Discussion Paper

Big Data Masterclass and DataCamp 2015

2016 | 15

Piet J.H. Daas
Barteld Braaksma
Robin Aly
Yuri Engelhardt
Djoerd Hiemstra
Raul Zurita Milla

Content

Summary

This documents describes the two-stage approach in Big Data training developed at Statistics Netherlands. The first stage is following a Masterclass on Big Data. This class provides the theoretical basis and consists of a one-day set of lectures on big data-related topics. The second stage is a DataCamp which enables participants a 'hands on' expertise through five days of practical work. This ensures a 'learning-by-doing' approach which is considered essential for Big Data training. Both stages and The results obtained in the 2015 DataCamp are included in this document.

Keywords

Big Data, training, data science, experimental approach

1. Introduction

Big Data has great potential for official statistics. In a world in which increasing amounts of data are being produced, more and more data may become available that contain traces of information relevant for official statistics production. However, this information can only be used if it is identified and extracted. It is for this reason that employees of National Statistical Institutes need to develop the skills to work with Big Data. Because of this need, it was decided by Statistics Netherlands to develop a two-stage approach in Big Data training. The first stage is following a Masterclass on Big Data. This class provides the theoretical basis and consists of a one-day set of lectures on big data-related topics. The second stage is a DataCamp which gives 'hands on' expertise through 5 days of practical work, since 'learning-by-doing' is considered an essential part of the training. Both stages will be described in this document. In addition, an overview is provided of the results obtained during the first DataCamp held in November 2015.

1.1 Masterclass

To properly introduce Statistics Netherlands employees to topics relevant for Big Data, a comprehensive one-day program was created. An overview of the program is shown in Table 1. The program allowed sufficient room for discussion(s) and primed the participants for working with Big Data. Prior to the start of the Big Data expert track, managers from the statistical subject matter departments of Statistics Netherlands selected those people of which they believed that they would benefit

Table 1. Overview of the lectures of the Masterclass Big Data

Start - end time	Time	Topics	Teacher(s)
9:30 – 10:00	30'	Opening and introduction to Big Data	Piet Daas
10:00 – 10:30	30'	Access to Big Data and privacy aspects	Peter Struijs
10:30 – 10:45	15'	Short break	
10:45 – 11:30	45'	Big Data and hard & software	Marco Puts
11:30 – 12:15	45'	Starting with Big Data: data exploration and visualisation	Edwin de Jonge / Piet Daas
12:15 – 13:15	60'	Lunch break	
13:15 – 14:00	45'	Editing of Big Data	Marco Puts
14:00 – 14:45	45'	Big Data estimation	Joep Burger
14:45 – 15:00	15'	Short break	
15:00 – 15:30	30'	Big Data and populations: mobile phone and traffic loops	Martijn Tennekes
15:30 – 16:00	30'	Big Data processes: using traffic loop data	Marco Puts
16:00 – 16:15	15'	The Big Data roadmap	Peter Struijs
16:15 – 16:30	15'	End of the day, time for additional questions & discussion	Piet Daas (e.a.)

from attending the Big Data Masterclass. No participants from Methodology, IT and supporting departments like HRM and Finance were admitted to avoid too much heterogeneity in knowledge, experience and interests. The Masterclass was given twice, in Heerlen on 18 June 2015 and in The Hague on 25 June 2015. Participants were able to attend the Masterclass at their preferred location. In total, 46 Statistics Netherlands employees attended the Big Data Masterclass and six teachers from the Methodology and Process Development department delivered the lectures.

2. DataCamp

In addition to the Big Data Masterclass, which basically trained people in a passive way by giving lectures, a clear need was felt for more practical training through a learning-by-doing format. The format was developed in close cooperation with the University of Twente and called a DataCamp. Preparations started in August 2015 and were conducted, mostly through Skype meetings, by a team consisting of two Statistic Netherlands employees and four senior researchers from the University of Twente.

The camp itself lasted 5 working days, from 23 to 27 November 2015, and was held in a hotel at the campus of the University of Twente. The overall schedule is shown in Table 2. The location was chosen outside the Statistics Netherlands premises to prevent that Statistics Netherlands employees were disturbed by any 'statistical issues of the day'. During the DataCamp, ten Statistic Netherlands (SN) employees and nine employees from the University of Twente (UT) participated. The SN-employees were mostly selected from the Masterclass participants. The UT-employees were PhD-students and post-doctoral researchers from three departments; i) the Faculty of Electronic Engineering, Mathematics and Computer Science (EWI), ii) the Faculty of Behavioural, Management and Social Sciences (BMS), and iii) the Faculty of Geo-Information Science and Earth Observation (ITC). The participants formed teams (see below) and obtained hands-on experience by working with Big Data and making use of an infrastructure specifically set-up for Big Data analysis, including a SPARK cluster.

The availability of data was essential. Therefore the following data sources were placed on the secure Hadoop cluster of University of Twente:

1) Collection of tweets provided by Twiqs.nl (from December 2010 onwards)
2) Processed and aggregated road sensor data for all highways in the Netherlands (from 2010 till 2015)
3) Anonymized AIS data (GPS signals) from ships in the Dutch waterways for a period of 1 month (July 2015)
4) Wi-Fi sensor data of the municipality of Assen (from February 22th until March 1st 2015) including road sensor data for that period
5) List of company websites and their content from Data provider

In addition, several Coosto data accounts were available to enable the study of messages on various social media platforms, such as Twitter and Facebook, collected by this commercial firm. Depending on the needs of the participants other data were downloaded during the camp. The sizes of these data sets were usually quite small.

Table 2. Agenda of the DataCamp 2015

Monday 23-11-2015			
When	**What**	**Where**	**Speaker/Chair**
11:00	Welcome CBS	Drienerburght D	Organizers
11:15	Big data processing principles	Drienerburght D	D. Hiemstra & R. Aly (UT)
12:30	Lunch	Drienerburght D	
13:45*	Research questions (w students)	Design Lab	P. Daas (SN)
14:00*	Break		
14:15*	Dutch tweets (w students)	Design Lab	E. Tjong Kim Sang (Meertens Institute) / D. Hiemstra (UT)
15:30	Break / walk to Drienerburgth		
15:45	Self-presentation and match making	Drienerburght D	Participants /Y. Engelhardt (UT)
16:30	Hands on (connection setup)	Drienerburght D	
18:00	Data dinner (with talk)	Drienerburght D	P. Daas (SN)
19:00	Evening session (open end)	Drienerburght D	

Tuesday 24-11-2015			
When	**What**	**Where**	**Speaker/Chair**
08:30	Morning session	Drienerburght D	
11:30	Team status updates	Drienerburght D	Participants/Organizers
12:00	Lunch	Drienerburght D	
13:00	Afternoon session	Drienerburght D	
17:00*	Public status updates	Drienerburght D	Participants/Organizers
18:00	Data dinner (with talk)	Drienerburght D	D. Hiemstra (UT)
19:00	Evening session (open end)	Drienerburght D	

Wednesday 25-11-2015			
When	**What**	**Where**	**Speaker/Chair**
08:30	Morning session	Drienerburght D	
11:30	Team status updates	Drienerburght D	Participants/Organizers
12:00	Lunch	Drienerburght D	
13:00	Afternoon session	Drienerburght D	
17:00*	Public status updates	Drienerburght D	Participants/Organizers
18:00	The Banquet	Faculty club	
19:00	Evening session (open end)	Drienerburght D	

Thursday 26-11-2015			
When	**What**	**Where**	**Speaker/Chair**
08:30	Morning session	Drienerburght E+F	
10:00	Walk to design lab		
10:45*	Visualizing (Geospat) Data (w students)	Design Lab	D. González (Vizzuality) /R. Zurita Milla & Y. Engelhardt (UT)

11:30*	Break		
11:45*	Data Visualization at CBS (w students)	Design Lab	M. Tennekes (SN) /R. Zurita Milla & Y. Engelhardt (UT)
12:30	Walk to Hotel	Drienerburght	
12:45	Lunch	Drienerburght E+F	
13:30	Afternoon session	Drienerburght E+F	
17:00*	Public status updates	Drienerburght E+F	Participants/Organizers
18:00	Participant evaluation of DataCamp	Drienerburght E+F	M. Puts (SN)
19:00	Evening session (open end)	Drienerburght E+F	

Friday 27-11-2015

When	What	Where	Speaker/Chair
08:30	Morning session	Drienerburght E+F	
11:30	Finish practical work & move to	Ravelijn	
12:00	Lunch and welcome guests	Ravelijn 1501	B. Braaksma (CBS) & R. Aly (UT)
12:20	Participants result presentation	Ravelijn 1501	Participants / B. Braaksma (SN) & R. Aly (UT)
13:30	An outsiders view on the data camp	Ravelijn 1501	F. Gromme (Goldsmiths, Univ. London) / B. Braaksma (SN)
13:45	Data camp highlights and summary	Ravelijn 1501	P. Daas (SNS) & D. Hiemstra (UT)
14:00	Signing of collaboration agreement	Ravelijn 1501	T. Tjin-a-Tsoi (SN) & V. van der Chijs (UT) / S. van Tongeren (UT)

* means open to the general public, more info http://bit.ly/DataCamp2015

Prior to the DataCamp, a number research questions were generated based on the data sets available. The main purpose of these questions was to act as examples for the participants. At the start or during the camp, participants were free in adjusting these questions or to propose their own; for instance as the result of the outcome of their (first) analysis. The following research questions were given as examples at the start of the camp:

1) Day trips of Dutch people: What information does social media provide on day trips of people living in the Netherlands?
2) Tourist movements: What places in the Netherlands do tourist visit?
3) Selectivity of people producing GPS containing tweets: What are the characteristics of Dutch people active on Twitter that create such tweets?

4) Selectivity of people participating on online campaigns: What are the characteristics of Dutch people active on Twitter that participate in online campaigns?
5) Economic activity of companies: Can the economic activity of companies be classified by data available on their website?
6) Economic growth: Do changes in traffic intensity indicate changes in economic growth?
7) Map tick bite risks: Can we find a way to “normalize” the amount of tick bites according to the “population at risk”?
8) Map phenological development Find species that behave in a similar way or try to identify similar periodical trends in species development?
9) Freelancers/Independent contractors (‘ZZP’): Can social media be used to identify and determine the number of freelancers/independent contractors ("ZZPers") in the Netherlands?
10) Location data of ships: What information does AIS data provide on travel and harbour visits of ships?
11) Social media usage in the Netherlands: How active are people really on the various social media platforms available in the Netherlands?

Since the focus was on obtaining hands-on experience, the first day of the camp paid special attention to making sure all pre-requisites were met for learning these skills. These pre-requisites were:

i) Formation of groups. In advance, all participants were asked to introduce themselves through a PowerPoint poster.
ii) Selection of research task and accompanying data set by each group. A gross list of tasks and data sources to choose from had been prepared by the organizing team
iii) Making sure every participant had a working computer and all accounts provided access to the backend infrastructure and data
iv) Introducing the attendees to working with the Big Data infrastructure
v) Providing examples of Big Data research to stimulate innovative ways of thinking (‘thinking outside the box’)

To maximize the time spent on doing hands on work during the DataCamp, Statistics Netherlands employees spend the night in the hotel on the campus. Breakfast was at the hotel and lunch and dinner were served in front of the meeting room where the analysis was performed. This enabled employees to spend as much time as needed on Big Data analysis. In addition, the room remained open until 22:00 in the evening. On Tuesday, Wednesday and Thursday - at 11 and at 17 o’clock - short work status updates were giving by each group so progress could be checked upon. This was also the period during which groups could provide feedback to one another and when external experts, such as experts from the University of Twente and Statistics Netherlands, could provide feedback.

Experienced senior researchers of University of Twente and Statistics Netherlands were continuously available during the DataCamp to assist the groups whenever needed. Their experience included use of Big Data infrastructures, Big Data analytics

and visualization methods, among other things. Djoerd Hiemstra, Robin Aly, Yuri Engelhardt and Raul Zurita Milla were available all week on behalf of the University of Twente and Piet Daas was available all week on behalf of Statistics Netherlands. On Wednesday and Thursday, Martijn Tennekes and Marco Puts from Statistics Netherlands provided additional support.

Three external guests visited the DataCamp. On Monday Erik Tjong Kim Sang from the Meertens Institute came along and gave a presentation on the Twitter data he has been collecting for over five years. The data is available on Twiqs.nl; a website that can be used for searching in billions of Dutch tweets. He presented the methods used for collecting, processing and storing the tweets, and explained how to visualize search results with examples. On Wednesday David González from Vizzuality attended the camp and gave a presentation on the benefits of using visualizations for (Big) data analytics, including an introduction to the CartoDB software package. He also assisted several groups in creating some nice map-based visualizations and animations. On Thursday and Friday Francisca Gromme from Goldsmiths, University of London visited the camp to study how the participants cooperated and interacted with the organizers. She gave a presentation on her findings during the closing meeting on Friday.

Thursday afternoon/evening was spend on creating slides and finalizing the analysis, so the results could be presented on the last day of the camp (Friday) to representatives of the higher staff of the University of Twente and of Statistics Netherlands. Among those were the Director General of Statistics Netherlands and the President of the Executive Board of the University of Twente. In the next sections the findings of each group attending the DataCamp are briefly described.

2.1 Netherlands in Bloom: A spatio-temporal model of the onset of spring

By: Maaike Hersevoort (SN) and Hamed Mehdipoor (UT-ITC)

Research question

The spatio-temporal mapping of the onset of flowering of different plants and trees is a subject that is of interest to scientists such as biologists, phenologists and climatologists. A valuable source of information is found in volunteered observation and open meteorological data. During the DataCamp we investigated how volunteered observations on the onset of flowering can be used to dynamically map the onset of spring in the Netherlands.

Data sources

Three different datasets were available. The first is a dataset of reports by volunteers recorded via the Dutch initiative ´Natuurkalender´. The volunteers reported the location and time of blooming onset of different plants and trees. We used data of 2003 to 2015 on the wood anemone (in Dutch: ´bosanemoon´) for our analysis. The second dataset is available at the website of the Royal Dutch Meteorological Institute

(in Dutch: KNMI) and consists of average daily temperatures on a 1x1 km grid covering the Netherlands. This data is also available for multiple years, but only data from 2013 and 2014 were used in our analysis. The third dataset was the collection of tweets from Twiqs.nl.

Methodology

The spatio-temporal mapping was done using the Growing Degree Days model (GDD). According to literature the date of blooming onset is highly correlated with the cumulative temperature prior to the date. Cumulative temperature for a given point on a given day is calculated as the addition of the average daily temperature for that point; starting date was January 1st. Negative temperatures are disregarded, as these days do not contribute to plant development. For the wood anemone the cumulative temperature threshold at which the blooming onset happens was calculated extracting and averaging the cumulative temperatures on the basis of the dataset of ´Natuurkalender´, using data from 2003 to 2015.

Using the big data cluster and the program RSpark (the statistical package R in a version adapted for efficient use of a big data cluster), the meteorological dataset of average daily temperatures was converted in two steps. First the cumulative temperature was calculated for each point of the 300 by 350 km grid for each day of the years 2013 to 2014, with starting point the first of January. In a second step these 2 times 365 matrices of cumulative temperatures were converted into binary matrices, noting for each point for each day if the cumulative temperature had reached the blooming onset threshold of the wood anemone.

Results

As an example figure 1 shows the binary maps for 2013 and 2014 for four different dates. The blooming onset does not happen all at once over the whole of the Netherlands. Instead, there is a clear pattern in space and time, which varies for 2013 and 2014. In 2014 the onset of flowering starts earlier then in 2013, the spread is oriented from west to east as opposed to south to north in 2013 and it takes about three weeks as opposed to a month in 2013 from the first flowering in the west of the Netherlands till the whole of the Netherlands is in bloom. 2013 was a very cold year, with a freezing period in March, which explains the late onset and the longer duration in 2013. The difference in spatial pattern is harder to explain. It is known that 2013 had more days with a northern wind than normal for the Netherlands (where normally the wind direction is predominantly southwest-northeast), which might explain the pattern.

Correlation with twitter findings

As a first validation of our results we analysed tweets of 2013 and 2014 mentioning the word bosanemoon in a variety of spellings. Even though the number of tweets mentioning ‘bosanemoon’ is very small (about 250 for each year) it is clearly visible that in 2013 people tweeted later about ‘bosanemoon’ than in 2014 (figure 2). The highest number of tweets about ‘bosanemoon’ in 2013 was during the second half of April, as in 2014 it was during the second half of March. Due to the low numbers, also some artefacts are visible, for example the relatively high number of tweets in

November, which on further analysis turned out to be related to a primary school named ´bosanemoon´.

Conclusion and further research

This first analysis shows that even for a region as small as the Netherlands a spatio-temporal modelling of spring onset is possible. Using this method and analysing multiple years and multiple species can contribute to the research on climate change.

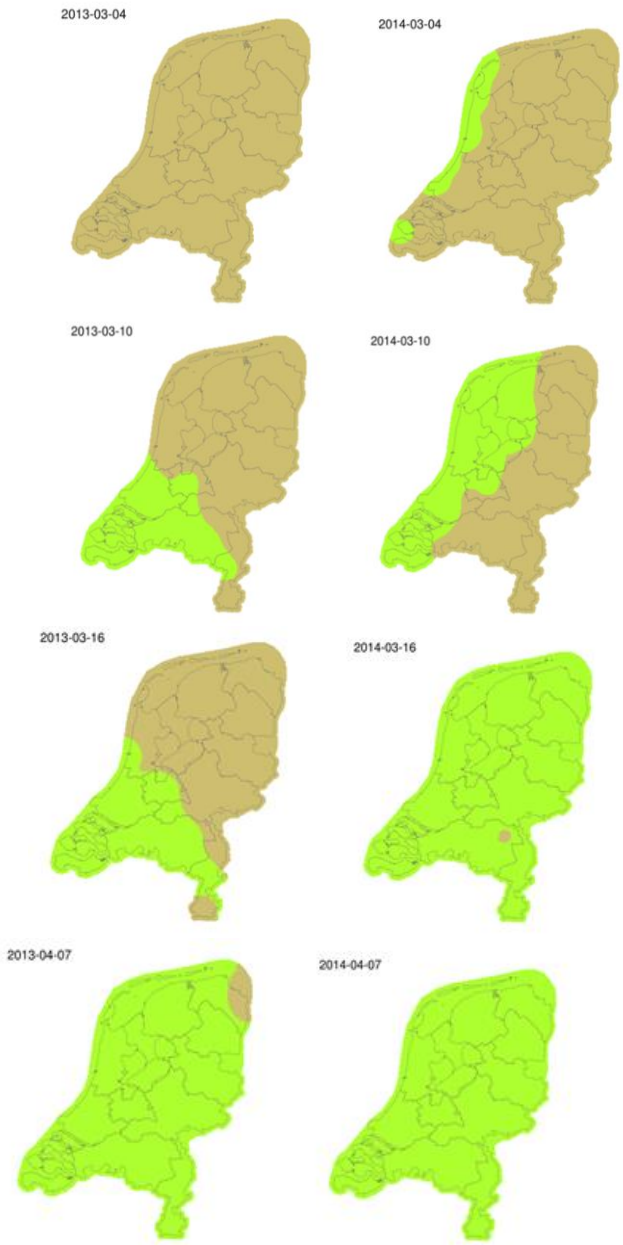

Figure 1 The onset of flowering of the wood anemone in the Netherlands, 2013 and 2014.

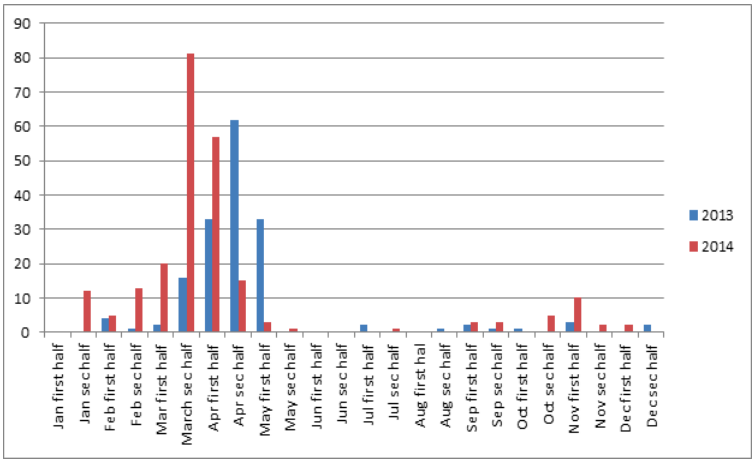

Figure 2. Number of tweets containing the word 'bosanemoon', 2013 and 2014.

This analysis might also be useful for volunteer-based networks such as ´Natuurkalender´ uses as it could help them direct their volunteers to pay attention to certain phenomena in certain regions or periods. In a broader context, the same methodology could be used on other data that has a spatio-temporal component and a cut-off point. An example is flu measurements ('de grote griepmeting') where volunteers spread out over the Netherlands are asked regularly if they are experiencing flu like symptoms. Another area might be the movement of people, for example tourists.

2.2 Mobility of people during a festival in the centre of Assen

By: Mitra Baratchi (UT-EWI), Egbert Hardeman (SN) and Maarten Pouwels (SN)

Research question

The work focussed on an exploratory study of the mobility of people during a festival in the centre of a city. A number of research questions emerged, such as:

- What kinds of people attend the festival?
- How do people move across the city?
- What locations are most visited and at what times?
- Do people from camping sites also attend the festival?
- How many people visit the festival on multiple days?
- How long do people stay, on average, in the city?

Used sources

The most important source is a set of data collected by a series of Wi-Fi sensors distributed throughout the centre of Assen during the Tourist Trophy (TT) festival; 24 until 26 of February 2015. The sensors registered all 'pings' from devices with Wi-Fi capability during a period of 8 days (2 days before, 3 days during and 3 days after the festival). For each 'ping' a number of items was stored, these were:

- Sensor id
- Timestamp (Unix)
- Signal strength

- Mac-address of the device
- OUI-code (Organizationally Unique Identifier) of the producer of the device

A number of additional sources were available covering the same period in time. For example: data on the number of vehicles detected by road sensors on the highways close to Assen. Due to lack of time, these were not used during the study.

Approach used
The research was predominantly exploratory and we were particularly focussed on low hanging 'fruit'. None of the persons who did the work were familiar with the data set. As a result, the initial work focussed on accessing and getting to know the data. We tried Spark-R and Pig-Latin for this. Since using Pig-Latin was successful, we did not try any other language. Pig-Latin was used to study the data set in more detail. Based on this experience, the AIS-position data of ships was also studied (see section 2.3).

Results
The first challenge was converting the Unix Timestamp into a more 'normally' used Coordinated Universal Time (UTC) stamp. Code was generated that was able to convert Unix Timestamp into dates and times (hours and seconds).
The second challenge was counting the number of devices detected per hour by all Wi-Fi sensors. We were successful in this after many hours of work.
The next step was counting the number of unique devices for each Wi-Fi sensor. For each sensor the geo-coordinates of its location were obtained from a reference dataset. Based on this the first result that could be visualized was obtained: a plot over time of the change of the number of devices detected by each Wi-Fi sensor during the festival in the centre of Assen. The assistance of David González from Vizzuality helped the group tremendously in creating the visualization in the web-based tool CartoDB.

Two important adjustments were made to improve the findings of the study:

1 A device can be detected multiple times per hour by the same Wi-Fi sensor; i.e. multiple 'pings' can be received from the same device. The code was adjusted to make sure each device was only included once every hour. Because of this adjustment, the total number of devices detected more approximated the total number of people at the location of the Wi-Fi sensor.
2 The program of the festival was used to check during which period at which sensor specific bands were performing on stage. This enabled us to additionally include the link between band playing and the number of people in the area.

Finally, the number of OUI-codes of the devices detected was checked. The OUI-code is a number that can be used to uniquely identify the manufacturer of a device. However, it is important to realize that a manufacturer can use multiple OUI-codes. From the internet a list was obtained including all OUI-codes and its accompanying

manufacturers. From the list of all connections per OUI-code, the most popular devices per manufacturer were obtained.

Pictures

This link shows the visualization of the number of devices per hour during the period investigated. The darker and larger the circles the more devices were detected. A black dot is used to indicate a location were a band played. In Figure 3 on overview is given of the Wi-Fi sensors detected. Remarkably, the figure reveals that the Wi-Fi sensors detected the presence of other sensors; 8% in Figure 3.

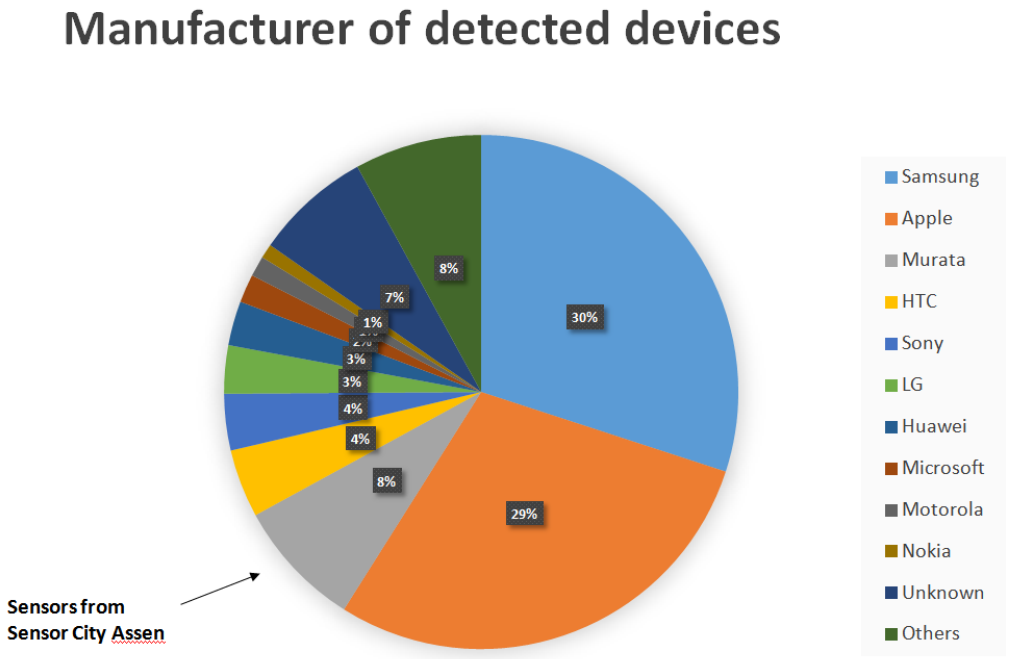

Figure 3: Circle diagram showing the number of devices detected per manufacturer.

2.3 Exploring the potential of AIS-data

By: Maarten Pouwels (SN), Egbert Hardeman (SN) and Mitra Baratchi (UT-EWI)

Research question

The work is focussed on exploring the potential of Automatic Identification System (AIS) data. This is a very large data set composed of GPS-coordinates of ships for which there is ample experience. Aim of this work was to get insights on the structure of the data and its potential applications. There is a special interest in tracking ships over time, e.g. for transport purposes, and use the data to determine the 'crowdedness' of the water ways. Particular attention was paid to the accuracy of the GPS-signals and the coverage of the country.

Sources used

Statistics Netherlands had anonymized an AIS-data set for the data camp. All research was performed on this data set. The file consisted of 8 so-called position files and 8 time files. The position files each contained about 150 million records. Each record contained the following variables:

- Latitude
- Longitude
- Speed over ground
- Course over ground
- Rotation
- Navigation status
- Ship id
- Line number

During the data camp the majority of the work focussed on analysing position files. Linking the position and time files is something that will be looked at in future research.

Approach followed

Since the research was predominantly of an exploratory nature, a large part of the work focussed on counting records and checking the coverage of variables. Amongst others, the number of unique ships and the occurrence of particular statuses were determined. For these tasks Spark SQL and Pig-Latin were used. Spark SQL was found to be particularly useful to count the numbers of records and Pig-Latin was used to process large files.

Results

Since it was not easy to add a time to the data in the position files, it was checked if the sequence of rule numbers sufficed to track a ship over time. This was found to be the case. By visualising the sequence of subsequent records produced by a single ship in CartoDB the various phenomena included in the dataset were revealed. This demonstrated how a ship traversed through various locks and revealed points in time were the ship was loaded and unloaded. However, the value of the navigation status variable of moving ships on its one did not provide enough information to explain all phenomena observed. Speed data needs to be included to obtain a more complete picture; e.g. ships that obviously dock still indicate that they are on their way. For example: http://bit.ly/1Z328Vh

In addition, an overview was made of the locations of all ships. This provided insight in the coverage of the whole data set. This suggested that the data set covered the target population rather well. The picture produced (Figure 4) revealed nearly every waterway in the Netherlands including the harbour of Antwerp and parts of the river Rhine in Germany.

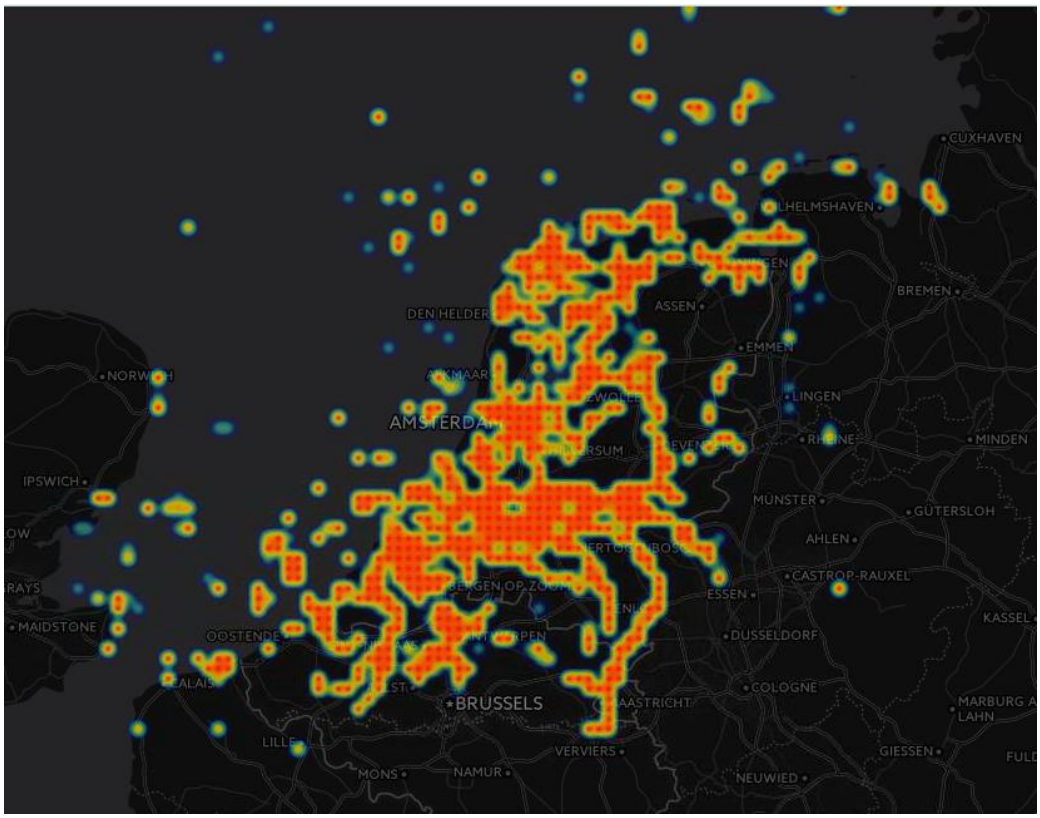

Figure 4. Overview of the positions of all ships included in the AIS-data set studied

2.4 Who Can Attract More Attention to Your Campaign?

By: Anna Priante (UT-BMS), Jair Santanna (UT-EWI) and Kenneth Chin-A-Fat (SN)

During the DataCamp Twitter data was investigated to check which persons gained the most attention during the Movember campaigns of 2012, 2013 and 2014. The first thing that became apparent during the studies was the fact that the total number of Movember tweets decreased rapidly over time (from 8912 in 2010, till 3262 in 2014), while the number of people on Twitter increased during that period (from 667 in 2012 till 2262 in 2014). To determine the most influential tweet or user, we have defined a number of measures described below. The first three are included in the Twitter data studied:

- Number of users that follow an account (followers)
- Number of users followed by an account (following)
- Number of messages created (Tweets)
- Number of messages forwarded (Retweets)

Our starting point is the fact that Twitter users with most number of retweets gained most of the attention. Twitter users with the largest number of followers are apparently 'listened' but are not retweeted often. Twitter users that are retweeted often must also create many tweets. This not only makes them influential but also actively involved in the Movember campaign. After studying the Twitter accounts it

appeared that, as was expected, users that gained most of the attention were predominantly famous Dutch people. More details in Figure 5 below.

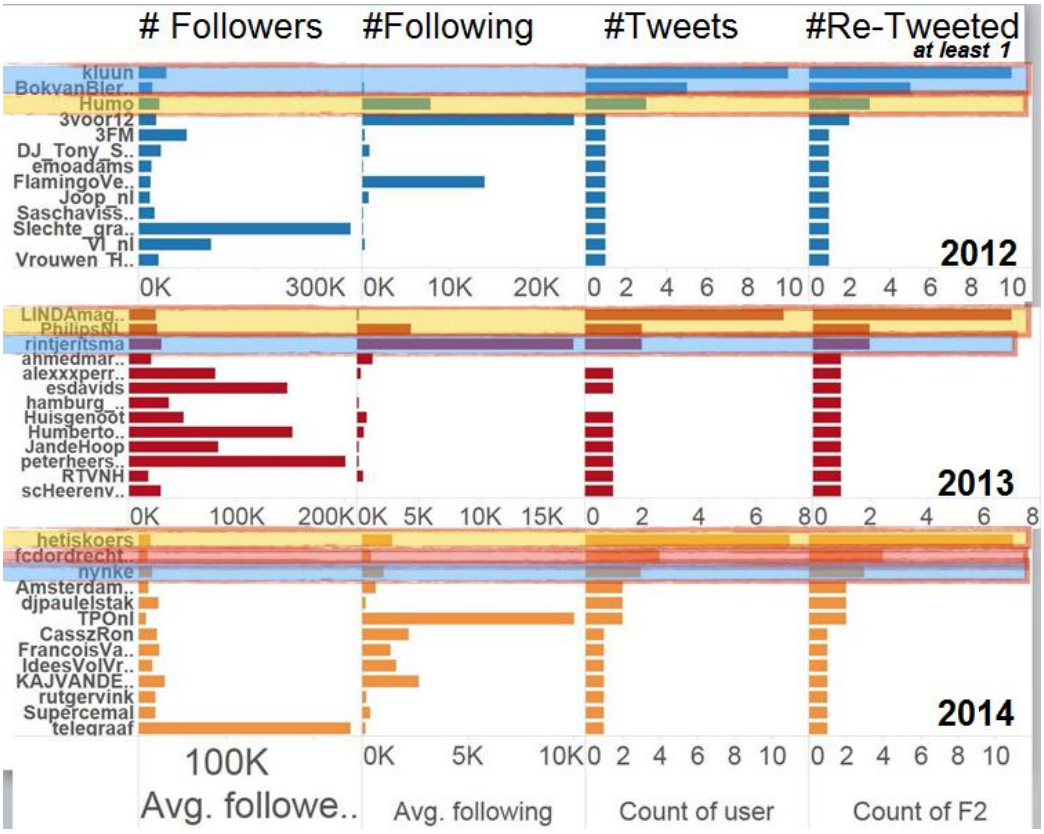

Figure 5 Characteristics of the most popular Twitter accounts in the 2012, 2013 and 2014 Movember campaign.

2.5 Extracting Tourist information from Twitter data

By: Irene Garcia-Martí (UT-ITC), Gustavo García-Chapeton (UT-ITC) and Danny Pronk (SN)

For the research the following question was the starting point: "Can a spatio-temporal footprint of tourist density/activities be created using Twitter data?". In other words: can we extract information from Twitter about the number of tourists in the Netherlands? And can we say something about the time of arrival throughout the year; were they stay and their activities?

Statistics Netherlands publishes tourism statistics based on monthly data collected for accommodation statistics. Amongst others, the number of foreign visitors and guests on camping grounds are obtained from this. When a relation is found between Twitter data and, for instance, the number of guests on camping grounds, official statistics could benefit from this. For instance, by producing more detailed regional information (from the number of tourist per province to detailed data per 5 x 5 km square) or by speeding up the publication. Sending out the survey and waiting for the

response takes a total of 2 months. Based on Twitter-data, tourism findings could potentially be published a few days after the end of the reporting period. It is expected that Twitter data will still be an addition to official statistics though. Completely replacing the latter would be extremely difficult, predominantly because of the selectivity of the data. Not every person in the Netherlands is active on Twitter and it is expected that some age specific groups will use this medium more than others.

Approach used

For the study, a data set with Dutch Twitter messages collected by twiqs.nl was used. It contained about 2.7 billion messages covering the period 2010-2015. Some messages contained geo-coordinates (combination of longitude and latitude) indication the location from which the message was send. A user has to select this option first in the Twitter app, and many just don't. We found that only 2% of the messages in the twiqs.nl data set had geo-coordinates included. As a first step, only those messages were selected. Messages were selected with Spark. Spark enables selecting subsets from data with SQL-like statements, such as: "select * where location is not null".

A Python script was used to subsequently extract the coordinates from the subset of Twitter-messages selected. Despite the fact that all messages were classified as Dutch, many of the messages were found to be positioned outside the Netherlands when plotted. Many of these messages were also found to be written in another language than Dutch. Therefore, all geo-coordinated messages were first checked it they originated from within the Netherlands. This was done by checking if the geo-coordinates of those messages lay within a rectangular square (a 'bounding box') drawn around the Netherlands. As a result half the messages remained.

Next, a set of words was created that could be used to identify Twitter messages originating from camping grounds. The following words were found to be particularly useful: 'tent', 'slaapzak (sleeping bag)', 'camping (camp ground)', 'kamperen (camp)', 'caravan', etc. With this combination of words between 300 till 2000 Tweets per month were selected for the 2012 data set.

Next a spatial analysis study was performed using ArcGIS. A grid of 5 x 5 km was created for the Netherlands. On this, for each cell and each month, the number of Tweets related to camping grounds was plotted. The number of Tweets per cell were aggregated for 2012 and normalized

Visualization

Finally an interactive visualization was created which was composed of three parts: A chart composed of coloured squares (from green, via yellow, to red) indicating the relative number of camping related tweets during a specific month. Findings for the month of August 2012 are shown below. In the first figure, large number of tweets can be discerned in typical vacation areas such as the sea-side in Zeeland and the Frisian islands ('Waddeneilanden'). The second figure shows data for the month of May. Especially in an area close to Heerlen large numbers of camping ground

messages are produced. This very likely resulted from visitors of the Pinkpop festival. Both maps were created with the 'Leaflet' tool.

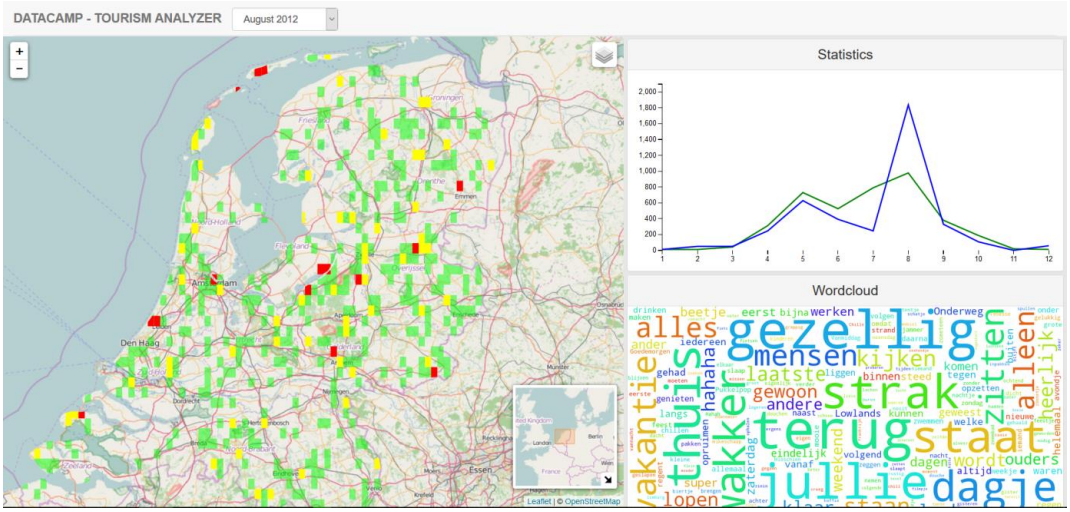

Figure 6 Location and frequency of camping word containing tweets for the Netherlands in 2012. In the top right insert the relative frequency over time compared to the official statistical data for that year is shown. The bottom right insert shows the relative frequencies of the other words included in the tweets.

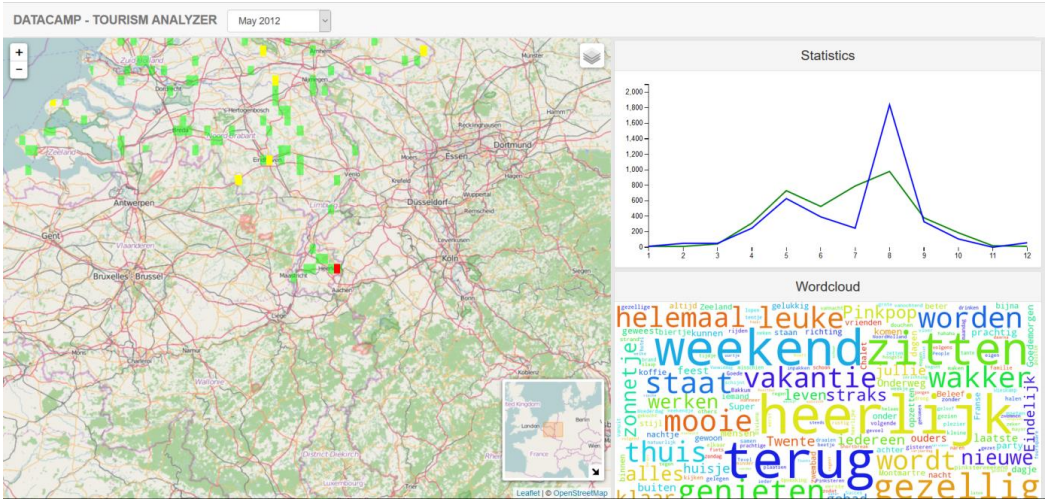

Figure 7 Location and frequency of camping word containing tweets for the South of the Netherlands in 2012. The top and botto right inserts are as described in the previous figure.

In the top right of the visualization, the change in the number of Tweets is shown for the month of the year and the number of visitors according to the official statistics. Both plots reveal a similar pattern, with a small peak in the month of May (May holiday) and a large one in August (summer holiday).

In the bottom right a word cloud is displayed based on the frequency of the words included in the Tweets selected. This reveals the topics people are tweeting about; obviously the words used to select the tweets, such as 'tent' and 'camping', were excluded. Camping-tweets most often contain the word 'cosy' ('gezellig'). Many of

the words included have a positive meaning. The word ‘shoarma’ is also often included.

Conclusion
The results obtained lead to two important conclusions. The first is that -during 2012- the Tweets displayed a similar pattern over time as the official statistical findings. The second is that the Tweets enable a much higher spatial granularity compared to the official statistical findings.
During the DataCamp only data for 2012 were studied. At a later date, Irene and Gustavo will be checking the data for the period 2013-2015. It will be of interest to see of the findings described above are corroborated by the data covering those years.

2.6 Social media usage

By: Rogier van Berlo (SN) and Jacqueline van Beuningen (SN)

Research question
Aim of the research was gaining insight in the usability of social media messages, such as those collected by twiqs.nl and Coosto, to supplement regular statistics production. More specifically, interest focussed on the availability of information relevant for the ICT survey for persons and households which includes questions on whether people use social media platforms such as Twitter or Facebook. No further questions on the extent of usage of social media are included.
The research question is composed of three parts:

- Can a Hadoop cluster be used to monitor social media messages?
- Can groups be discerned by using classification methods?
  - At the population level (all inhabitants of the Netherlands)
  - To discern various parts of the population (such as males and females)
- Can social media data be used to monitor the impact of various (internet related) events on society?

For all statistics on the Dutch population it is essential that inhabitants of the Netherlands can be identified with adequate accuracy. Three considerations are import for this:

- Selecting messages from Dutch inhabitants: language ‘NL’ to select messages written in the Dutch language (including Flemish people and the Dutch in the Caribbean islands and excluding any Dutch inhabitants that have written non-Dutch messages), location in their profile (self-reported, not always correct or very specific), GPS-location from which messages are sent (only available for a small part of the messages). Every method has its pros and cons and none is sufficient by itself.
  For the research described here, we used the language setting to select messages in Dutch. This method is not exhaustive since 1) only the language filter is used and not necessarily all Dutch tweets are included and 2) accounts from

non-Dutch inhabitants are included. It is likely that a combination of the above mentioned methods produces the best results, for example a selection of messages for which the language is 'NL' excluding any accounts with a location outside the Netherlands in their profile.

- Selecting profiles of actual persons: social media also includes messages created by companies or other kinds of profiles, such as bots (which for instance post weather forecasts). For official statistics it is essential to discern between accounts of persons and companies on the one hand and other kinds of profiles on the other hand.

- Converting messages to unique persons: the number of messages posted per person fluctuates tremendously. This indicates that persons posting large amounts of messages may more strongly affect the findings than persons that post few messages. Apart from that, some persons are active on multiple platforms or with several accounts on a single platform. Currently it is assumed that we are dealing with unique individuals in each of these cases. Any overlap within or between these platforms should be determined to prevent people from being included more than once (this is a complex task). Also, only one message could be selected per person for analysis. It depends on the statistic and its foreseen use which choice has to be made. For instance, for statistics on social media usage any selection may not be needed, while for statistics based on sentiment it might be.

Sources

The following sources were used:

- Twiqs.nl (www.twiqs.nl): a selection of Twitter messages posted between 2010 and 2015
- Coosto: all public social media messages posted on various platforms from 2009 onwards.

Approach followed

Qualitative research methods were predominantly used. These were partly quantitatively validated. Since the work performed is predominantly of an explorative nature, any methods developed will need to be further refined.
As a first step, the use of social media was mapped by studying twiqs.nl data on the Hadoop cluster. Spark-SQL was used in combination with python. The plots were produced with the ggplot package in R.

Subsequent studies were performed with search queries in Coosto. Here, words specific for the profiles of companies were investigated in an attempt to discern profiles of companies and persons. Apart from that, search queries were used to check if and what topics are discussed and what words are used by males and females on social media. Finally, it was checked if the impact of internet failures on society can be measured.

Results

The use of Twitter can be determined by analysing the creation of messages included in the twiqs.nl data set on the cluster. Figure 8 shows the temporal pattern in social media activity per hour and day of the week for week days. It reveals that between 17:00 and 20:00 hours messages are most often send with the exception of Friday. During that day the peak of major activity is earlier on the day. At night hardly any one tweets. Weekend days reveal another pattern: people start later and end earlier, resulting in an overall pattern in which fewer messages are sent compared to the number produced during week days.

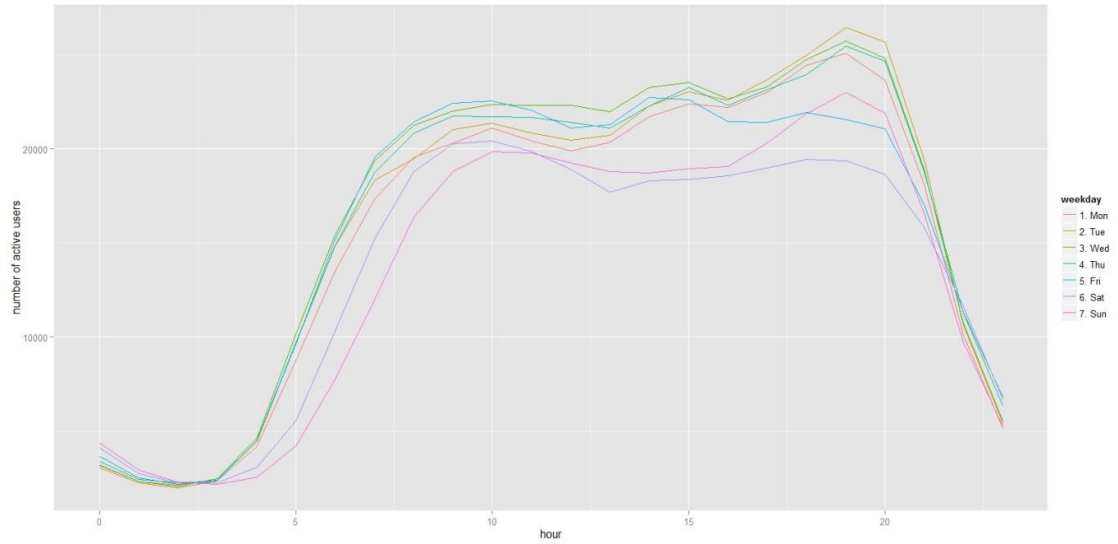

Figure 8. Dutch tweets per hour per day of the week in 2015

Develop classification methods

It is unclear why fewer messages are sent during the weekend as people in general have more time available during those days (with the exception of students with weekend jobs). A possible explanation might be the difference in the behaviour on Twitter between persons and companies. This indicates that it is important to discern between the profiles of these types. This urged the need to create a set of words that can be used to identify company profiles. This was done by taking a sample of 1000 profiles that were manually classified as persons, companies or others. The most often used and most distinctive words included in profiles of companies were determined. This resulted in a conservative method describing companies. After removal of stop words and other non-informative words, the following set of words remained to discern company profiles from those of persons: web care (webcare), company (bedrijf), organization (organisatie), office (bureau), we/wij (we), us (ons), ours (onze), you/your (u/uw), work days (werkdagen), Mon / Fri (ma t/m vrij), opened (geopend), official (officieel), questions (vragen), services (diensten), complaints (klachten) and 'find here' ("vind je hier"). This word set was verified with the word cloud from a second sample of company profiles which were previously coded. Several words overlapped, but not all. The results are shown as a world cloud in figure 9. The decision was made to focus on the words that were particularly relevant for companies.

Figure 9. Word cloud of the most often occurring words in company profiles on Twitter (in cooperation with Piet Daas).

Using this selection of words, the query shown in Figure 8 was repeated. It was found that spaces around the words from the selection needed to be included in the queries to correctly identify words in the profiles. In Figure 10 the differences between profiles identified as companies and persons are shown. Companies (in red) are more active on Twitter in the early evening between 17:00 and 20:00 hours. A possible explanation for this is the fact that companies feel the urge to react to questions or complaints of consumers posted at the end of the working day. As a final step, a selection of profiles selected by these criteria was manually checked. It was found that a very large part of the profiles selected indeed originated from companies. The method needs to be refined. Although the method correctly identified companies, it did not classify all company profiles as companies. That is, based on the selection not many people were in the company group, but still many companies were in the people group. It will be impossible to clean these groups completely due to for example sole proprietors. What matters is to be able to correctly discern different patterns of results for these groups as is clear from figure 10.

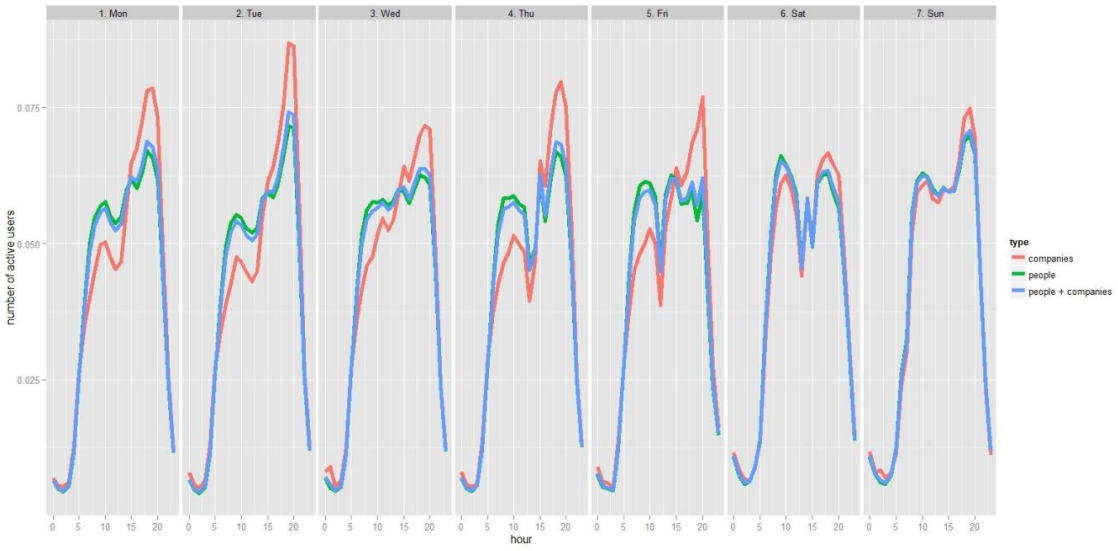

Figure 10. Dutch tweets per hour according to the day of the week for companies (red), persons (green) and both (blue) in 2015

We also looked at differences between men and women. A method has been developed at Statistics Netherlands that determines the gender of Twitter users based on the combination of first name, profile content and profile picture. This method could be expanded by including the content of the messages posted. Despite the fact that man and woman discuss different topics on social media, it is to be expected that these topics change over time making to topics on their own not very useful. However, compared to man, woman will more likely use words describing emotions. This difference may form the basis for a more stable and useful form of classification. This still has to be validated. In addition, men and women may use social media platforms differently: e.g. men are more active on Twitter and women are more active on Facebook. Hence, postings from these platforms are more likely to come from men or women respectively. Based on the combination of different types of information such as profile, name, picture and text information a predictive model to determine gender can be developed. The cut-off point for a satisfactory model still needs to be determined.

Measuring the impact of internet failures

The direct impact of any given failure or break-down on the internet is currently not measures in the ICT survey. This survey relates to an extended period of time (i.e., questions are asked regarding the previous 12 months) and needs to be planned well in advance. For example, respondents are asked to report any problems encountered in the last 12 months which induces memory effects. No details on any specific issues can be asked. Using social media we can see spikes in messages following internet failures such as online banking from a major bank being down on any given day. Here it was found that the effect of problems with online banking could be related to social media messages. It was found that during days on which online banking problems occurred a peak in the number of messages on this topic occurred. This was partly caused by messages produced by the bank where the banking problem occurred. However, when messages produced by companies were removed from these messages, one was able to study messages at the regional level and, in this way, determine the area where most people suffered from online banking problems. A similar approach could also be applied to study other events and sentiment related measurements (for instance attitude related questions in the Statistics Netherlands surveys). Word clouds of word frequencies or the number of hashtaged words could assist the interpretation of the sentiment in these studies. It is essential to cover the target population as correctly as possible.

Conclusion and discussion

The results described above reveal that it is possible to monitor activity on social media and extract statistical information from it. However, it should be noted that the classification methods used were not yet sufficiently developed. The research described here indicates that there are leads to improve on such methods. All analyses were performed on Twitter data but can be expanded to other social media platforms for which data are available. It is important to notice that (user) information (such as profile information) available may differ for each platform and may require adjusting some of the methods developed.

It has been demonstrated that methods can be developed to discern or predict profiles of companies, persons and others. An initial set of words has been developed that requires some additional validation and fine tuning. This method could additionally be expanded by including information provided by names and profile pictures (e.g. a picture with no face increases the chance that it is a company profile). The determination of individual characteristics may also benefit from this work.
It is important to notice that a classification method based on profiling or use of specific language characteristics may never be 100% accurate. However, this does not have to be a huge problem for Big Data based analysis as methods with a low accuracy may provide sufficient information to discern groups from the trends and patterns obtained. More research is needed to demonstrate the level of accuracy needed to provide reliable and stable results. It is to be expected that in the near future statistics can be produced on the use of social networks and social media in general for the Dutch population. In addition, in the meantime Coosto data on social media has already been used in combination with survey data in the book ICT, Knowledge and Economy (in Dutch "ICT, kennis en economie") 2016 (p.97) to show that most people on social media mostly read posts rather than post messages themselves.

2.7 Tourism in Tweets by Text Mining

By: Mena B. Habib (UT-EWI) and Nynke Krol (SN)

Research question
Analysis of non-Dutch tweets originating from within the Netherlands. Which locations are visited by tourists in the Netherlands?, both at the city and attraction level. And how do they move across the Netherlands? After some initial work the topic was adjusted to: Dutch locations were English tweets referred to.

Used sources

Data

- Dutch Twitter data for the year 2014, collected by Erik Tjong Kim Sang from the website www.twiqs.nl.
- GeoNames, geographical database on www.geonames.org

Software

- Scala API for Spark
- Java, in particular the Stanford Named Entity Recognizer (NER)
- R and various packages

Approach used
A selection of tweets from the twiqs.nl data set was used, namely all non-Dutch tweets. This selection was performed using Spark (Scala). The selected data set (in JSON format) was further analysed via an algorithm implemented in Java; the Stanford NER text mining algorithm. From the subset of Tweets written in the English

language, locations were selected through text mining. Based on the structure of the texts, any location referred to was selected.

The locations were subsequently linked to a set of Dutch location names in the GeoNames database. Apart from the extracted geolocation this provided some additional information, such as the type of location (park, zoo etc.). Next, monthly aggregates were determined per geolocation. R was used to remove some inconsistencies and to produce a number of visualizations. Final products were 3 plots (gif files). A bubble plot showing the monthly number of English tweets referring to specific locations in the Netherlands, a plot specifically tailored to parks and a plot focused on locations in Amsterdam. In the latter plot the locations were plotted as texts.

Results

The vast majority of the English Tweets refer to Amsterdam. Zooming in greatly reduces the frequencies of the locations. Determining the exact coordinates of a location name was not always trivial: does a tweet that mentions Beerse/Beerze refer to Beerse in Belgium or the park Beerze in Twente?

Visualizations

As described above, below plots are shown that indicate the number of English Tweets referring to locations in the Netherlands (Fig 11) and locations in Amsterdam (Fig 12).

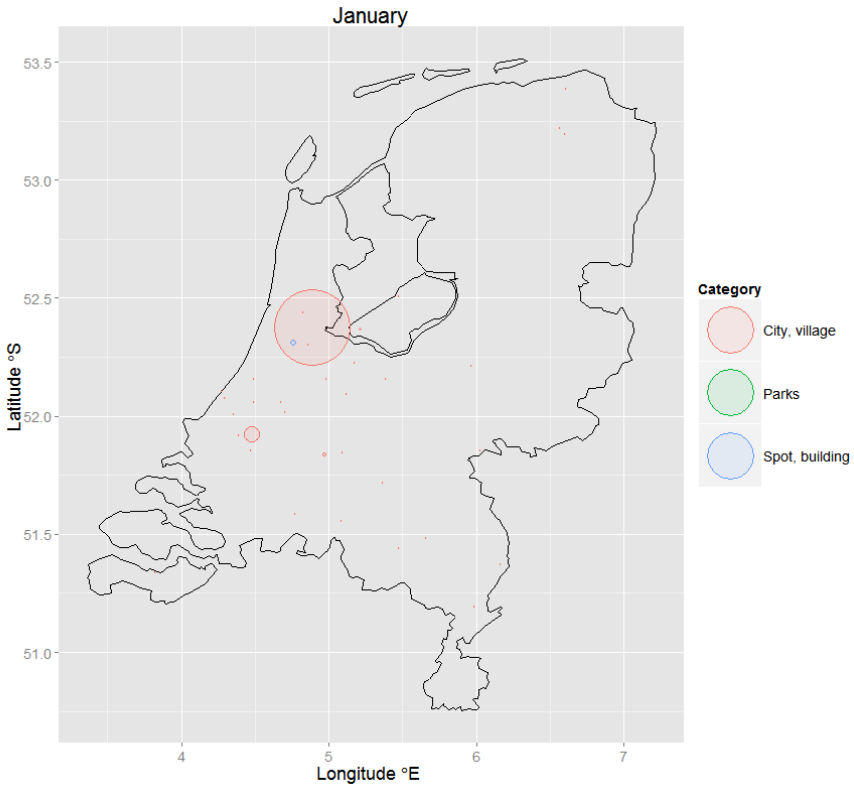

Figure 11. English tweets referring to locations in the Netherlands

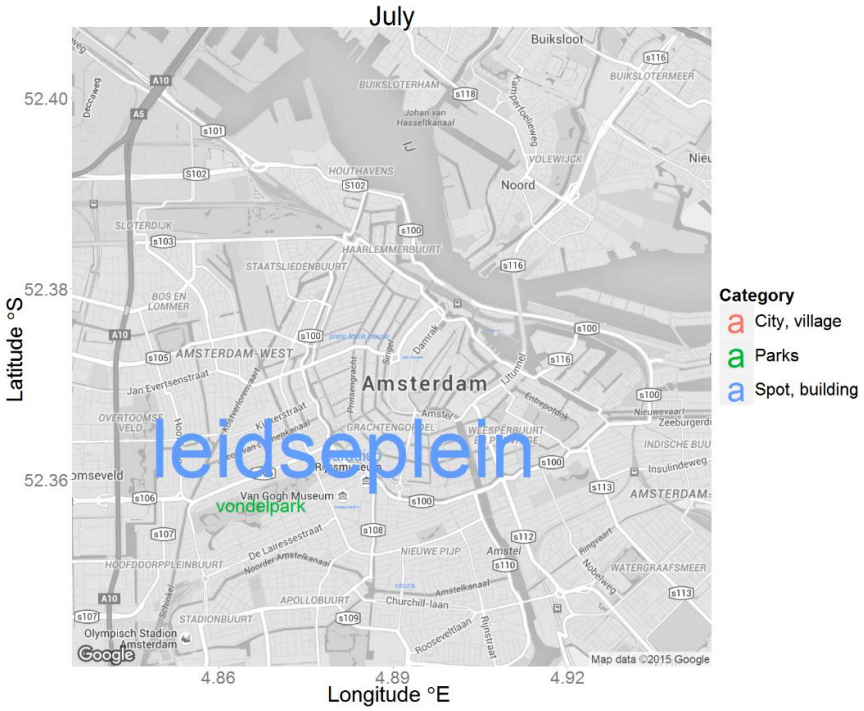

Figure 12. English tweets referring to locations in Amsterdam

2.8 Relation between economic growth and traffic intensity

By Ronald van der Stegen (SN) and Dan Ionita (UT-EWI)

Research question

Is there a relation between economic growth and changes in traffic intensity? This can be checked in a number of ways:

- Macro-level:
  - What is the relation between traffic intensity over a longer period of time (for instance quarterly data) with the economic growth in the Netherlands?
- Detailed over time:
  - Can traffic intensity be used to more accurately obtain working day patterns, thereby improving the correction for working days and improving the estimation of economics growth?
  - Is there a relation between experimental monthly economic growth and traffic intensity?
- Detailed according to region:
  - Is there a relation between regional economic growth and regional changes in traffic intensity?

Data sources

Two data source were available for this work.

- Data from national accounts as published on the CBS –website. The majority of the data is available from 1996 onwards.

- Traffic loop data from NDW in 2 versions. The selected detailed data and corrected daily aggregates per road sensor on the Dutch highways. The data does not discern between various vehicle length categories and is available for the period 2010-2014.

Data processing

The work started with the aggregated data set. The major issue with traffic loop data is that the data is not available for all days: not all sensors provide data on a daily basis. This was solved by calculating the average number of vehicles registered by the sensors on a particular highway. Subsequently, the highways were selected that provided data for more than 95% of the days. Especially in the first two quarters of 2010, this was a problem.

Next data was aggregated per month or quarter. To enable a proper comparison of these aggregates, seasonal correction was performed. Advantage of using monthly data is that these can be aggregated further in various ways; for instance creating quarterly aggregates starting at month 2 etc. This may affect the correlation. Other factors, such as the weather, may disturb the relation. The original detailed (and not aggregated) data was not used in this study because of time constraints. Advantage of the aggregated dataset is that a lot of calculations could be performed on a standard laptop in R.

Results

The research revealed the following findings:

- Macro-level: GDP growth correlated very well with an increase in traffic intensity in the previous quarter. A correlation coefficient between 80 and 90% was found, depending on the period selected (Figure 13).

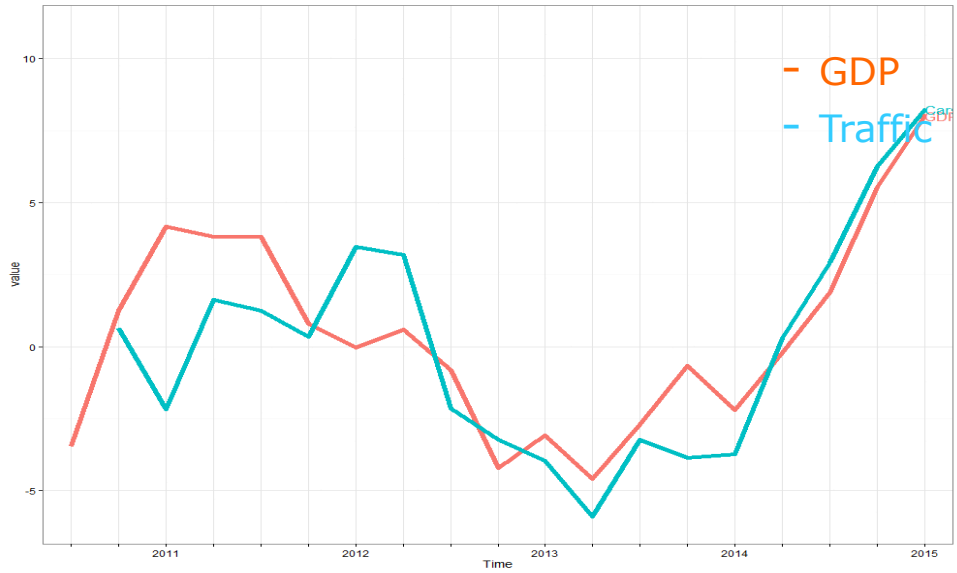

Figure 13. Relation between quarterly Traffic intensity and GDP

- Details over time: An effect caused by weather and working days is observed in the data. This seriously reduces the correlation between monthly GDP and monthly traffic data. Because of time constraints, this topic was not investigated further.
- Details over regions: Because the geo-coordinates are known for all road sensors, the data can be plotted on a map (Figure 14).

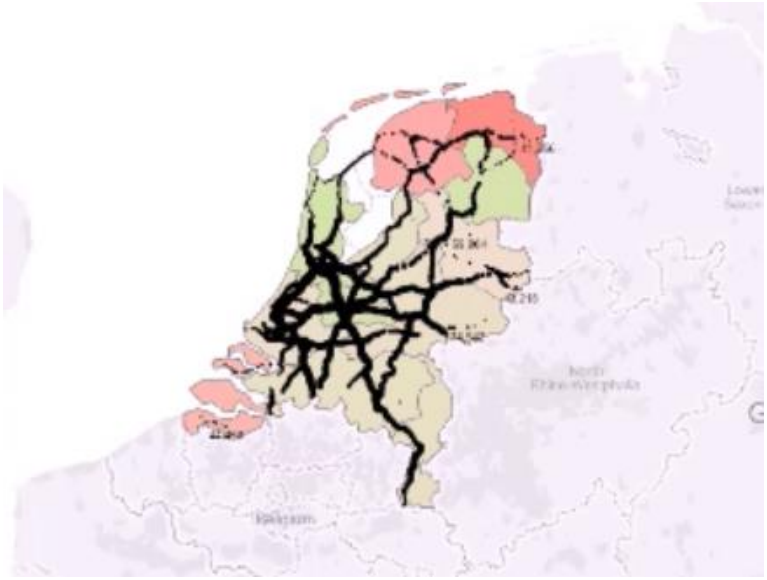

Figure 14. Plot of changes in regional GDP in the Netherlands and traffic intensity on the highways in these regions.

2.9 Website tells it all: we know your business

By: Marko Roos (SN) and Mohammad Khelghati (UT-EWI)

Research question
Can the information on a website of a company be used to automatically determine the Standard Industrial Classification (SIC, in Dutch: SBI) code for that company?

Data sources
- List of company names and associated websites from Dataprovider
- List of Standard Industrial Classification (SIC) and description of activities Included
- Content of websites from companies (a subset of the Dataprovider list)
- List of Chamber of Commerce (CoC) numbers for Dutch companies

Data processing
The list provided by Dataprovider contained a total of 2 million websites. Of these websites 500.000 included a Chamber of Commerce (CoC) number. This was essential to enable the identification of a particular company as the owner of the website. A total of 470.000 websites had, in addition to the CoC-number, also a zip code

assigned to the web site. The latter enabled linking the company to a specific area in the Netherlands. This formed the basis of the first quality checks that were performed. Comparing the number of unique 3-digit (1145) and 2-digit (100) SIC-codes in the dataset revealed that assigning 2-digit SIC-codes was a good starting point for the work performed during the DataCamp.

Results

First the 470.000 websites with a CoC-number and Post Code were plotted on maps of the Netherlands. For each 2-digit SIC-code, a separate plot was created revealing the distribution of companies over the Netherlands. The results are shown in Figure 15. Clearly for some 2-digit SIC-codes large numbers of companies are included in the data set of Data provider; for example codes 46, 47 and 93. This suggests that - certainly in those cases- sufficient data is available in the Dataprovider set to already produce potential interesting statistics.

Next it was checked how well SIC-codes were assigned to companies by comparing the relation between SIC-codes and CoC-numbers in the Dataprovider data set and the combination of those data available at Statistics Netherlands. This check was performed in the secure environment of Statistics Netherlands and revealed huge differences between various 2-digit SIC-codes. Figure 16 shows that SIC-code 46 and 47 matched very well, while SIC-code 93 did clearly not.

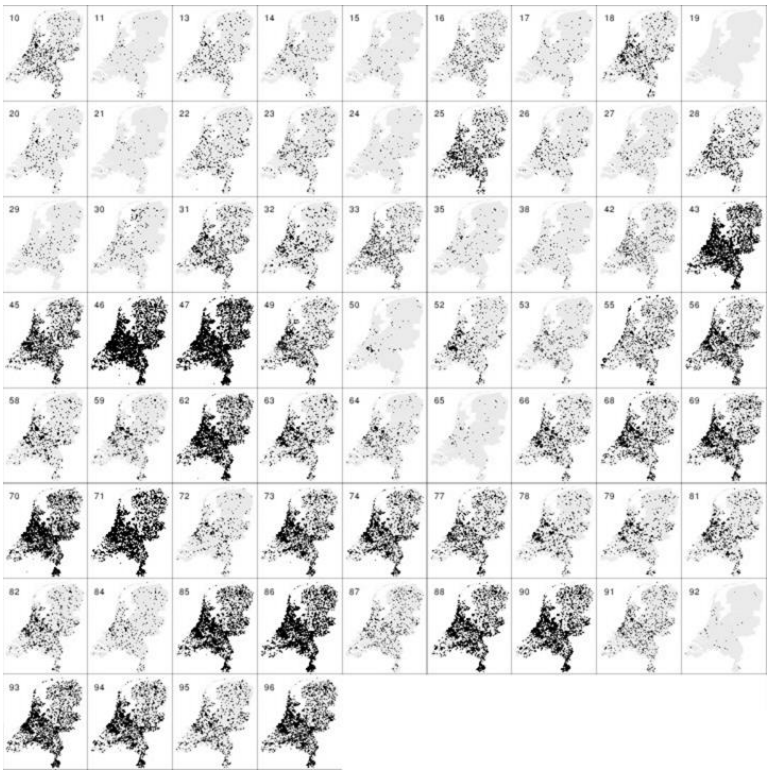

Figure 15. Plots of companies per 2-digit SIC-codes and their location in the Netherlands.

The subsequent step was the development of a classifier that is able to derive the correct SIC-code from the text and keywords provided by the web site data of a company. For this a training set of companies is needed to which the text and keywords on the website and the (100%) correct SIC-codes are assigned. This proofed difficult as no such set was available during the DataCamp. Since construction of such a set is time consuming, it was decided to use the correctly assigned data as a training set to test the principle of the approach. A naïve Bayes classifier was trained with the data available. Results were promising as shown in Figure 17, with the exception of some SIC-code such as 95. Downside was also that many abbreviations occurring on web sites of companies were used as is and were not replaced by the original wordings. Creating a 100% correct training set should be the start of future studies.

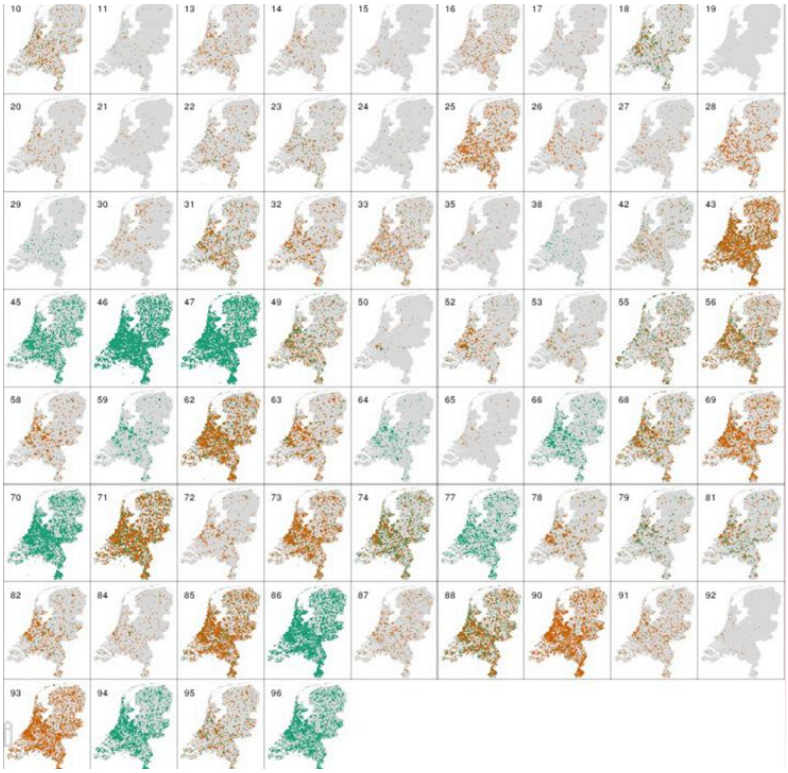

Figure 16. Plots of companies per 2-digit SIC-codes, their location in the Netherlands and the result of the comparison between Chamber of Commerce and Statistics Netherlands assigned codes. Green indicates identical results, orange indicates non-identical results, and black indicates no match was found.

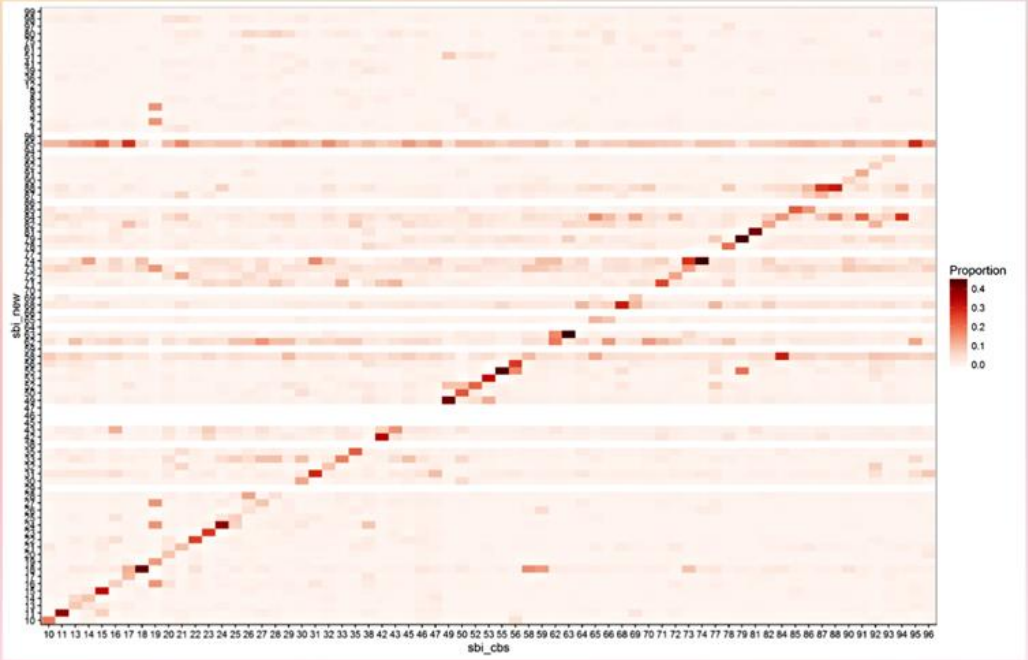

Figure 17. Portion of correct SIC-codes assigned by to companies by Statistics Netherlands and the Chamber of Commerce data. The straight line between both codes indicates that a large part is assigned correctly. The straight line at SIC code 95 on the y-axes, however, reveals a serious issue.

3. Discussion

After the DataCamp was finished and evaluated the following observations were made:

- It is important to start with a more theoretical introduction followed by a practical approach as the latter benefits from the change in mindset introduced by the former (which is needed for Big Data analysis).
- It really helps to learn new things when full attention can be paid to such tasks during an undisturbed period.
- Many people stated that they quite easily picked-up analysing Big Data on the Twente infrastructure. SPARK, especially Spark SQL, Pig, Python, R and Tableau were found to be unbelievably helpful.
- It is important to filter Big Data, i.e. extracting information relevant for the research question at hand. Creating visualizations really helps in this process.
- Participants were surprised about the interesting findings many Big Data sources provided. A considerable part of these findings were unexpected.
- The participants appreciated that the DataCamp did not focus on the Big Data hype but on proper down-to-earth data analysis. They learned a lot from each other.
- The application oriented research questions were very much appreciated as was the goal-driven way of working during the camp.
- The main goal of the camp was learning, but many results were received enthusiastically and considered worth following up.
- All Statistics Netherlands participants have become Big Data ambassadors in their own departments.
- The DataCamp itself was covered on the corporate news site of Statistics Netherlands and was later followed by the topic ‘on at the arrival of spring’.
- In a considerable number of Big Data presentations by Statistics Netherlands examples of DataCamp findings have been included.

The enthusiasm of the participants, the lessons learned and the interesting results all support the idea to continue with the DataCamp. This will indeed be the case, the next DataCamp will be held from the 6th until the 9th of December 2016 at the same location as the previous camp. The group of participants will be expanded though. From Statistics Netherlands not only people from the Statistical divisions may participate but also employees of the Process development, IT and Methodology division can join. In addition, PhD-students from other Universities may also participate. Because it was found that the camp put a considerable burden on the senior experts involved, ways to include previous participants in the new DataCamp are currently being discussed.

Acknowledgements

The authors gratefully acknowledge the support and expertise from Marco Puts, Martijn Tennekes, Erik Tjong Kim Sang, David González and Francisca Gromme to the DataCamp. Their contribution greatly increased the quality of the research performed and findings obtained. Joep Burger, Bart Buelens, Peter Struijs and Edwin de Jonge are gratefully acknowledged for their contribution to the Masterclass Big Data. We also like to thank the DataCamp participants Rogier van Berlo, Jacqueline van Beuningen, Kenneth Chin-A-Fat, Egbert Hardeman, Maaike Hersevoort, Nynke Krol, Maarten Pouwels, Marko Roos, Ronald van der Stegen (from Statistics Netherlands) and Mitra Baratchi, Mena B. Habib, Dan Ionita, Mohammad Khelghat, Hamed Mehdipoor, Anna Priante and Jair Santanna (from University of Twente) for their contribution, enthusiasm and hard work. The Big Data coordinating group of Statistics Netherlands is gratefully acknowledged for their approval to organize a Big Data Masterclass and DataCamp.

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2015–2016	2015 to 2016 inclusive
2015/2016	Average for 2015 to 2016 inclusive
2015/'16	Crop year, financial year, school year, etc., beginning in 2015 and ending in 2016
2013/'14–2015/'16	Crop year, financial year, etc., 2013/'14 to 2015/'16 inclusive

Due to rounding, some totals may not correspond to the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands, Studio BCO

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70
Via contact form: www.cbsl.nl/information