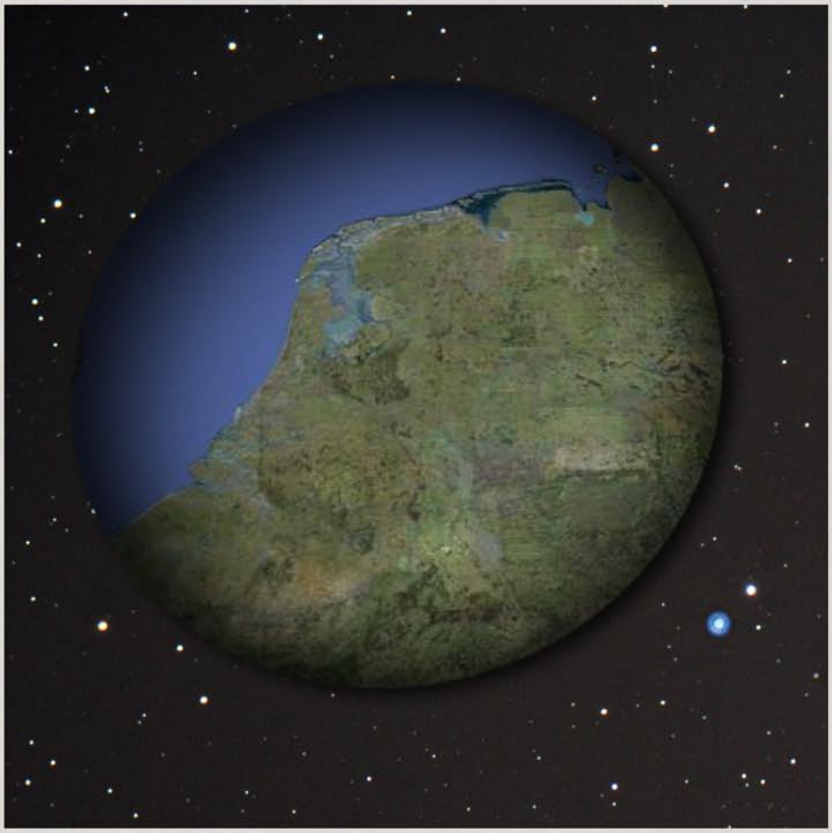
Nieuws op microblogs

Een automatisch recommender system voor Nederlandstalig nieuws op basis van Twitter



Rob van der Goot

01-07-2012

Nieuws op microblogs:

Een automatisch reccomender system voor Nederlandstalige nieuws op basis van twitter.

Rob van der Goot
s1915770
R.M.van.der.Goot@student.rug.nl

Rijksuniversiteit Groningen
Groningen, Nederland
Faculteit der Letteren
Bachelorscriptie Informatiekunde
Begeleider: Johan Bos

01-07-2012

Voorwoord

Deze scriptie is het resultaat van het onderzoek dat ik heb gedaan om mijn bachelorstudie informatiekunde aan de Rijksuniversiteit Groningen af te ronden. Het meeste werk heb ik natuurlijk zelf gedaan, maar er zijn toch een paar mensen die mij enorm hebben geholpen. Ten eerste Rinke Beimin, waarmee ik samen het begin van het programma heb geschreven. Verder waren natuurlijk alle mensen van de Rug die mij hebben geholpen om alle benodigde vaardigheden en theorieën te leren erg belangrijk. In het speciaal waren Johan Bos en Erik Tsjong Kim Sang belangrijk vanwege de begeleiding van de scriptie. Ook wil ik mijn moeder bedanken, zij heeft de paper gecontroleerd op spelfouten.

Voor de implementatie van het programma in Java was het boek Java Programming from the Beginning (King 2000) de belangrijkste bron. Verder wil ik nog de schrijvers van de software die ik heb gebruikt om mijn eigen software te maken bedanken, dit zijn de makers van Java, Eclipse, LaTeX, Ubuntu, Kate, en natuurlijk de server 'Siegfried' van de RUG.

Als laatste wil ik de mensen bedanken die het geduld konden opbrengen om 100 nieuwsberichten te annoteren. En hierbij in het speciaal Lars van der Goot, omdat hij uiteindelijk het dubbele aantal nieuwsberichten heeft geannoteerd.

Verder rest mij niks anders dan u veel plezier te wensen met het lezen van mijn scriptie.

P.S. Er is voorlopig een demo van mijn programma beschikbaar op: `http://siegfried.let.rug.nl/s1915770`

Inhoudsopgave

B **27**

Samenvatting

Op twitter staat enorm veel informatie over een breed gebied van onderwerpen door een grote groep van gebruikers. In deze paper ga ik deze informatie gebruiken om informatie/data over nieuwsberichten te vergaren. Hiervoor wordt als ruwe data een groot gedeelte van alle Nederlandse tweets gebruik. Het eindproduct is een website die alle nieuwsberichten (met bijbehorende reacties) van de huidige dag toont op volgorde van populariteit. Om dit probleem op te lossen heb ik een programma geschreven dat bestaat uit vier onderdelen, te weten:

- Het zoeken van nieuws: Hiervoor heb ik twee methodes ontwikkeld. Een die alle tweets van 15 al bestaande nieuwsmedia als nieuws kenmerkt, en een die nieuws zoekt op basis van "#nieuws".
  Resultaat: De beide methodes wijken niet van elkaar af op basis van precisie (hoeveel % is ook echt nieuws), het verschil zit echter in het aantal en de kwaliteit van de berichten. De officiële twitteraccounts van nieuwsmedia delen minder (dubbel) nieuws, maar de syntax is wel gestandaardiseerd. Mijn voorkeur is daarom naar deze methode uitgegaan. Om dubbele nieuwsberichten te herkennen is de Jaccard Coëfficiënt een goede meeteenheid.

- Het zoeken van reacties op nieuws: Ook hier heb ik twee methodes ontwikkeld, de eerste methode zoekt nieuws op basis van de gehele nieuwstitel. De andere methode verwijderd eerst de stopwoorden uit de nieuwstitel en zoekt dan naar tweets die alle overgebleven woorden bevatten.
  *Resultaat:*De alternatieve methode vindt gemiddeld meer dan 30% meer reacties, terwijl de relevantie hoog blijft. Er moet echter wel gebruik gemaakt worden van een minimum aantal kernwoorden, anders kan de relevantie wel significant dalen.

- Het sorteren van de nieuwsberichten: De nieuwsberichten kunnen op dezelfde manier gesorteerd worden als de reacties, namelijk met een score. De variabele die gebruikt kunnen worden zijn hier echter anders, en zijn afhankelijk van de definitie van 'populair'.
  Resultaat: In mijn applicatie heb ik ervoor gekozen om de IDF-scores van alle reacties op de nieuwsberichten bij elkaar op te tellen en het restultaat hiervan als totaalscore te gebruiken voor het nieuwsbericht. Hierdoor zijn de scores gebasseerd op wat mensen over het nieuwsbericht zeggen.

- Het sorteren van de reacties op de nieuwsberichten: Ook bij de reacties worden scores berekend per item. Hier licht alleen de nadruk anders, voor reacties is de uniekheid belangrijk. De score van een bericht wordt dus bepaald aan de hand van de IDF-scores van de woorden die erin voorkomen.
  Resultaat: De beste methode is om alle IDF-scores van de woorden bij elkaar op te tellen en het resultaat hiervan als score voor de tweet te gebruiken.

Uiteindelijk zijn er dus meerdere manieren om een goede nieuwswebsite op basis van twitter te maken. Het gaat erom waar de zwaartepunten gelegd wordt. Omdat de data door gebruikers is gecreëerd en de website ook als doelgroep heel Nederland heeft, heb ik het zwaartepunt vooral op die gebruiker gelegd. Maar voor een specifiekere nieuwswebsite of een nieuwswebsite voor een bepaald publiek (geslacht, leeftijd enz.) zullen de methodes wellicht beter iets aangepast kunnen worden.

Hoofdstuk 1

Inleiding

1.1 Probleemstelling

Omdat er een overvloed aan nieuwsberichten op twitter staat is het moeilijk voor de gebruikers om te filteren wat belangrijk, leuk of interessant nieuws is om te lezen.

Het probleem wat in deze paper behandeld wordt is dus de chaos en ongestructureerdheid van nieuwsberichten op twitter.

1.2 Doelstelling

Het doel is om een manier te vinden om Nederlandse nieuwsberichten automatisch van twitter te extraheren, en deze vervolgens met bijbehorende meningen op een gebruiksvriendelijke manier te tonen. Het resultaat wordt dus een HTML-bestand(website), waar elke Nederlander die geïnteresseerd is in het Nederlandse nieuws terecht kan. Door dit te implementeren in de programmeertaal Java wordt duidelijk of de gebruikte methode werkt, en waar deze nog verbeterd kan worden.

1.3 Opbouw

Om de belangrijkste nieuwsberichten op een overzichtelijke manier voor de gebruiker te presenteren zijn meerdere acties nodig:

1. Zoeken naar nieuwsberichten op twitter.
2. Zoeken naar reacties op nieuwsberichten op twitter.
3. Sorteren van de reacties op de nieuwsberichten van twitter.
4. Sorteren van nieuwsberichten op basis van populariteit.
5. De resultaten overzichtelijk en aantrekkelijk weergeven voor de gebruiker.

Voor al deze onderdelen geldt dat eerst de meest simpele uitvoering wordt geïmplementeerd en uitgevoerd. Vanaf dit uitgangspunt kan gekeken worden waar verbeteringen mogelijk zijn. Vervolgens wordt getracht om deze punten te verbeteren, en achteraf moet duidelijk worden of dit is gelukt. Na dit proces kan er weer opnieuw begonnen worden, deze productiecyclus gaat door tot het systeem naar behoren werkt.

1.4 Vergelijkbaar werk

Op het moment dat ik begon met het ontwikkelen van het systeem (maart 2012) bestonden er nog geen Nederlandse nieuwswebsites volledig gebasseerd op social media, voor het Engelse nieuws bestonden er al een tijdje vergelijkbare websites, zoals `http://storyfull.com` (wordt nog door een redactie gepubliceerd), `http://www.onlyrealnews.com` (volledig automatisch) en `http://digg.com` (ook volledig automatisch).

In april 2012 kwamen er echter ook twee Nederlandse alternatieven, te weten: `http://sociaalnieuws.nl` en `http://poplinks.nl`. Mijn ontwikkelde systeem onderscheidt zich echter door niet alleen te kijken naar het aantal 'likes' of retweets, maar vooral te kijken naar wat mensen over een nieuwsbericht te zeggen hebben. Een ander groot verschil is de snellere verversing, op de 2 alternatieven blijven nieuwsberichten soms een hele week of zelfs meerdere maanden in de top 5 nieuwsberichten staan. Omdat mijn nieuwswebsite elke dag alle oude nieuwsberichten verwijderd is er meer reden voor de gebruiker om regelmatig terug te komen naar de website. Ook zijn op mijn website direct de tweets/meningen te lezen die reageren op het nieuwsbericht.

Hoofdstuk 2

Methode

In dit hoofdstuk gaan we proberen om een zo goed mogelijke methode te maken om populaire nieuwsberichten te vinden en te presenteren. De verschillende onderdelen worden onderscheiden in secties. Binnen deze secties worden verschillende methodes vergeleken op volgorde van complexiteit. In het begin van elke sectie gaan we uit van de meest simpele implementatie.

2.1 Data

Om een automatisch systeem te maken dat Nederlands nieuws verzameld hebben we allereerst natuurlijk data nodig. In dit geval wordt twitter gebruikt als bron, dus zijn er veel Nederlandse tweets nodig die laten zien wat Nederland bezig houdt. Omdat we een goed data-corpus tot onze beschikking hebben op de Siegfried server van de RUG die elk uur Nederlandse tweets verzameld (voor meer informatie: (Tjong-Kim-Sang 2011)), gebruiken we deze tweets. De tweets in dit corpus worden per uur ververst, de tweets staan dus ook per uur opgeslagen. Het nadeel van deze data is dat het meestal een of twee uur achterloopt. Een alternatief is de API van twitter te gebruiken, maar voor dit onderzoek is dat overbodig. De achterstand is niet hinderlijk.

Het programma loopt via een timer, dit is een losse applicatie die het hoofdprogramma aanroept. Deze timer kijkt regelmatig of er nieuwe tweets te vinden zijn, is dit het geval dan wordt het hoofdprogramma geactiveerd. Er is dan weer voor een uur aan tweets, hierin worden eerst de nieuwe nieuwsberichten gezocht. Als dit klaar is dan wordt voor alle nieuwsberichten van de gehele dag gezocht naar reacties hierop. Als dit klaar is dan kunnen de scores voor de reacties en daarna ook voor de nieuwsberichten berekend worden en kan de data gepubliceerd worden in html. Voor de duidelijkheid wordt hieronder het systeem nog eens schematisch weergegeven:

```
voor elk uur{
   als het geen nieuwe dag is (24 uur){
      zoek naar nieuwsberichten
         (sla alle nieuwsberichten op)
      zoek naar reacties op alle nieuwsberichten
         (sla deze op bij de nieuwsberichten)
      publiceer alle data in html
   anders (het is wel een nieuwe dag){
      verwijder alle nieuwsberichten en reacties
```

2.2 Nieuwsberichten zoeken

Om nieuwsberichten aan te bieden aan een gebruiker hebben we natuurlijk als eerste een programma nodig dat nieuwsberichten kan verzamelen. Hieronder worden eerst verschillende methodes op volgorde van complexiteit besproken met hun voor- en nadelen. Vervolgens wordt

er nog een probleem besproken dat bij alle methodes voorkomt, namelijk het verzamelen van dubbele nieuwsberichten.

2.2.1 Baseline

De baseline is eigenlijk bedoeld om alle tweets als nieuws te classificeren; de precisie bedraagt dan echter maar ongeveer 0,33% (op basis van een willekeurige steekproef van 300 tweets waaronder zich 1 nieuwstweet bevond). Dit is natuurlijk lang niet goed genoeg, dus ben ik snel verder gaan zoeken naar andere manieren om nieuwsberichten te verzamelen.

2.2.2 #nieuws

Een groot voordeel van twitter als bron is dat de gebruikers zelf hun tweets al categoriseren op basis van hashtags. Een hekje gevolgd door een woord in een twitterbericht, geeft dus het onderwerp/de gedachte van de tweet aan. We kunnen er dus van uitgaan dat alle tweets waarin "#nieuws" voorkomt een nieuwsbericht bevatten. Dit zou echter erg naïef zijn. Omdat iedereen een tweet kan delen waar "#nieuws" in voorkomt zonder dat het gaat om een echt nieuwsbericht, zal de verkregen data niet voor honderd procent uit nieuwsberichten bestaan. Wat hier als eerst opvalt is dat er erg vaak hetzelfde bericht wordt gevonden. Voor een nieuwswebsite is het natuurlijk interessant om te weten hoe vaak een nieuwsbericht wordt getweet, op dit moment zijn we echter alleen nog maar geïnteresseerd in de nieuwsfeiten. Daarom moeten de duplicaten samengevoegd worden. Zie hiervoor de laatste sectie van dit hoofdstuk, hierin zal ik bespreken hoe dubbele nieuwsberichten herkend kunnen worden.

2.2.3 Nieuwsaccounts volgen

Een andere manier om nieuwsberichten te vinden is het volgen van verschillende twitteraccounts van nieuwsmedia. We hebben ervoor gekozen om zowel de traditionele media (kranten), als de modernere media (nieuwswebsites) te volgen. Om te kiezen welke media we moeten volgen hebben we besloten om te kijken naar de tien kranten met de grootste oplage (Wikipedia 2012) en de vijf betrouwbaarste nieuwswebsites op basis van de vertrouwensindex van Newcom Research & Consultancy (Newcom 2011). Het eerste probleem is echter dat sommige media (parool, spits en nu.nl) meerdere twitteraccounts hebben, dit is echter makkelijk opgelost door alle accounts van deze media toe te voegen. Ook met deze methode worden er veel dubbele nieuwsberichten gevonden. Veel media tweeten namelijk over hetzelfde nieuws, maar het komt ook voor dat nieuwsmedia meerdere keren hetzelfde bericht plaatsen. Dit kan met verschillende accounts maar het gebeurt ook dat een tweet meerdere keer wordt gedeeld. Dus ook voor deze methode is het nodig om dubbele nieuwsberichten te verwijderen.

2.2.4 Dubbele nieuwsberichten herkennen

Met beide bovenstaande methodes zullen in meer of mindere mate dubbele nieuwsberichten voorkomen. Voor het goed functioneren van de rest van het programma zal dit herkend moeten worden. Hiervoor is een normalisator en daarna een vergelijker nodig. De methodes worden getest in hoofdstuk 3.1.1.

Door het bestuderen van deze data blijkt dat het onmogelijk is om dit perfect te doen. Veel dezelfde nieuwsberichten hebben een titel die taalkundig niet overeenkomt en gaan toch over hetzelfde. Soms is het ook juist andersom. zie de onderstaande voorbeelden:

> NOS Lagere straf voor 'crèchemoord'
> telegraaf 18 jaar voor moord crèche http://t.co/cgd8kUvv

> Dlaatstenieuws De huidige problemen op de #huizenmarkt raken vooral huiseigenaren jonger dan 35 jaar. Bekijk de 'Feitenkaarten': http://t.co/WCsfHX0m #FD
> NUnl_economie Vooral risico's voor jongere huiseigenaren: http://t.co/ROcEaMpK

> NUnl_buitenland Protesten op Dag van de Arbeid http://t.co/71uVb5yS
> NUnl_beurs Beurzen vieren Dag van de Arbeid http://t.co/X9OLhT1k

Maar ook berichten als:

> NUnl 'We moeten elkaar niet de maat nemen': http://t.co/dHSOJOrE

Zijn moeilijk te matchen met andere berichten zonder de website te bezoeken. Ook al is het dus onmogelijk om alle dubbele nieuwsberichten te verwijderen, het blijft belangrijk om zoveel mogelijk van de dubbele nieuwsberichten te verwijderen.

De normalisator verwijdert alle leestekens en zet hoofdletters om naar kleine letter voor een makkelijkere/efficiëntere vergelijking, verder zijn de gebruikersnaam en de URL niet van toegevoegde waarde voor het herkennen van duplicaten, dus ook deze worden verwijderd. Een tweet gaat nu van:

> aldomb #Nieuws! Obama spendeert 21 miljoen dollar aan webadvertenties voor verkiezingen: http://goo.gl/y8nvd #USA

naar:

> nieuws obama spendeert 21 miljoen dollar aan webadvertenties voor verkiezingen usa

Na het testen van de methode op verschillende data, valt het vooral op dat er 2 manieren zijn waarop nieuwsaccounts hun nieuws tweeten.

> <naam account> <nieuwsTitel> <link>
> <naam account> <nieuwsTitel> <begin nieuwsbericht> <link>

Met vaak nog ergens hashtags ertussen. Maar er zijn ook vaak varianten in leestekens of zelfs woordkeuze. Daarom is er een robuustere methode nodig, hiervoor gebruik ik de Jaccard Coëfficiënt (Naumann and Herschel 2010). De Jaccard Coëfficiënt berekend de taalkundige afstand tussen twee zinnen. Dit is precies wat we nodig hebben, we willen namelijk weten hoeveel de verschillende nieuwsitems op elkaar lijken. Er zijn echter wel wat testen nodig om de juiste grenswaarde te bepalen (hoeveel % moeten de zinnen overeen komen voordat ze over hetzelfde nieuwsbericht gaan), zie hiervoor het hoofdstuk resultaten).

Omdat gelijkheid op verschillende manieren gedefinieerd kan worden, kan de Jaccard Coëfficiënt op verschillende manieren gemplementeerd worden. In mijn implementatie gebruik ik woorden als eenheden, en wordt er dus naar de overeenkomende woorden van twee zinnen gekeken. Er kan ook naar n-grammen gekeken worden, maar dit is veel tijdrovender, omdat er dan elke keer een lijst met n-grammen aangemaakt moet worden en er ook veel meer vergelijkingen per zin uitgevoerd moeten worden.

Ik heb de berekening van de Jaccard Coëfficiënt geïmplementeerd in Java, hieronder staat de psuedo code van mijn implementatie:

```
aantalGelijkeWoorden = 0
voor (alle woorden in zin1)
  als zin2.bevat(woord)
      aantalGelijkeWoorden +1
jaccard = aantalGelijkeWoorden / (aantalGelijkeWoorden + zin1.length + zin2.length)
```

Vervolgens wordt de Jaccard waarde vergelijken met een nader te bepalen constante (2.4.1) om te kijken of de twee zinnen overeenkomen. Is dit het geval dan hoeft de nieuwe zin niet toegevoegd te worden aan de bestaande nieuwsfeiten.

2.3 Zoeken van reacties op nieuwsberichten

Nu de nieuwsberichten zijn gevonden, kunnen we de bijbehorende reacties van gebruikers zoeken. Deze reacties kunnen later gebruikt worden om de populariteit van de nieuwsberichten te berekenen. Maar zijn ook interessant voor de gebruiker om te lezen, zo wordt meteen duidelijk wat twitteraars van het nieuws vinden. Voor het zoeken naar reacties heb ik twee methodes ontwikkeld die hieronder uitgelegd worden.

2.3.1 Reacties zoeken op basis van nieuwstitel

Om deze methode te implementeren blik ik eerst terug op de methode om dubbele nieuwsberichten te herkennen. Hiervoor had ik een normalisator ontworpen die de titel van een nieuwsbericht uit een tweet kan extraheren. Deze heb ik als een Java methode ontwikkeld, die methode kunnen we nu dus makkelijk hergebruiken.

Als we titel los hebben gehaald van de tweet, kan het zoeken beginnen. Het programma gaat simpelweg alle tweets lezen om te kijken of de titel voorkomt in elk tweet. Is dit het geval dan wordt de reactie opgeslagen. Dit deel van het programma is het meest tijdrovend, omdat voor erg veel berichten gekeken moet worden of ze matchen met erg veel nieuwsberichten. Aan het begin van de dag zijn er nog niet zoveel nieuwsberichten, maar als ze van een paar uur verzameld worden, zijn er al snel een paar honderd. Om voor alle nieuwsitems in alle tweets te zoeken zijn twee algoritmes mogelijk. Hieronder zijn de algoritmes voor beide manieren te vinden.

```
voor elke tweet in het uur
   normaliseer deze tweet
   voor alle nieuwsberichten
      normaliseer nieuwsbericht
      als het nieuwsbericht overeenkomt met de tweet
         sla de tweet op

voor elke nieuwsbericht
   normaliseer nieuwsbericht
   zoek in het corpusbestand naar gelijkenissen (verwijder leestekens en ignore case)
   voor alle resultaat-tweets:
         sla de tweet op
```

Het is op basis van deze algoritmes onmogelijk om te zien welk algoritme het efficiëntst werkt, dit hangt van te veel factoren af (taal, hardware en data). Daarom ga ik beide methodes implementeren en voor een hele dag al het nieuws en de bijbehorende reacties laten zoeken, in hoofdstuk 3.2.1 staan hiervan de resultaten

2.3.2 Reacties zoeken op basis van kernwoorden

Als we ervan uit gaan dat een bericht met dezelfde kernwoorden als het nieuwsbericht een reactie op het nieuwsbericht is, dan kunnen we echter ruimer zoeken dan de methode die hierboven beschreven wordt. Omdat het systeem per dag werkt, zal deze assumptie vaak gemaakt kunnen worden. Als er op eenzelfde dag namelijk meerdere berichten zijn die taalkundig erg overeen komen, gaan deze vaak over hetzelfde onderwerp (zie voor hoe vaak: 3.1.1).

Deze methode werkt dus op basis van kernwoorden, maar de vraag is: hoe kunnen we de kernwoorden makkelijk/efficiënt vinden. Om dit te doen draai ik de situatie om, alles wat geen kernwoord is, is namelijk een stopwoord. Dus als we alle stopwoorden verwijderen, dan blijven vanzelf de kernwoorden over. Stopwoorden zijn makkelijker te herkennen dan kernwoorden, met behulp van een frequentielijst kunnen we snel de meestvoorkomende woorden vinden. Hieronder staat de methode die ik heb gebruikt om een frequentielijst te maken in de Linux commandline:

```
cut -d' ' -f2- * | sed 's/ /\n/g' | sed -r 's/^[^[:alpha:]]*|[^[:alpha:]]*$//g'
| sed '/^$/d' | tr '[[:upper:]]' '[[:lower:]]' | sort | uniq -c | sort -nr
```

Dit commando bevat ook normalisatie, alles wat geen 'Alpha' (alfabetisch) symbool is wordt verwijderd evenals het eerste woord. De links heb ik niet verwijderd omdat deze allemaal uniek zijn en dus toch onderaan komen te staan. Verder worden alle woorden naar kleine letters geconverteerd zodat "Nederland" en "Nederland" een type (woord) is. De uitvoer van dit commando wordt opgeslagen in een tekstbestand. Hierin kan men kijken wat stopwoorden zijn, en deze vervolgens doorgeven aan het programma. Om een goede lijst met stopwoorden te krijgen heb ik een frequentielijst gemaakt met al het nieuws van een hele maand (april 2012), en hier vervolgens gekeken vanaf welke frequentie er woorden voorkomen die iets over een nieuwsbericht zeggen. Dit bleek rond de frequentie van veertig te zijn. De eerste versie van de lijst met stopwoorden omvatte dus alle woorden die meer dan veertig keer voorkwamen.

Maar nieuws is erg tijd-specifiek, dus om te voorkomen dat sommige woorden in de lijst met stopwoorden staan alleen omdat het nieuws in april 2012 populair was is er meer nodig. Dus heb ik ook een frequentielijst gemaakt van de twitterdata van een half jaar voor april 2012, oktober 2011. Alle woorden die in beide maanden meer dan veertig keer voorkomen, zijn in bijna alle gevallen stopwoorden. Door de lijst op deze manier te verbeteren, heb ik nu een betrouwbare lijst met stopwoorden. (zie voor het resultaat sectie 3.2.2)

Nu de lijst met stopwoorden klaar is kan het programma geschreven worden. Voor elk nieuwsfeit zullen eerst de stopwoorden verwijderd moeten worden en vervolgens kan er naar berichten gezocht worden *alle* overgebleven kernwoorden van het nieuwsfeit bevatten. Als een nieuwstitel echter minder dan drie kernwoorden bevat, zijn dit er te weinig en wordt er teveel aan het toeval overgelaten. Als dit het geval is dan wordt de gehele nieuwstitel gebruikt zoals in de vorige sectie is beschreven.

2.4 Publiceren

Als de nieuwsberichten en de reacties op die nieuwsberichten zijn gevonden is de data voor een gewone gebruiker nog niet erg bruikbaar. Het doel van deze paper is om een methode te ontwikkelen die nieuws presenteert aan nieuws-geïnteresseerde Nederlanders. Er is echter enorm veel data, waarvan veel niet interessant is voor de meeste mensen. In dit hoofdstuk ga ik methodes ontwikkelen om nieuwsberichten te sorteren, maar ook om reacties te sorteren en accentueren. Zo is de kans groter dat de gebruiker meteen nieuws leest dat hem interesseert. De data kan echter ook voor onderzoeksdoeleinden worden gebruikt, dan kan het niet nodig zijn om deze scores te berekenen en kan dit gedeelte van het programma wellicht achterwege gelaten worden.

Om de data aan de gebruiker te tonen is gekozen voor HTML. Deze taal is vooral gekozen omdat hiermee het grootste deel van de doelgroep aangesproken kan worden met een (HTML) versie van de resultaten. Bijna alle apparaten die een scherm hebben en op het internet aangesloten zijn kunnen html weergeven. Het enige wat we nodig hebben zijn links, zodat de nieuwstitels naar de bijbehorende berichten en reacties kunnen verwijzen. Een alternatief zou HTML met PHP zijn, dit zou beter kunnen samenwerken in combinatie met een database. En het heeft als voordeel dat het dynamischer is, maar mijn applicatie krijgt maar een keer in het uur een update, dus de website hoeft maar een keer per uur geüpdatet worden.

2.4.1 Ordenen van reacties

Nadat alle reacties zijn verzameld valt het op dat er veel reacties zijn die alleen maar het nieuwsbericht retweeten (herhalen). Deze berichten zijn oninteressant om te lezen omdat ze allemaal precies hetzelfde zijn. Omdat de berichten in chronologische volgorde worden geschreven, zijn ze nu nog gesorteerd op tijd. Dit ziet er als volgt uit:

Voor de bezoeker van een nieuwswebsite heeft dit weinig toegevoegde waarde, daarom is er een methode nodig die interessantere reacties hoger plaatst en oninteressante reacties lager. Hiervoor maak ik een systeem dat elke reactie een score geeft, als al deze scores zijn berekend kunnen de reacties hierop gesorteerd worden. Er zijn echter meerdere manieren om zo'n score te berekenen, een aantal hiervan komen hieronder aan bod.

Toen ik begon aan dit onderzoek leek het me een geschikte methode om alles wat geen nieuwstitel is als werkelijke reactie te zien. Maar omdat ik voor de methode heb gekozen die reacties zoekt op basis van kernwoorden uit de titel van het nieuwsbericht werkt deze methode niet meer, niet alle berichten bevatten namelijk nog de nieuwstitel. Daarom moet er dus een slimmere methode komen.

De grote vraag voor dit stuk van het programma is: wat is er interessant om te lezen? Zoals hier boven al is beschreven is het niet interessant om telkens hetzelfde te lezen, het is veel interessanter om steeds nieuwe/verschillende berichten te lezen. De scores die de berichten krijgen moeten dus weergeven hoe uniek de berichten zijn.

De methode die mij hier geschikt voor lijkt is het berekenen van de IDF (Inverse Document Frequency) (Manning, Raghavan, and Schitze 2008). De IDF wordt normaal gesproken gebruikt in zoekmachines om uit te zoeken welke term(en) unieker zijn, deze term(en) worden vervolgens zwaarder meegeteld in de resultaten. Ik pas de IDF hier iets anders toe, omdat het doel ook anders is. Ik reken voor elk woord in een tweet de IDF uit, dus elk woord krijgt een score die weergeeft hoe uniek het woord is. Nadat dit is gebeurd kan de sortering op twee manieren plaats vinden. De eerste manier is om te tellen hoeveel woorden in een tweet uniek zijn, dit zijn alle berichten die een IDF waarde hebben die groter is als een later te bepalen constante. En vervolgens de tweets te sorteren op aantal unieke woorden. Een andere manier is om de tweets een totaalscore te geven. Om deze te berekenen tel ik alle scores van de woorden in een tweet bij elkaar op en zo krijgt de tweet een totaalscore. Deze totaalscore laat zien hoe uniek de tweet is op basis van de woorden die in de tweet voorkomen.

Als de reacties goed gesorteerd zijn, staat er alsnog veel overbodige informatie bij de bovenste tweets. In bijna alle tweets wordt namelijk de volledige titel herhaald, om dit eruit te filteren kan er voor gekozen worden om woorden met een IDF score van boven een nader te bepalen constante te accentueren. Om achter deze constante te komen heb ik de IDF uitgerekend voor een woord dat in een van de tien zinnen voorkomt ($\log(10/1)$). Toen kwam ik op de waarde 2.302585092994046. Maar een iets ruimere marge is misschien beter, bovendien kunnen we een klein snelheidsvoordeel krijgen als we met een integer vergelijken, dus we ronden dit af naar 2.

Als al deze woorden vetgedrukt zijn, is de pagina veel drukker (zie: 3.3.1). Er is veel afwisseling tussen dikgedrukte en normale letters. Dit komt mede doordat veel mensen eerst het nieuwsbericht herhalen en daarna in een hele zin hun mening delen. In deze 'meningszin' komen soms een paar stopwoorden voor of woorden uit het nieuwsbericht. Deze woorden worden daarom weergegeven in normale letters. Dit is op te lossen door woorden die omsingeld zijn door diktedrukte woorden ook dikgedrukt te maken, zo ziet het resultaat er iets netter en overzichtelijker uit (zie: 3.3.1).

2.4.2 Ordenen van nieuwsberichten

Omdat ons systeem honderden nieuwsberichten per dag vindt, is het ook handiger voor de gebruiker als ze de populairdere/interessantere nieuwsberichten eerst kunnen zien voordat de minder populaire nieuwsberichten aan bod komen. Hiervoor gebruik ik een zelfde soort systeem als bij het ordenen van de reacties. Er wordt een score berekend per nieuwsbericht en op basis van deze scores worden de nieuwsberichten gesorteerd. Voor een relatief simpele implementatie van dit systeem maken we de assumptie dat als er veel getweet wordt over een bericht dat dit een interessant bericht is. De implementatie van deze methode in Java is vanzelfsprekend. De data van de reacties is onder de titel van het nieuwsbericht in hetzelfde tekstbestand geschreven, dus als we de regels van de bestanden tellen dan weten we meteen hoeveel reacties er zijn op

het nieuwsbericht. Na het sorteren staan de tweets met de meeste reacties dus bovenaan.

We kunnen naar wens meer variabelen toevoegen en deze een gewicht geven. Het gaat erom dat populaire nieuwsberichten die veel aandacht trekken een hogere score krijgen. De scores van de berichten zijn daar een goede graadmeter voor, als veel mensen allemaal verschillende dingen over een nieuwsbericht tweeten dan kunnen we stellen dat het nieuwsbericht belangrijker/populairder is als een nieuwsbericht dat gewoon geretweet wordt. Dus kunnen we in plaats van het aantal reacties tellen ook de scores van de reacties bij elkaar optellen en hierop sorteren.

Er zijn nog meer factoren die we mee kunnen laten tellen. We kunnen bijvoorbeeld kijken hoe populair het twitteraccount is dat het oorspronkelijke nieuwsbericht heeft getweet. Of we kunnen hier nog een stap verder gaan en de assumptie maken dat accounts die vaak wat interessants delen en dus veel volgers hebben, deze keer weer wat interessants delen. Dan kunnen we alle het aantal volgers van alle accounts die een reactie hebben geplaatst bij elkaar optellen. Voor deze informatie hebben we echter wel de data nodig, deze worden door twitter zelf openbaar gedeeld in XML bestanden (bijvoorbeeld: https://twitter.com/users/show.xml?screen_name=nos). Hier kunnen we in het XML-veld "<followers_count>" makkelijk vinden hoeveel volgers het account heeft en dus hoe het account gewaardeerd wordt door medetwitteraars. Het probleem hier is echter dat er door twitter een limiet is gesteld aan hoe vaak deze informatie opgevraagd mag worden. op het moment van schrijven is die limiet 150 keer of 350 keer met een "Whitelisted account'(Twitter 2010). Dit is beide niet genoeg voor mijn programma, er worden regelmatig meer tweets gevonden. Deze methode kan dus niet getest worden.

Ook kan bij het detecteren van dubbele nieuwsberichten zoals gebeurd in hoofdstuk 2.2.4 meteen een aparte score bijgehouden worden die het aantal nieuwsmedia bijhoudt dat het nieuwsbericht tweet. Als het doel is om een systeem te maken dat alleen betrouwbare nieuwsberichten publiceert, kan er ook voor een minimum aantal nieuwsmedia worden gekozen dat het nieuwsbericht heeft getwitterd.

2.4.3 HTML schrijven

Nu we de data allemaal gesorteerd hebben is het tijd om de werkelijke website te schrijven. Hiervoor moet de data omgezet worden in HTML-bestanden. Op de homepage staat een lijst met nieuwstitels gesorteerd op de uitgerekende scores. Wie op de scores klikt gaat naar het complete nieuwsbericht. Om dit voor elkaar te krijgen gebruik ik een eerder ontwikkelde Java-methode die de titel van een nieuwsbericht extraheert. Verder ontwikkel ik een nieuwe Java-methode die de link naar het bericht kan vinden. Hier ga ik ervan uit dat de eerste link van de tweet de link naar het nieuwsbericht is. Achter de nieuwstitels komt een link naar de reacties, zodat makkelijk en snel de bijbehorende tweets bekeken kunnen worden. Verder staan er nog wat links naar andere dingen die de gebruiker misschien interesseert. Er is een 'over ons' pagina die uitlegt wat voor website dit is en hoe deze werkt, maar er zijn ook links naar de 5 hoogst scorende berichten van de gehele week, zodat er ook nog op nieuws teruggeblikt kan worden.

Voor elk nieuwsbericht wordt een aparte pagina gemaakt. Bovenaan staat groot de titel die uit de Tweet geëxtraheerd is met een link naar het originele artikel op de bron-nieuwswebsite. Voor deze titel staat het logo van het nieuwsaccount dat de tweet gedeeld heeft. Dit logo wordt verkregen door het twitter XML-bestand van de gebruiker (het veld "<profile_image_URL>".

Hoofdstuk 3

Resultaten

3.1 Zoeken van nieuws

Voor het zoeken naar nieuws zijn als testdata de tweets een volledige dag gebruikt, het gaat hier om een willekeurige woensdag. De twee manieren beschreven in het hoofdstuk 2.2 worden allebei uitgevoerd. Vervolgens wordt eerst getest hoe het beste dubbele nieuwsberichten verwijderd kunnen worden. Hiervoor vergelijken we drie methodes, te weten: niks verwijderen, identieke berichten verwijderen, op basis van de Jaccard Coëfficiënt verwijderen.

Als de dubbele nieuwsberichten verwijderd zijn kan er onderzocht worden welke van de twee methodes om nieuws te zoeken het beste werkt. Hiervoor heb ik negen verschillende mensen 50 willekeurige nieuwsberichten van beide methodes laten annoteren. Het is van belang om meerdere mensen te vragen omdat iedereen een eigen definitie van 'nieuws' heeft. Aan de hand van deze steekproef zal blijken welke methode het best werkt.

3.1.1 Dubbele nieuwsberichten verwijderen

Om dit te testen laat ik het programma zes keer nieuwsberichten zoeken. Namelijk, voor alle drie methodes om dubbele te herkennen twee keer (zoeken op #nieuws en voor de nieuwsbronnen). Er zijn hier twee dingen van belang: hoeveel dubbele nieuwsberichten worden er gevonden en hoe efficiënt is het algoritme. Zie de onderstaande tabel voor de resultaten:

Tabel 3.1: Resultaten van het testen van verschillende methodes om dubbele nieuwsberichten te detecteren. Tijd is in minuten en seconden.

Methode		#nieuws methode	15 nieuws-bronnen methode
Alles	aantal tweets % tijd	3982 100% 00:06	305 100% 02:10
Identieke berichten verwijderd	aantal tweets % tijd	3609 91% 08:24	286 94% 02:20
Op basis van Jaccard Coëfficiënt verwijderd	aantal tweets % tijd	2287 57% 05:57	268 88% 02:18

Uit tabel 3.1 wordt zichtbaar dat op basis van de Jaccard Coëfficiënt meer dubbele nieuwsberichten worden gedetecteerd. Dit is geen verrassing omdat met de Jaccard Coëfficiënt alle

identieke berichten worden herkend (Jaccard Coëfficiënt heeft dan de waarde 1), en daarbovenop nog meer berichten als hetzelfde ziet. Wat echter wel opvallend is, is dat de Jaccard Coëfficiënt ook nog efficiënter werkt.

3.1.2 15 Nieuwsbronnen vs. #nieuws

Nu de dubbele nieuwsberichten zijn verwijderd is het belangrijk om te bekijken hoeveel procent van de gevonden nieuwsberichten ook echt nieuws bevatten. Hiervoor is een steekproef genomen uit de data en zijn er 9 verschillende personen gevraagd om 50 berichten te annoteren als 'nieuws' of als 'niet nieuws'. Er is de proefpersonen nadrukkelijk gevraagd om persoonlijke interesses zoveel mogelijk buiten spel te houden, en objectief naar de data te kijken. Hieronder de resultaten:

Tabel 3.2: De scores van beide manieren om nieuws te vinden op basis van de geannoteerde data. De cijfers geven het aantal % tweets aan dat volgens de proefpersonen nieuws bevatte.

persoon	#nieuws	15 nieuwsbronnen
1	52	42
2	68	66
3	74	68
4	68	52
5	54	42
6	48	38
7	32	32
8	64	76
9	82	80

Om te kijken of er een significant verschil is tussen de twee methodes gaan we de statistische gepaarde t-toets gebruiken. Hiervoor moeten we eerst controleren of voor beide methodes de datapunten normaal verdeeld zijn, zie hiervoor figuren A.1 en A.2 in de bijlage. De normaalplotten laten zien dat de data bij benadering normaal verdeeld is. Hieronder een samenvatting van de data:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Hashtag	9	32	82	60.22	15.213
Nieuwsbronnen	9	32	80	55.11	17.751
Valid N (listwise)	9				

Figuur 3.1: Descriptive statistics uit SPSS

Figuur 3.1 laat zien dat de data van de twee methodes niet veel van elkaar verschillen. Het gemiddelde van de hashtag methode is iets hoger, maar de standaard afwijking van de andere methode is iets groter. Vanuit deze samenvatting kunnen we echter nog geen conclusies trekken. De toets moet dus nu uitgevoerd worden. De bijbehorende hypotheses zijn:

H_0: Uh = Un
H_a: Uh $\neq$ Un

Oftewel:

Nulhypothese: de gemiddelde scores van beide methodes zijn hetzelfde.
Alternatieve hypothese: de gemiddelde scores van beide methodes verschillen significant.

Ik kies voor een betrouwbaarheidsniveau van 95% (p=0,05) zodat we redelijk zeker kunnen zijn van de uitkomst. Na de test uit te voeren in SPSS krijg ik de volgende resultaten:

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Hashtag - Nieuwsbronnen	5.111	8.313	2.771	-1.279	11.50	1.844	8	.102

Figuur 3.2: Resultaten gepaarde t-toets uit SPSS

De sigma heeft een waarde van 0,102. Dit betekend dat de nulhypothese wordt aangenomen op basis van p = 0,05. De alternatieve hypothese wordt dus verworpen. De conclusie is dan dat de scores van de twee methodes niet significant van elkaar verschillen en beide methodes dus even goed zijn.

Na het lezen van de nieuwsberichten van beide methodes heb ik ervoor gekozen om voor de rest van deze Paper de methode die gebruik maakt van de twitteraccounts van 15 grote nieuwsbronnen te gebruiken. Deze methode zorgt voor nieuwsberichten met een meer gestandaardiseerde vorm, daardoor zijn minder dubbelen en zijn de nieuwstitels beter te herkennen.

3.2 Zoeken van reacties op nieuwsberichten

Een methode die reacties van mensen zoekt moet aan twee eisen voldoen; ten eerste moeten er zoveel mogelijk (relevante) berichten gevonden worden. En ten tweede moet het systeem zo efficiënt/snel mogelijk werken.

3.2.1 Reacties zoeken op basis van nieuwstitel

De reacties van deze methode zijn altijd relevant omdat ze de titel altijd helemaal identiek voorkomt in de tweet, relevantie hoeven we dus niet te testen. Het is echter wel van belang zo snel mogelijk zo veel mogelijk reacties te vinden, hiervoor heb ik het systeem weer voor dezelfde dag laten draaien als ik heb gedaan om het zoeken naar nieuws te testen. Om de resultaten zo realistisch mogelijk te maken laat ik het programma ook weer opnieuw naar nieuws zoeken voor elk uur: zo zoekt het systeem niet naar reacties op nieuws dat nog niet plaatsgevonden heeft. Hiervoor gebruikt het programma de methode van de 15 nieuwsbronnen gecombineerd met de Jaccard coëfficiënt om dubbele nieuwsberichten te vermijden. De tijd die het programma hiervoor nodig had was 2 minuten en 18 seconden, de tijd om de reacties te zoeken is dus de totale tijd minus 2 minuten en 18 seconden. Deze methode is met twee verschillende algoritmes geïmplementeerd, zie hoofdstuk methode. De resultaten van deze 2 implementaties worden hieronder getoond:

Tabel 3.3: Vergelijking tussen 2 verschillende algoritmes die reacties zoeken op basis van exacte nieuwstitels

methode	Algoritme 1	Algoritme 2
Resultaten	3342	3183
Relevantie	100%	100%
Tijd	04:59	23:28

Algoritme 1 werkt dus duidelijk sneller, deze methode doorloopt het corpus maar 1 keer, dit blijkt dus de bepalende factor te zijn (in combinatie met de hardware van de server en Java).

3.2.2 Reacties zoeken op basis van kernwoorden

Het resultaat van het programma om stopwoorden te zoeken staat hieronder (figuur 3.3).

de	op	voor	in	van	het	bij	niet
een	audio	gt	video	en	met	naar	fd
aan	is	geen	wil	uit	te	meer	zijn
over	tegen	nog	heeft	dat	om	na	
als	moet	gaat	ook	er	onder	wordt	
uur	terug	nieuws	veel	die	nu	je	
kan	we	komt	vkopinie	zich	maakt	hebben	
dan	tot	wel	al	zo	verder	of	
doet	wat	maar	weg	ik	dit	door	

Figuur 3.3: Lijst met stopwoorden verkregen uit nieuwstitels

Om te kijken of de meer gecompliceerde methode die deze stopwoorden gebruikt om reacties op nieuwsberichten te vinden beter werkt dan de simpelere methode zal ik deze vergelijken met het beste algoritme van de vorige paragraaf (3.2.1). Hierbij is het belangrijk om vooral de focus te leggen op de berichten die eerst niet werden gevonden, maar met de nieuwe methode wel worden gevonden. De vraag is: is het wel de moeite waard om het gecompliceerdere algoritme te gebruiken? Om deze vraag te beantwoorden moeten we een kosten/baten vergelijking maken. Om te beginnen de kosten, hoeveel extra tijd kost het om de reacties te vinden, en hoeveel leveren we in qua relevantie? Om dit duidelijk te krijgen heb ik het programma met beide methodes voor vier willekeurige dagen verspreid door het jaar alle reacties laten zoeken.

Bij de nieuwe resultaten heb ik handmatig voor een steekproef van 500 tweets gekeken hoeveel procent van de reacties relevant is. Bij de vorige methode is dit zoals eerder gemeld niet nodig, de reacties zijn allemaal relevant ze bevatten immers de gehele nieuwstitel. Tijdens het bekijken van de resultaten kwam ik erachter dat er bijna geen irrelevante resultaten worden gevonden. Er was een nieuwsbericht over “Rutte”(de naam van een voetbaltrainer van PSV en de minister-president) dat verantwoordelijk was voor 5 irrelevante reacties, daarbuiten waren er nog maar 3 andere reacties gevonden die niet over het nieuwsbericht gingen. Hieronder staan de volledige resultaten van deze test:

Tabel 3.4: Vergelijking tussen de 2 methodes om reacties op nieuwsberichten te vinden

methode	Identieke titels	Kernwoorden
Aantal gevonden berichten	11160	14939
Relevantie	100%	98,4%
Tijd	15:25	17:41

Er worden dus veel meer reacties gevonden met de nieuwe methode, hiertegenover staat een kleine verhoging van de tijd. Verder viel me tijdens het lezen van de reacties op, dat met de nieuwe methode meer informatieve reacties worden gevonden. Dit is een gevolg van de mildere filtertechniek, omdat niet meer de exacte titel gevonden hoeft de worden zijn er nu ook berichten die van de titel afwijken maar wel over hetzelfde onderwerp gaan. De relevantie daalt echter maar nauwelijks met deze methode.

3.3 Publiceren

In deze sectie zijn wat voorbeelden te vinden van de verschillende onderdelen binnen dit deel van het programma, er is geen absolute kwantitatieve variabele om te beoordelen wat de beste methodes zijn. Maar door te vergelijken hoe de resultaat HTML pagina's eruitzien kan ik toch redelijk beoordelen welke methodes het beste werken.

3.3.1 Het ordenen van de reacties

Het eerste onderdeel van het ordenen van reacties is het sorteren van de reacties. Hiervoor heb ik twee varianten van het programma gemaakt die beide hun score uitrekenen op basis van IDF, de eerste methode telt alleen hoeveel woorden uniek zijn terwijl de tweede (gecompliceerdere) methode een score voor een zin uitrekent. In deze paragraaf vergelijk ik de resultaten van de twee programma's op basis van drie verschillende nieuwsberichten en ga ik vervolgens in op het accentueren van woorden.

De resultaten van het eerste nieuwsbericht staan in figuren 3.4 en 3.5. Hier valt het vooral op dat de drie tweets van DTNNetherlands die hetzelfde zijn lager komen te staan met de complexere berekening (figuur 3.5), maar ook andere tweets die weinig informatie bevatten komen met deze methode net iets lager te staan.

Nataschakossen Ik heb net Veel dieren dood door brand kinderboerderij **via #ADnl gelezen. Vind ik echt heel zielig:-(**
MuseumJoCas Zo'n 150 dieren dood bij brand in kinderboerderij Middelharnis **http://t.co/bBdavIoJ via @NOS *waar blijvende brandvoorschriften* #politiek**
Tijgernest Middelharnis: veel dieren dood bij brand. **Geen "voormalige" kinderboerderij dus. Arme dieren. nu.nl: WEL slachtoffers! http://t.co/DwuJvbpY**
Nataschakossen Ik heb net Veel dieren dood door brand kinderboerderij **via #ADnl gelezen. http://t.co/8Lw4N6Ta**
joepie47 "@telegraaf: Veel dieren dood door brand kinderboerderij **http://t.co/TU336ld8"@ iemand hekel aan de paashaas???**
FranssenFrans âœž 150 dieren dood door brand kinderboerderij **http://t.co/Pw3TbI3g via #Groeiendverzetâ˜® http://t.co/g84H5VIy**
DTNNetherlands DTN The Netherlands Veel dieren dood bij brand: AMSTERDAM - Bij een brand op een kinderboerderij in Middelharnis... **http://t.co/5RoMg4z7**
DTNNetherlands DTN The Netherlands Veel dieren dood door brand kinderboerderij: Door een brand in een kinderboerderij in Middel... **http://t.co/couFmoCH**
DTNNetherlands DTN The Netherlands Veel dieren dood door brand kinderboerderij: Door een brand in een kinderboerderij in Middel... **http://t.co/Gsn0Uh7s**
nieuws_nunl [NEWS] Dieren dood door brand: Kinderboerderij **afgebrand. http://t.co/caURskV4 [/NEWS]**
112today Veel dieren dood door brand in kinderboerderij: Middelharnis - In een kinderboerderij **voor verstandelijk gehandi... http://t.co/6Kbt7Hj0**
nieuws_nunl [NEWS] Veel dieren dood bij brand: AMSTERDAM - Bij een brand op een kinderboerderij in Middelharnis zijn... **http://t.co/QnUlg1hf [/NEWS]**
provincienieuws - Provincie Nieuws - Veel dieren dood door brand kinderboerderij: MIDDELHARNIS - Door een brand in... **http://t.co/X9NKMOm9 #Noordholland**
112today Veel dieren dood door brand in kinderboerderij: **http://t.co/SGGflb5E via @AddThis**
MerijnActueel Bron:Jeugdjournaal// Dieren dood door brand: Kinderboerderij **afgebrand. http://t.co/awsPuuMP**

Figuur 3.4: (Nieuwsbericht 1) De 15 reacties met de hoogste score op basis van de methode die unieke woorden telt.

Nataschakossen Ik heb net Veel dieren dood door brand kinderboerderij **via #ADnl gelezen. Vind ik echt heel zielig:-(**
MuseumJoCas Zo'n 150 dieren dood bij brand in kinderboerderij Middelharnis **http://t.co/bBdavIoJ via @NOS *waar blijvende brandvoorschriften* #politiek**
Tijgernest Middelharnis: veel dieren dood bij brand. **Geen "voormalige" kinderboerderij dus. Arme dieren. nu.nl: WEL slachtoffers! http://t.co/DwuJvbpY**
Nataschakossen Ik heb net Veel dieren dood door brand kinderboerderij **via #ADnl gelezen. http://t.co/8Lw4N6Ta**
joepie47 "@telegraaf: Veel dieren dood door brand kinderboerderij **http://t.co/TU336ld8"@ iemand hekel aan de paashaas???**
FranssenFrans âœž 150 dieren dood door brand kinderboerderij **http://t.co/Pw3TbI3g via #Groeiendverzetâ˜® http://t.co/g84H5VIy**
112today Veel dieren dood door brand in kinderboerderij: Middelharnis - In een kinderboerderij **voor verstandelijk gehandi... http://t.co/6Kbt7Hj0**
provincienieuws - Provincie Nieuws - Veel dieren dood door brand kinderboerderij: MIDDELHARNIS - Door een brand in... **http://t.co/X9NKMOm9 #Noordholland**
jari4now Veel dieren dood door #brand kinderboerderij **http://t.co/c41CNcJ8 #brandpreventie # brandveiligheid**
nieuws_nunl [NEWS] Dieren dood door brand: Kinderboerderij **afgebrand. http://t.co/caURskV4 [/NEWS]**
DTNNetherlands DTN The Netherlands Veel dieren dood bij brand: AMSTERDAM - Bij een brand op een kinderboerderij in Middelharnis... **http://t.co/5RoMg4z7**
DTNNetherlands DTN The Netherlands Veel dieren dood door brand kinderboerderij: Door een brand in een kinderboerderij in Middel... **http://t.co/couFmoCH**
DTNNetherlands DTN The Netherlands Veel dieren dood door brand kinderboerderij: Door een brand in een kinderboerderij in Middel... **http://t.co/Gsn0Uh7s**
Leeuwenheart Scheisse.....Veel dieren dood door brand kinderboerderij - **AD.nl: http://t.co/yOE4rqRM**
Speurtraining 100 Ã 150 dieren dood door brand op een kinderboerderij. **Bah!**

Figuur 3.5: (Nieuwsbericht 1) De top 15 reacties gebaseerd op een complexere berekening met de IDF van alle woorden.

In de figuren 3.6 en 3.7 hieronder staan de resultaten van het tweede nieuwsbericht. Hierin is weer hetzelfde te zien als bij het vorige nieuwsbericht, de tweets die minder unieke woorden bevatten staan net iets lager bij de complexere methode (figuur 3.7).

Figuur 3.6: (Nieuwsbericht 2) De 15 reacties met de hoogste score op basis van de methode die unieke woorden telt.

Figuur 3.7: (Nieuwsbericht 2) De top 15 reacties gebaseerd op een complexere berekening met de IDF van alle woorden.

Om de werking van de verbeterde methode beter te illustreren heb ik bij het derde nieuwsbericht een woord uitgezocht dat wel als uniek wordt beoordeeld (de IDF-waarde is hoger dan 2), maar toch een lage score heeft. Zie voor de resultaten figuur B.1 en B.2 in de bijlage (Hoofdstuk A, blz.25). Hier is ook te zien dat bijna alle berichten die dit woord bevatten hetzelfde zijn. Deze reactie hoort dus relatief laag te staan. Wat ook hier het geval is als de complexere methode wordt gebruikt.

Behalve het sorteren van de reacties, heb ik ook besloten dat de reacties geaccentueerd moeten worden. Hiervoor wordt in 2.4.1 dezelfde data gebruikt als voor het sorteren (namelijk de IDF), ik heb echter ook een applicatie ontwikkeld die woorden die tussen 2 unieke woorden staan ook vetgedrukt maakt. Hieronder staan de resultaten voor 5 willekeurige reacties die hiervoor in aanmerking kwamen:

> **Tijgernest** Middelharnis: veel dieren dood bij brand. **Geen "voormalige"** kinderboerderij **dus. Arme** dieren. **nu.nl: WEL slachtoffers! http://t.co/DwuJvbpY**
> **Tijgernest** Middelharnis: veel dieren dood bij brand. **Geen "voormalige" kinderboerderij dus. Arme dieren. nu.nl: WEL slachtoffers! http://t.co/DwuJvbpY**

> **RazioPuntNl Nu.nl** - Tech - Vodafone werkt nog aan herstel netwerk **http://t.co/n391MbNt : zo ontmoet je** nog **eens nieuwe mensen!**
> **RazioPuntNl Nu.nl** - Tech - Vodafone werkt nog aan herstel netwerk **http://t.co/n391MbNt : zo ontmoet je nog eens nieuwe mensen!**

> **MartineBoerkamp Wat enorm heftig:** Het ongeluk op de A35 heeft nog twee levens **geëist.** De **moeder** en **21-jarige dochter uit** Zoetermeer **zijn ook overleden** #a35

MartineBoerkamp Wat enorm heftig: Het ongeluk op de A35 heeft nog twee levens **geëist. De moeder en 21-jarige dochter uit Zoetermeer zijn ook overleden #a35**

Janzeist Elfstedentocht: **IJs** is **nog niet overal dik genoeg** - Vandaag kan het wel eens de beslissende dag worden of er een **16e... http://t.co/2OxWFMew**
Janzeist Elfstedentocht: IJs is nog niet overal dik genoeg - Vandaag kan het wel eens de beslissende dag worden of er een **16e... http://t.co/2OxWFMew**

conavanderhorn reageert op **#NUjij:** Dries Roelvink zoekt ruzie met Frans Bauer op Twitter - **Kan** er **geen verbod** op **die riool... http://t.co/fKmcNP86**
conavanderhorn reageert op #NUjij: Dries Roelvink zoekt ruzie met Frans Bauer op Twitter - **Kan er geen verbod op die riool... http://t.co/fKmcNP86**

3.3.2 Het ordenen van de nieuwsberichten

Voor het sorteren van nieuwsberichten is een formule nodig die verschillende variabele over de data combineert tot één score. Welke formule hiervoor gebruikt wordt hangt af van het precieze doel van de resultaten die gepresenteerd moeten worden. De formule kan handmatig gemaakt worden door te beoordelen welke variabelen belangrijk zijn om de populariteit te berekenen.

Om tot een formule te komen kunnen echter ook verschillende tests gebruikt worden. Voor een aantal verschillende formules kunnen dan verschillende websites gemaakt worden, waarbij bijvoorbeeld op elke website scores uit een enqute worden bijgehouden of er wordt gekeken naar op hoeveel nieuwsberichten een gebruiker klikt of hoelang een bezoeker op een bepaalde pagina van de website blijft. Zie voor meer methodes om verschillende versies van een website te testen (Rogers, Sharp, and Preece 2002).

Deze methodes zijn echter buiten het bereik van dit onderzoek. Om deze testmethodes uit te voeren moet de website langere tijd operatief zijn en moeten er veel mensen de website bezoeken.

Hoofdstuk 4

Conclusie

Om te laten zien hoe we het beste een totaal systeem kunnen maken om de data over nieuwsberichten op twitter automatisch te organiseren gaan we de verschillende onderdelen die het programma heeft bij langs. De interface is hierbij buiten beschouwing gelaten omdat hier maar een standaard model voor is gebruikt en het ontwerpen van een interface niet een doel is van het onderzoek.

4.1 Nieuwsberichten zoeken op twitter

Hiervoor kan er het best een selectie gemaakt worden van twitteraccounts van betrouwbare nieuwsbronnen. Deze methode tweet de nieuwsberichten het duidelijkst, waardoor de rest van het programma beter werkt. Er is niet gebleken dat een van de twee geteste methodes beter werkt (zie sectie 3.1.2). Deze methode werkt ook veel sneller, er zijn echter wel beduidend minder resultaten gevonden. Dit is een gevolg van het volgen van minder accounts en het beter kunnen herkennen van dubbele nieuwsberichten. Ik neem echter aan dat alle (voor de gemiddelde Nederlandse nieuwslezer) echt belangrijke en interessante nieuwsberichten wel door één van de nieuwsbronnen wordt getweet. En dit zijn juist de berichten waar we naar op zoek zijn.

De Jaccard Coëfficiënt is de beste methode om dubbele nieuwsberichten te herkennen. Hiermee worden met de bovenstaande methode 6% meer dubbele nieuwsberichten herkend en dit zonder de precisie te verlagen. Verder kost het het programma niet veel meer moeite, in onze test was deze methode zelfs sneller omdat er minder vergelijkingen hoeven plaats te vinden.

4.2 Reacties op nieuwsberichten vinden op twitter

Om zoveel mogelijk relevante reacties te vinden is een betere methode nodig dan een systeem dat alleen kijkt in welke tweets de exacte nieuwstitel voorkomt. De relevantie moet echter hoog blijven. Na het testen van een methode die naar berichten zoekt die alle kernwoorden uit het nieuwsbericht bevatten bleek dat dit erg goed en robuust werkt. Het percentage relevante berichten blijft hoog en het aantal reacties stijgt behoorlijk (zie sectie 3.2.2). Hieruit kunnen we de conclusie trekken dat dit de beste methode is voor dit deel van het programma.

4.3 Sorteren van reacties op nieuwsberichten

Zoals beschreven in hoofdstuk 3.3.1 werkt de methode die de IDF-scores bij elkaar optelt in plaats van het aantal unieke woorden (IDF-Score > 2) telt net iets beter. Met deze methode worden reacties die meerdere keren voorkomen beter herkend en krijgen ze dus een lagere score.

4.4 Sorteren van nieuwsberichten

Zoals beschreven in 3.3.2 zijn de testmethodes voor verschillende formules buiten het bereik van dit onderzoek, er is namelijk geen sprake van een "juiste" manier om de nieuwsberichten te sorteren. Daarom heb ik ervoor gekozen om de IDF-scores van alle reacties op een bepaald nieuwsbericht bij elkaar op te tellen, en het resultaat van deze berekening te gebruiken om dat bepaalde nieuwsbericht een score te geven, vervolgens kunnen de nieuwsberichten hierop gesorteerd worden.

Door de som van alle IDF-scores van de reacties te gebruiken wordt een nieuwsbericht hoger ingeschat als veel verschillende mensen er veel verschillende dingen over te zeggen hebben. Dit is dus een goede graadmeter voor een website die als doelgroep alle Nederlanders die geïnteresseerd zijn in nieuws heeft. Is het echter belangrijk om betrouwbare nieuwsberichten of alleen echt belangrijke nieuwsberichten te publiceren dan is het wellicht beter om naar het aantal bronnen en/of het aantal reacties te kijken.

Hoofdstuk 5

Toekomstig werk

Ik denk dat er veel potentie in dit concept zit. Op de manier waarop mijn applicatie nu werkt is het al voor veel mensen interessant om de nieuwsberichten en de bijbehorende reacties te lezen. Het baseren van een nieuwsdienst op social media is een goed idee omdat normaal gesproken redacties op basis van gevoel beslissen wat nieuwslezend Nederland interessant vindt. Maar wie weet dit nu beter dan die mensen zelf? Er zijn echter nog veel uitbreidingen mogelijk. Relaties op social media kunnen bijvoorbeeld worden gebruikt om nog specifieker persoonlijk relevante nieuwsberichten te vinden. Of de nieuwsberichten kunnen nog gecategoriseerd worden zodat mensen berichten kunnen lezen die hun persoonlijk interesseren.

Een probleem is echter de copyright en privacy van de nieuwsberichten en tweets. Ik mag niet zomaar automatisch de gehele nieuwsberichten (en plaatjes) van een site kopiren zodat ze op mijn site te lezen zijn. Verder is het maar de vraag of het een goed idee is om tweets van iedereen op te slaan in verband met privacy. Zo blijft de website zelf redelijk oppervlakkig over de nieuwsberichten.

Verder is het sneller om de data op te slaan in het werkgeheugen of in een database. Ik heb ervoor gekozen om alles telkens als tekstbestanden op de harde schijf op te slaan. Zie voor een andere versie van hetzelfde concept (Beimin 2012).

Er is voorlopig een demo beschikbaar op: `http://siegfried.let.rug.nl/s1915770`

Referenties

Beimin, R. (2012). Het laatste nieuws als eerst. implementatie van een recommender system voor nederlandse nieuwsberichten op twitter.

King, K. N. (2000). *Java Programming: From the Beginning*. W.W. Norton & Company.

Manning, C. D., P. Raghavan, and H. Schitze (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.

Naumann, F. and M. Herschel (2010). *An Introduction to Duplicate Detection*. Morgan and Claypool Publishers.

Newcom (2011). vertrouwensindex 2011. Technical report, Newcom Research and Consultancy.

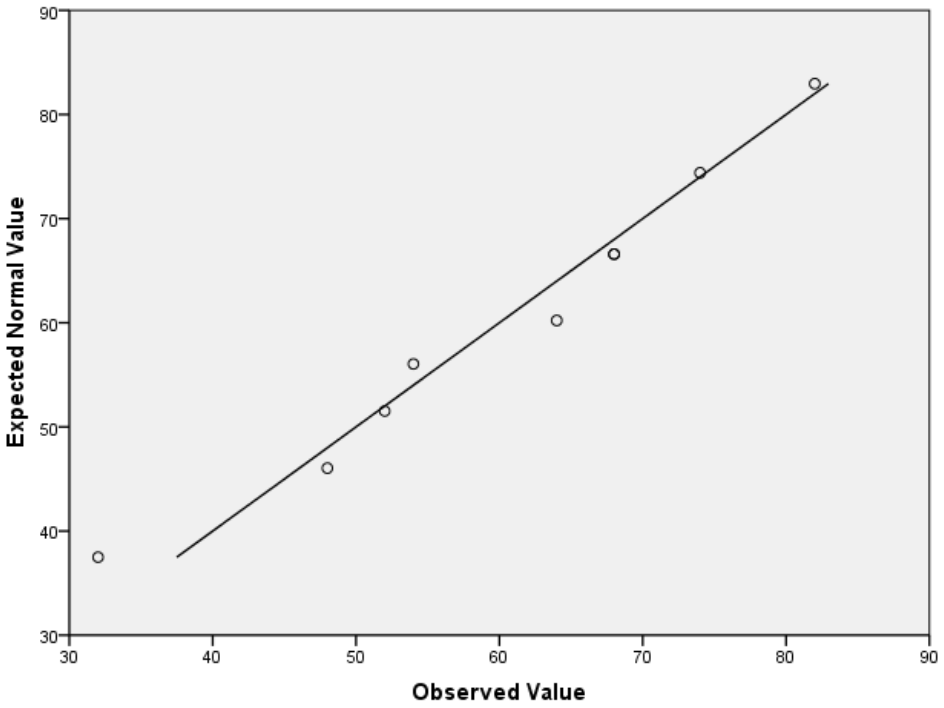
Rogers, Y., H. Sharp, and J. Preece (2002). *Interaction Design: Beyond Human-Computer Interaction*. Chichester, West Sussex, England: John Wiley & Sons.

Tjong-Kim-Sang, E. (2011). Het gebruik van twitter voor taalkundig onderzoek. *TABY: Bulletin voor Taalwetenschap 39*(1/2), 62–72.

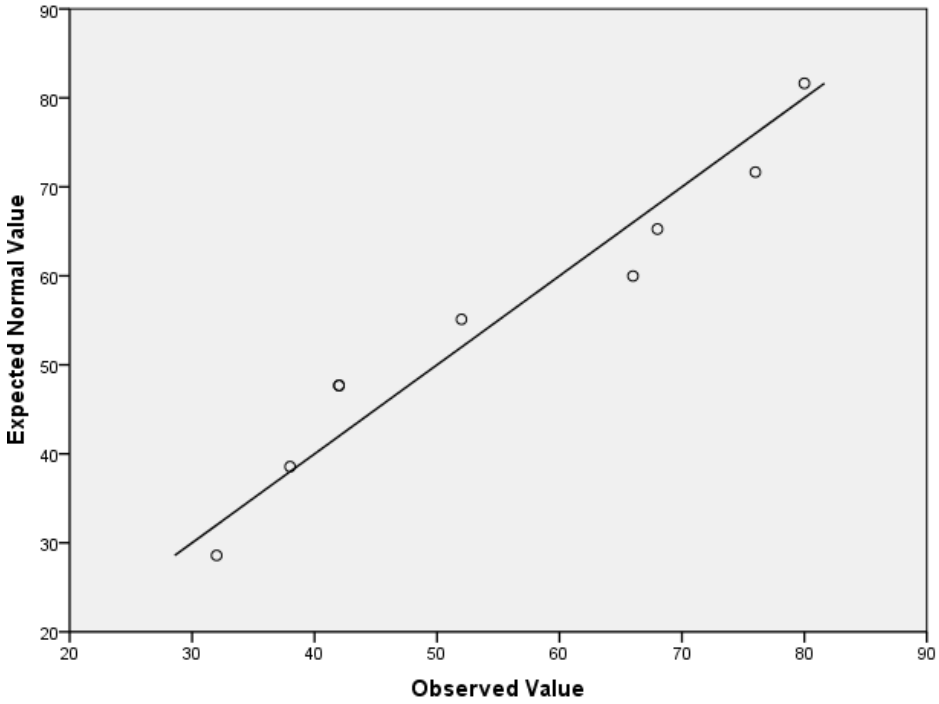
Twitter (2010). How do i get whitelisted. `http://support.twitter.com/entries/160385-how-do-i-get-whitelisted`. [Geraadpleegd 27-05-2012].

Wikipedia (2012). Lijst van dagbladen. `http://nl.wikipedia.org/wiki/Lijst_van_dagbladen`. [Geraadpleegd 27-05-2012].

Bijlage A



Figuur A.1: Normaalplot voor de hashtag resultaten2.



Figuur A.2: Normaalplot voor de nieuwsbron resultaten.

Bijlage B


L1vR Brandweer redt kind uit auto in water http://t.co/UFAQhUAr via @nunl #tJa Gordels, meestal levens reddend. zou een touwtje aan moeten zitten
lauradaphneee Brandweer redt kind uit auto in water http://t.co/9baqcKCi. Kindje zat nog vast door gordels autostoeltje.. Ja heel veilig hoor,nachtmerrie!
carolienebeltje Brandweer redt kind uit auto in water http://t.co/If8S10Kh ik zou denk ik wel in het water zijn gesprongen om hun te gaan helpen
brwroermond Maar goed dat er duikers in de buurt zijn!! Brandweer redt kind uit auto in water in Weert Â» 112 Limburg: http://t.co/LeyOCw3K via @AddThis
Tijgernest Weert: brandweer redt kind uit auto te water, omstanders was dat niet gelukt http://t.co/kJP2PBoH complimenten dat men hulp verleende!
Limburgreporter Brandweer Weert redt kind uit auto te water: WEERT â€“ De politie wil langs deze weg haar waardering uitspreken vo... http://t.co/tekdhDPW
RSSjoeri Brandweer redt kind uit auto in water: Een auto met drie inzittenden is zondagmiddag in het kanaal bij de Indust... http://t.co/bs5MIj58
maastrichtinfo Brandweer redt kind uit auto in water: Het kind moest worden gereanimeerd en is in zorgwekkende toestand naar he... http://t.co/rMHPFRa7
Rallyrobert Ik heb net " Brandweer redt kind uit auto in water" via #ADnl gelezen. http://t.co/HehxUrt4
112LimburgTwitt Nu online: Brandweer redt kind uit auto in water in Weert. Heeft een FOTO Raportage Kijk op http://t.co/z1PS0t6x #Weert #Industriekade
VeiligNL Brandweer redt kind uit auto in water: Twee andere inzittenden, de vaderâ€¦ http://t.co/rvTs3xkz â–¸ #Veilig #Nederland
MijnRoermond Brandweer redt kind uit auto in water: Een auto met drie inzittenden is zondagmiddag in het kanaal bij deâ€¦ http://t.co/Yq125hg6
112LimburgTwitt Brandweer redt kind uit auto in water in Weert. Heeft een NIEUWS Update. Kijk op http://t.co/z1PS0t6x #Weert #Ongeval #Industriekade
sjaqweer Goede zaak, brandweerduikers : (via @NUnl) Brandweer redt kind uit auto in water: http://t.co/Is0TP8qi
Limburgreporter Video: Brandweer Weert redt kind uit auto te water | Limburgreporter: http://t.co/bcAf4LwV via @AddThis
Limburgreporter Video: Brandweer Weert redt kind uit auto te water | Limburgreporter: http://t.co/SxViMwjs via @AddThis
WatGebeurtErNu Nu op de wereld: Brandweer redt kind uit auto in water: De brandweer heeft in Weert een kind ger... http://t.co/TqkB6y16 #watgebeurternu
feedNL FOK! â–¸ Brandweer redt kind uit auto in water http://t.co/E5y8luOS #columns #forum #NL
DTNNetherlands DTN The Netherlands Brandweer redt kind uit auto in water: De brandweer heeft zondagmiddag in Weert een kind ger... http://t.co/SBhm1fFJ
nieuws247 [anp] Brandweer redt kind uit auto in water http://t.co/0npeZAaE [nederland nieuws]
nieuws_nunl [NEWS] Brandweer redt kind uit auto in water: WEERT - De brandweer heeft zondagmiddag in Weert een kind ... http://t.co/d3nALxi2 [/NEWS]
jeadvocaat helden bestaan! â€œ@NUnl: Brandweer redt kind uit auto in water: http://t.co/jECDYT09â€
inmijneendje Fok!! Nieuws: Brandweer redt kind uit auto in water http://t.co/jeKxUhsC #inmijneendje
Amsterdam_Nu #amsterdam Brandweer redt kind uit auto in water http://t.co/yuzXha6t #nieuwsamsterdam #nieuws
JasonDoorson â€œ@NUnl: Brandweer redt kind uit auto in water: http://t.co/npOKtVE8â€ toch weer duikers brandweer.
provincienieuws - Provincie Nieuws - Brandweer redt kind uit auto in water: http://t.co/FL2ssoDO #nieuws #limburg
newsbinnenland #binnenlandsnieuws Brandweer redt kind uit auto in water http://t.co/n3WPZimv
nieuwsheadline Brandweer redt kind uit auto in water http://t.co/w2Zo9q66 #nieuws #nederland
supernuttig Brandweer redt kind uit auto in water: WEERT - De brandweer heeft zondagmiddag in Weert een kind ge... http://t.co/VD5SrmDy #supernuttig
Nieuws_alerts Brandweer redt kind uit auto in water: De brandweer heeft in Weert een kind gered uit een auto die in het kanaal... http://t.co/rZTXoC6s
NieuwsfeedNL Brandweer redt kind uit auto in water http://t.co/LTIWexpm @volkskrant
NLtijdschrift #Libelle Brandweer redt kind uit auto in water: WEERT - De brandweer heeft zondagmiddag in Weert een kind gered ... http://t.co/UjHMfhED
RTLNieuwsnl Brandweer redt kind uit auto in water: De brandweer heeft in Weert een kind gered uit een auto die in het kanaal... http://t.co/lKKj8UhW
nieuwsgaring Brandweer redt kind uit auto in water: De brandweer heeft in Weert een kind gered uit een auto die in het kanaal... http://t.co/UNnAKqmS
headlinesnieuws Brandweer redt kind uit auto in water: De brandweer heeft in Weert een kind gered uit een auto die in het kanaal... http://t.co/yaDza34G
Hetnieuws24 Brandweer redt kind uit auto in water: De brandweer heeft in Weert een kind gered uit een auto die in het kanaal... http://t.co/nMQA9qWs
WatIsErNu Brandweer redt kind uit auto in water: De brandweer heeft in Weert een kind gered uit een auto die in het kanaal... http://t.co/B3EWlV1F
nieuwsflitser Brandweer redt kind uit auto in water http://t.co/TeiLaKUZ #Nieuwsflitser
dirk18041965 Brandweer redt kind uit auto in water http://t.co/MttpbFNs via @nunl
bogobogo_nieuws Brandweer redt kind uit auto in water: De brandweer heeft zondagmiddag in Weert een kind gered ... http://t.co/qctycBUO #nieuws #actueel
MyNewsGathering #MNG Brandweer redt kind uit auto in water http://t.co/Byj69jUe
NL_actueel #NL_actueel Brandweer redt kind uit auto in water - WEERT - De brandweer heeft zondagmiddag in Weert een kind gered ... http://t.co/j94bh9eK
Onlinejournaal New: Brandweer redt kind uit auto in water - Brandweer redt kind uit auto in water WEERT De brandweer heeft zondagm... http://t.co/pcVMpINs
HetSportForum Brandweer redt kind uit auto in water: De brandweer heeft zondagmiddag in Weert een kind gered ui... http://t.co/gjpi5CFE #nieuws #sport
roynefs Brandweer redt kind uit auto in water #nuiphone #in http://t.co/LSaofVft
WilbertTintel Brandweer redt kind uit auto in water http://t.co/bsDmthGC #Weert #Limburg

Figuur B.1: (nieuwsbericht 3) Een hele pagina van resultaten (simpelere methode) met alle voorkomens van het woord 'kanaal' ingekleurd.

Figuur B.2: (nieuwsbericht 3) Een hele pagina van resultaten (complexere methode) met alle voorkomens van het woord 'kanaal' ingekleurd.