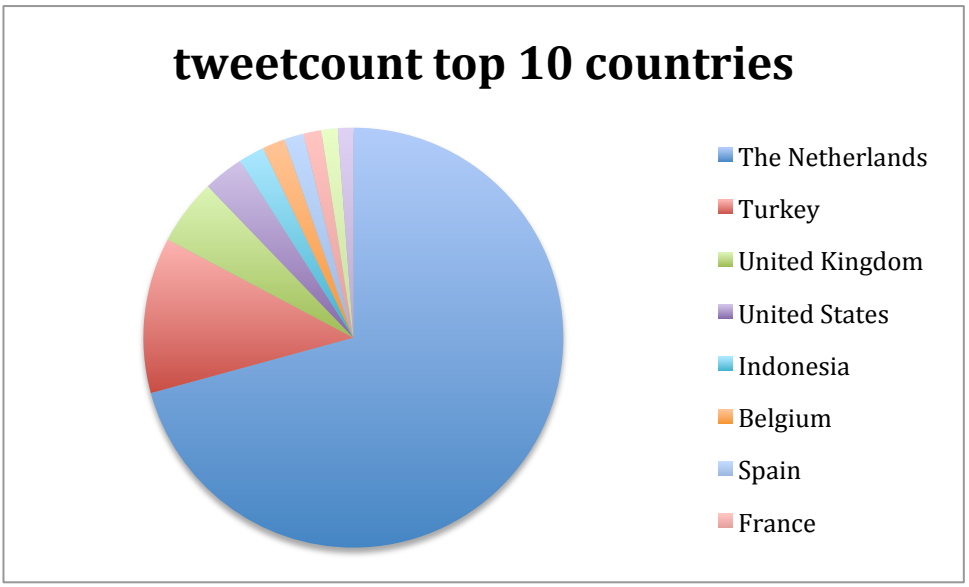
Table 2

We made a pie chart (pie chart 2) of the ten countries that tweeted with our hashtags most. The country with the next highest amount of tweets (not on the chart) is Italy with 161 tweets. Just as in the pie chart we used to give an impression of the tweets per city, this chart cannot be used as an exact reference to the number of tweets per country. The reason the pie chart is useful though is that it is easy to estimate the distribution in a glace. For the exact number of tweets per country consult Table 2, of see Bar Chart 1, generated from

the exact same data as pie chart 2, but excluding the Netherlands for scaling purposes.



Bar chart 1



Pie chart 2

Appendix

All code (and data and visualization files) are available in the public GitHub repository TM-Project.

Appendix A: Keywords and their frequencies

Keyword	Hashtag Count	Total Count
Amsterdam	18341	646399
koning	46848	187450
koningin	23630	84329
Holland	2875	83462
Beatrix	9443	79667
Maxima	19672	70361
Nederland	4542	66957
koninginnedag	20445	47630
Coronation	173	27703
Queensday2013	85	27703
Oranje	2763	25803
kroning	11193	19605
koningslied	8175	17546
balkon	1240	17450
Queensday	2885	16236
abdicatie	10202	15810
inhuldiging	6019	15757
prinses	641	13019
Troon	7991	9769
koningsdag	2871	8826
koningspaar	408	5505
inauguration	297	5405
toespraak	292	4954
Kingsday	460	3343
de dam	42	3213
Kroon	643	2331
Koninginnendag	347	2076
Feestje	159	1962
paleis	111	1618
volkslied	66	1116
majesteit	75	989

wilhelmus	220	603
koets	13	457
Troonwisseling	114	441
koninginnedag2013	346	354
Mokum	90	319
koningsmantel	29	174
Willemalexander	52	120
Wimpel	2	37
zonderbaardgeenkoning	14	14
Koningwillem	10	10
Kingsday2013	3	5
Goudenkoets	3	4
Koninginnendag2013	3	3
Oranje-nassau	0	0
Oranjenassau	0	0
zbgk	0	0

Appendix B: WordCloud Code

Code for the word clouds

```
import nltk

f=open("myfile.txt")
text=f.read().lower()
f.close()

print "finished reading"

tokens=nltk.word_tokenize(text)

wordfile=open("wordfile.txt",'w')

print "sdfsd"

fd=nltk.FreqDist(tokens)
for token in fd:
      if fd[token]>50 and token not in
nltk.corpus.stopwords.words('dutch') and token not in
nltk.corpus.stopwords.words('english') and token.isalpha():
            wordfile.write(token.encode("utf-8") + ':' +
str(fd[token]))

wordfile.close()
```

Appendix C: Geographic Analysis

50 largest municipalities and generated number of tweets

Dutch cities alphabetically	tweets per city
Alkmaar	116
Almelo	54
Almere	185
Amersfoort	180
Amstelveen	118
Amsterdam	6360
Apeldoorn	524
Arnhem	232
Breda	210
Delft	100
Den Bosch	131
Den Haag	4
Deventer	126
Dordrecht	221
Ede	95
Eindhoven	339
Emmen	124
Enschede	234
Gouda	115
Groningen	241
Haarlem	603
Heerlen	76
Helmond	105
Hengelo	81
Hilversum	224
Leeuwarden	84
Leiden	84
Leidschendam-Voorburg	40
Lelystad	86
Maastricht	93
Nijmegen	168
Oss	70
Purmerend	61
Roosendaal	60

Rotterdam	847
Schiedam	106
Sittard-Geleen	81
Tilburg	183
Utrecht	488
Venlo	76
Vlaardingen	103
Westland	138
Zaanstad	115
Zoetermeer	118
Zwolle	264

```
tweet_fullname={}

for post in dbnp.tweets.find({"place": {"$exists": True}})[:100000]: #be sure
to take this 100 out later!
        if post["place"] is None:continue
        tweetcity=post["place"]["full_name"]
        incity=False
        try:
                for city in citylist:
                        if city.decode('latin1') in tweetcity.decode('latin1'):
                                incity=True
                                break
                if incity==False:continue
                if city in tweet_fullname:
                        tweet_fullname[city]+=1
                else:
                        tweet_fullname[city]=1
        except UnicodeEncodeError:continue

tweet_fullname
```

Code to generate City Frequency Counts

Countries and their Tweet-Frequencies in our dataset.

Country	tweetcount in descending order
The Netherlands	24762
Turkey	4195
United Kingdom	1778
United States	1088
Indonesia	700
Belgium	610

Spain	525
France	483
Germany	442
Brazil	416
Italy	161
Malaysia	130
Canada	116
Ireland	101
Mexico	89
Chile	71
Latvia	69
Austria	51
Switzerland	50
South Africa	43
Sweden	42
Argentina	37
Denmark	35
Greece	34
Marokko	22
Poland	21
Russia	17
Japan	15
Norway	15
Serbia	14
Ukraine	13

```
tweet_country={}

for post in dbnp.tweets.find({"place": {"$exists": True}}): #be sure to take this
100 out later!
        if post["place"] is None:continue
        country=post["place"]["country"] # shows the country
        if country in tweet_country:
                tweet_country[country]+=1
        else:
                tweet_country[country]=1


tweet_country
```

Code to generate Country Frequency Count