Twiqs.nl: Searching in billions of Dutch tweets

Erik Tjong Kim Sang
Meertens Institute, Amsterdam, The Netherlands

23 November 2015

Erik Tjong Kim Sang

- MSc from Delft, PhD from Groningen
- computer linguist, specialized in machine learning
- working with social media data since 2010
- 2011: prediction of Dutch elections with tweets
- 2012-2013: built website twiqs.nl for researching Dutch tweets

Meertens instituut

Twitter

Twitter is a microblog website started in 2006 by the American company Obvious (Jack Dorsey)

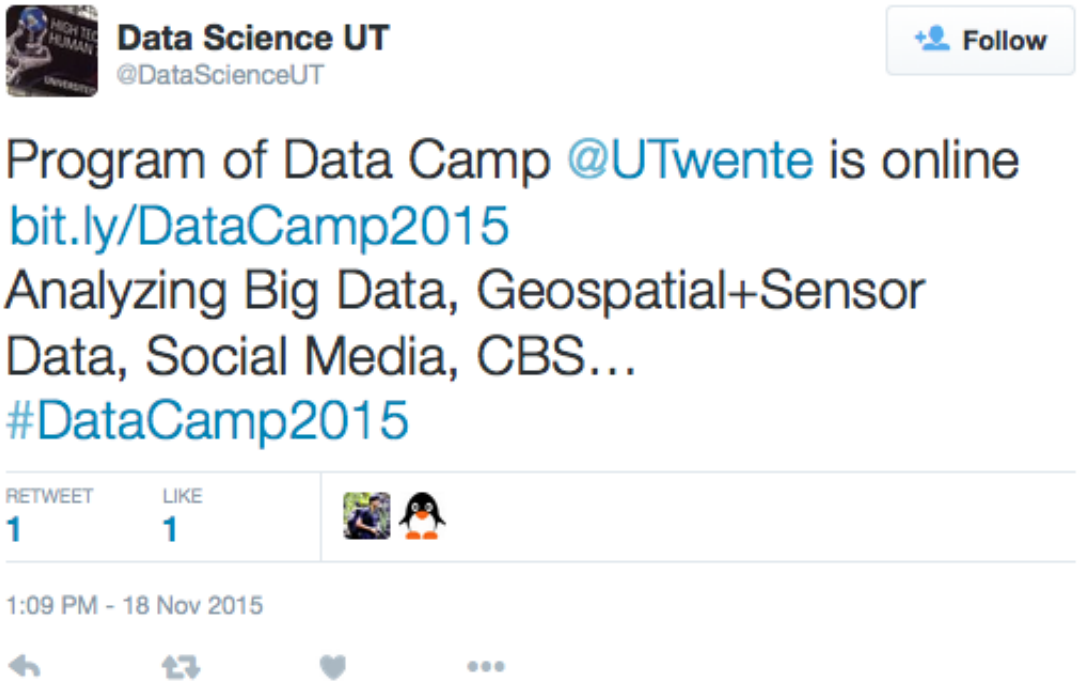
Users broadcast messages with a maximum length of 140 characters

Users can subscribe to messages of other users, in which case they *follow* the other users

Relations on Twitter are asymmetrical: you can follow somebody but that user does not have to follow you

The most followed Twitter users can be found on `twitaholic.com/`

What does a tweet look like?



Meertens instituut

Behind the screen (1)

{"filter_level":"medium","retweeted":false,"in_reply_to_screen_name":null,"possibly_
sensitive":false,"truncated":false,"lang":"nl","in_reply_to_status_id_str":null,"id":
405643870785904640,"in_reply_to_user_id_str":null,"in_reply_to_status_id":null,"created_at":
"Wed Nov 27 10:26:49 +0000 2013","favorite_count":0,"place":null,"coordinates":null,
"twinl_source":["track"],"text":"Wat is big data en hoe gebruiken bedrijven dit? Deze
video geeft antwoorden op deze vragen. Heel interessant! \nhttp:\/\/t.co\/UNl4CTqcLj",
"contributors":null,"geo":null,"entities":{"symbols":[],"urls":[{"expanded_url":
"http:\/\/bit.ly\/1jJhxnc","indices":[111,133],"display_url":"bit.ly\/1jJhxnc","url":
"http:\/\/t.co\/UNl4CTqcLj"}],"hashtags":[],"user_mentions":[]},"source":"web","favorited":
false,"in_reply_to_user_id":null,"retweet_count":0,"id_str":"405643870785904640","user":
{"location":"","default_profile":true,"statuses_count":68,"profile_background_tile":false,
"lang":"en","profile_link_color":"0084B4","id":183749925,"following":null,"favourites_
count":0,"protected":false,"profile_text_color":"333333","description":"Dit jaar wordt
er een symposium georganiseerd over BigData! Kijk op http:\/\/www.exa-it.nl voor meer
informatie.","verified":false,"contributors_enabled":false,"profile_sidebar_border_color":
"C0DEED","name":"SNiC Symposium","profile_background_color":"C0DEED","created_at":"Fri
Aug 27 20:15:17 +0000 2010","default_profile_image":false,"followers_count":49,"profile_
image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/378800000605291918\/
21cf5531df36b98396bac82d358dede7_normal.jpeg","geo_enabled":false,"profile_background_
image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_

Meertens instituut

Behind the screen (2)


```
{   "lang":"en",
    "created_at":"Wed Nov 18 13:09:21 +0000 2015",
    "text":"Program of Data Camp @UTwente is online
            bit.ly/DataCamp2015 Analyzing Big Data,
            Geospatial+Sensor Data, Social Media,
            CBS #DataCamp2015",
    "hashtags":[..."DataCamp2015"...],
    id_str":"667087250103123968",
    "user":{"name":"Data Science UT",
             followers_count":21, ...}
}
```

Research based on tweets

An increasing number of students and researchers wants to study tweets

Tweets cover a wide variety of topics and are written in many different languages and dialects

Researchers and students often lack the technical knowledge to collect, store and analyze tweets

For this reason the Netherlands eScience Center has developed the website `twiqs.nl` for supporting research involving tweets written in Dutch

twiqs.nl

Twiqs.nl can be used for searching in Dutch tweets written in the time frame December 2010 until today

Tweets can be selected based on keywords or based on metadata, for example the age of the person that wrote the tweet

Search results do not include tweets but only summaries of search results, for example word frequencies, maps, related words and aggregated information about users

The tweet collection used for the searches is incomplete: it contains about 40% of the tweets written in Dutch since December 2010

How do we collect tweets?

We use the Twitter streaming API for continuously searching for new tweets which contain at least one of a list of 229 frequent words

We collect about one million tweets per day of which about 800,000 are identified by our software as written in Dutch

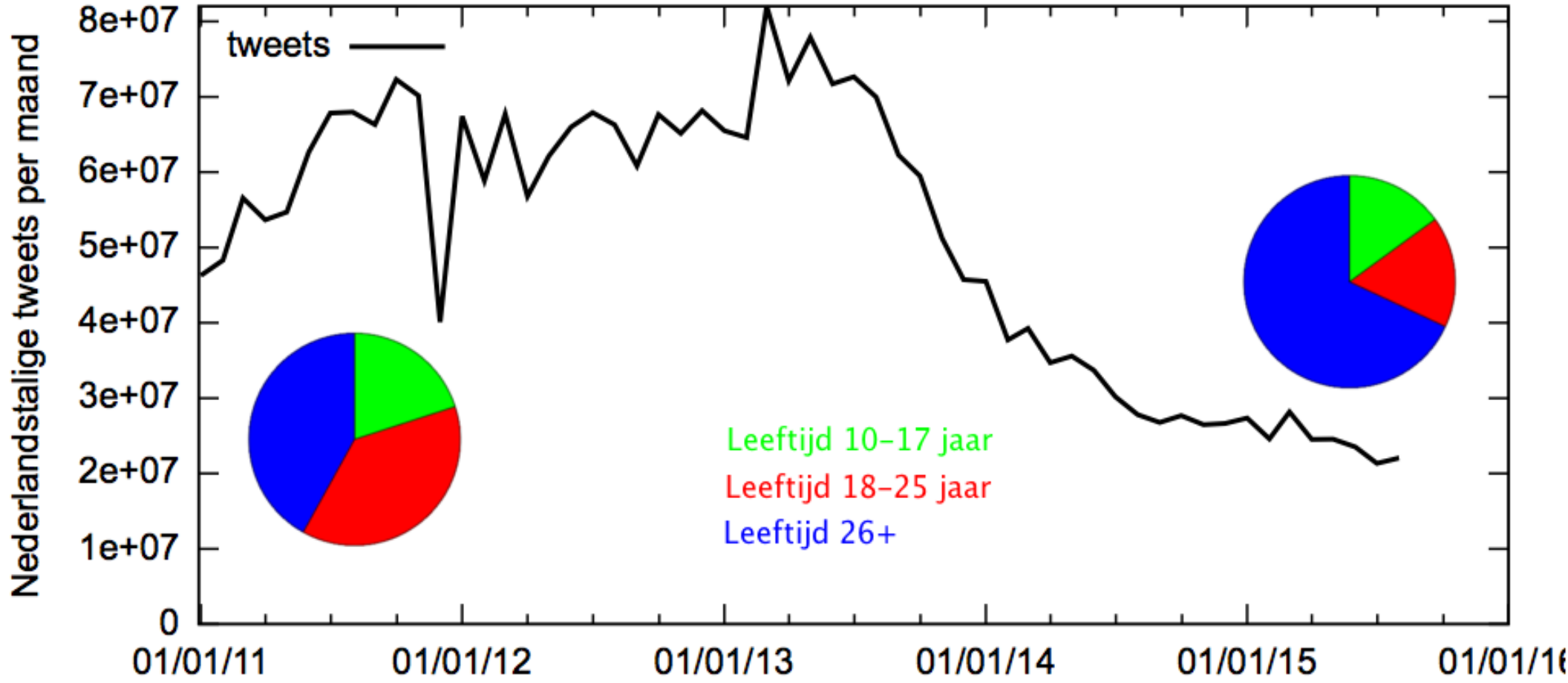
The tweets are stored in compressed files, each of which contains one hour of tweets and their metadata encoded in the data format JSON

The tweet collection contains 4.3 billion tweets of which 2.9 billion are written in Dutch

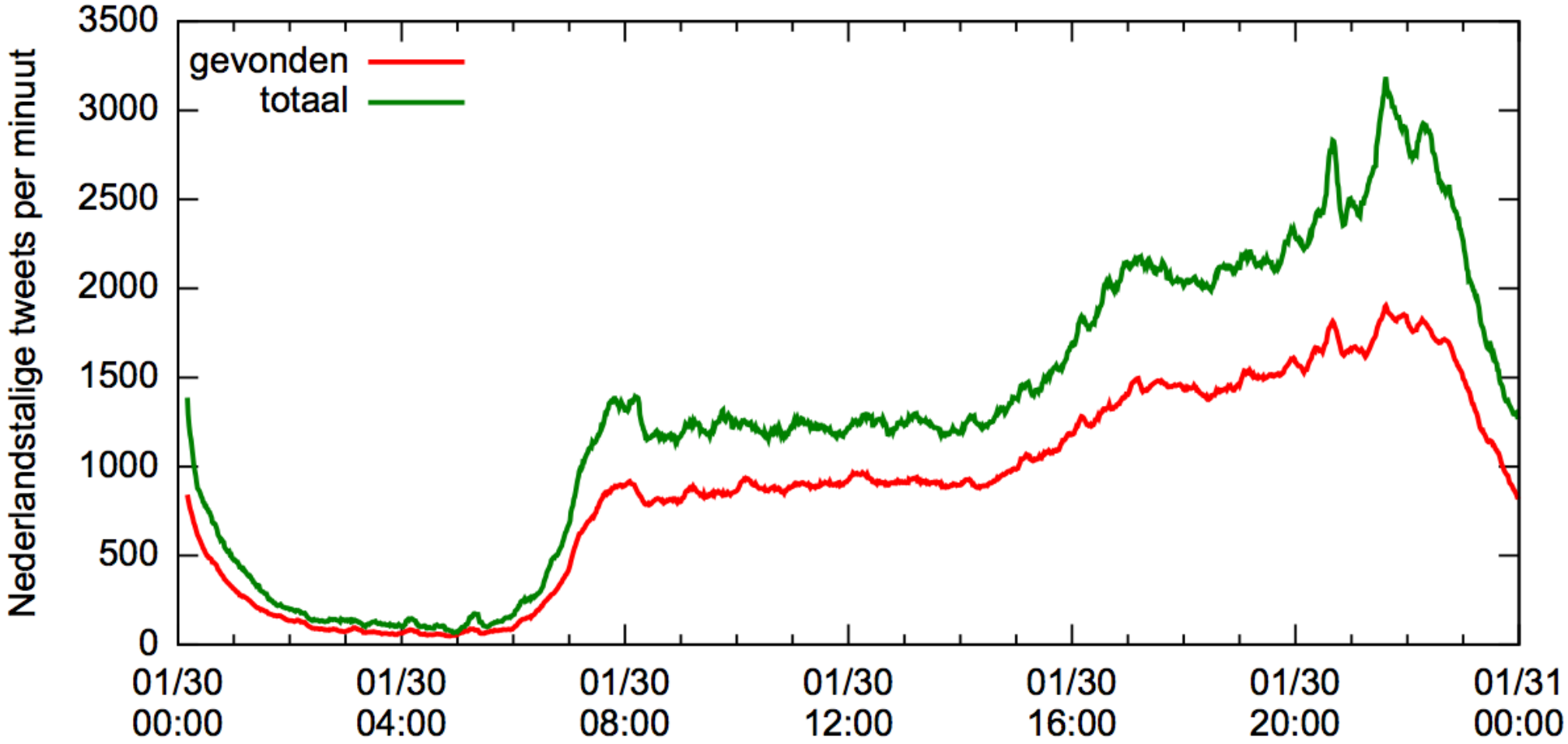
Twiqs.nl: Searching in billions on Dutch tweets 23 November 2015

Twiqs.nl tweet counts per year

Year	Dutch	Other	Remarks
2010	17,204,599	3,785,448	starts at 16-12-2010
2011	706,908,122	248,700,184	
2012	774,926,259	510,549,954	
2013	795,202,317	401,535,733	
2014	392,039,290	147,422,993	
2015	258,541,864	20,926,460	until 20-11-2015
Total	2,944,822,451	1,332,920,772	


Nederlandstalige tweets per maand
8e+07
7e+07
6e+07
5e+07
4e+07
3e+07
2e+07
1e+07
0
tweets
Leeftijd 10-17 jaar
Leeftijd 18-25 jaar
Leeftijd 26+
01/01/11
01/01/12
01/01/13
01/01/14
01/01/15
01/01/16

Gevonden 1.279.263 van 1.869.116 tweets (68%).



The challenge: searching in the collection of tweets

Question: how can we find tweets in this large collection?

Answer: use an information retrieval system

But...

- we have a really large collection of data
- each second new tweets are added to the collection
- we want to search in both tweets and in metadata

Solution: use a computer cluster (cloud)

We use a computer cluster to process search queries

Every search query is split in several smaller queries, each of which processes one file with tweets from the same one hour

These smaller queries can be processed in parallel

The search software of twiqs.nl runs on a Hadoop system, which is based on Google's MapReduce algorithm

The search software runs on the experimental Hadoop cluster of SURFsara (170 nodes each with 8 cpus; 2.3 petabyte storage). This cluster is available for all researchers in The Netherlands

Searching with MapReduce: Map task

The map task selects the relevant tweets per hour:

```
for (each tweet in data file) {
   if (query matches tweet) {
      process tweet
      put results on output with data key (minute)
   }
}
```

By developing the (Java) search software ourselves, we have full control over the search process

Example Reduce task

The Reduce task combines the results of the various Map tasks:

```
combine Map results of the current data key (minute)
put results per data key on the output
```

The Reduce task sorts the output of the Map tasks based by the data key values

Advantages of MapReduce

- Full control over search and data processing
- Searching in several hours of data is almost as fast as searching in one hour of data

Disadvantages of MapReduce

- It is not as fast as when using information retrieval
- Processing might have to wait on jobs of other users

Zoek in tweets:

Zoeken

6 juni 2013 -

Op twiqs.nl kan je in Nederlandstalige tweets vanaf december 2010 zoeken. De gevonden tweets kan je dan op Twitter bekijken. De zoekdatabase is niet compleet. Contactpersoon voor deze website is Erik Tjong Kim Sang <erikt(at)xs4all.nl>

Voorbeelden:

- "griep" in januari/maart 2013 (grafiek)
- "carnaval" in februari 2013 (kaart)
- "ajax,feyenoord" op zondag 28 oktober 2012 (grafiek)
- "vaatstra" op maandag 19 november 2012 (gebruikersinformatie)
- "nait" in november 2012 (kaart voor dialectwoord)
- "fail" op maandag 3 december 2012, 08:00-08:59 (woordenwolk)

Created at: 6 June 2013 15:10:49. erikt(at)xs4all.nl (clusterstatus)

TOPICS ON TWITTER

When are certain topics discussed on Twitter?

With twiqs.nl we can determine how popular a certain topic is on Dutch Twitter in a certain time period

We count how often tweets mention a certain topic in a time frame and put the counts in a graph

But during the night fewer tweets are written and we try to keep that effect out of the graphs

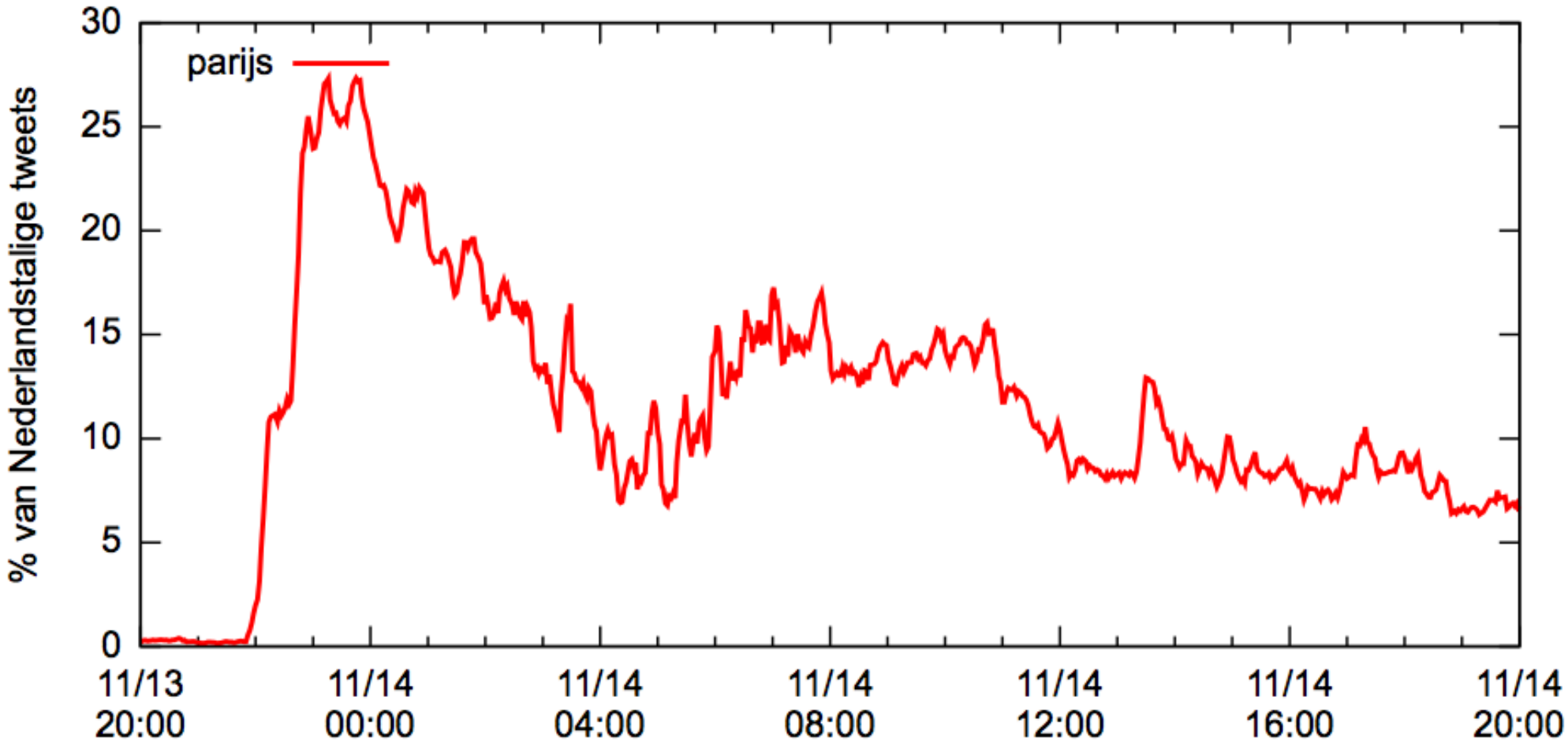
Therefore we do not make graphs with tweet totals but graphs with percentages, for example the percentage of Dutch tweets that contain the word *Parijs*

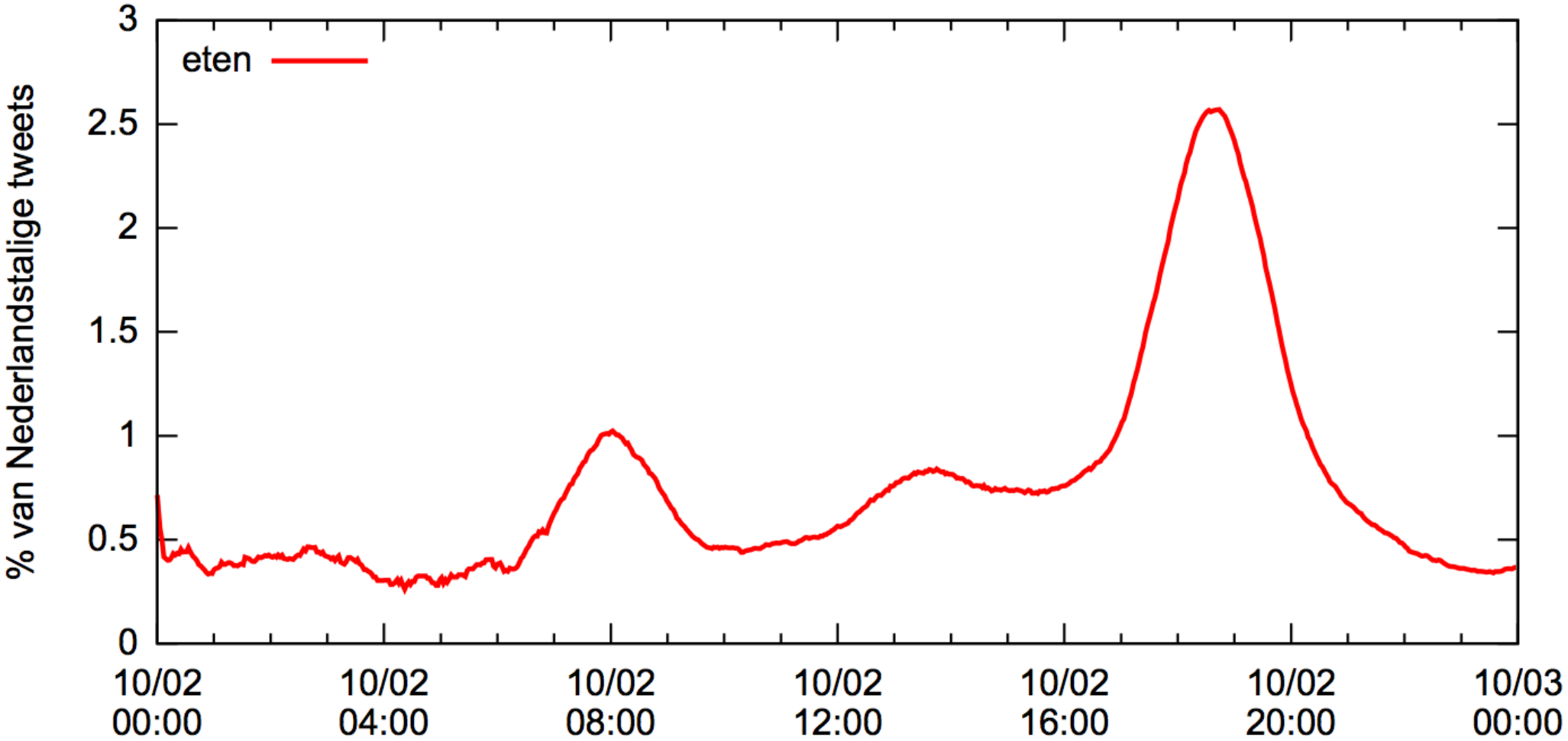
Graphs of daily word usage

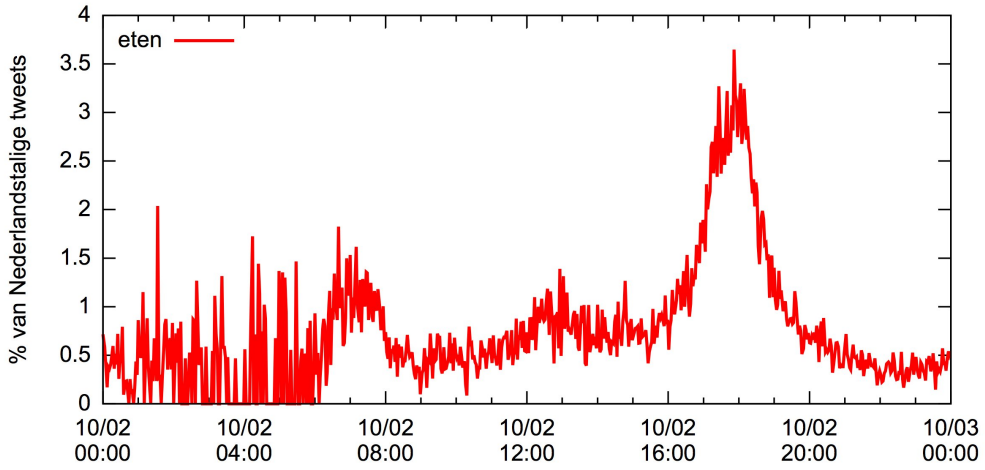
With `twiqs.nl` you can make graphs of the frequencies of words used on Dutch Twitter

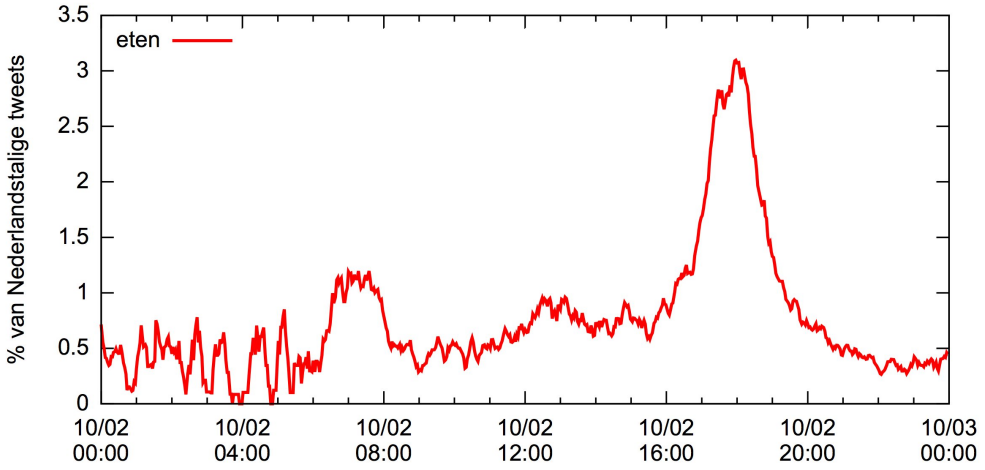
The graphs will present you the frequencies of words on Dutch Twitter almost live

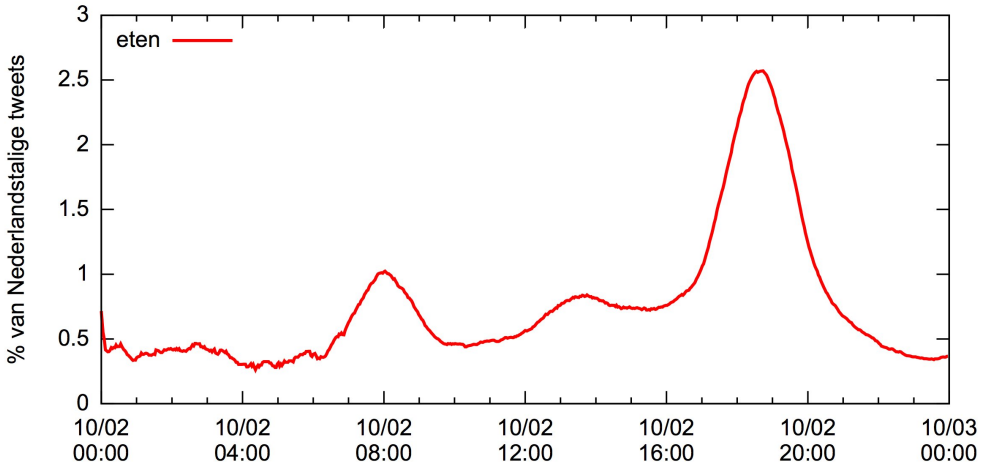
You can also request graphs for other days than today, for example for Friday 13 November 2013 (Paris attacks)

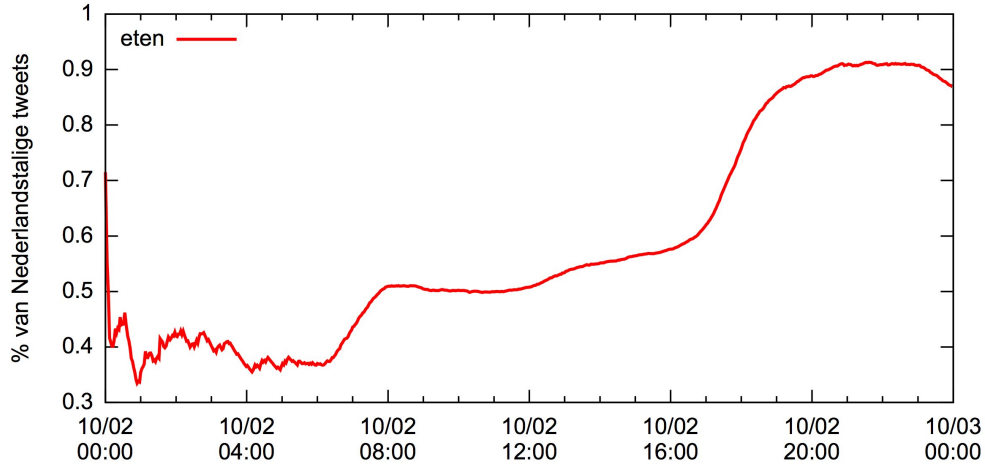

parijs
% van Nederlandstalige tweets
30
25
20
15
10
5
0
11/13
20:00
11/14
00:00
11/14
04:00
11/14
08:00
11/14
12:00
11/14
16:00
11/14
20:00

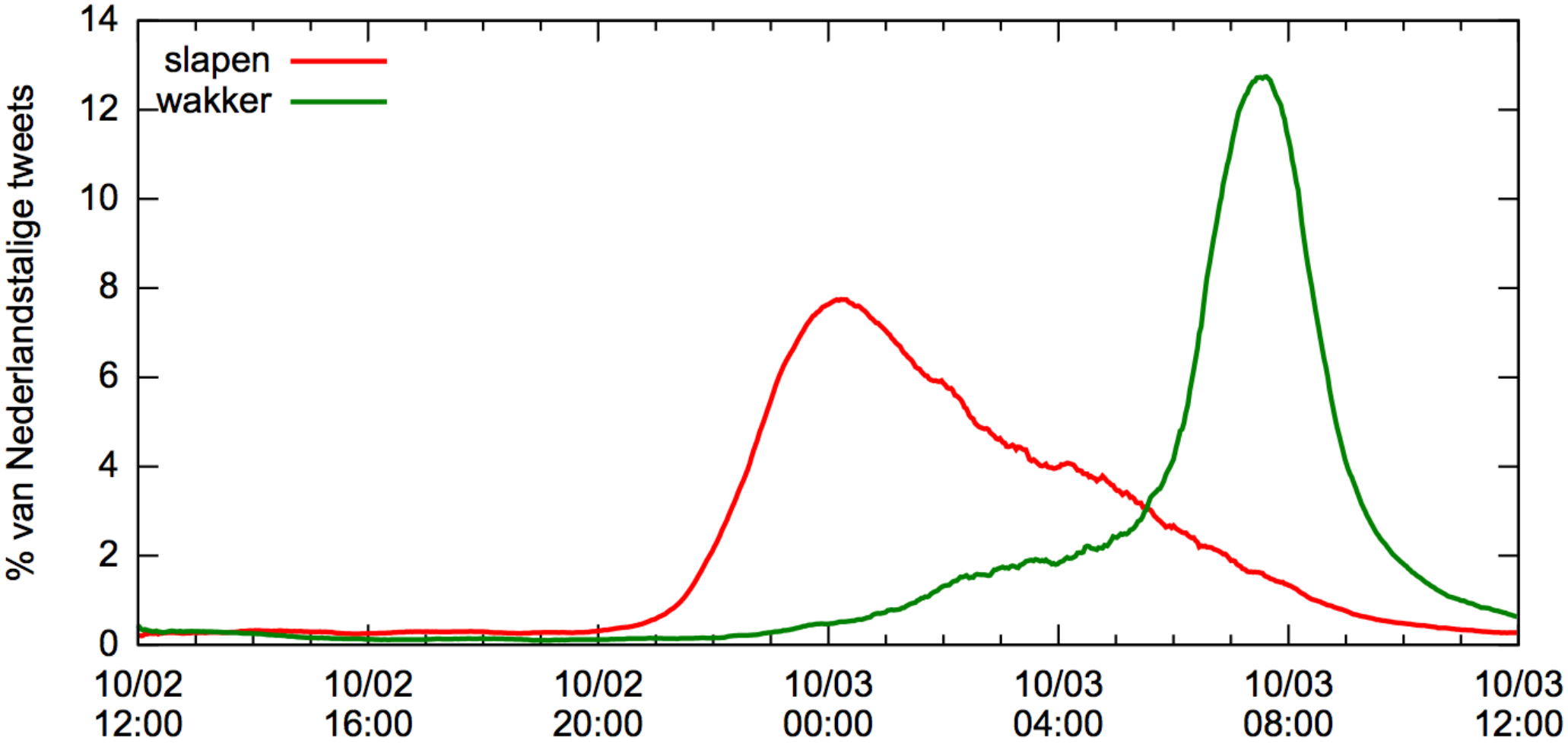

eten
% van Nederlandstalige tweets
3
2.5
2
1.5
1
0.5
0
10/02 00:00
10/02 04:00
10/02 08:00
10/02 12:00
10/02 16:00
10/02 20:00
10/03 00:00

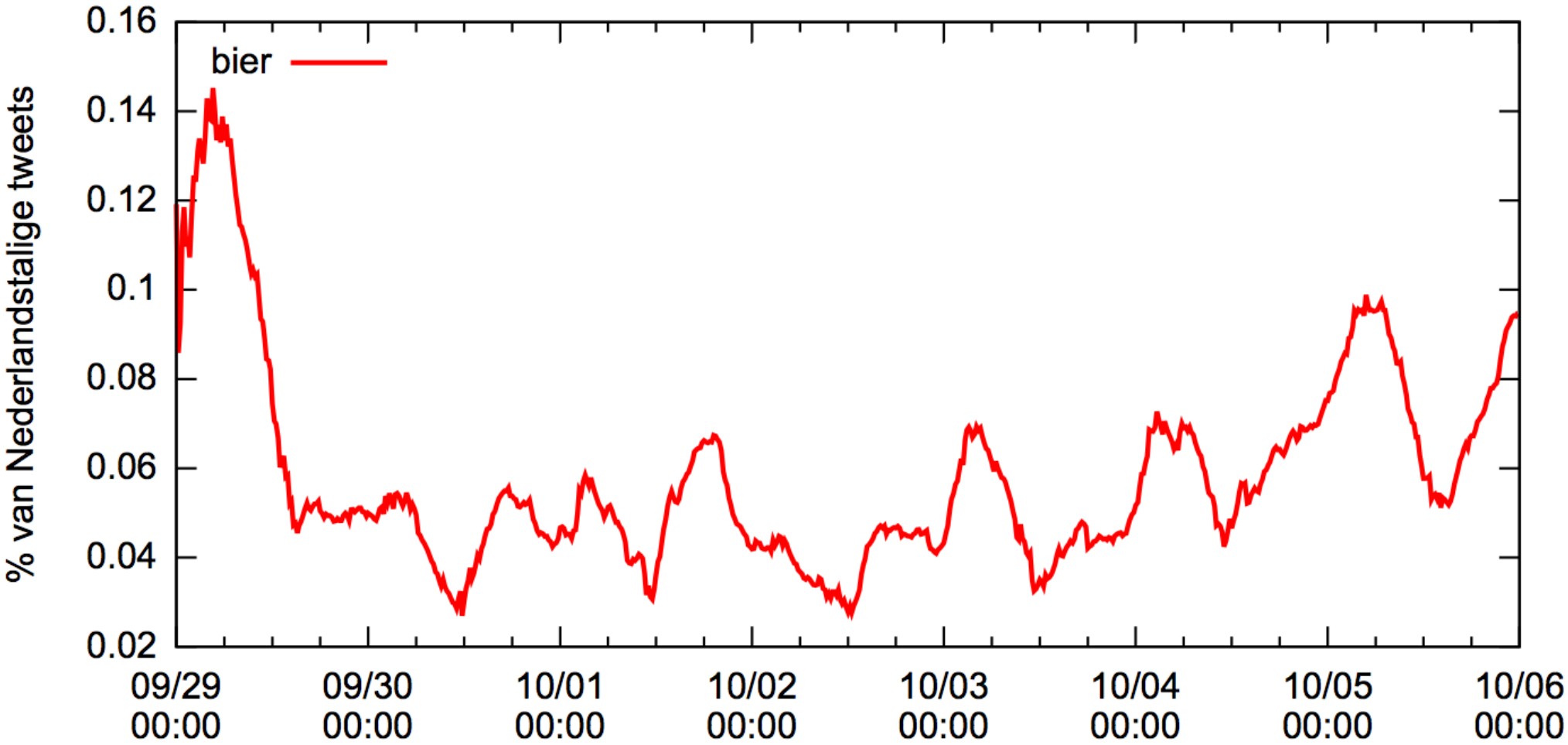

eten
% van Nederlandstalige tweets
4
3.5
3
2.5
2
1.5
1
0.5
0
10/02 00:00
10/02 04:00
10/02 08:00
10/02 12:00
10/02 16:00
10/02 20:00
10/03 00:00

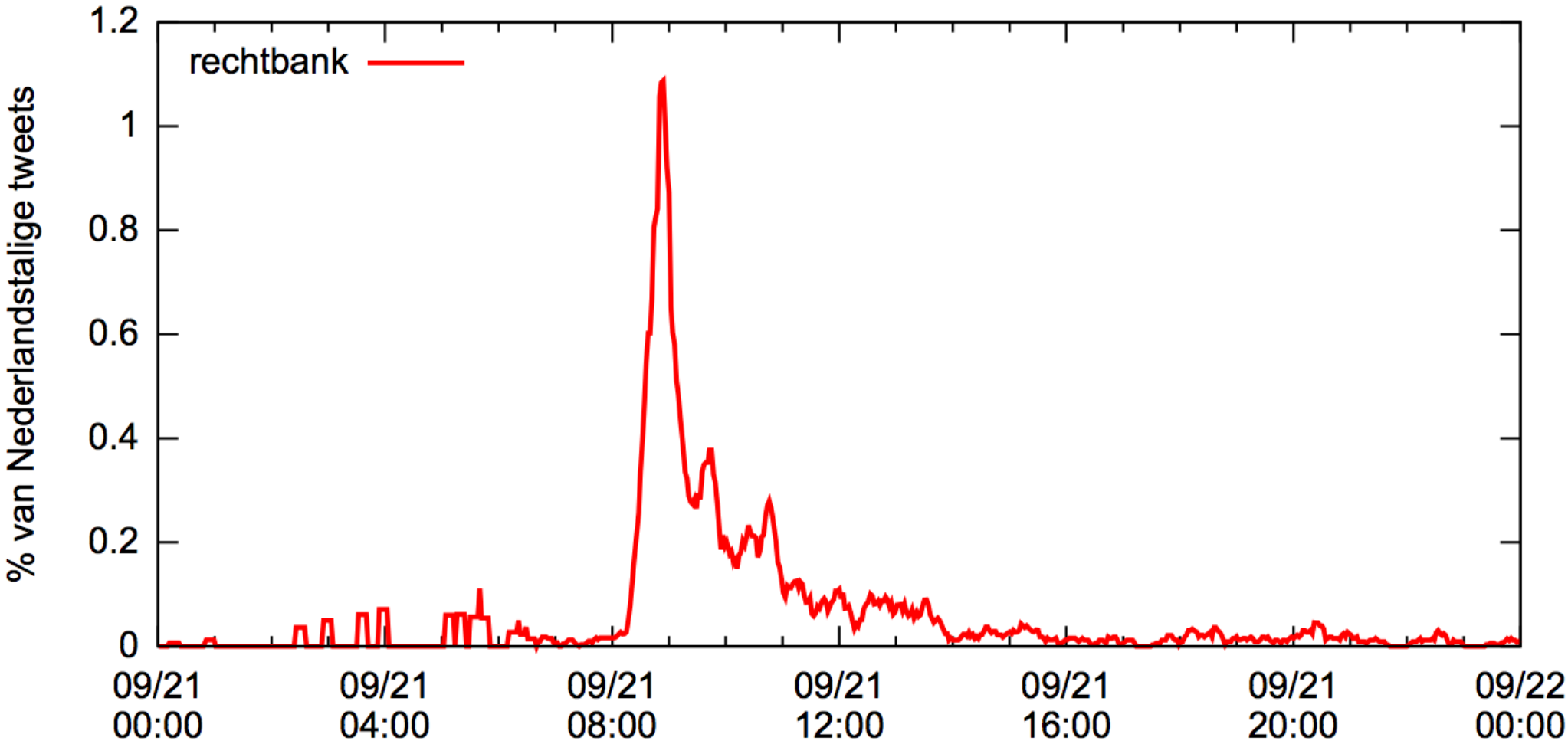

eten
% van Nederlandstalige tweets
3.5
3
2.5
2
1.5
1
0.5
0
10/02 00:00
10/02 04:00
10/02 08:00
10/02 12:00
10/02 16:00
10/02 20:00
10/03 00:00

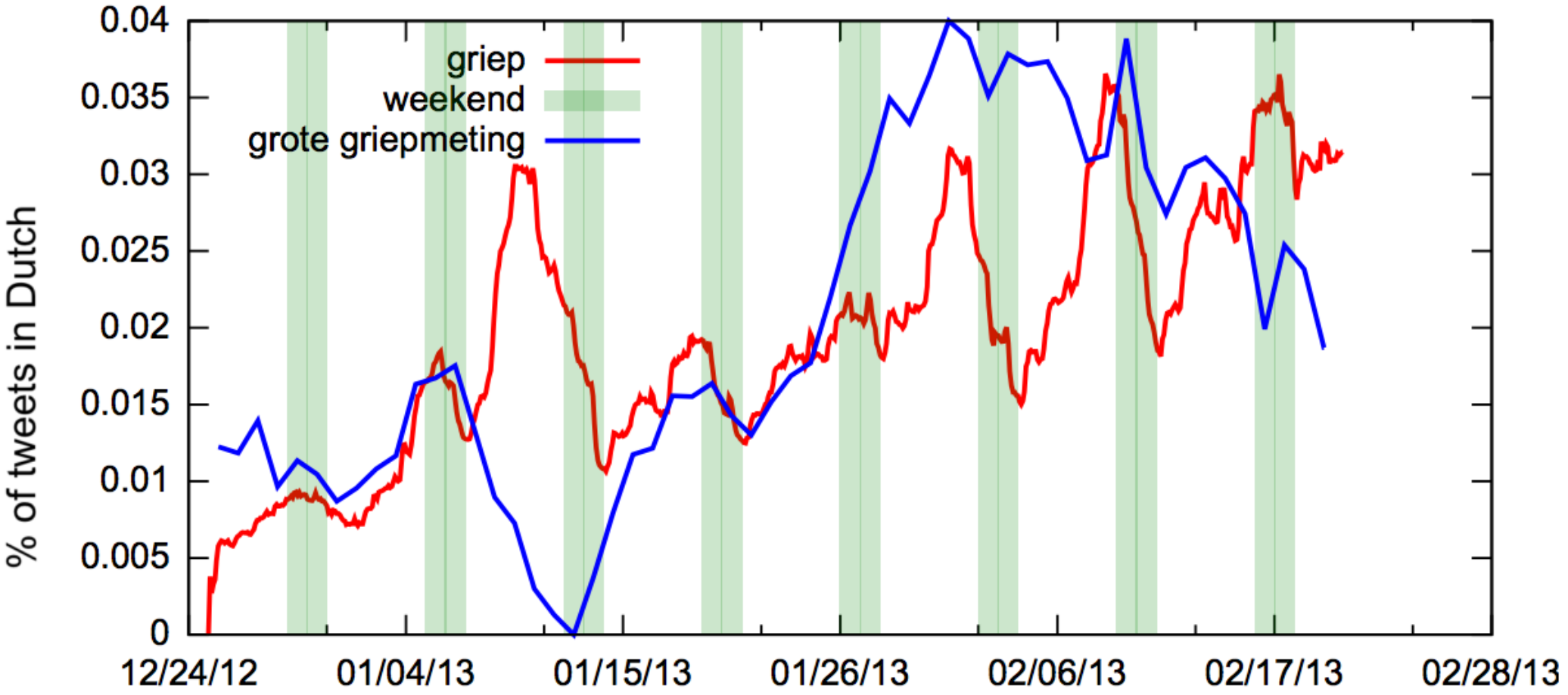

eten
% van Nederlandstalige tweets
3
2.5
2
1.5
1
0.5
0
10/02 00:00
10/02 04:00
10/02 08:00
10/02 12:00
10/02 16:00
10/02 20:00
10/03 00:00


eten
% van Nederlandstalige tweets
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
10/02 00:00
10/02 04:00
10/02 08:00
10/02 12:00
10/02 16:00
10/02 20:00
10/03 00:00


slapen
wakker
% van Nederlandstalige tweets
0
2
4
6
8
10
12
14
10/02 12:00
10/02 16:00
10/02 20:00
10/03 00:00
10/03 04:00
10/03 08:00
10/03 12:00


bier
% van Nederlandstalige tweets
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
09/29 00:00
09/30 00:00
10/01 00:00
10/02 00:00
10/03 00:00
10/04 00:00
10/05 00:00
10/06 00:00


rechtbank
% van Nederlandstalige tweets
1.2
1
0.8
0.6
0.4
0.2
0
09/21 00:00
09/21 04:00
09/21 08:00
09/21 12:00
09/21 16:00
09/21 20:00
09/22 00:00


griep
weekend
grote griepmeting
% of tweets in Dutch
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
12/24/12
01/04/13
01/15/13
01/26/13
02/06/13
02/17/13
02/28/13

Summary

We can draw graphs of word frequencies on Dutch Twitter

The graphs show which topics are popular on Twitter at a certain time

This provides insights in what topics an average user of Dutch Twitter is interested in

Noise can be removed from the graphs by increasing the smooth factor

PREDICTING ELECTION RESULTS

Predicting election results

Can we use Twitter for predicting the results of an election?

Tumasjan et al (2010): yes, this is possible:

1. count the number of times that a tweet mentions a political party
2. convert the tweet counts to seat counts
3. done!

Twiqs.nl: Searching in billions on Dutch tweets 23 November 2015

Prediction an election by counting tweets

Party	Short name	Long name	Total	Seats Twitter	Seats PB	Seats MdH	Average polls
PVV	2226	1	2227	18	12	12	12
VVD	1562	0	1562	13	14	16	15
CDA	1504	0	1504	12	9	10	9.5
PvdA	1056	1	1057	9	13	13	13
SP	839	0	839	7	8	7	7.5
GL	243	505	748	6	5	3	4
D66	610	0	610	5	6	5	5.5
CU	159	79	238	2	3	3	3
PvdD	103	51	154	1	1	1	1
SGP	139	0	139	1	2	2	2
50+	6	43	49	0	1	2	1.5
OSF	-	-	-	1	1	1	1
			offset	21	4	4	-

The bad news

Nobody has been able to reproduce the excellent prediction results of Tumasjan (2010)

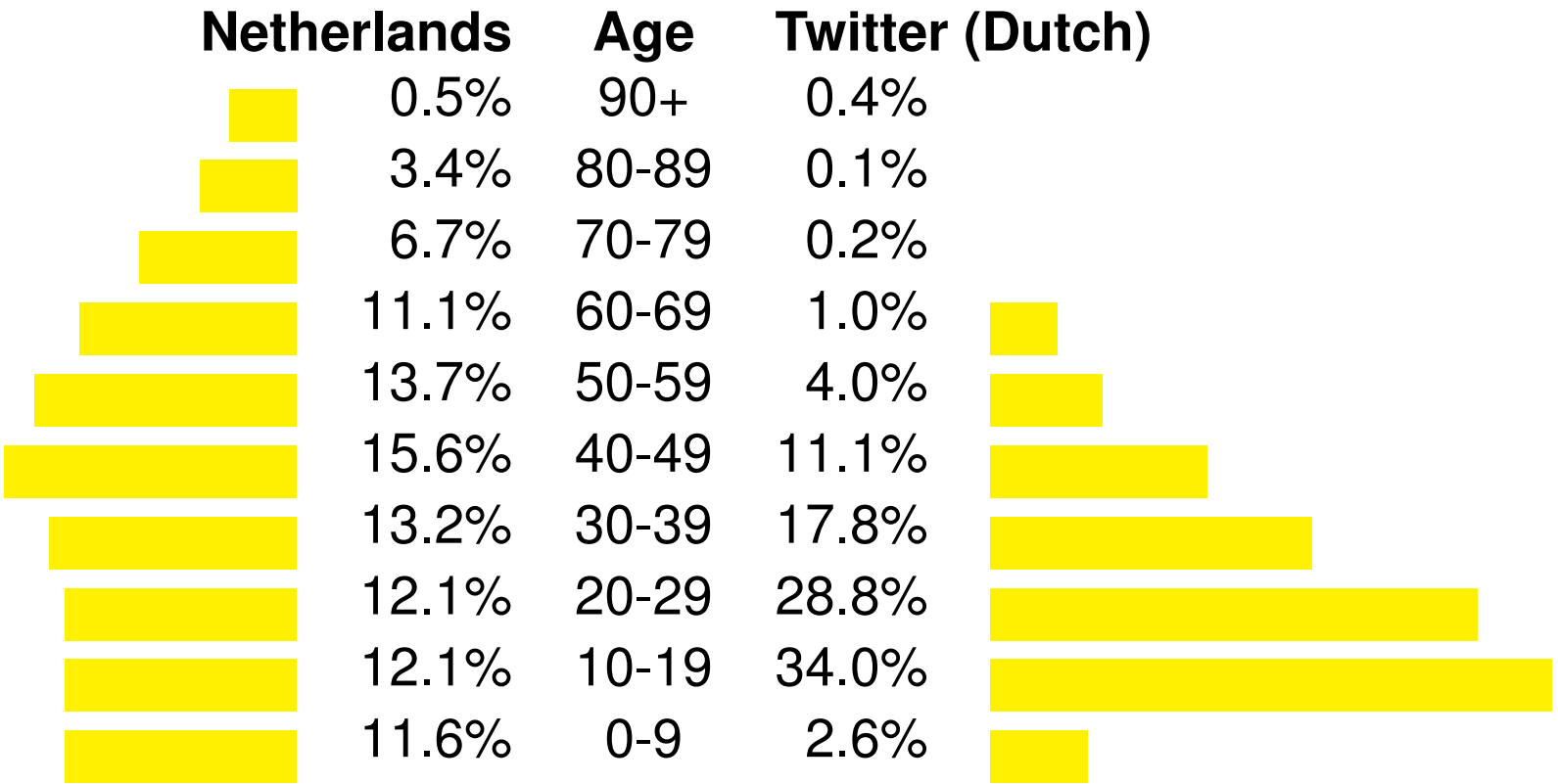
Problems of this approach:

- not all tweets mentioning a party are positive
- sometimes one person will mention a party many times on Twitter
- Twitter users are not a representative sample of the Dutch population

Examples of negative tweets

- De **VVD** en de **PvdA** hoeven slechts Uw centjes naar Brussel over te maken en voor de rest kunnen ze (met U) doen en laten wat ze willen.
  VVD *and* **PvdA** *only need to send your money to Brussels and otherwise they can do as they please*
- Mijn pappie wordt nog steeds gepest door dat **#pvv** schorem op Twitter! Morgen ga ik aangifte doen, moet toch pas 's middags werken.
  My dad is still being harassed by those **PVV** *bastards on Twitter. I'm going to press charges tomorrow, I only working in the afternoon.*
- **SP** heeft nog steeds niet door dat alleen dure woningen worden gebouwd, omdat dit meer geld in de Gemeentekas brengt http://t.co/0tvpeX8a
  SP *still does not understand that only expensive houses are built because this earns the municipality more money*

Ages of Twitter users



Meertens instituut

Method for predicting elections

1. count the number of tweets that mention a political party
2. subtract the negative tweets
3. link the counts to a reliable political poll of the same time
4. repeat steps 1 and 2 right before the elections
5. determine an election prediction based on steps 1–4

Example: PvdA was polled at 12 seats and at the same time was mentioned in 20% of the tweets. If the party is mentioned in 25% of the tweets right before the election, we estimate it to win 25/20 * 12 = 15 seats.

Twiqs.nl: Searching in billions on Dutch tweets 23 November 2015

Predicting elections (Dutch Senate 2011)

Party	Result	Seats PB	Seats MdH	Seats Twitter
VVD	16	14	16	14
PvdA	14	12	11	16
CDA	11	9	9	8
PVV	10	11	12	10
SP	8	9	9	6
D66	5	7	5	8
GL	5	4	4	3
CU	2	3	3	3
50+	1	2	2	2
SGP	1	2	2	2
PvdD	1	1	2	2
OSF	1	1	0	1
offset	-	14	14	18

Twiqs.nl: Searching in billions on Dutch tweets 23 November 2015

Predicting elections without using tweets

Party	Result	Seats PB	Seats MdH	Seats Twitter
VVD	16	14	16	16
PvdA	14	12	11	13
CDA	11	9	9	10
PVV	10	11	12	12
SP	8	9	9	7
D66	5	7	5	5
GL	5	4	4	4
CU	2	3	3	3
50+	1	2	2	1
SGP	1	2	2	2
PvdD	1	1	2	1
OSF	1	1	0	1
offset	-	14	14	8

Meertens instituut

Summary

Tweets are not useful for predicting elections, at least not this time

The demographic features of Twitter users are different than those of the population of The Netherlands

It is important to compare the performances of complex computer models with simple models, the so-called baselines (here: the prediction that did not use tweets)

WHAT IS THE HAPPIEST LOCATION IN THE NETHERLANDS?

What is the happiest location in The Netherlands?

We start with making this research question more specific:

We want to use tweets for answering the question so a more concrete version is: which location in The Netherlands broadcasts the highest percentage of positive tweets?

location is a very general concept

We need areas and therefore we examine provinces and municipalities

Sentiment analysis

Determining if a text is positive, negative or neutral is called sentiment analysis

The best way to perform this task is to have people read the texts

They can assign a human emotion to the text: happy, unhappy, confused, angry, aggressive, ...

We want to process millions of texts (tweets) and therefore use an automatic solution, without determining topics

Automatic sentiment analysis

We use two word lists: one with positive and one with negative words

When a tweet contains one positive words, we classify it as positive, and when it contains one negative word, we classify it as negative

Whenever a tweet contains more than one word of the two word lists, we only consider the final relevant word

The words *geen* (*no*) and *niet* (*not*) reverse the sentiment value of the following word

Tweets without words from the sentiment word lists will be classified as neutral

Determining the location of a tweet

A small percentage of tweets contains position information in the shape of a longitude and a latitude coordinate, for example 6.568027,53.219353

Borders of regions are defined as sequences of such coordinate pairs (lines)

But how do we determine which region contains a coordinate point?

Solution: the algorithm point in polygon

Computing a sentiment score of a region

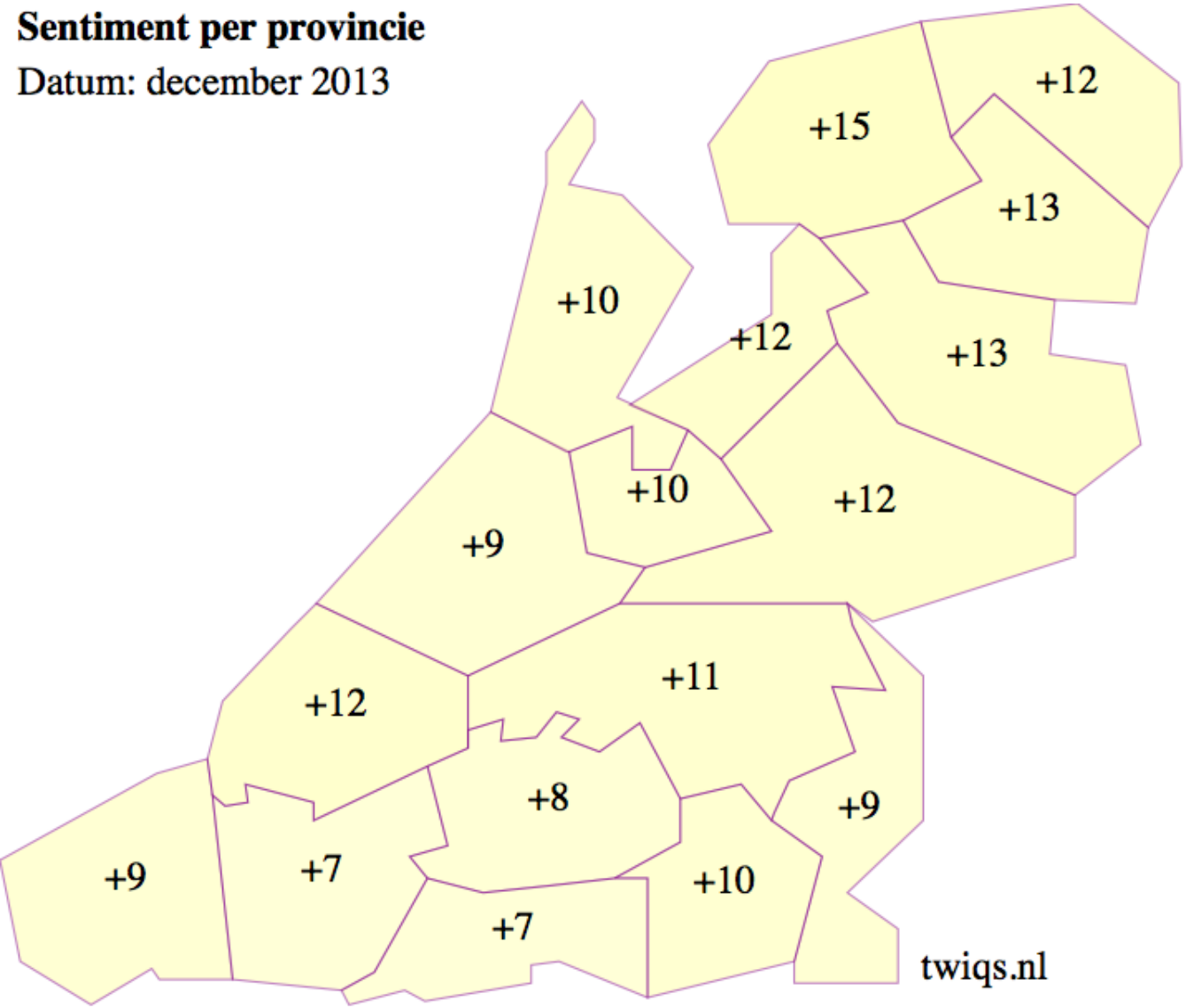
1. find tweets that have been sent from the region
2. determine the sentiment score of each of the relevant tweet
3. subtract the percentage of negative tweets from the percentage of positive tweets

The result is a score between -100% (only negative tweets) and +100% (only positive tweets)

Next: display the sentiment scores of all regions on a map

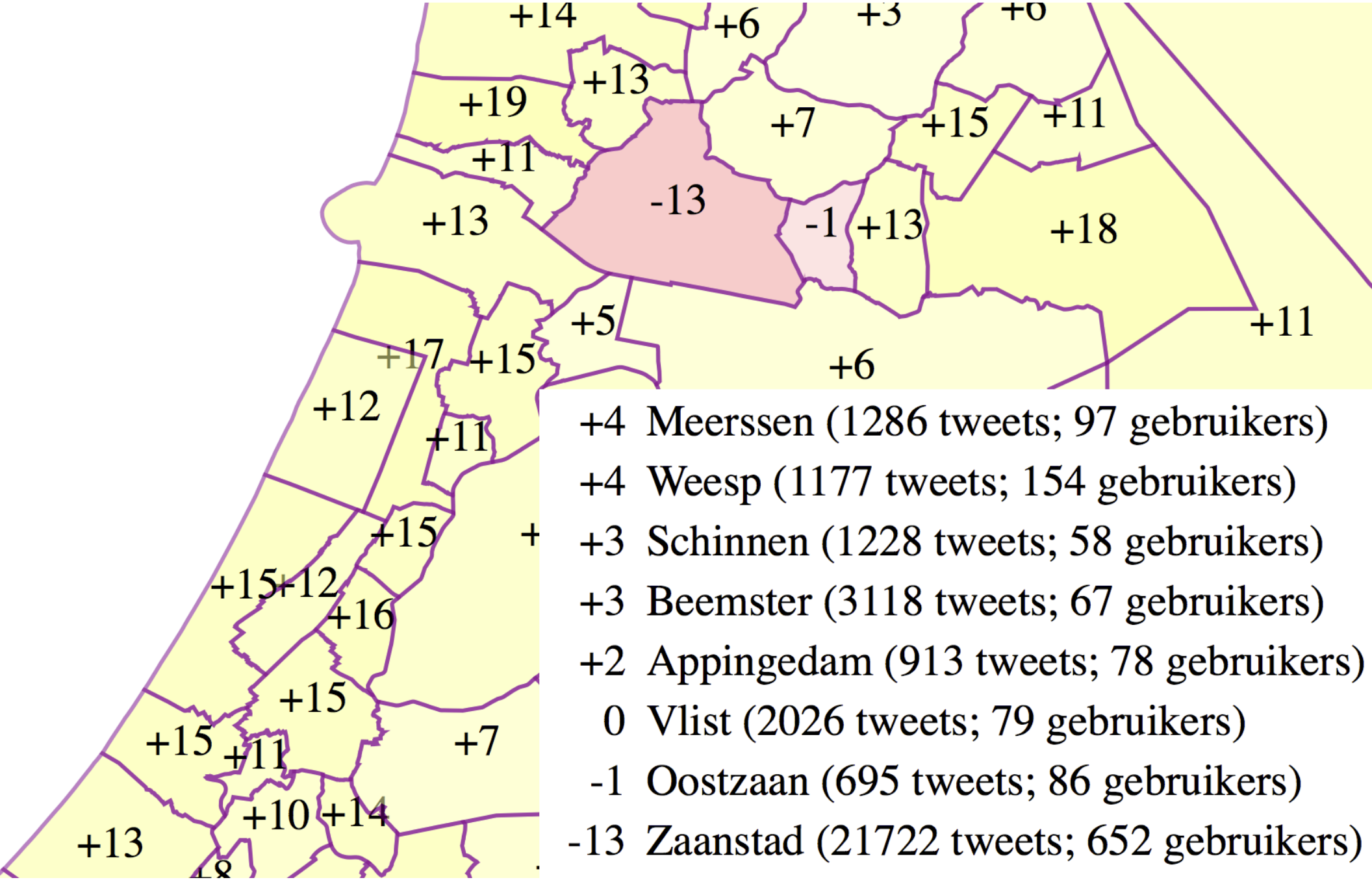
Sentiment per provincie

Datum: december 2013



twiqs.nl

+15 Friesland (79122 tweets)
+13 Drenthe (50702 tweets)
+13 Overijssel (148889 tweets)
+12 Gelderland (231153 tweets)
+12 Groningen (64111 tweets)
+12 Zeeland (42544 tweets)
+12 Flevoland (42641 tweets)
+11 Noord-Brabant (251010 tweets)
+10 Rest van de wereld (96067 tweets)
+10 Utrecht (160789 tweets)
+10 Noord-Holland (261698 tweets)
+10 Limburg BE (28379 tweets)
+9 West-Vlaanderen (66458 tweets)
+9 Zuid-Holland (335579 tweets)
+9 Limburg NL (85853 tweets)
+8 Antwerpen (113438 tweets)
+7 Oost-Vlaanderen (62003 tweets)
+7 Vlaams-Brabant (45096 tweets)



+4 Meerssen (1286 tweets; 97 gebruikers)
+4 Weesp (1177 tweets; 154 gebruikers)
+3 Schinnen (1228 tweets; 58 gebruikers)
+3 Beemster (3118 tweets; 67 gebruikers)
+2 Appingedam (913 tweets; 78 gebruikers)
0 Vlist (2026 tweets; 79 gebruikers)
-1 Oostzaan (695 tweets; 86 gebruikers)
-13 Zaanstad (21722 tweets; 652 gebruikers)

The most negative user on Dutch Twitter...

Meertens instituut

Solution

Now every user can only contribute one score to a region: his or her average score

This makes the region scores more fair: one person - one vote

Results of December 2013:

1.	+30	Ameland (FR)
2.	+28	De Marne (GR)
3.	+27	Ubbergen (GE)
	+27	Rucphen (NB)
5.	+26	Reusel-De Mierden (NB)

	+9	Uithoorn (NH)
413.	+8	Albrandswaard (ZH)
414.	+7	Renswoude (UT)
	+7	Hillegom (ZH)
	+7	Beek (LI)

Meertens instituut

Summary

A small percentage of tweets contains location information which can be used for mapping tweet data

While determining the sentiment of a region it is important to maintain the *one person - one vote* principle

On the average tweets written in Dutch are positive!

THE END

Literature

- Erik Tjong Kim Sang and Antal van den Bosch, Dealing with Big Data - The Case of Twitter. In: *Computational Linguistics in The Netherlands Journal*, volume 3, 2013. http://ifarm.nl/erikt/papers/clin2013.pdf
- Erik Tjong Kim Sang and Johan Bos, Predicting the 2011 Dutch Senate Election Results with Twitter. In: *Proceedings of the SASN Workshop*, EACL2012, Avignon, France, 2012 http://ifarm.nl/erikt/papers/sasn2012.pdf
- Andranik Tumasjan, Timm Sprenger, Philipp Sandner and Isabell Welpe, Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In: *Proceedings of the Fourth AAAI conference on Weblogs and Social Media*, pages 178-185, 2010 http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1833192
- Erik Tjong Kim Sang, *Het gebruik van Twitter voor Taalkundig Onderzoek*, In: TABU magazine, 2011. http://ifarm.nl/erikt/papers/tabu2011.pdf

Bonus: Animation of Twitter data

Pulse of the Nation is an animation which combines tweets with time, tweet volume and sentiment

You will see a map of the United States where:

- states become larger as the tweet volume increases
- states change color as the sentiment changes (green vs red)

Address: `http://www.youtube.com/watch?v=ujcrJZRSGkg`