Project Text Mining and Collective Intelligence

“*Title*”

Hester Bartelsman, Eddie Stok, Jana Wagemaker, Leonard Wein

Manos Tsagkias

Text Mining and Collective Intelligence

Amsterdam University College

May 2013

**Introduction**

The inauguration of the King the 30th of April this year, for all of us the first abdication and coronation we would witness in person, seemed a fitting topic for our final project for Text Mining and Collective Intelligence; it was a big event happening in our city, and, as we found out, alive all over the world.

No More Inspiration

**Word Clouds**

Word clouds are visual representations of text data; tweets in our case. The importance of each word is based on the number of times it occurred in all the tweets crawled on Queens day and is shown in the word clouds by font size. The colour and position is random.

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



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



As you can see, this one reveals quite a bit more details than the previous one. We can see that the words are mainly English, which is interesting, since most of the key words we selected the tweets on were in Dutch. ‘Orange’ occurred 322007 times, where ‘oranje’ occurred only 69124 times. This could suggest that Dutch people prefer tweeting in English. It could also suggest that the few English hash tags we used for selecting tweets brought in a lot of English tweets not necessarily related to the coronation. EDDIE, maybe you could add some stuff here, because I don’t know how many of the tweets were from the Netherlands or from English speaking countries, but you do of course!

This next word cloud is one of all key terms we used for crawling all the tweets. Since ‘Amsterdam’ occurs 642855 times and the word next in line (‘koning’) occurs 102501 times, this word cloud has pretty odd proportions. For this purpose, there is another word cloud below it with all key terms except ‘Amsterdam’. As you can see, the results are quite different.





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

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





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