Twitter and the European Hyperagora: What can the Twittersphere Tell us about Political Deliberation and Opinions in Europe?

M. Callaghan, V. Niberg

November 14, 2015

Data Gathering

Due to the large amount of data we process, we ran the data gathering and cleaning in the background on a server using the prefix setsid.

Tweets

We used a modified version of GetOldTweets (Henrique 2015), a Java program that scrapes data from twitter search. The file getting_tweets/input.txt contains a list of search terms related to the Greek crisis in three periods, each comprising some weeks before and after the negotiation and signing of the memoranda. The search terms were collected using an adapted form of snowball sampling (Biernacki and Waldorf 1981), searching an initial list and recursively adding related terms found in the results. By running

```
sudo setsid ./compile_run.sh ../getting_tweets/input.txt
```

from the GetOldTweets folder, we ran through each search term and each period, searched twitter, and saved the results as a txt file in the data folder. After an initial assessment of the results, we refined our search terms and ran GetOldTweets again with /getting_tweets/input2.txt. A third file (getting_tweets/input3.txt) aims to return a time-inpedependent list of tweets in order to control for the growth of Twitter over time.

We end up with a long list of files in the data/GOToutput folder, which in the data cleaning process will be merged into one corpus file.

Users

We found the unique users in our corpus of tweets and used the TwitteR package (Gentry 2015) to gather richer data about each user. TwitteR uses the twitter API and gives the opportunity to collect all information twitter has abou the user. Where a users's last tweet was geocoded, we took the latitude and longtitude. We end up with the file data/user_info.csv

Many users do not geotag their tweets, instead stating their location, and we used APIs from MapQuest and Google to geocode user-reported location, giving us the file places.csv.

Data Cleaning

The txt files containing the tweets for each query and period are merged into a corpus file. This corpus file was merged with the user_info file, which in turn was merged with the places file. We end up with a large file containing tweets for our queries in each period with elaborete user information.

Some of the queries we defined returned irrelevant data, due to their ambiguity. We identified these by selecting random tweets from the search queries, reading the tweets, and checking for relevance to the topic. For example, the query "bailout", although certainly relevant for our topic, was insufficiently precise and returned a lot of data about the banking bailouts, especially in the 2010 period.

The following list summarizes the queries which we excluded.

- athens
- bailout
- 2-pac
- 3-pac

Data Analysis

Descriptive Statistics

The word cloud gives us an overall picture of the words used in the collected tweets connected to the European public-debt crisis.[1]

The following tables and bar charts give a first overview over the collected data. The first table shows absolute numbers and relative distributions of specific query returns. The second table describes distribution of specific tweets over time. The bar chart describes geographical distribution of tweets for those tweets that we have at the time of writing this been able to obtain information on location.[2]

query	n	percent
#aGreekment	6103	1.13
athens+berlin	1263	0.23
austerity+greece	12362	2.29
economic+adjustment	283	0.05
eucrisis	260	0.05
eurocrisis	6449	1.20
eurogroup	3397	0.63
eurogroup+agreement	748	0.14
eurosummit	17780	3.29
euro+summit	22450	4.16
germany+greece	12757	2.36
greece+crisis	12540	2.32
greece+reforms	15497	2.87

[1] to a certain degree in this case a word cloud is redundant, because it mainly returns our query turns. However, it also indicates the prevalence of specific terms, and further, which other terms are mentioned regularly in those tweets.

[2] API from Google restricts requests to a daily maximum and the API from MapQuest restricts requests to a monthly maximum.

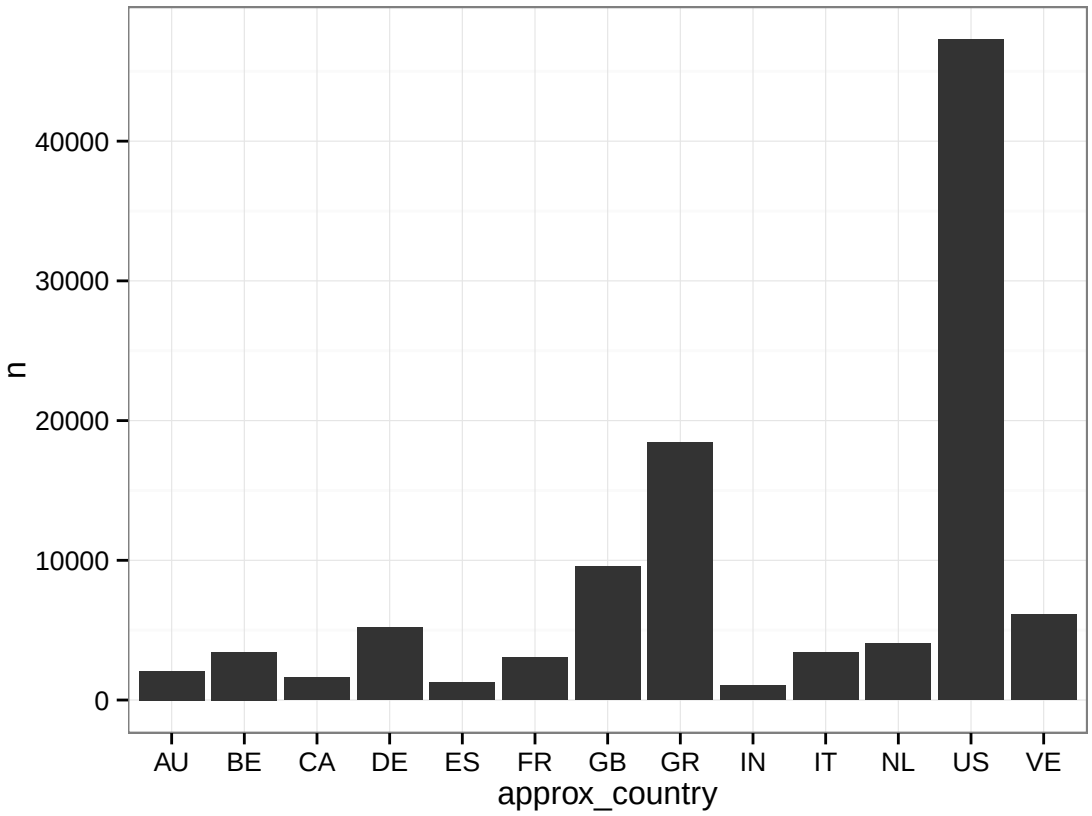
query	n	percent
grexit	15017	2.78
imf+greece	9842	1.82
memorandum+greece	1165	0.22
merkel+greece	16478	3.05
mou+greece	958	0.18
#nai	9721	1.80
#nai+greece	323	0.06
notmyeurope	395	0.07
#oxi	4940	0.92
rescue packages	155	0.03
schäuble+greece	5949	1.10
syriza	6945	1.29
tax+evasion+greece	1059	0.20
#thisisacoup	18322	3.40
#thisisnotacoup	778	0.14
tsipras	18684	3.46
varoufakis	31630	5.86
bailout+eurozone	19674	3.65
bailout+greece	28117	5.21
bailout+greek	8559	1.59
efsm+greece	673	0.12
efsm+greek	181	0.03
ems+greece	90	0.02
esm+greek	862	0.16
greece+debt	25472	4.72
greek+crisis	84478	15.66
greek+debt	26343	4.88
greek+reforms	11587	2.15
memorandum+greek	479	0.09
six-pack+greece	14	0.00
six-pack+greek	39	0.01

The results that our queries returned differ substanially in size. While some queries (e.g. merkel+greece) return over 15000 tweets, others (e.g. greece+reforms) return only around 3000 tweets in the specified time periods. We therefore think it is for analytical reasons useful to differentiate between high-return and low-return queries.

period	control	n	percent
2010	0	28165	7.10
2010	1	8875	11.26
2012	0	78049	19.67
2012	1	65123	82.65
2015	0	290508	73.23
2015	1	4799	6.09

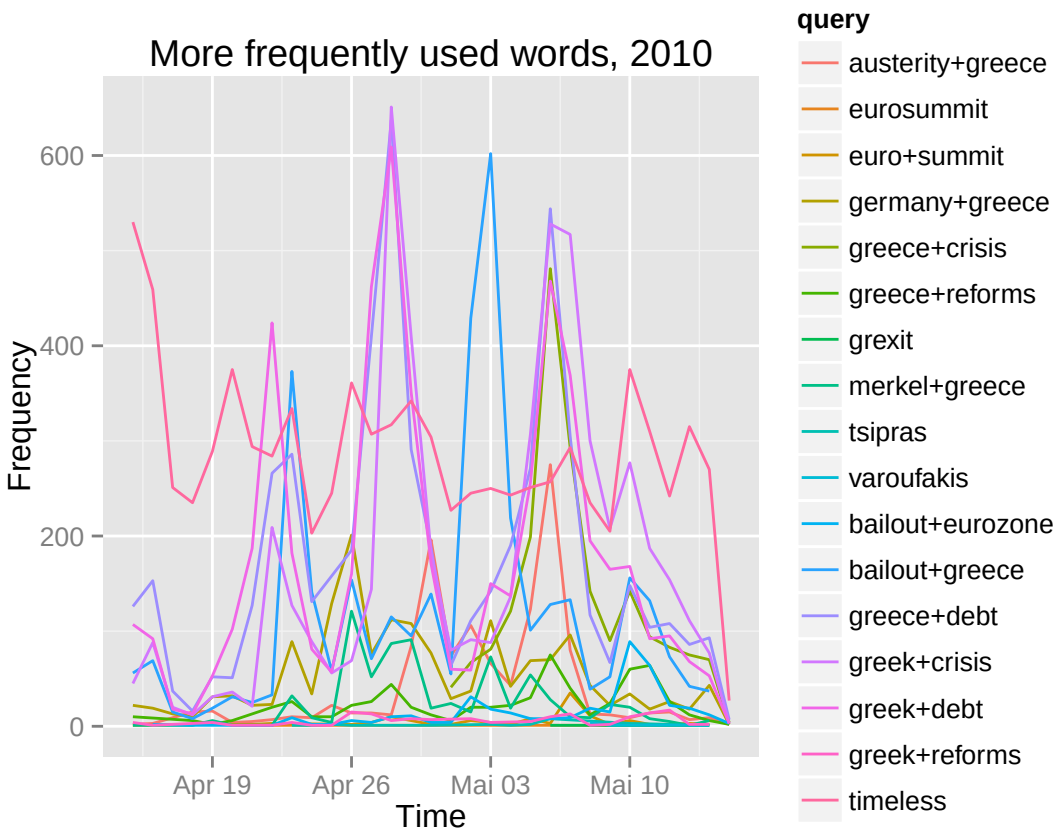
The results show that twitter coverage of the Greek bailouts has increased heavily over time. However, since the shown results are not normalized, it is difficult to say how much of the growth of the population of tweets can be attributed to an increase in twitter usage or to an increase of interest in the topic. As a control query, we included "timeless" in our search. However, this query does not behave the way we had expected it to. For the finalized version of the project, we will have to identify a more appropriate control query. When

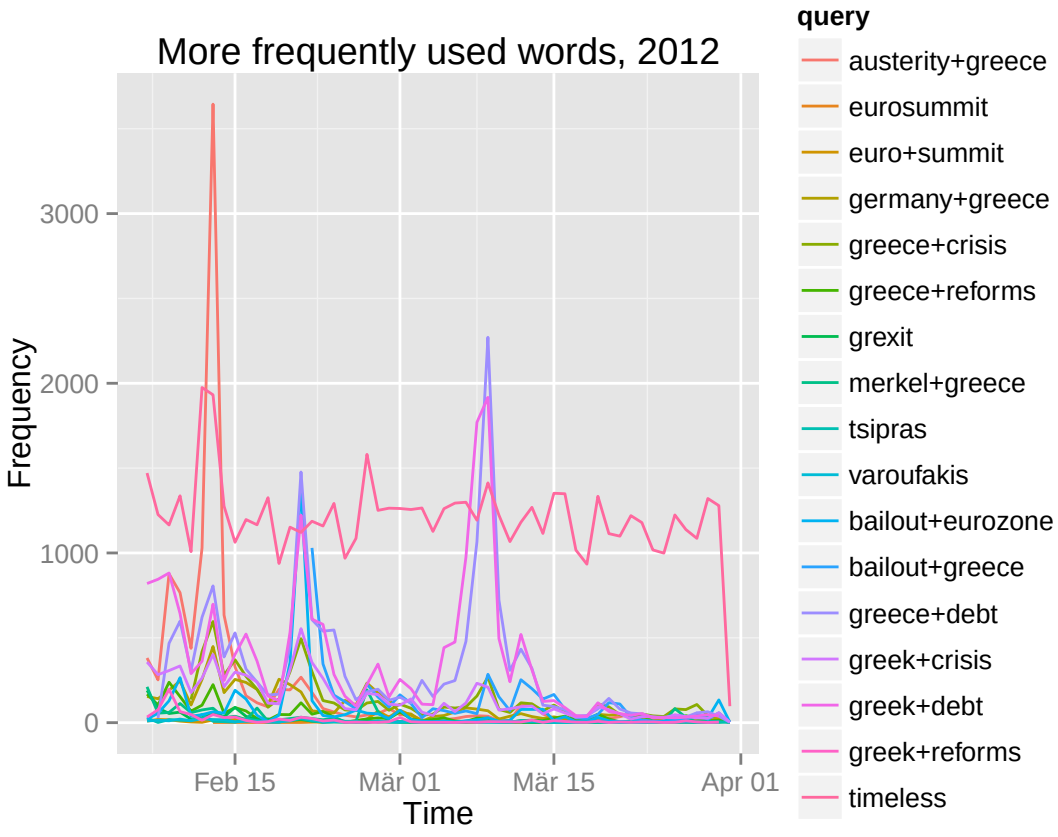
looking at the development of different queries over time, an increase can be found for all queries except for for imf+greece, which decreased in 2012 and increased in 2015 again.

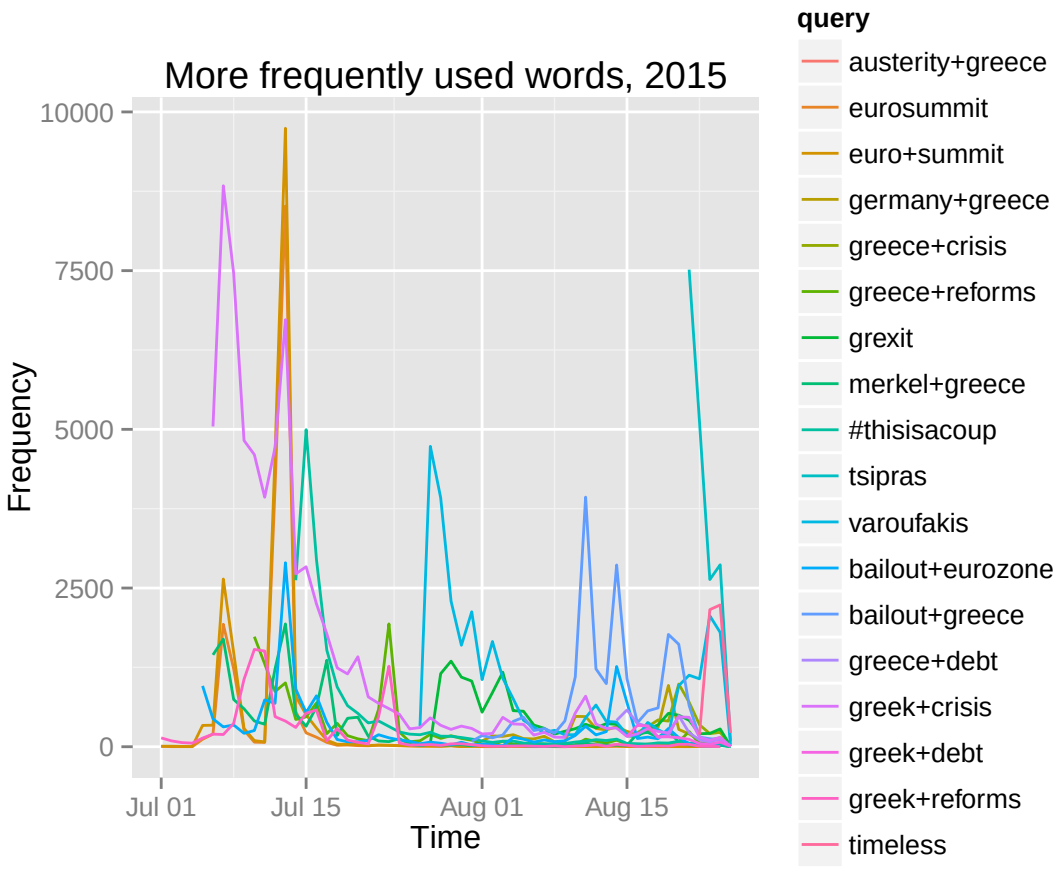


This bar chart shows the distribution of all tweets over countries filtering those with a significant amount of tweets. This group seems to be a mix of high twitter user numbers and affectedness/involvement/relation to the events in Greece. Most probably reflecting high user numbers in the US, the US seems to be the origin of most tweets regarding the European sovereign-debt crisis. However, when controlling for period, the US resembles an atypical case. While for most other countries, tweet number on the topic increase over the years, they decrease substantially in the US.

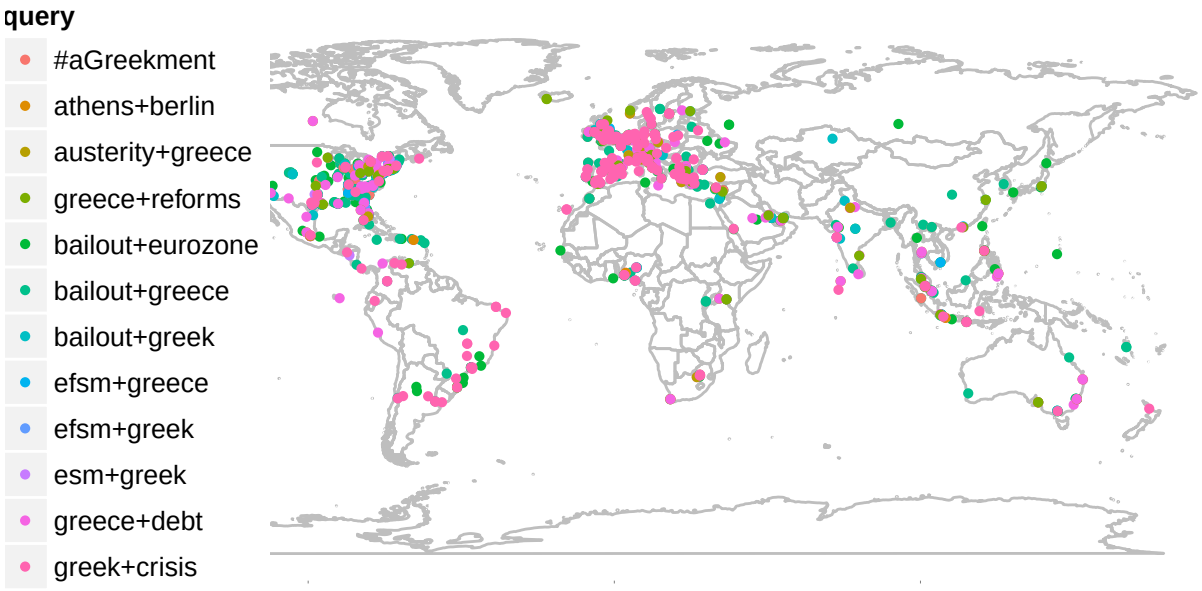
We argued earlier that it would be helpful to distinguish between two groups of queries/tweets by high and low return rates in order to understand the behaviour of the queries. The following charts report the developments over time for the high return rates as this group creates more intelligle results. Extreme points in these charts point to certain events. Also, the emergence of certain hashtags can be observed.


More frequently used words, 2010
Frequency
600
400
200
0
Apr 19
Apr 26
Mai 03
Mai 10
Time
query
austerity+greece
eurosummit
euro+summit
germany+greece
greece+crisis
greece+reforms
grexit
merkel+greece
tsipras
varoufakis
bailout+eurozone
bailout+greece
greece+debt
greek+crisis
greek+debt
greek+reforms
timeless

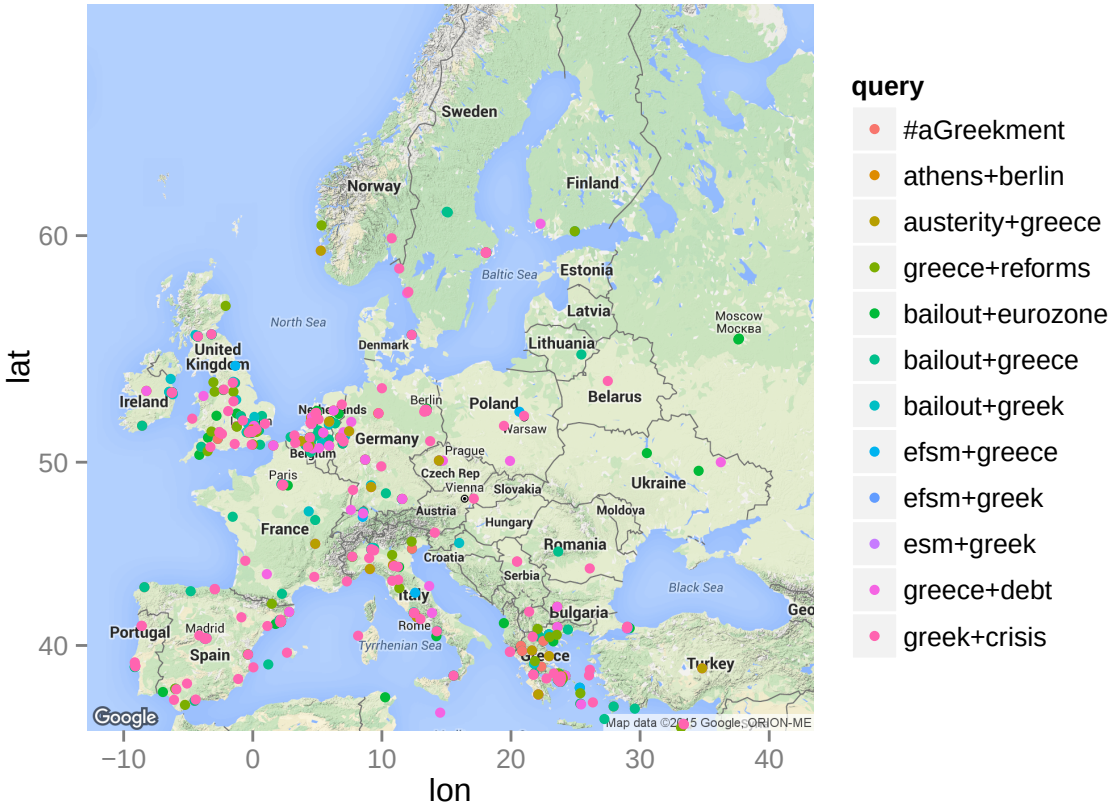

More frequently used words, 2012
Frequency
3000
2000
1000
0
Feb 15
Mär 01
Mär 15
Apr 01
Time
query
austerity+greece
eurosummit
euro+summit
germany+greece
greece+crisis
greece+reforms
grexit
merkel+greece
tsipras
varoufakis
bailout+eurozone
bailout+greece
greece+debt
greek+crisis
greek+debt
greek+reforms
timeless


More frequently used words, 2015
Frequency
10000
7500
5000
2500
0
Jul 01
Jul 15
Aug 01
Aug 15
Time
query
austerity+greece
eurosummit
euro+summit
germany+greece
greece+crisis
greece+reforms
grexit
merkel+greece
#thisisacoup
tsipras
varoufakis
bailout+eurozone
bailout+greece
greece+debt
greek+crisis
greek+debt
greek+reforms
timeless

As our data includes information on location and our research question includes a comparative aspect, in the following section we will describe our data by locational information.



This map shows the distribution of individual queries over different locations worldwide. Colour indicates type of query. It shows "twitter hubs" on this issue in Greece, Benelux-states an GB.



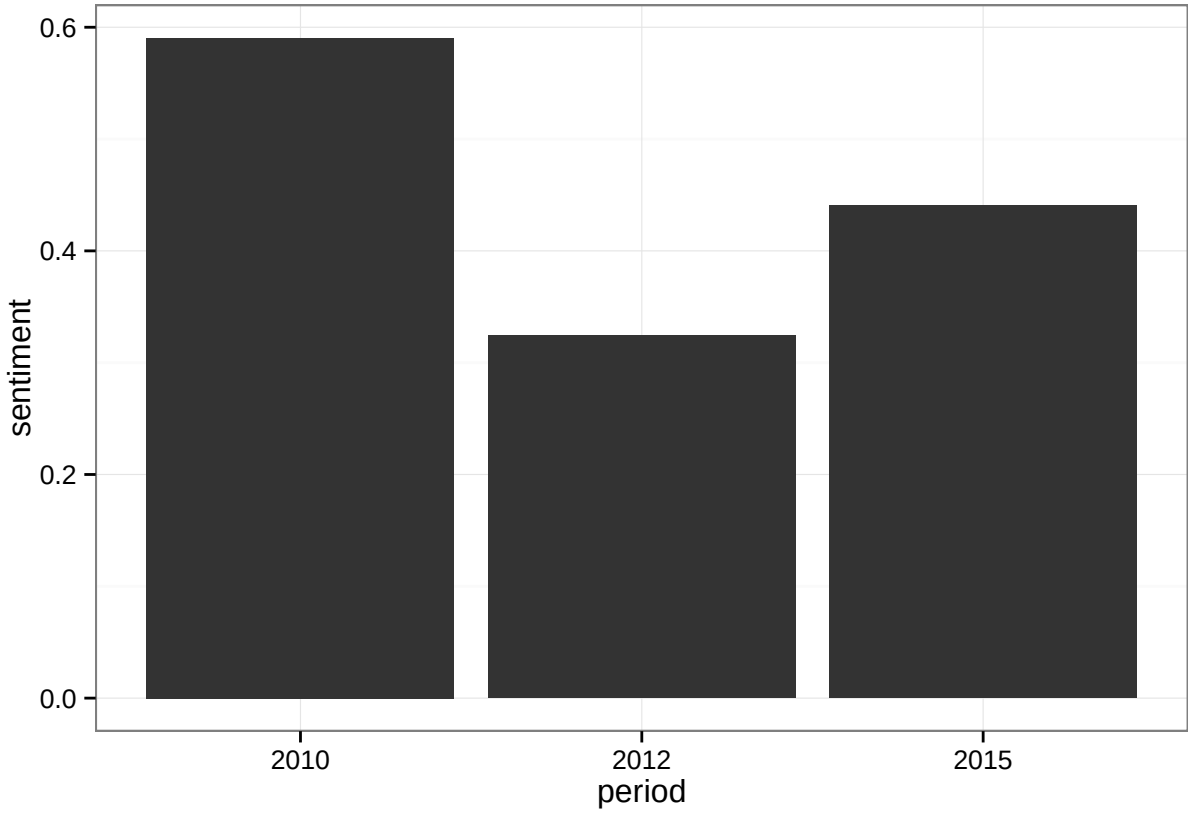
The following maps show the distribution of individual queries over different locations Europe-wide. Colour indicated type of query.

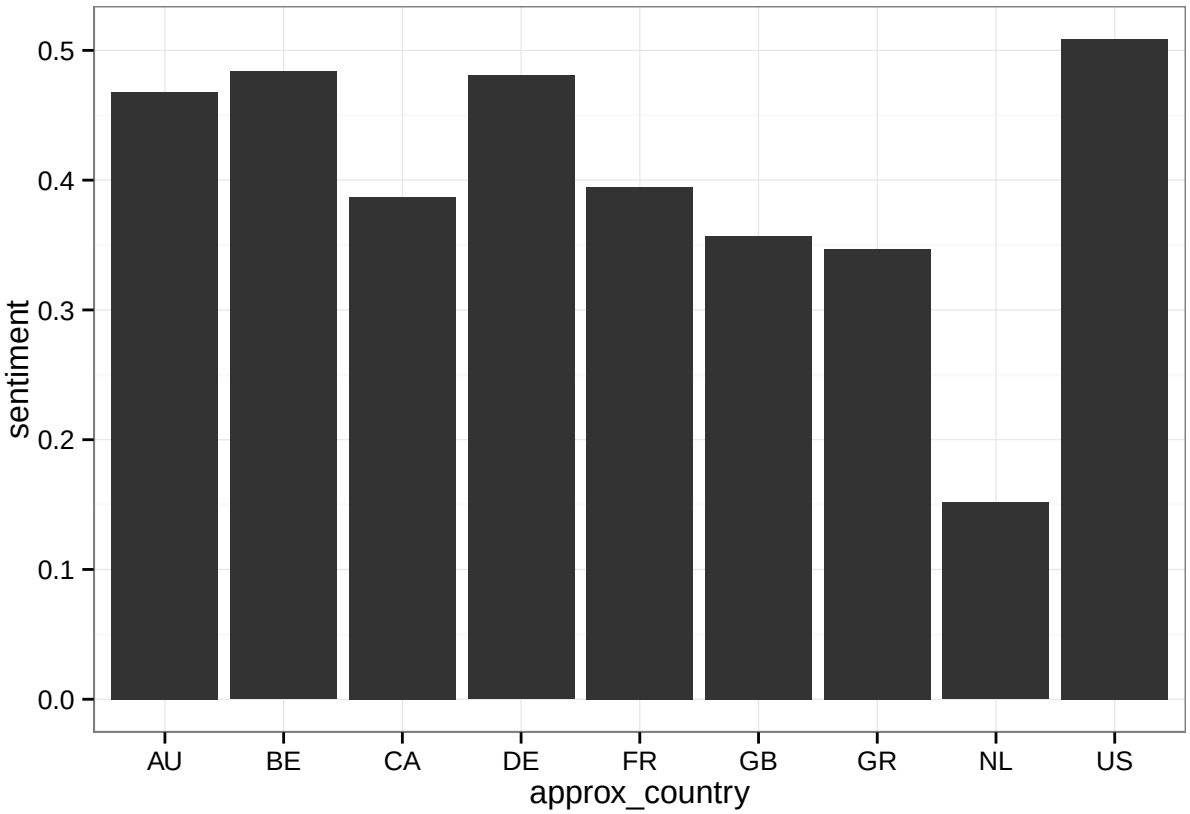
Inferential Statistics: Sentiment Analysis

For sentiment analysis, we used the R packages tm.lexicon.GeneralInquirer (Theussl, Hofmarcher, and Hornik 2015) and tm.plugin.sentiment (Annau 2014). For those tweets for which we have at the time of writing this document already had downlaoded location information we calculated the sentiment from the text corpora based on simple word counts. The package already includes a sentiment dictionary. The sentiment analysis returned a number in between -3 and 3 for all tweets. Low scores indicate negative sentiment, high scores indicate positive sentiment.

We looked at individual tweets and compared the results to our perception of sentiment of the tweet. While there seemed to be some problems, the sentiment analysis scores did seem reasonable.

However, when calculating the sentiments for queries, the returned results did not match our expectations. Below we report the resulted for the query "merkel+greece". We report the sum of the positive scores divided by the number of tweets. The figures indicate that the sentiments were very positive in most of the European countries and very similar.





We conclude from this that our current sentiment analysis does not return reasonable results. We will in the next stage of the project try to create a model that returns more accurate results.

References

Annau, Mario. 2014. *Tm.plugin.sentiment: Text Corpus Sentiment Analysis*. http://R-Forge.R-project.org/projects/sentiment/.

Biernacki, Patrick, and Dan Waldorf. 1981. "Snowball Sampling: Problems and Techniques of Chain Referral Sampling." *Sociological Methods & Research* 10 (2). SAGE Publications: 141–63.

Gentry, Jeff. 2015. *TwitteR: R Based Twitter Client*. http://CRAN.R-project.org/package=twitteR.

Henrique, Jefferson. 2015. "Get Old Tweets Programmatically." Accessed October 21. https://github.com/Jefferson-Henrique/GetOldTweets.

Theussl, Stefan, Paul Hofmarcher, and Kurt Hornik. 2015. *Tm.lexicon.GeneralInquirer: General Inquirer Categories*.