Dealing with Big Data: the Case of Twitter

Erik Tjong Kim Sang (Netherlands eScience Center, Amsterdam)
Antal van den Bosch (Radboud University, Nijmegen)

netherlands eScience center

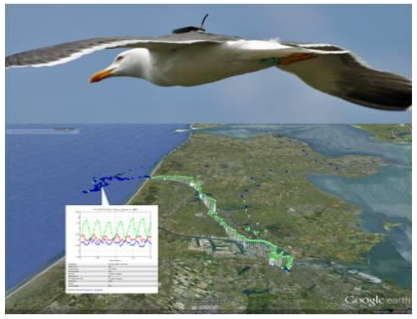
by SURF & NWO

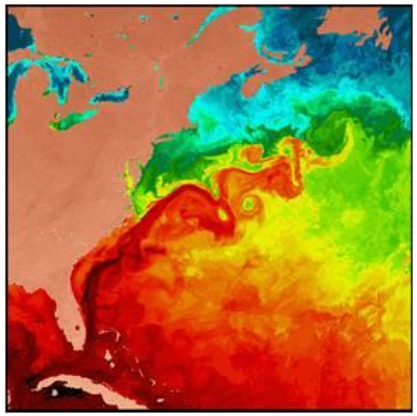
Netherlands eScience Center

We support and reinforce multidisciplinary and data-intensive research through creative and innovative use of information and communication technology

We distribute project calls which offer both money and expertise

Currently we are involved in 21 projects dealing with various science areas





The project TwiNL

Twitter: a social network on which participants communicate by exchanging messages of up to 140 characters

With about 8 million messages in Dutch per day covering different topics, this is an interesting data source for researchers

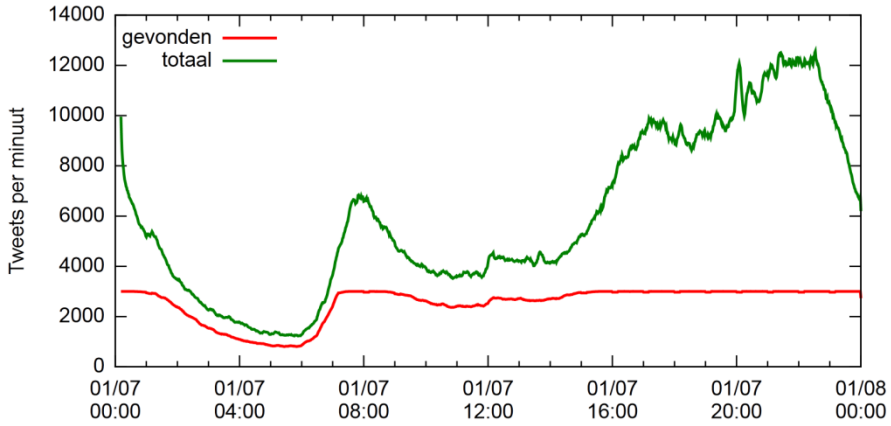
Project task: collect Dutch tweets, offer a search facility for the data and provide different views on search results

Collecting the data

We continuously search for common Dutch words on Twitter (keyword tracking)

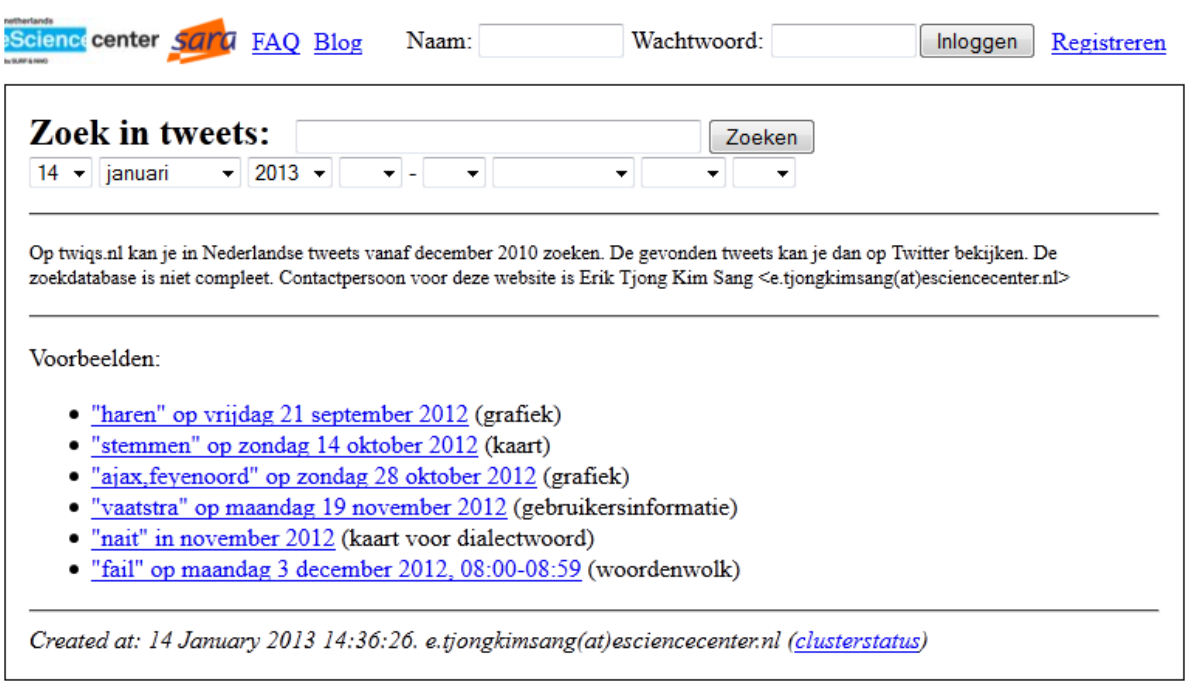
Additionally we collect all the tweets of most frequent 5000 users in Dutch of the previous month (user tracking)

This generates close to 4 million tweets per day



Searching the data

A web interface has been built for searching the data: http://twiqs.nl



Challenges and solutions

We collected two years of Dutch tweets: about 2 billion (2,000,000,000) tweets with metadata taking up about 5 terabytes (5,000,000,000,000) of disk space

The collection is updated continuously

We want to be able to search both tweet text and the associated metadata

We divided the data in files containing one hour of tweets

Parallel processing on a 530-node Hadoop cluster accessible via Google's MapReduce framework, is used for limiting the search times

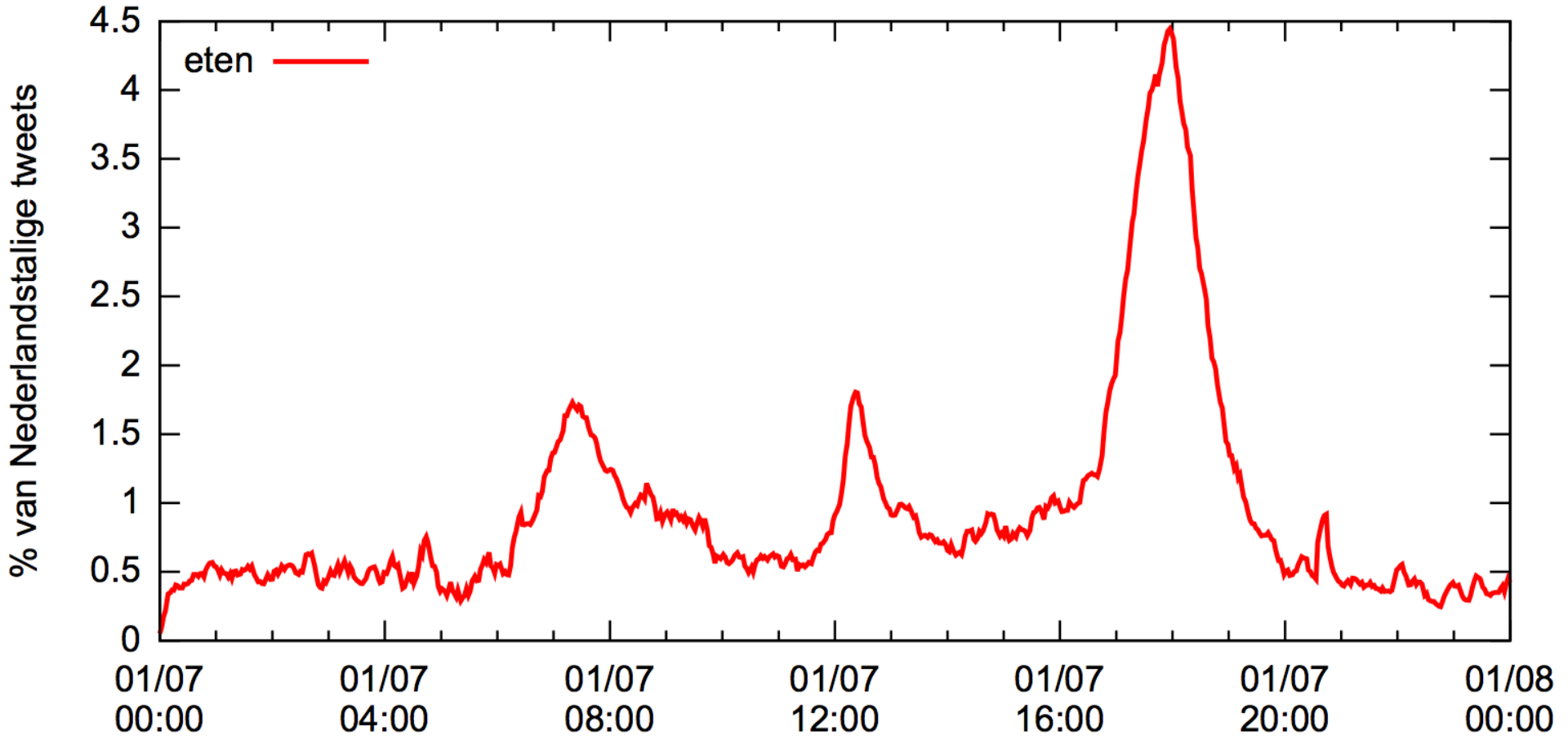
Results of earlier searches are kept in a cache

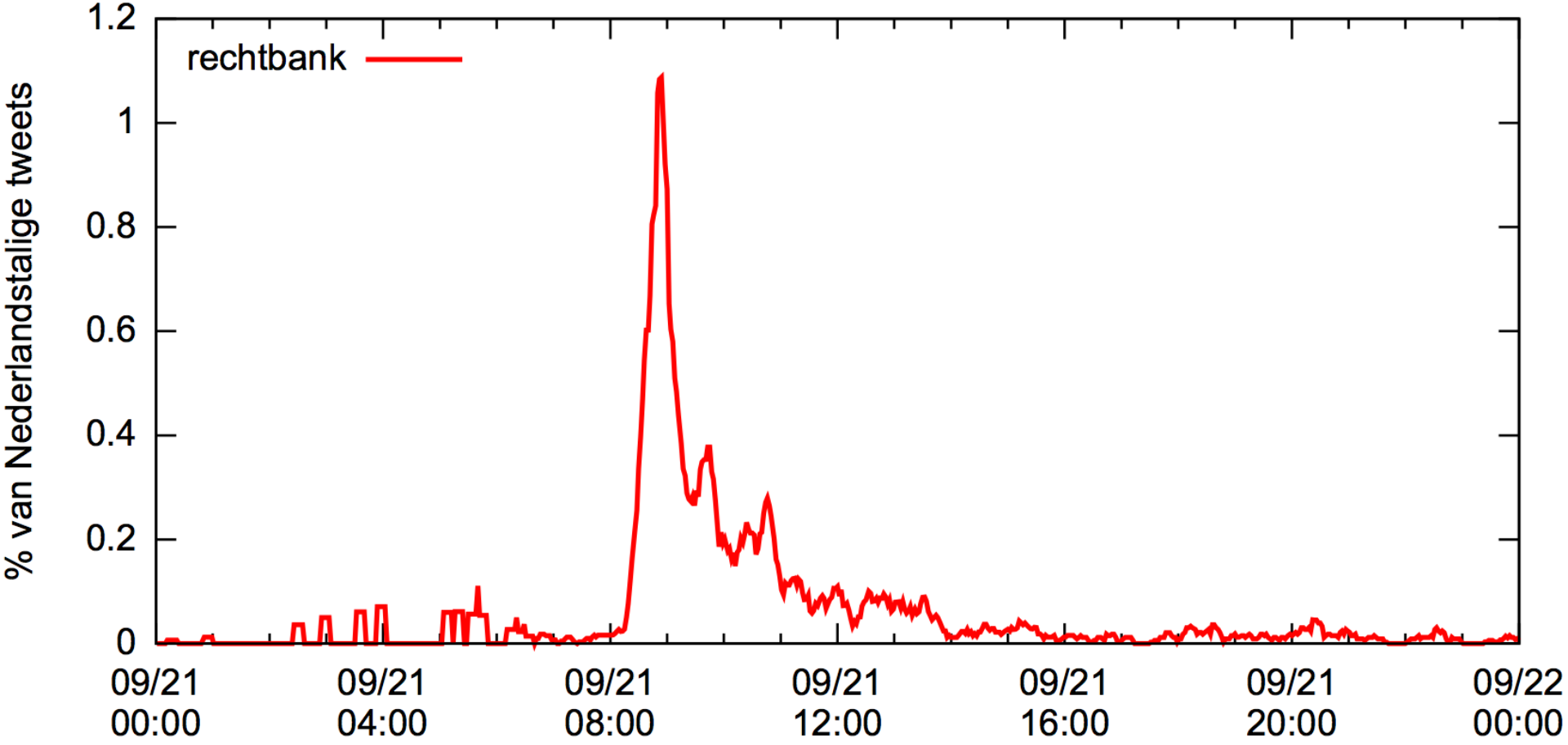
Views on search results

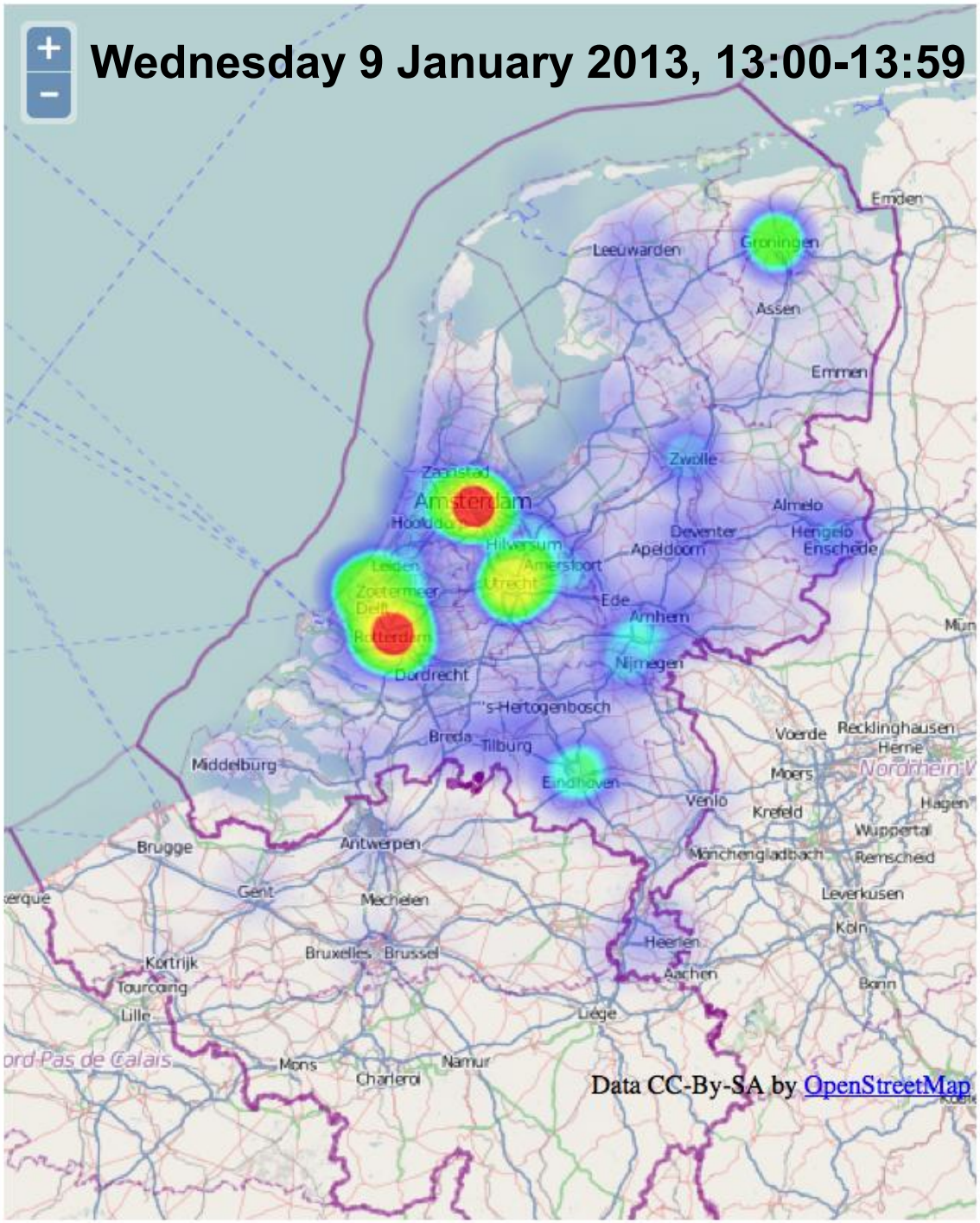
We offer five different views on search results:

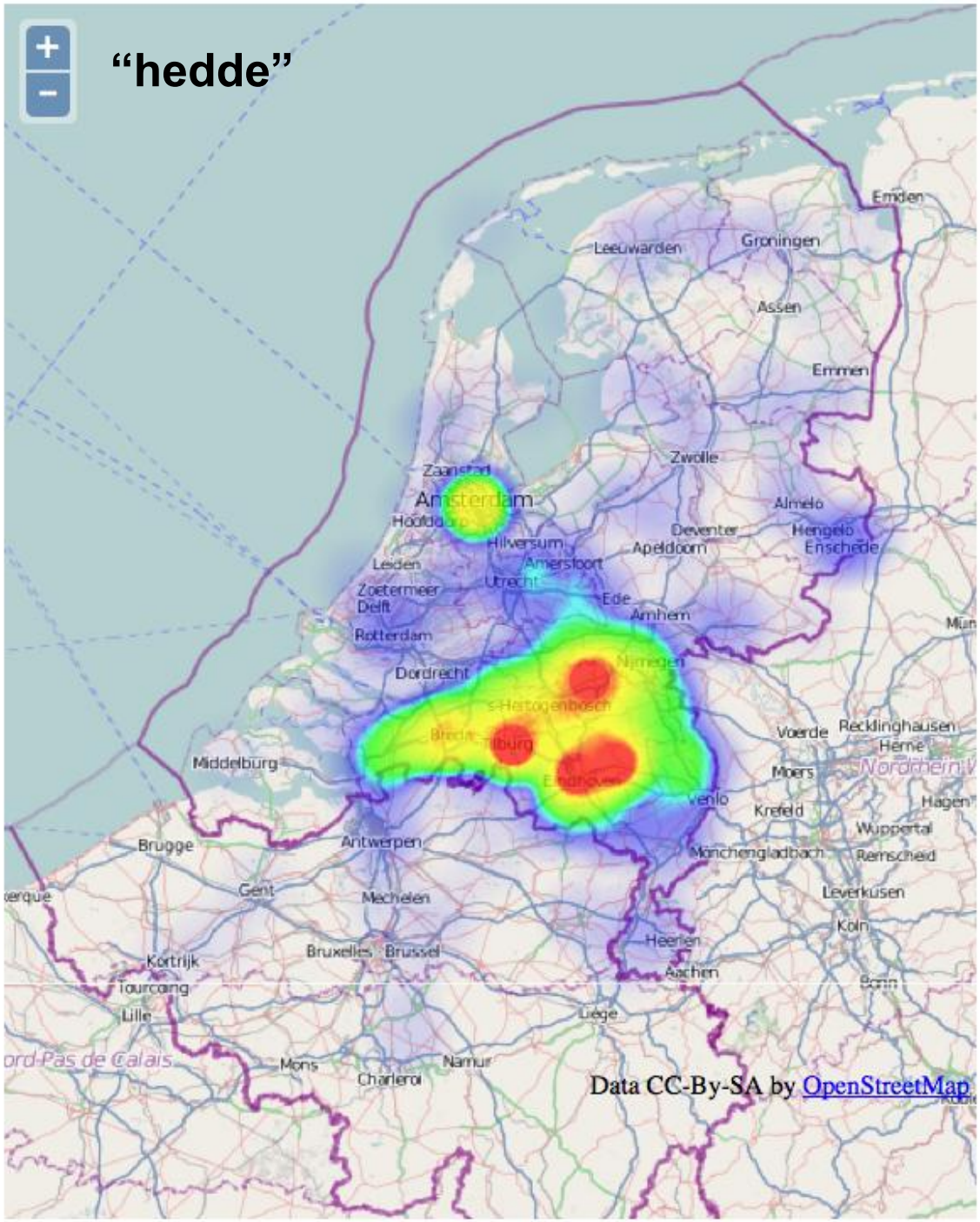
- Tweets (restricted)
- Keyword frequency graph
- Keyword usage map
- Vocabulary overview, including sentiment analysis
- User analysis

All numbers used for creating the views are available for download

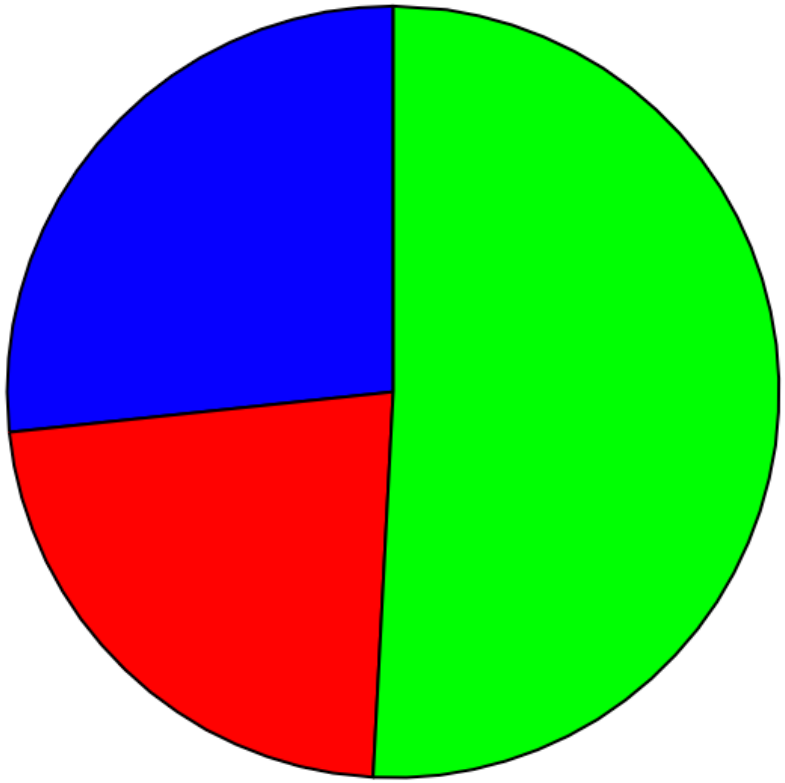

eten
% van Nederlandstalige tweets
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
01/07 00:00
01/07 04:00
01/07 08:00
01/07 12:00
01/07 16:00
01/07 20:00
01/08 00:00


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


Wednesday 9 January 2013, 13:00-13:59
Emden
Leeuwarden
Groningen
Assen
Emmen
Zwolle
Zaanstad
Amsterdam
Almelo
Hengelo
Enschede
Deventer
Apeldoorn
Hilversum
Amersfoort
Leiden
Zoetermeer
Utrecht
Delft
Rotterdam
Ede
Arnhem
Nijmegen
Dordrecht
's-Hertogenbosch
Breda
Tilburg
Middelburg
Eindhoven
Venlo
Voerde
Recklinghausen
Herne
Moers
Krefeld
Hagen
Wuppertal
Mönchengladbach
Remscheid
Leverkusen
Köln
Bonn
Brugge
Antwerpen
Gent
Mechelen
Bruxelles
Brussel
Kortrijk
Tourcoing
Lille
Heerlen
Aachen
Liège
Mons
Charleroi
Namur
Data CC-By-SA by OpenStreetMap


“hedde”
Emden
Leeuwarden
Groningen
Assen
Emmen
Zwolle
Zaanstad
Amsterdam
Almelo
Hengelo
Enschede
Deventer
Apeldoorn
Hilversum
Amersfoort
Leiden
Utrecht
Zoetermeer
Delft
Ede
Arnhem
Rotterdam
Dordrecht
Nijmegen
s-Hertogenbosch
Breda
Tilburg
Eindhoven
Middelburg
Venlo
Voerde
Recklinghausen
Herne
Moers
Krefeld
Hagen
Wuppertal
Remscheid
Mönchengladbach
Leverkusen
Köln
Bonn
Brugge
Antwerpen
Gent
Mechelen
Heerlen
Aachen
Kortrijk
Bruxelles
Brussel
Tourcoing
Lille
Liège
Mons
Charleroi
Namur
Data CC-By-SA by OpenStreetMap

All tweets of Tuesday 4 December 2012

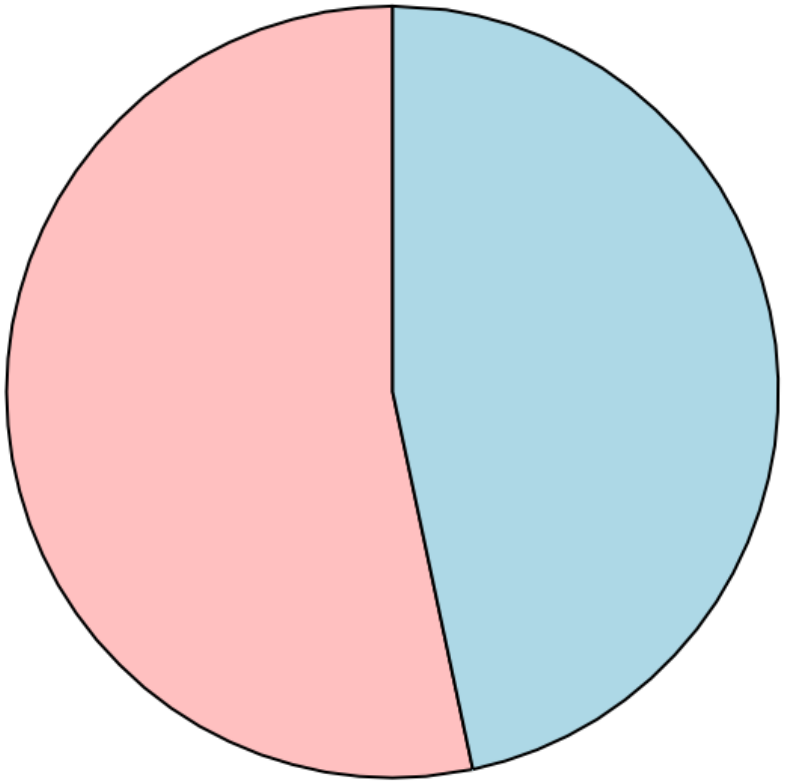


Leeftijd

< 18 jaar: 50.7%

18-25 jaar: 22.4%

> 25 jaar: 26.8%

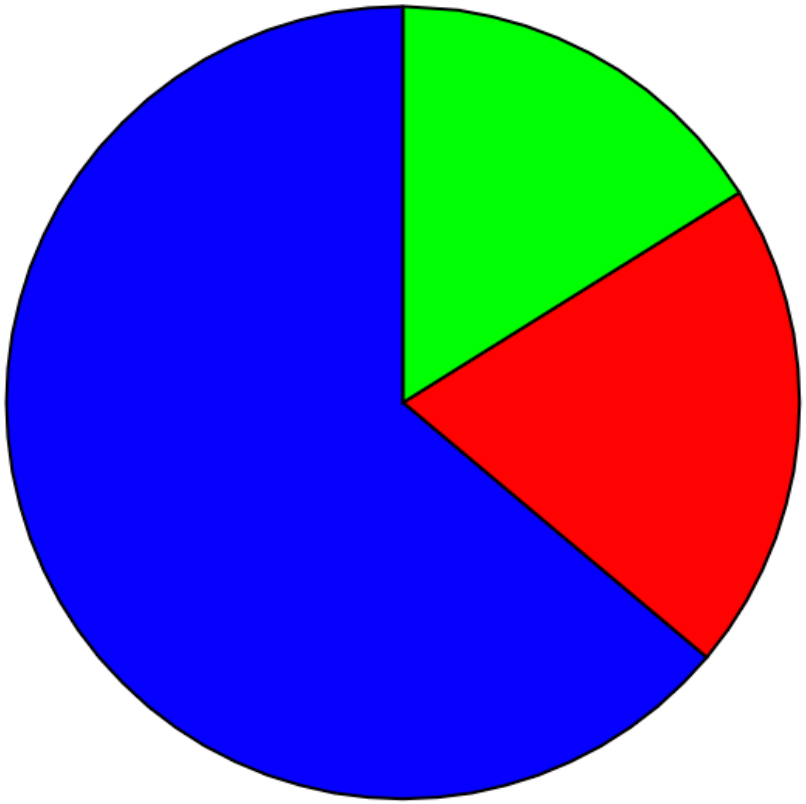


Geslacht

man: 46.6%

vrouw: 53.3%

“bultrug”

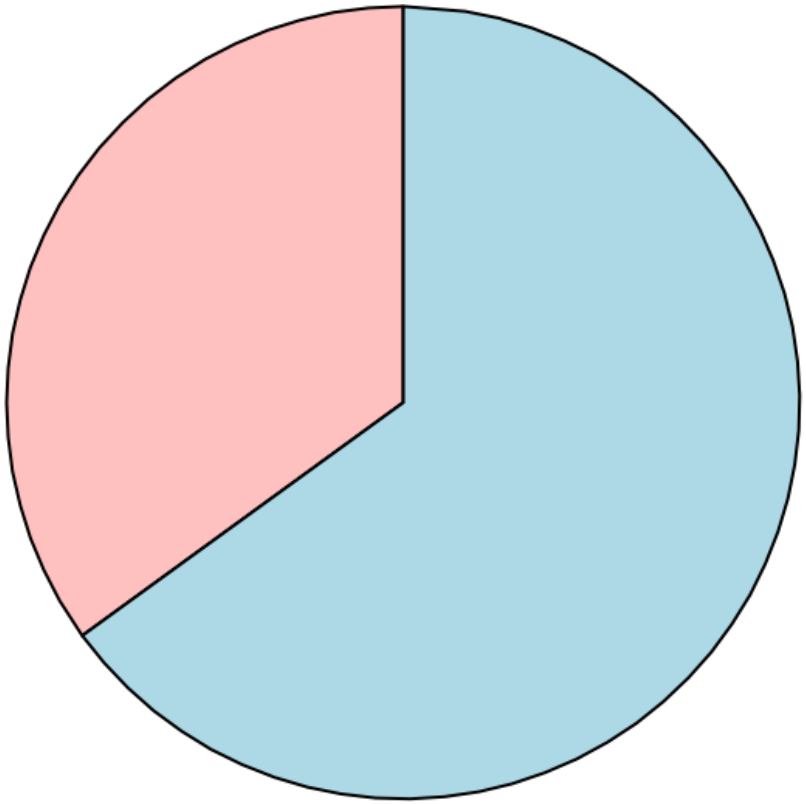


Leeftijd

< 18 jaar: 16%

18-25 jaar: 20%

> 25 jaar: 64%



Geslacht

man: 65%

vrouw: 35%

zweden: +49

duitsland: +8

belgie: +6

nederland: +3

mali: -1

sylvie: -23

rafael: -8

armstrong: -4

oprah: -1

leona: +37

wijsheid: +17

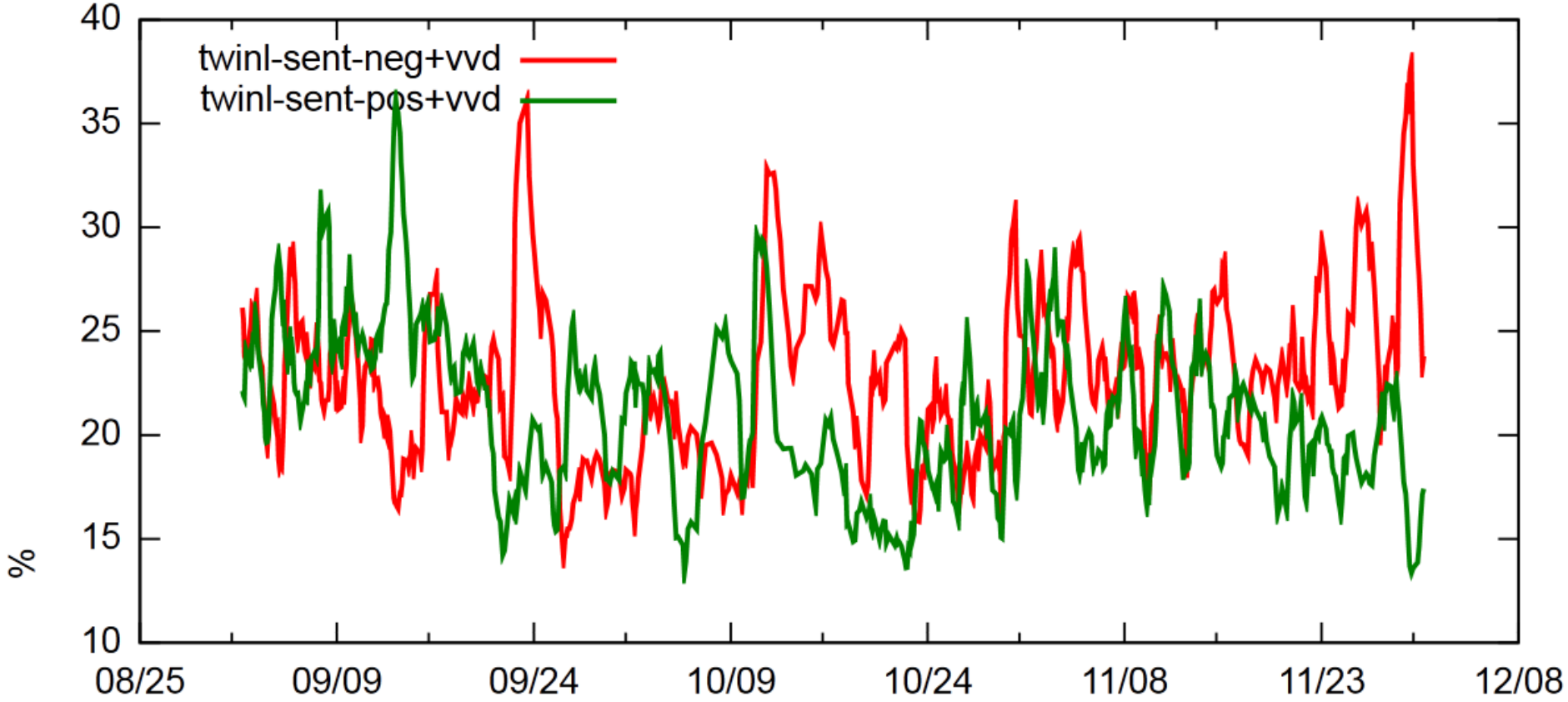
macht: +13

bier: +12

iphone: +10

geld: -1

Sentiment for “VVD” in september-november 2012



Concluding remarks

We have build a web interface for searching Dutch tweets: http://twiqs.nl

We use a parallel architecture (Hadoop) rather than an index for storing the underlying data

The interface can be used by university staff and students for exploring Twitter conversations, for example for topic discussions, dialect usage and product sentiment

Case studies relying on this data resource will be presented after this talk