

Twiqs.nl: Zoeken in miljarden Nederlandstalige tweets Erik Tjong Kim Sang Meertens Instituut 19 februari 2014	Twiqs.nl: Zoeken in miljarden Nederlandstalige tweets 19 februari 2014 Erik Tjong Kim Sang • gestudeerd in Delft, gepromoveerd in Groningen • computertaalkundige, gespecialiseerd in machine learning • werkt sinds 2010 met socialmediadata • 2011: voorspelling provinciale verkiezingen met Twitter • 2012-2013: website twiqs.nl gebouwd voor onderzoek met Twitter
Twiqs.nl: Zoeken in miljarden Nederlandstalige tweets 19 februari 2014 Twitter Twitter is een microblogwebsite gestart in 2006 door het Amerikaanse bedrijf Obvious (Jack Dorsey) Gebruikers zenden berichten van maximaal 140 tekens uit Gebruikers kunnen zich abonneren op de berichten van andere gebruikers: dit heet "iemand volgen" Anders dan op Facebook zijn relaties op Twitter asymmetrisch: je kan iemand volgen maar die hoeft jou niet te volgen De meestgevolgde gebruikers staan op twitaholic.com/	Twiqs.nl: Zoeken in miljarden Nederlandstalige tweets 19 februari 2014 Hoe ziet een Twitterbericht eruit?  SNiC Symposium @SNiC_Symposium Volgen Wat is big data en hoe gebruiken bedrijven dit? Deze video geeft antwoorden op deze vragen. Heel interessant! bit.ly/1jJhxcn Beantwoorden Retweeten Favoriet Meer
Twiqs.nl: Zoeken in miljarden Nederlandstalige tweets 19 februari 2014 Achter de schermen (1) <pre>{ "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", "text": "Wat is big data en hoe gebruiken bedrijven dit? Deze video geeft antwoorden op deze vragen. Heel interessant! \nhttp://t.co/UNl4CTqCj", "contributors": null, "geo": null, "entities": {"symbols": [], "urls": [{"expanded_url": "http://bit.ly/1jJhxcn", "indices": [[11, 133], "display_url": "bit.ly/1jJhxcn", "url": "http://t.co/UNl4CTqCj"}]}, "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, "favorites_count": 0, "protected": false, "profile_text_color": "333333", "description": "Dit jaar wordt er een symposium georganiseerd over BigData! Kijk op http://www.exe-it.nl voor meer informatie.", "verified": false, "contributors_enabled": false, "profile_sidebart_color": "CODERED", "name": "SNiC Symposium", "profile_background_color": "CODERED", "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/37880000605291818/21c5331df3b98396bac82d58dede7_normal.jpeg", "geo_enabled": false, "profile_background_image_url": "http://abs.twimg.com/images/themes/theme1/bg.png", "profile_background_</pre>	Twiqs.nl: Zoeken in miljarden Nederlandstalige tweets 19 februari 2014 Achter de schermen (2) <pre>{ "lang": "nl", "created_at": "Wed Nov 27 10:26:49 +0000 2013", "text": "Wat is big data en hoe gebruiken bedrijven dit? Deze video geeft antwoorden op deze vragen. Heel interessant! \nhttp://t.co/UNl4CTqCj", "hashtags": [], "id_str": "405643870785904640", "user": { "name": "SNiC Symposium", "followers_count": 49, ... } }</pre>
Twiqs.nl: Zoeken in miljarden Nederlandstalige tweets 19 februari 2014 Onderzoek doen met tweets Steeds meer studenten en onderzoekers willen tweets bestuderen In tweets wordt een grote variatie aan onderwerpen besproken. Ze zijn geschreven in vele talen en dialecten Het ontbreekt onderzoekers en studenten vaak aan de technische kennis of tweets te verzamelen en te analyseren Het Netherlands eScience Center heeft daarom de website twiqs.nl ontwikkeld ter ondersteuning van onderzoek naar Nederlandstalige tweets	Twiqs.nl: Zoeken in miljarden Nederlandstalige tweets 19 februari 2014 twiqs.nl Met Twiqs.nl kan worden gezocht naar Nederlandstalige tweets in de periode december 2010 tot nu. Tweets kunnen worden geselecteerd op basis van steekwoorden of aan de hand van metadata, bijvoorbeeld gebruikersleeftijd De zoekresultaten bevatten geen tweets maar samenvattingen, bijvoorbeeld frequenties, kaarten, verwante woorden en gebruikersinformatie De tweetcollectie is niet compleet: zij bevat ongeveer 40% van de Nederlandstalige tweets geschreven sinds december 2010

Hoe worden de tweets verzameld?

We gebruiken de Twitter Streaming API om continu te zoeken naar nieuwe tweets die een van 229 veelvoorkomende Nederlandse woorden bevatten

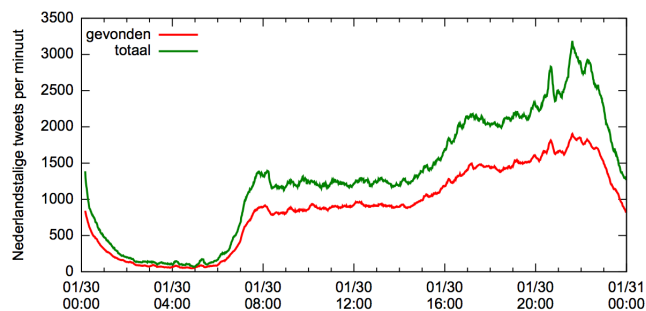
Hiermee verzamelen we 2 miljoen tweets per dag van welke 1,3 miljoen als Nederlands worden herkend door een taalrader

De tweets worden opgeslagen in gecomprimeerde bestanden die elk 1 uur bevatten aan tweets inclusief de metadata (in JSON-formaat)

De collectie bevat op dit moment ongeveer 1,3 terabytes (2,0 miljard tweets)

8

Gevonden 1.279.263 van 1.869.116 tweets (68%).



De uitdaging: zoeken in de tweetverzameling

Vraag: Hoe kunnen we tweets vinden in deze grote verzameling?

Antwoord: gebruik een informationretrievalstelsysteem

Maar...

- we hebben echt veel data
- er komen elke seconde nieuwe tweets bij
- we willen zoeken in zowel data als in metadata

10

Oplossing: gebruik een computercluster (cloud)

We gebruiken een computercluster om zoekopdrachten uit te voeren

Iedere zoekopdracht wordt gesplitst in verschillende opdrachten die elk naar tweets zoeken in een bestand van 1 uur

De deelopdrachten kunnen dan parallel worden uitgevoerd

De zoeksoftware is geschreven voor systeem Hadoop dat gebaseerd is op MapReduce van Google

Het draait op het experimentele Hadoopcluster van SURFsara (90 nodes; 630 Tb), dat overigens voor alle Nederlandse onderzoekers beschikbaar is

11

Zoeken met MapReduce: Maptaak

De maptaak selecteert de relevante tweets per uurbestand:

```
voor (iedere tweet in databestand) {
  als (zoekopdracht matcht tweet) {
    verwerk tweet
    plaats resultaten op de uitvoer per datasleutel
  }
}
```

Door de zoeksoftware zelf te schrijven (in Java), hebben we de volledige controle over het zoekproces en het verwerkingsproces.

12

Voorbeeld Reducetaak

De Reducetaak combineert de resultaten van de verschillende Maptaken:

```
// verwerk de resultaten per datasleutel (minuut)
combineer Mapresultaten van de huidige datasleutel
plaats resultaten per datasleutel op de uitvoer
```

De Reducetaak sorteert de uitvoer van de Maptaken op basis van de datasleutels

13

Voordelen van MapReduce

- Volledige controle over zoeken en verwerken van gegevens
- Het zoeken in meerdere uren aan data is ongeveer even snel als het zoeken in 1 uur aan data

Nadelen van MapReduce

- Geen snelle resultaten zoals bij information retrieval
- Verwerking moet soms wachten op andere zoektaken

14

Science center SURF SARA FAQ Blog Naam: Wachtwoord: Inloggen Registreren

Zoek in tweets:

6 | juni | 2013 |

Op twigs.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 <erik(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, erik(at)xs4all.nl ([clusterstatus](#))

ONDERWERPEN OP TWITTER

Wanneer wordt een onderwerp besproken op Twitter?

We kunnen bepalen hoe populair een onderwerp is op Twitter in een bepaalde tijdsperiode

We tellen hoeveel tweets over een onderwerp worden verstuurd in bepaalde tijdsperiodes en plaatsen de aantallen in een grafiek

Maar 's nachts worden minder tweets verstuurd dan overdag en dat effect proberen we uit de grafieken te houden

We gebruiken daarom geen totalen in de grafiek maar *percentages*, bijvoorbeeld het percentage Nederlandstalige tweets met daarin het woord "Ireen"

17



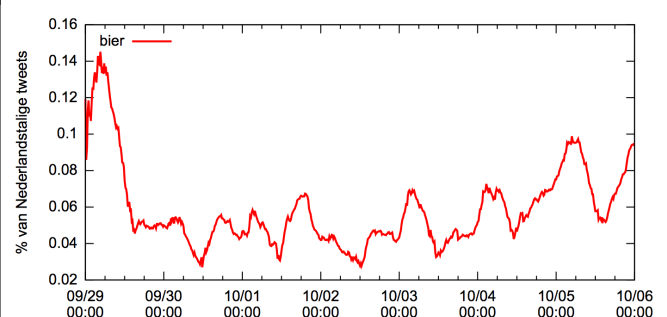
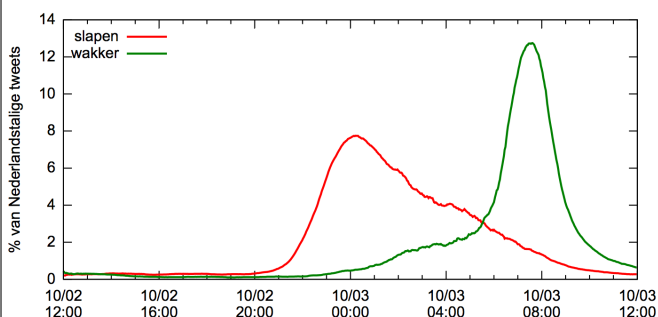
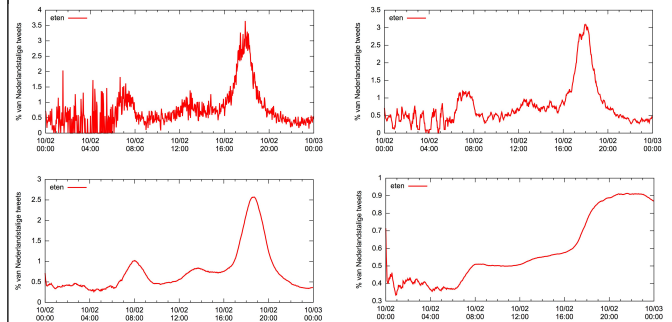
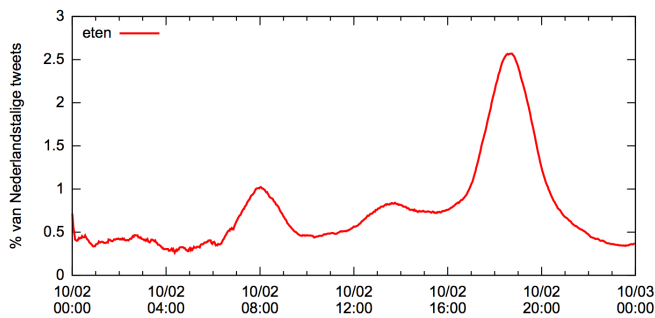
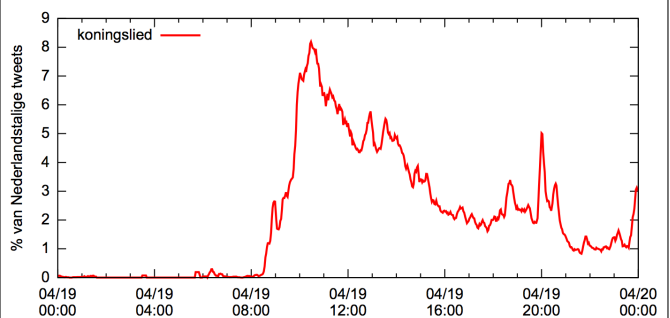
Grafieken voor dagelijks woordgebruik

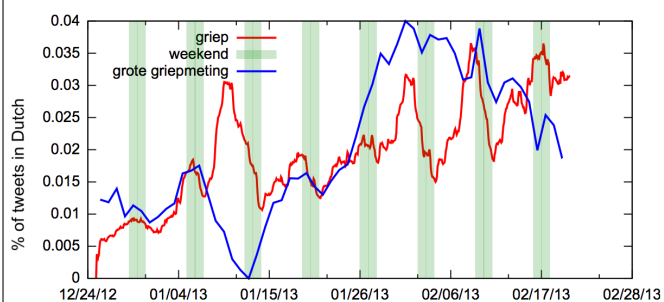
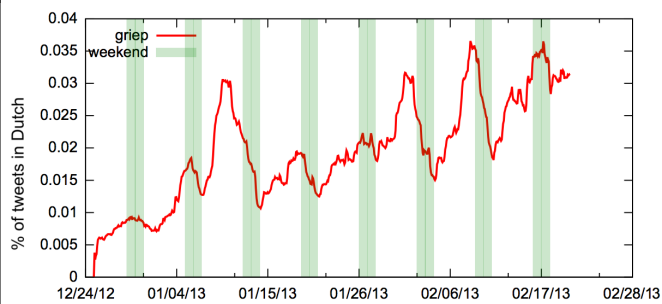
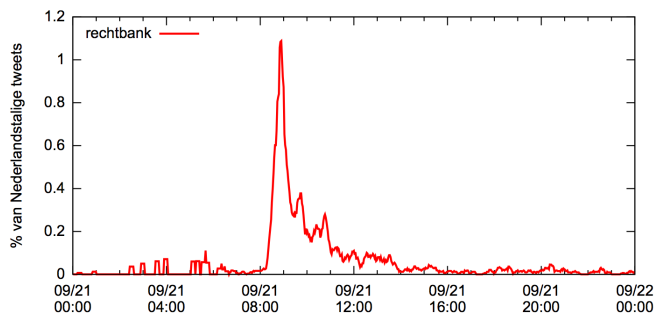
Met `twigs.nl` kan je zelf grafieken laten tekenen van het gebruik van Nederlandse woorden op Twitter

De grafiek geeft bijna live de frequenties van het gebruik van woorden in de afgelopen 24 uur

Je kan ook een grafiek van een andere dag opvragen, bijvoorbeeld voor vrijdag 19 april 2013 (presentatie Koningslied)

18





Twigs.nl: Zoeken in miljarden Nederlandstalige tweets

19 februari 2014

Samenvatting

We kunnen grafieken maken van het woordgebruik op Twitter door de tijd heen

Via de grafieken kunnen we zien wanneer welk onderwerpen populair zijn op Twitter

Daardoor krijgen we inzicht in wat de gemiddelde Twittergebruiker bezighoudt

Oneffenheden in de grafiek kunnen worden gladgestreken door de smoothfactor te verhogen

27



Twigs.nl: Zoeken in miljarden Nederlandstalige tweets

19 februari 2014

Voorspellen van verkiezingen

Zouden we de uitslag van een verkiezing kunnen voorspellen met Twitter?

Tumasjan et al (2010): ja, dit kan:

1. tel hoe vaak elke partijnaam voorkomt in een tweet
2. zet de tweetaantallen om naar zetelaantallen
3. klaar!

29



VERKIEZINGEN VOORSPELLEN

Twigs.nl: Zoeken in miljarden Nederlandstalige tweets

19 februari 2014

Voorspelling door het tellen van 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	-

30



Twigs.nl: Zoeken in miljarden Nederlandstalige tweets

19 februari 2014

Het slechte nieuws

Niemand heeft ooit de goede resultaten van Tumasjan (2010) kunnen reproduceren

Problemen met de aanpak:

- niet alle tweets over een partij zijn positief
- soms komen veel tweets van dezelfde persoon (Twittergebruiker)
- Twittergebruikers zijn geen goede representatie van de Nederlandse bevolking

31



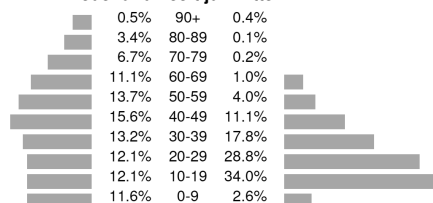
Voorbeelden van negatieve tweets

- 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.
- Mijn pappie wordt nog steeds gepest door dat **#pvv** schorem op Twitter! Morgen ga ik aangifte doen, moet toch pas 's middags werken.
- SP** heeft nog steeds niet door dat alleen dure woningen worden gebouwd, omdat dit meer geld in de Gemeentekas brengt <http://t.co/0tvpex8a>
- CDA** bedacht de langstudeerboete, voerde hem in, schafte hem af en weigert nu mee te werken aan het afschaffen. Is dat stabiel bestuur?
- 461 inwoners (stemmers **D66**) schrijven de wet voor aan de rest van **#Westvoor**ne
- De **christenunie** vind dat alloctonen kinderen niet t land uit mogen worden gezet..... BULLSHIT, OPROTTEN!

32

Leeftijden van Twittergebruikers

Nederland Leeftijd Twitter



33

Aanpak voorspellen verkiezingen

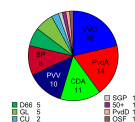
- tel aantal keren dat elke partij wordt genoemd op Twitter
- trek de negatieve tweets daarvan af
- koppel de aantallen aan een betrouwbare peiling
- herhaal stap 1+2 vlak voor de verkiezingen
- bepaal een verkiezingsvoorspelling op basis van stappen 1-4

Voorbeeld: de PvdA had 12 zetels in de voorspelling en 20% van de politieke tweets. Als de partij een week later in 25% van de tweets wordt genoemd dan komt dat overeen met $12 \cdot 25 / 20$ is 15 zetels.

34

Voorspellen van verkiezingen (Eerste Kamer 2011)

Partij	Resultaat	Zetels PB	Zetels MdH	Zetels 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



35

Voorspelling zonder gebruik van tweets

Party	Result	Seats Twitter	Population weight
VVD	16	16	2.23
PvdA	14	13	1.93
CDA	11	10	1.41
PVV	10	12	1.78
SP	8	7	1.11
D66	5	5	0.82
GL	5	4	0.59
CU	2	3	0.45
50+	1	1	0.22
SGP	1	2	0.30
PvdD	1	1	0.15
OSF	1	1	-
offset	-	8	

36

Samenvatting

Tweets helpen niet bij het voorspellen van verkiezingen (althans deze keer niet)

De demografische gegevens van gebruikers van Twitter zijn anders dan die van de bevolking van Nederland

Het is belangrijk om de prestatie van complexe computermodellen te vergelijken met eenvoudige modellen (baselines, hier: de voorspelling zonder tweets)

37

WAT IS DE GELUKKIGSTE PLEK IN NEDERLAND?

Wat is de gelukkigste plek in Nederland?

We beginnen met het concreter maken van onze onderzoeksvraag:

We willen tweets gebruiken dus kunnen we beter vragen: van welke plek in Nederland worden percentueel gezien de meeste gelukkige tweets verstuurd?

"plek" is een erg algemeen concept. We hebben gebieden nodig en kijken daarom naar provincies en gemeenten

39

Sentimentanalyse

Het bepalen of een tekst positief of negatief is over een bepaald onderwerp heet sentimentanalyse

De beste manier om dit te doen is door mensen de teksten te laten lezen

Zij kennen dan een menselijke emotie toe aan de teksten: tevreden, ontevreden, verward, boos, agressief, ...

Wij willen miljoenen teksten (tweets) analyseren en kiezen daarom voor een automatische oplossing zonder bepaling van het onderwerp

40

Automatische sentimentanalyse

We gebruiken twee woordenlijsten: een met positieve en een met negatieve woorden

Als een tweet 1 positief woord bevat is het positief, als het 1 negatief woord bevat dan is het negatief

Bevat het meerdere woorden uit de woordenlijsten dan gebruiken we alleen het laatste woord

De woorden *geen* en *niet* voor een sentimentwoord draaien de sentimentswaarde ervan om

Tweets zonder woorden uit de woordenlijsten noemen we neutraal

41

Positiebepaling

1 à 2 % van de tweets bevat een positie in de vorm van lengtegraad en breedtegraad, bijvoorbeeld: 6.568027,53.219353

Grenzen van provincies en gemeentes zijn gedefinieerd als reeksen van zulke posities

Maar hoe weten we binnen welk gebied een positie valt?

Oplossing: het programma point in polygon

42

Berekening van het sentiment van een gebied

Bepaal het sentiment van de tweets in het gebied

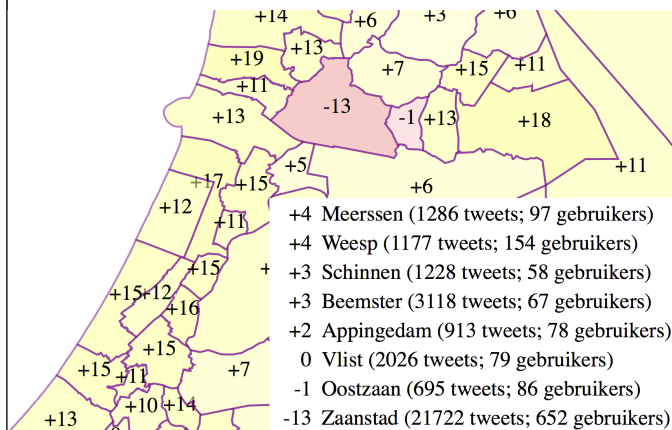
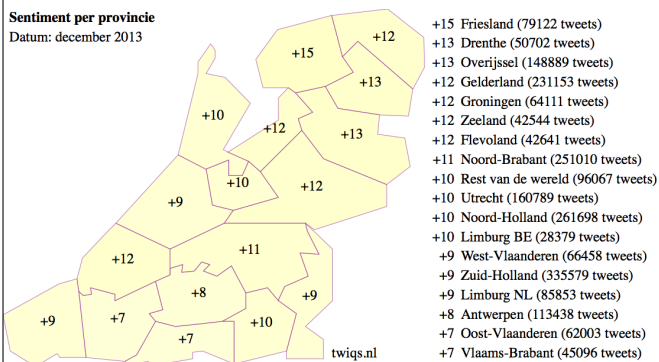
Trek het percentage negatieve tweets af van het percentage positieve tweets.

Het resultaat ligt tussen +100% (alleen maar positieve tweets) en -100% (alleen maar negatieve tweets)

Plaats de resultaten op een kaart met alle gebieden

43

Sentiment per provincie
Datum: december 2013



De meest negatieve twitteraar...



MeteoWestzaan.nl

@MeteoWestzaan



Baro 1025,0mb-Stijgt langzaam. Temp 16,8c (-0,4). Hum 44%. Rain last 24h 0,0mm. Wind 0,2kph WNW / gust 2,2kph. UV 0.

[#hetweerinwestzaan](#)

Reply Retweet Favorite More

8:00 AM - 6 Oct 13 from Zaanstad, North Holland

46

Oplossing

We gebruiken nu het gemiddelde van de sentimenten per gebruiker

Hierdoor heeft elke gebruiker evenveel invloed op het sentiment van het gebied

Resultaten december 2013:

1. +30 Ameland (FR)	+9 Uithoorn (NH)
2. +28 De Marne (GR)	413. +8 Albrandswaard (ZH)
3. +27 Ubbergen (GE)	414. +7 Renswoude (UT)
+27 Rucphen (NB)	+7 Hillegom (ZH)
5. +26 Reusel-De Mierden (NB)	+7 Beek (LI)

47

Samenvatting

Een klein percentage van tweets bevat locatie-informatie die je kan gebruiken in kaarten

Bij het bepalen van het sentiment van een groep tweets, is belangrijk om het principe *one man, one vote* te gebruiken

Nederlandstalige tweets zijn gemiddeld gezien positief!

EINDE

48

Literatuur

- 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>

50

Bonus: Animatie van Twitterdata

Pulse of the Nation is een animatie die tweets koppelt aan tijd, tweetvolume, locatie en stemming

In de animatie zie je een kaart van de Verenigde Staten waarin:

- delen groter en kleiner worden naarmate het tweetvolume toeneemt
- delen verkleuren naarmate de stemming verandert

Adres: <http://www.youtube.com/watch?v=uJcrJZRSgkg>

51