Leefpatronen in Twitterberichten van actieve gebruikers

C.W.T.P. (Christiaan) Thijssen
e-mail: christiaan.thijssenstudent.ru.nl

Begeleider: prof. dr. ir. Th.P. (Theo) van der Weide

25 juni 2011

Inhoudsopgave

Abstract

Twitter is een zeer populair medium dat nog steeds een groeiend aantal gebruikers kent. Mensen kunnen via dit medium kleine berichten publiceren en berichten van anderen bekijken. Al deze berichten zijn (in beginsel) volledig openbaar. Tegenwoordig beschikken de meeste smartphones ook over de functionaliteit om Twitterberichten te publiceren. Hierdoor gaan vooral jongeren steeds meer berichten op Twitter plaatsen en wordt er veel persoonlijke informatie verstrekt. In dit document presenteer ik mijn vermoeden dat je uit de verzameling van twitterberichten van een gebruiker zijn leefpatronen af kunt leiden en (delen van) de agenda van deze gebruiker kunt terugzien. Ik formaliseer dit vermoeden tot een theorie en ga vervolgens bekijken aan de hand van een gedownloade dataset van Twitterberichten of er voldoende informatie over leefpatronen te vinden zijn in de berichten. Vervolgens ga ik kijken welke oplossing het beste gekozen kan worden om deze leefpatroongegevens geautomatiseerd te extraheren uit de set van berichten van een bepaalde gebruiker.

1 Inleiding

Twitter [15] is een van de populiarste microblogging netwerken. Mensen kunnen via dit netwerk korte berichten van maximaal 140 tekens publiceren op internet onder hun Twitter gebruikersnaam maar vaak vermelden gebruikers ook hun volledige naam in hun profiel. Uit het jaarlijkse onderzoek genaamd The state of the Twittersphere [3] door Kathryn Corrick blijkt dat in begin 2011 Twitter ongeveer 200 miljoen gebruikers kent. Dagelijks worden er door al deze gebruikers 110 miljoen berichten geplaatst. In principe zijn alle berichten die je via Twitter deelt te bekijken door alle gebruikers van Twitter. Daarnaast zijn alle Twitterberichten van een bepaalde persoon te vinden op internet door met een zoekmachine te zoeken naar de naam van deze persoon. Een selectie van de meest recente publicaties van deze persoon verschijnen dan als een van de eerste zoekresulaten. Je publiceert al je berichten dus echt voor de hele wereld, ze zijn voor anderen makkelijk te vinden en de berichten zijn volledig openbaar. Het is wel mogelijk om je berichten te beveiligen, zodat alleen mensen waar jij toestemming voor hebt gegeven je publicaties kunnen bekijken, maar deze functie wordt enkel door een minderheid gebruikt. De gebruiker kan vervolgens medegebruikers aan zijn Twitteraccount toevoegen waarvan hij hun berichten op de hoofdpagina van Twitter kan lezen. De gebruiker volgt op deze manier alle berichten die de andere gebruikers delen. Alle berichten van deze medegebruikers die de gebruiker dan volgt verschijnen in chronologische volgorde in deze zogenaamde timeline. Op deze manier kunnen gebruikers informatie over dagelijkse dingen met elkaar delen. Maar ook nieuwsfeiten (zowel lokaal als landelijk) kunnen via dit medium snel gedeeld worden. Het hoofdzakelijke idee achter Twitter is dat mensen korte stukjes informatie delen over wat ze op een bepaald moment aan het doen zijn, waardoor kennissen en vrienden meer te weten kunnen komen over elkaars hobby's en interesses. Het is ook mogelijk om berichten enkel aan bepaalde gebruikers te richten. Op deze manier is het mogelijk gesprekken via dit medium te voeren. Deze gerichte berichten zijn nog steeds publiek, maar de persoon aan wie je ze richt krijgt hier een extra notificatie van.
Al met al lijkt Twitter dus een zeer krachtig communicatiemiddel. Maar is het publiceren van dergelijke kleine berichten zo ongevaarlijk als wij denken? Of moeten we toch nog steeds in ons achterhoofd blijven houden dat de berichten volledig publiek zijn en moeten we uitkijken welke informatie we met de rest van de wereld delen?

2 Vermoeden

Twitter word ook steeds populairder onder jongeren. Steeds meer jongeren beschikken over een smartphone waarmee ze overal toegang hebben tot het internet. Dit zorgt er niet alleen voor dat ze overal op de hoogte kunnen zijn van nieuwsfeiten of Twitterberichten van vrienden, het zorgt er ook voor dat ze steeds makkelijker informatie delen over wat ze op een bepaald moment aan het doen zijn. Waar je vroeger moest wachten met het plaatsen van een Twitterbericht totdat je weer achter de computer zat, je nu direct met de mobiele telefoon een bericht kunt publiceren. [1] Het is dan ook steeds meer een trend aan het worden om maar zoveel mogelijk Twitterberichten te plaatsen op internet over de meest uiteenlopende zaken. Hierdoor zijn de berichten vaak minder doordacht en lijken de berichten ook veel minder informatie te bevatten. Waar voorheen alleen een Twitterbericht werd geplaatst over wat gebruikers de hele dag gedaan hadden, wordt

[1] Uit de door mij verzamelde dataset met 1.066.082 Twitterpublicaties blijkt dat ongeveer 40% van de berichten afkomstig was van een mobiel apparaat.

nu voor vrijwel elk moment van de dag een bericht geplaatst. De informatie in de Twitterberichten worden hierdoor steeds trivialer. De gebruikers laten bijvoorbeeld weten dat ze gaan eten, slapen of gaan werken. Dit lijkt heel erg onschuldig en nietszeggend, maar wellicht geeft dit een gigantische hoeveelheid informatie over iemands leven wanneer je deze nietszeggende berichten combineert en de tijd waarop ze zijn geplaatst in achting neemt. Wanneer gebruikers van Twitter namelijk op een dergelijke precisie (van enkele tientallen minuten) publiceren wat ze aan het doen zijn, zijn ze eigenlijk een soort logboek aan het bijhouden. In dergelijke logboeken kan er vervolgens gezocht worden naar patronen. Op basis van deze patronen is het dan ook mogelijk om een voorspelling te doen van een toekomstige weekplanning van een Twittergebruiker. Veel Twittergebruikers zullen zich er dan ook niet van bewust zijn dat ze zoveel informatie prijsgeven over hun dagindeling en leefgewoonten. Wanneer je deze mensen ook zou vragen om hun agenda openbaar te maken zouden ze dit in veel gevallen absoluut niet doen. Toch doen ze dit misschien via Twitterberichten onbewust al. Ik vermoed dus dat er bij veel actieve Twittergebruikers veel informatie over hun leefgewoonten en dagindeling zichtbaar zal worden wanneer je geautomatiseerd de Twitterberichten analyseert en vervolgens een weekplanning hiervan genereert.

2.1 Onderzoeksvraag

De onderzoeksvraag die in mijn onderzoek centraal zal staan is:
Op welke manier kun je een tool ontwerpen die geautomatiseerd leefpatronen uit een verzameling Twitterpublicaties van actieve gebruikers kan halen?

2.2 Variabelen

- **ontwerp van een tool**
  Dit is een afhankelijke variabele van mijn onderzoeksvraag. Dit onderzoek zal een blauwdruk van een tool geven waarmee leefpatronen in Twitter te analyseren zijn. Daarnaast zullen deelconcepten van de blauwdruk verder uitgewerkt zijn (met proeven aan de data) waarmee ik de waarschijnlijkheid van de werking van deze concepten zal schetsen. Bij de afweging of het daadwerkelijk de moeite waard is om een dergelijke tool te ontwerpen zal ik kijken naar de verwachte prestatie maar ook naar de verwachte executietijd. Wanneer de tool namelijk dagen bezig is om voor een gebruiker een conclusie te vormen of dat de conclusie niet duidelijk genoeg kan zijn op basis van de beschikbare informatie, is het waarschijnlijk niet de moeite waard om de tool verder te ontwikkelen.

- **Leefpatronen**
  Met leefpatronen bedoel ik gebeurtenissen die veelvuldig in het dagelijkse leven van iemand voorkomen. Dit kan bijvoorbeeld zijn op welke tijdstippen iemand meestal eet, op welke tijdstippen een persoon meestal gaat sporten of gaat werken enz. Deze patronen zal de software waar ik een blauwdruk voor wil gaan geven als resultaat opleveren. De tool zal in het beste geval een weekplanning van een gebruiker weergeven die opgebouwd is uit de informatie uit een verzameling Twitterberichten van deze gebruiker. Het doel is dan om deze tool zo te ontwerpen dat de gegeven weekplanning zo veel mogelijk overeen komt met de daadwerkelijke agenda van de gebruiker.

- **Twitterpublicaties**
  Dit is een onafhankelijke variabele in mijn onderzoeksvraag. Het betreft hier de berichten die een Twittergebruiker plaatst op zijn microblog over de gebeurtenissen die hij of zij meemaakt. Deze berichten zullen per Twittergebruiker verschillen en daarom zullen we onze analyse ook uit moeten voeren op meerdere Twitterpublicaties van meerdere actieve Twittergebruikers.

- **Actieve gebruikers**
  Dit is ook een onafhankelijke variabele in mijn onderzoeksvraag. Zoals hierboven al aangegeven, moeten we voor meerdere personen kijken welke leefpatronen we in de verzameling Twitterberichten van deze persoon kunnen vinden. Dit omdat we anders moeilijk kunnen generaliseren naar een grotere groep. We willen echter alleen kijken naar Twitterberichten in het Nederlands. Deze beperking moeten we onszelf opleggen omdat we bij dit onderzoek aspecten moeten toepassen die met taalanalyse te maken hebben. Wanneer we het onderzoek toe willen passen op gebruikers die verschillende talen gebruiken moeten we ook een soort vertaallaag in de tool inbouwen en dit valt buiten de scope van het onderzoek. Deze afbakening houdt dus ook in dat we maximaal kunnen generaliseren naar de groep van Nederlandse Twittergebruikers. Daarnaast kijken we hier expliciet naar actieve gebruikers. Dit houdt in dat

de gebruikers zeer veel Twitterberichten moeten publiceren om in aanmerking te komen voor ons onderzoek. We willen namelijk de gebruikers gaan bekijken die veel Twitterberichten plaatsen die weinig inhoudelijke informatie lijken te bevatten, maar wanneer gecombineerd juist veel informatie bevatten omdat de Twitterberichten een hoge dichtheid in tijdsopeenvolging hebben. De gebruikers waar wij ons op willen richten, sturen gemiddeld elk half uur een bericht over wat zij aan het doen zijn. Onze ultieme testpersonen hebben dus een mobiele telefoon waarmee ze de meeste dagelijkse gebeurtenissen gemiddeld twee keer per uur verwoorden in Twitterberichten.

2.3 Methode

Eerder heb ik (in mijn onderzoeksplan) gesteld dat ik een tool ga ontwerpen en implementeren waarmee we geautomatiseerd leefpatronen kunnen ontdekken in een verzameling van Twitterberichten van een bepaalde Twittergebruiker. Hierbij maken we gebruik van de informatie in een Twitterbericht en de tijd waarop het Twitterbericht is gepubliceerd. Wanneer de implementatie van deze tool was afgerond, zou ik de kwaliteit van de uitvoer van de tool gaan analyseren door middel van een enquêtering onder Twittergebruikers waarvan een weekagenda is gegenereerd met behulp van de ontwikkelde tool. Toen ik aan het onderzoek begon bleek echter al snel dat de in het onderzoeksplan gestelde doelstellingen toch te hoog waren voor de beschikbare tijd en dat er nog heel wat keuzes gemaakt moeten worden tijdens een vooronderzoek voordat we aan een implementatie van de beoogde tool kunnen beginnen. Dit document kan dus als een verslag van een vooronderzoek beschouwd worden dat eventueel uitgebreid kan worden in een vervolgonderzoek. Ik zal mij in dit document richten op de vraag of Twitterberichten voldoende informatiewaarde met betrekking tot leefpatronen hebben en hoe een tool ontworpen kan worden om leefpatronen te ontdekken in Twitterberichten. Wanneer uit dit document blijkt dat het nog steeds interessant lijkt om door te gaan met het ontwikkelen van een dergelijke tool, kan in een implementatie van de tool plaatsvinden en een vervolgonderzoek gestart worden op de manier zoals deze voorheen geschetst is in het onderzoeksplan. Deze afweging is dan ook gemaakt in de conclusie in sectie 11 op pagina 39.

In dit document zullen we daarom allereerst kijken naar de relevantie van het onderzoek doen naar geautomatiseerd leefpatronen ontdekken in zijn algemeenheid. Vervolgens zal ik kort bekijken wat voor relevante onderzoeken er al gedaan zijn met betrekking tot het Twitter medium en het vermoeden dat ik eerder heb gepresenteerd formaliseren tot een theorie. Met deze theorie als leidraad zal ik daarna twee aspecten gaan bekijken in dit vooronderzoek. Namelijk of de informatie over leefpatronen te vinden is in verzamelingen Twitterberichten van Twittergebruikers (Leefpatrooninformatie in Twitterberichten, sectie 9 op pagina 18) en hoe we deze informatie over leefpatronen geautomatiseerd kunnen verkrijgen uit deze verzameling Twitterberichten (Semantische Analyse, sectie 10 op pagina 27). Verschillende tests zullen gedaan worden aan de hand van een verzameling Twitterberichten van verschillende Twittergebruikers. Informatie over deze set en over de manier van verzamelen van deze informatie wordt behandeld in sectie 8 op pagina 16.

3 Verantwoording

3.1 Aanleiding

De reden dat ik bovenstaande wil onderzoeken is dat ik vermoed dat wanneer je van een gebruiker de informatie uit alle Twitter berichten combineert, je een grote hoeveelheid informatie over die persoon verkrijgt. Ik zie steeds vaker dat jongeren zo veel mogelijk Twitterberichten willen publiceren en deze berichten hebben dan vaak een geringe diepgang. Ze lijken daarom niets of weinig prijs te geven van het leven van de gebruiker, maar ik verwacht dat dit juist veel informatie verschaft over de persoon wanneer je deze berichten combineert en in verband brengt met het tijdstip waarop het bericht is gepubliceerd. Ik vrees dat de meeste Twittergebruikers hier niet bij stilstaan wanneer ze dergelijke berichten met geringe diepgang herhaaldelijk publiceren. Toch kan het zo zijn dat ze hiermee onbewust hun eigen privacy ondermijnen.

3.2 Relevantie

3.2.1 Privacy

Veel gebruikers van Twitter plaatsen berichten over de meest uiteenlopende zaken, vaak zonder dat ze er bij stil staan dat deze informatie openbaar is. De berichten zijn afzonderlijk misschien niet van grote pri-

vacygevoelige waarde, gecombineerd kunnen ze dit misschien wel zijn. Het is ook niet gebruikelijk dat mensen berichten verwijderen. Wanneer ze geplaatst zijn, worden ze vergeten en weet de gebruiker vaak niet meer wat hij ooit gepubliceerd heeft. Op deze manier kun je uit een reeks berichten van iemand wellicht privacygevoelige patronen halen waarvan de gebruikers zich niet van bewust waren dat ze deze informatie publiceerden. Wanneer iemand bijvoorbeeld regelmatig bericht wanneer hij naar zijn werk gaat met de bus, of in de file staat met de auto naar zijn werk, kan door het combineren van al deze berichten en met de tijd wanneer deze Twitterberichten geplaatst zijn, wellicht bepaald worden wanneer deze persoon aan het werk is. Op deze manier zou dus het hele werkrooster van een persoon achterhaald kunnen worden. Dit terwijl de persoon in kwestie waarschijnlijk bezwaar zou maken als iemand zijn werkrooster op internet zou publiceren. Met andere woorden: Het lijkt erop dat gebruikers van Twitter er zich niet goed van bewust zijn welke informatie ze in totaliteit publiceren. Ze bouwen echter zelf aan een uitgebreide profilering van hun leven. Met de beoogde tool kun je hier dan ook meer duidelijkheid over verschaffen. Wanneer blijkt dat er meer informatie uit de Twitterberichten te halen valt dan men aanvankelijk dacht, kan dit een impuls voor de gebruikers zijn om beter stil te staan bij wat ze publiceren en kunnen ze overwegen hun Twitterberichten te beveiligen. Wanneer de gebruikers de Twitterberichten beveiligen zijn ze niet meer openbaar. Alleen gebruikers waaraan toestemming is gegeven kunnen de berichten dan nog inzien.

Het is belangrijk dat mensen zich realiseren dat alles wat ze op internet publiceren en openbaar is ook gebruikt kan worden door kwaadwillenden. Martin Bryant publiceerde op zijn blog [1] redenen waarom je geen geografische informatie moet toevoegen aan je Twitterberichten. De belangrijkste redenen waren dat het mogelijk is dat je Twitterberichten publiceert met de strekking dat je een tijd op vakantie bent en je in hetzelfde (of in een ander) Twitterbericht de geografische ligging van je huis hebt geplaatst (door een bericht te plaatsen wanneer je thuis bent en geografische gegevens meestuurt). Het is dan mogelijk dat criminelen bij je thuis inbreken omdat ze uit je Twitterberichten hebben kunnen halen waar je woont en wanneer je niet thuis bent. Dit is mede omdat je niet kunt bepalen wie je geografische gegevens wel en niet mogen inzien. Wanneer je er dus voor kiest niet al je Twitterberichten te beveiligen, kan ook iedereen je geografische locatie inzien.

3.2.2 Profileringen

Naast het belang om te onderzoeken of gebruikers van Twitter onbewust hun eigen privacy aantasten, is er nog een mogelijk wetenschappelijk belang bij de resultaten. Als ik duidelijke patronen kan ontdekken in leefpatronen van mensen, kan dit namelijk ook in andere onderzoeken gebruikt worden. Het is bijvoorbeeld goed mogelijk dat ik een duidelijk patroon vindt voor de vrijetijdsbesteding van Twitter gebruikers. Wanneer de Twittergebruikers die ik onderzocht heb divers genoeg zijn, kunnen we generaliseren naar een grotere groep mensen (Nederlanders bijvoorbeeld). De leefpatronen die ik uit de Twitterberichten zou kunnen halen zijn een goede grondslag voor een profilering van mensen. Deze informatie kan vervolgens van belang zijn bij een onderzoek van de sociale wetenschappen naar de gemiddelde dagindeling of leefgewoonten van mensen. De resultaten kunnen ook interessant zijn voor de commercie. Het is voor de commercie interessant om te weten wat mensen op welk tijdstip vaak doen, zodat ze hierop in kunnen spelen met hun marketingstrategieën. De tools die ik ontwikkel om resultaten te verkrijgen zal ik publiceren tezamen met een verantwoording hoe ik tot de implementatie van die tools gekomen ben. Op deze manier kunnen deze tools ook voor commerciële doeleinden gebruikt worden.

3.2.3 Forensisch onderzoek

De tool waarmee leefpatronen van Twittergebruikers te ontdekken zijn kan ook nuttig zijn voor forensisch onderzoek. Wanneer verdachten of slachtoffers van een misdrijf ook Twitter gebruiken om informatie te verstrekken over wat er gaande is, kan de recherche wellicht informatie halen uit de leefpatronen van deze personen en hier aanwijzingen uit halen die het onderzoek naar het misdrijf kunnen ondersteunen. Op basis van de patronen die gevonden kunnen worden in Twitterberichten, kan dan voorspeld worden wat de personen in kwestie normaal gesproken aan het doen waren op het moment dat het misdrijf plaatsvond. Aangezien de Twitterberichten volledig openbaar zijn, mag de politie deze ook zonder toestemming doorzoeken. Het zal niet zo zijn dat er veroordelingen op basis van Twitterberichten plaats zullen gaan vinden, aangezien niet te bewijzen is dat de verdachte Twittergebruiker ook daadwerkelijk het bericht geplaatst heeft, maar het kan de politie wellicht wel aanwijzingen verschaffen. Op deze manier is wellicht meer inzicht te krijgen in de aanleiding en situatie van het misdrijf.

4 Eerder onderzoek naar Twitter

Er wordt op dit moment zeer veel onderzoek gedaan naar Twitter. Dit is waarschijnlijk omdat Twitter een relatief nieuw medium is en omdat het momenteel zeer sterk in populariteit toeneemt. Daarnaast is Twitter een voorbeeld van sociale media en dus op het moment erg populair.

4.1 Gebruik van Twitter

Uit een onderzoek van Bernando A. Huberman en zijn team [10] uit 2008 blijkt namelijk onder andere dat gebruikers gemiddeld 255 Twitterberichten geplaatst hebben sinds de ingebruikname van de microblogging dienst. Gemiddeld was dit gedurende een tijdsbestek van 206 dagen. Dit houdt dus in dat gebruikers gemiddeld net iets meer dan een bericht per dag plaatsen op Twitter. Echter, dit onderzoek is in 2008 gehouden, wanneer Twitter lang nog niet zo populair was onder jongeren en smartphones nog in opkomst waren. Ik verwacht dan ook dat dit gemiddelde sterk is gestegen sinds dit onderzoek. Een ander gegeven dat blijkt uit dit onderzoek is dat grofweg een derde van de gebruikers enkel Twittergebruiker is om berichten van anderen te lezen en zelf nooit berichten op hun microblog plaatst. Ook was ongeveer 25 procent van de Twitterberichten direct gericht aan een andere gebruiker (waardoor niet iedereen dit bericht gelijk te zien krijgt). Een vierde van de Twitterberichten wordt dus gebruikt voor directe communicatie tussen Twittergebruikers. Een van de belangrijkste vindingen van Huberman[10] is dat het aantal gebruikers dat de berichten van een persoon volgt de belangrijkste drijfveer is om Twitterberichten te publiceren. Iemand met veel volgers is ook extra actief met het plaatsen van berichten.

4.2 Doelen van Twitterpublicaties

Akshay Java en zijn team onderzochten in 2007 [11] hoe en waarom mensen microblogdiensten gebruiken en dit onderzoek was vooral toegespitst op Twitter. Java en zijn team ontdekten dat het soort Twitterberichten onder te verdelen valt in vier groepen met een verschillend doel, De doelen zijn:

- Rapportage over dagelijkse gebeurtenissen
- Voeren van gesprekken
- Delen van informatie (zoals url's)
- Nieuwsfeiten verspreiden

Hierbij was de eerste groep (met berichten over dagelijkse gebeurtenissen) veruit het grootst. De meeste Twittergebruikers berichten dus over dagelijkse dingen. Dit is voor mijn onderzoek dan ook de belangrijkste groep. Voor de beoogde tool moet er uit de verzameling van Twitterberichten juist die berichten gefilterd worden die over dagelijkse gebeurtenissen gaan. Hierbij moeten dus de berichten die tot de groep van het delen van informatie en nieuwsfeiten verspreiden behoren worden uitgesloten. Mogelijk zijn de Twitterberichten die mensen aan andere gebruikers direct richten (behorende tot de groep "voeren van gesprekken")ook bruikbaar. Gesprekken kunnen namelijk ook betrekking hebben op iets waar mensen mee bezig zijn of wat ze in de nabije toekomst zullen ondernemen. Daarnaast lijkt het nodig dat minder relevante berichten over bijvoorbeeld gedachten of gevoelens worden gefilterd. Het filteren van de bruikbare informatie zal de grootste uitdaging bij het ontwerpen van de beoogde tool zijn, aangezien je op taalkundig niveau niet makkelijk een onderscheid kunt vinden tussen nieuwsberichten en algemene berichten en berichten die de Twittergebruiker zelf verzonnen heeft.

4.3 Twittergebruikers als sociale sensors

Takenshi Sakaki, Makoto Okazaki en Yutaka Matsuo hebben in augustus 2010 een ontwerp gepubliceerd [14] voor het gebruiken van het medium Twitter als waarschuwingssysteem voor natuurgeweld. Zij richtten zich vooral op een aardbeving en een tyfoon omdat deze in hun leefgebied (Japan) redelijk frequent voorkomen. Hun idee is als volgt:

- Wanneer er een Tyfoon of aardbeving wordt geconstateerd door mensen, zullen sommigen hierover een bericht op Twitter plaatsen.
- Er zal een toename zijn van berichten die betrekking hebben op aardbevingen of Tyfonen.

- Geautomatiseerd kunnen Twitterberichten real-time geanalyseerd worden door te zoeken naar woorden als "Aardbeving", "schudden" en "Tyfoon". Op basis van een semantische analyse kan er dan bepaald worden of het daadwerkelijk over natuurgeweld gaat. Wanneer er een grote toename van berichten over aardbevingen of tyfonen in een bepaald gebied zichtbaar is, kan er geconcludeerd worden dat er natuurgeweld ontstaan is. In combinatie met de GPS coördinaten die bij dit bericht zitten kan een ruimtelijk model gemaakt worden en op basis daarvan kan berekend worden welke gebieden in de nabije toekomst aangetast zullen worden door dit natuurgeweld.
- In de gebieden waar verwacht wordt dat het natuurgeweld daar naar toe zal trekken worden mensen gealarmeerd zodat ze voorzorgsmaatregelen kunnen treffen tegen het natuurgeweld (schuilplaats opzoeken of iets dergelijks).

Met dit idee, dat ze hebben uitgewerkt tot een algoritme, hebben ze aangetoond dat je Twittergebruikers kunt zien als sensoren en op basis van het uitlezen van gegevens van de sensoren (het analyseren van Twitterberichten) je verschillende real-time gegevens kunt verkrijgen. Omdat mensen die berichten op sociale netwerken plaatsen hier gezien worden als de sensoren in een meetsysteem, noemen de auteurs dit gegevens uitlezen van sociale sensoren.

5 Theorie

Hieronder zal ik mijn vermoeden gaan formaliseren tot een theorie.

5.1 Verzamelingen

Ik zal binnen de formalisering uitgaan van verschillende verzamelingen. Hieronder zal ik definiëren welke verzamelingen ik gebruik.

- U: De verzameling van (actieve) Twittergebruikers.
- T: De verzameling van tijdsmomenten (in seconden).
- E: De verzameling van gebeurtenissen (die de huidige Twittergebruiker heeft meegemaakt).
- B: De verzameling van Twitterberichten.
- DB: De verzameling van alle Deelverzamelingen van Twitterberichten

5.2 Predicaten

- $occurence(e : E, t : T)$: Predicaat die waar oplevert wanneer de meegegeven gebeurtenis e voorkomt op meegegeven tijdstip t.
- $report(b : B, t : T)$: Predicaat die waar oplevert wanneer het meegegeven Twitterbericht b wordt geplaatst op meegegeven tijdstip t.
- $describes(b : B, e : E)$: Predicaat die waar oplevert wanneer het meegegeven Twitterbericht b de meegegeven gebeurtenis e omschrijft. Op dit punt komt het aan op semantische analyse van de berichten. Voor een mens is het relatief makkelijk te bepalen of een bericht behoort tot een onderwerp (en dus eventueel een leefpatroon) of niet, maar om dit geautomatiseerd te doen is een stuk lastiger. Hier komen we nog verderop in dit document op terug.

5.3 Functies

- $UserTweets(u : U, DB)$: Geeft de grootste verzameling van Twitterberichten terug die allemaal afkomstig zijn van de gegeven Twittergebruiker u.
- $addWeek(t : T)$: Deze functie voegt (ongeveer) een week in seconde toe aan t. Deze functie gebruiken we om aan te geven dat een gebeurtenis behorende tot een patroon wekelijks moet terugkeren.

- $patternCollection(e : E, DB)$: Deze functie geeft de grootste deelverzameling terug met Twitterberichten die behoren tot een Twitterpatroon van e (en afkomstig zijn van de te analyseren Twittergebruiker).
- $tweetTime(b : B)$: Deze Functie geeft het tijdstip terug waarop een Twitterbericht is geplaatst. Deze tijd zal relatief zijn aan het begin van de week waarin het Twitterbericht is geplaatst. Wanneer een bericht bijvoorbeeld is geplaatst in week 3 (sinds het eerste bericht uit de gedownloade verzameling van Twitterberichten van de huidige gebruiker) op Woensdag om 14:30, dan geeft de functie $tweetTime(b : B)$ een tijd in seconde terug die maandag, dinsdag en woensdag tot 14:30 bestrijkt. Op deze manier worden Twitterberichten die geplaatst zijn verdeeld over meerdere weken geprojecteerd op een tijdsbestek van een week.

5.4 Aannames

Aanname: Gebeurtenissen rapporteren op Twitter 1

$\forall u : U, \forall e : E, \exists t : T, occurence(e,t) \longrightarrow \exists b : userTweets(u, DB), \exists delay : T, report(b, t + delay) \wedge describes(b, e) \wedge delay < 1800$

Dit is een van onze basisaannames dat moet gelden voor onze tool om te kunnen werken. Deze aanname zegt dat voor alle gebeurtenissen die de Twittergebruiker meemaakt gerapporteerd zullen worden door de gebruiker door middel van een Twitterbericht. Natuurlijk is het een utopisch idee dat alle Twittergebruikers alles wat ze doen de hele dag zullen verwoorden in een Twitterbericht. Maar we gaan er hier vanuit dat er in ieder geval voor elke gebeurtenis die bij een leefpatroon hoort wel een bericht geplaatst wordt op Twitter. Bovenstaande aanname zegt dat voor elk voorkomen van een gebeurtenis op tijdstip t een bericht bestaat dat gerapporteerd is net na (een delay) het voorkomen van de gebeurtenis en dat het gerapporteerde een beschrijving van de gebeurtenis is. We willen natuurlijk dat het bericht over deze gebeurtenis zo snel mogelijk na het voorkomen ervan geplaatst wordt op Twitter, maar het is ook nog bruikbaar als hier een vertraging in zit (de delay). Daarom stellen we nog als laatste eis dat de vertraging niet meer mag zijn dan een half uur (1800 seconden). Op deze manier kunnen we op een half uur nauwkeurig bepalen wanneer een activiteit wordt uitgevoerd die behoort tot een bepaald leefpatroon.

Definitie: leefpatroon 1

$lifestylePatern(e) :: \exists t : T, occurence(e,t) \longrightarrow occurence(e, addWeek(t))$

Deze definitie van een wekelijks patroon van een gebeurtenis e geeft weer dat een patroon minimaal een opvolgende gebeurtenis moet hebben, ongeveer een week later. Dit is de minimale eis voor een patroon aangezien de regel ook geldt voor gebeurtenissen die meerdere keren voorkomen met eventueel een week tijd ertussen. Enkel gebeurtenissen die niet twee keer voorkomen met ongeveer een week tijd ertussen worden met deze regel uitgesloten. Ook sluit deze regel geen patronen uit die op meer dagen van de week voorkomen. Wanneer iemand elke week zowel op maandag en op woensdag gaat sporten, ziet deze regel de sportgebeurtenissen van maandag als een patroon en van woensdag als een (onafhankelijk) patroon. Merk op dat we bij deze definitie impliciet uitgaan van een persoon waar de gebeurtenissen uit E betrekking op mogen hebben. Het is namelijk niet de bedoeling dat we over het algemeen naar leefpatronen gaan zoeken bij de gebeurtenissen van alle gebruikers bij elkaar.

Definitie: Patroon in Twitterberichten 1

$Twitterpattern(e) :: b1 : B, b2 : B, describes(b1, e) \wedge describes(b2, e) \wedge \exists t : T, report(b1, t) \wedge report(b2, t + addWeek(t))$

Een Twitterpatroon betreffende gebeurtenis e definiëren we hierboven als volgt: Wanneer er twee berichten bestaan die de gebeurtenis e beschrijven, en deze Twitterberichten gepubliceerd worden met ongeveer een week tijd ertussen, dan is er een Twitterpatroon voor gebeurtenis e. Dit is ongeveer analoog aan onze definitie van een leefpatroon. Wanneer twee Twitterberichten die dezelfde gebeurtenissen beschrijven voorkomen, stellen we dat deze behoren tot een patroon behorende tot gebeurtenis e. Echter, we willen ons enkel richten op de patronen in Twitterberichten van een dezelfde gebruiker. We willen namelijk per gebruiker kunnen

bepalen welke patronen er in zijn of haar berichten verwerkt zitten. De definitie komt er dan als volgt uit te zien:

Definitie: Patroon in Twitterberichten 2

$Twitterpattern(u, e) :: b1 : userTweets(u, DB), b2 : userTweets(u, DB), describes(b1, e) \wedge describes(b2, e) \wedge \exists t : T, report(b1, t) \wedge report(b2, t + addWeek(t))$

5.5 Aannemelijk te maken

Aannemelijk te maken: Patronen in Twitterberichten omschrijven leefpatronen 1

$Twitterpattern(u, e) \longrightarrow lifestylePattern(e)$

We willen graag een recept geven om leefpatronen te extraheren uit een serie Twitterberichten van iemand. Daarvoor moeten we eerst gaan kijken of het op waarheid berust dat wanneer er een Twitterpatroon over de gebeurtenis e bestaat (dus een serie berichten die gebeurtenis e beschrijven), dat dit patroon van Twitterberichten dan ook een leefpatroon van de desbetreffende gebruiker beschrijft. Als er dus een Twitterpatroon van gebeurtenis e bestaat, bestaat er ook een leefpatroon behorende tot het leven van de geanalyseerde gebruiker waar gebeurtenis e toe behoort. Met andere woorden: Wanneer er een Twitterpatroon te vinden is waarbij elke week op een bepaald tijdstip een gebeurtenis wordt omschreven, is dit een leefpatroon van de gebruiker. Als we deze theorie vervolgens uitschrijven krijgen we het volgende:

Aannemelijk te maken: Patronen in Twitterberichten omschrijven leefpatronen 2

$\exists u : U, \forall e : E, (b1 : userTweets(u, DB), b2 : userTweets(u, DB), describes(b1, e) \wedge describes(b2, e) \wedge \exists t : T, report(b1, t) \wedge report(b2, t + addWeek(t))) \longrightarrow (\exists t : T, occurence(e, t) \longrightarrow occurence(e, addWeek(t)))$

Als deze theorie blijkt te kloppen, en er dus inderdaad redelijk goed leefpatronen geëxtraheerd kunnen worden uit een verzameling Twitterberichten van een bepaalde gebruiker, zouden we kunnen bepalen wanneer de geanalyseerde gebruiker gebeurtenissen meemaakt die te maken hebben met een bepaald leefpatroon.
Vervolgens kunnen we de plaatsingstijden in de week van Twitterberichten die bij een gebeurtenis horen verkrijgen door een nieuwe verzameling aan te maken. We gaan er hierbij even vanuit dat bovengenoemde eis dat de door de functie $patternCollection(e : E, DB)$ gebruikte Twitterberichten enkel tot een patroon van een gebeurtenis behoren van de geanalyseerde gebruiker door deze functie wordt afgedwongen. De functie zal dus enkel Twitterberichten opleveren die tot een patroon behoren dat betrekking heeft op de gebeurtenis e (als argument meegegeven).

Uiteindelijke doel: Verzameling van tijden die bij een bepaalde soort gebeurtenis horen 1

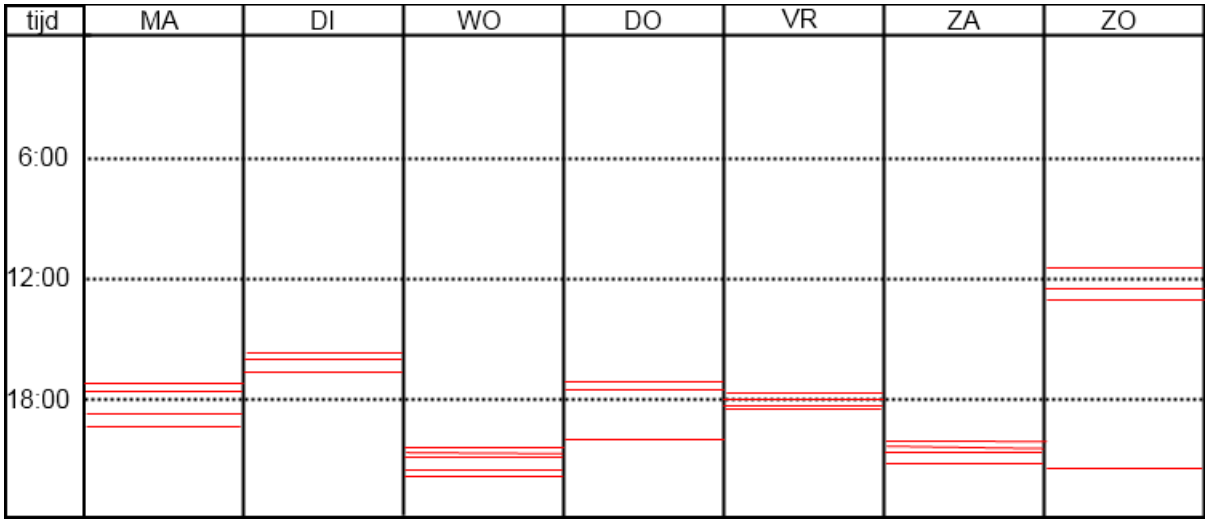
$tweetTimes(u, e) = \{tweetTime(patternMessage) | patternMessage \in patternCollection(e, DB)\}$

Wanneer we tweetTimes van de Twitterberichten, die volgens bovenstaande regels bij een gebeurtenis e behoren, vervolgens weer zouden geven in een weekrooster, dan zouden de patronen zichtbaar moeten worden. Zie het voorbeeld in Figuur 1 op pagina 12. Hierin is het patroon weergegeven dat te maken heeft met eten. Zoals je ziet zijn de gevonden rapportages van ongeveer vier tot vijf weken over elkaar heen geschoven en zie je per dag ongeveer vier keer een rapportage van gebeurtenissen die te maken hebben met het leefpatroon van etenstijden.

6 Deelvragen

Uit bovenstaande theorie volgen een aantal dingen die bekeken moeten worden voordat we kunnen besluiten om een dergelijke tool daadwerkelijk te implementeren of niet. De twee belangrijkste deelvragen die uit de theorie naar voren komen zijn de volgende:

- *Zijn leefpatronen daadwerkelijk goed zichtbaar in Twitterpublicaties van Twittergebruikers?*
  Deze vraag heeft betrekking op de implicatie "Patronen in Twitterberichten omschrijven leefpatronen" in sectie 5.5 op pagina 11 die wij aannemelijk moeten maken. Daarnaast gaat de functie



Figuur 1: Voorbeeld van een leefpatroon weergegeven in een weekrooster. Als voorbeeld gebruiken we hier het eetpatroon (hoofdmaal).

patternCollection$(e : E, DB)$ in de omschrijving van het uiteindelijke doel in de theorie er vanuit dat er op basis van Twitterberichten patronen gevonden kunnen worden die te maken hebben met leefpatronen van de gebruiker. We gaan kijken of het waarschijnlijk is dat bij actieve Twittergebruikers informatie over leefpatronen te vinden zijn. Dit punt is cruciaal voor de werking van de beoogde tool. Wanneer het namelijk zo is dat het toch niet waarschijnlijk genoeg is dat gemiddelde Twittergebruikers informatie over dagelijkse gebeurtenissen op Twitter plaatsen (redelijk snel nadat de gebeurtenis plaatsvond of tijdens het plaatsvinden van de gebeurtenis), zal er te veel ruis in de informatie zitten om er gestaafde conclusies over leefpatronen uit te kunnen trekken. We zullen een antwoord op deze vraag proberen te vinden in de sectie 9 op pagina 18 waar we zullen kijken naar de publicatiefrequentie van gebruikers, de gebruikte onderwerpen en we zullen een tal van case-studies uitvoeren.

- *Welke manier kunnen we het beste kiezen om relaties te leggen tussen berichten die te maken hebben met dezelfde soort gebeurtenis?*
  Deze vraag heeft in het formele model het meeste betrekking op de *describes*$(b : B, e : E)$ functie. Dit punt lijkt mij het lastigste om te automatiseren aangezien je hier semantische betekenis van berichten moet gaan parseren naar categoriën van leefpatronen. Het is niet zo dat we aan elk Twitterbericht een geschikte categorie kunnen toewijzen, er moet binnen de *describes*$(b : B, e : E)$ functie dus ook een filtermogelijkheid ingebouwd worden. We gaan daarom bekijken welke mogelijkheden er zijn om geautomatiseerd relaties te leggen tussen Twitterberichten die betrekking hebben op een bepaalde gebeurtenis.

Aan de hand van deze twee deelvragen wil ik in dit document gaan kijken of het eerder gepresenteerde idee van een tool met succes gerealiseerd zou kunnen worden in de toekomst.

7 Attributen Twitterbericht

We zullen beginnen met een ontleding van een Twitterbericht. Als je een Twitterbericht bekijkt via de webinterface [15] zie je namelijk alleen de hoofdtekst, gebruikersnaam en volledige naam van de gebruiker, maar er wordt nog veel meer informatie opgeslagen. Deze informatie is ook vrij toegankelijk wanneer een gebruiker zijn Twitterberichten niet heeft beveiligd.

7.1 Beschikbare attributen

Allereerst geef ik hieronder in Tabel 1 op pagina 14 weer welke attributen een Twitterbericht standaard heeft. In de derde kolom heb ik aangegeven of een attribuut altijd ingevuld is, of dat het ook niet ingevuld

kan zijn (en een null waarde bevat). Als laatste attribuut bevat het object van een Twitterbericht een object met attributen over de gebruiker. In de Tabel 2 op pagina 15 zie je hier een opsomming van. Naast de attributen die ik in de tabel heb opgenomen zijn er ook nog een aantal attributen opgenomen in het object waarin informatie staat opgeslagen over de opmaak van de webinterface van Twitter (kleurcode's en dergelijke). Aangezien dit voor ons niet interessant is, heb ik deze weggelaten. Elk object dat een Twitterbericht beschrijft, is ook nog een userobject opgenomen. Hierin staat alle informatie over de gebruiker van wie het Twitterbericht afkomstig is. In Tabel 2 heb ik de belangrijkste attributen uit dit object ook opgesomd. In de kolom "vereist" heb ik aangegeven of het verplichte informatie is die de Twittergebruiker met Twitter moet delen of dat ze kunnen kiezen deze informatie open te laten. In principe kunnen we voor ons beoogde doel alleen attributen gebruiken die vereist zijn.

7.2 Bruikbaarheid van attributen

Voor het bepalen van leefpatronen zijn twee attributen zeer belangrijk. Dit zijn namelijk de hoofdteksten en de tijd waarop dit bericht is gepubliceerd, respectievelijk te vinden in de attributen text en create_at. Het voordeel van deze twee attributen is dat ze niet kunnen worden leeg gelaten. Je kunt geen leeg Twitterbericht publiceren en de tijd waarop het gepubliceerd wordt, wordt door de servers van Twitter zelf toegevoegd. Ook het tijdstip kun je dus niet leeg laten of handmatig veranderen. Aangenomen dat de gebruikers steeds met een kernwoord publiceren wat ze elk moment van de dag aan het doen zijn, zou je gecombineerd met de tijd waarop dit geplaatst is exact de agenda van de gebruiker kunnen extraheren. Helaas is het niet zo dat de gebruiker echt op elk moment een bericht plaatst en zijn de berichten zeker niet in een vorm dat ze enkel bestaan uit een kernwoord.
Een ander attribuut dat bruikbaar is, is de bron waar het Twitterbericht vandaan komt. Hier wordt door de cliënt waarmee de gebruiker een bericht publiceert ingevoerd wat de naam van de cliënt is. Dit gegeven kun je gebruiken om de precisie van een bepaald bericht te bepalen. Je kunt op basis van deze bron namelijk een onderscheid gaan maken tussen of het bericht is geplaatst met een mobiele telefoon of een ander draagbaar apparaat, of dat het geplaatst is met een desktop computer. Gebaseerd op de aanname dat een gebruiker wel gedurende de hele dag zijn telefoon bij zich heeft en in staat is een Twitterbericht te plaatsen maar niet altijd in de buurt van een computer is, kun je stellen dat de berichten van een mobiele telefoon informatie verschaffen die meer real time is dan de informatie die geplaatst wordt via een desktop computer. Wanneer een Twitterbericht van een desktopcomputer afkomstig is, is de kans aanwezig dat het bericht terugblikt op een gebeurtenis verder in de geschiedenis of betrekking heeft op een groter tijdsbestek. Wellicht zal de gebruiker de informatie namelijk afstemmen op de frequentie waarop hij berichten plaatst. Wanneer een gebruiker minder frequent berichten plaatst (omdat hij dit alleen kan doen wanneer hij in de buurt van een computer is), zal zijn informatievoorziening globaler zijn.
Daarnaast kan de locatie gebruikt worden om te kijken of een gebruiker bezig is met een reguliere activiteit die vaker voorkomt in zijn weekplanning of niet. De meeste activiteiten die vaak terugkomen zullen namelijk steeds op dezelfde plaats zijn. Voor werken geldt vaak dat de locatie hetzelfde is (alhoewel dit afhankelijk is van het soort werk). Voor het leefpatroon sporten zal ook gelden dat de persoon steeds bij dezelfde sportvereniging zal gaan sporten. Over het algemeen kun je dus stellen dat activiteiten die vaker terugkomen, en dus bij een leefpatroon horen, ook vaak op dezelfde geografische locatie plaats zullen vinden. De geografische locatie zou dus kunnen ondersteunen in het bepalen of Twitterbericht wel of niet tot een bepaald leefpatroon behoort of niet. Helaas is het momenteel nog zo dat niet iedereen zijn GPS coördinaten toe laat voegen aan zijn Twitterberichten. Volgens Bryant [2] blijkt uit een onderzoek van het bedrijf Sysomos uit eind 2009 dat nog maar 0,23% van de Twitterberichten GPS coördinaten bevatte. Dit is dus nog niet een noemenswaardig percentage dus lijkt het erop dat het lastig wordt dit gegeven te gebruiken in een onderzoek naar leefpatronen. Misschien zal dit echter in de toekomst wel kunnen (wanneer dit percentage toe neemt).
Van het gebruikerobject is voor het analyseren van leefpatronen weinig interessant. Het enige wat van belang kan zijn is de tijdzone (time_zone) en de taal (lang). De tijdzone kan van belang zijn om de tijdstippen van gebeurtenissen te corrigeren met de tijdzone van degene die de uitkomst van de tool (het weekrooster) zou willen lezen om verwarring te voorkomen. Helaas is het niet verplicht om als gebruiker je tijdzone in te stellen en daarom hebben ook maar weinig mensen dit gedaan. Het taal attribuut kunnen we gebruiken om te kijken in welke taal we een semantische analyse moeten gaan doen (om het Twitterbericht te analyseren). Helaas is de taalinstelling enkel bedoeld om de juiste taal weer te geven in de webinterface van Twitter en is het alleen mogelijk om een kleine selectie van talen te selecteren. Enkel de talen die de webinterface ondersteunt kun je instellen. Dit zijn er helaas maar 7. Hierdoor stellen mensen meestal Engels in, terwijl dit niet betekent dat

Attribuut	Beschrijving	vereist
text	De hoofdtekst van het Twitterbericht zelf.	Ja
truncated	Boolean die aangeeft of het huidige Twitterbericht een deel van een opgesplitst bericht is (i.v.m. de maximale grootte van 140 tekens).	Ja
place	Object met informatie over de plaats waar de gebruiker het Twitterbericht heeft geplaatst (standaard NULL). Dit object kan ook coördinaten bevatten en heeft een eigen Unique ID. De overige attributen zullen we hier niet noemen.	Nee
favorited	Boolean of de auteur van het Twitterbericht dit bericht tot zijn favorieten vindt behoren of niet.	Ja
id_str	Unique Identifier voor het huidige Twitterbericht.	Ja
coordinates	De GPS coördinaten van de positie waarvandaan de gebruiker het Twitterbericht heeft gepubliceerd (meestal verkregen via zijn telefoon).	Nee
retweet_count	Hoe vaak dit Twitterbericht is herhaald door andere gebruikers.	Ja
source	Naam van de cliënt software waarmee het Twitterbericht is gepubliceerd.	Ja
in_reply_to_screen_name	Gebruikersnaam van de gebruiker waar het huidige Twitterbericht een antwoord op is.	Nee
in_reply_to_status_id_str	Unique Identifier van het Twitterbericht waar dit huidige Twitterbericht een antwoord op is.	Nee
geo	Object met geografische informatie over de plaats waar de gebruiker zich bevond tijdens het publiceren van het Twitterbericht.	Nee
created_at	Het tijdstip en de datum waarop het Twitterbericht is gepubliceerd.	Ja
contributors	Een vernoeming van gebruikers die de gepubliceerde informatie hebben aangedragen (gebruikt door commerciële Twitteraccounts).	Nee
retweeted	Boolean die aangeeft of dit bericht een herhaling is van een Twitterbericht van iemand anders (een retweet).	Ja
in_reply_to_user_id_str	De Unique Identifier van de auteur van het bericht waarop het huidige Twitterbericht een antwoord is.	Nee
user	Het gebruikers object waar informatie over de gebruiker van het huidige Twitterbericht in staat (zie het gebruikers object tabel (Tabel 2)).	Ja

Tabel 1: De attributen van een Twitterbericht

Attribuut	Beschrijving	Vereist
url	Webadres van de gebruiker (door de gebruiker zelf in te vullen, kan bijvoorbeeld verwijzen naar zijn blog of profielwebsite).	Nee
screen_name	Gebruikersnaam van deze gebruiker.	Ja
description	Korte beschrijving (biografie) van de gebruiker (door de gebruiker zelf in te vullen).	Nee
show_all_inline_media	Boolean die aangeeft of foto's en andere media direct moeten worden geopend en aan de gebruiker worden getoond.	Ja
lang	De taal van de gebruiker. Keuze uit: Italiaans, Spaans, Engels, Koreaans, Frans, Duits en Japans.	Ja
geo_enabled	Boolean die aangeeft of de gebruiker locatiegegevens wil gebruiken of niet.	Ja
time_zone	De tijdzone waarin de gebruiker leeft (Moet handmatig ingesteld worden door de gebruiker).	Nee
location	De locatie van de gebruiker (Meestal woonplaats, door de gebruiker in te vullen).	Nee
id_str	Unique Identifier van de gebruiker.	Nee
statusses_count	Aantal Twitterberichten de gebruiker in totaal heeft geplaatst op Twitter.	Ja
followers_count	Het aantal gebruikers dat de Twitterberichten van deze gebruiker heeft toegevoegd aan zijn Timeline.	Ja
created_at	Tijdstip en datum wanneer de gebruiker zich aangemeld heeft voor Twitter (en het gebruikersaccount heeft aangemaakt).	Ja
contributors_enabled	Boolean of deze gebruiker met bijdrager vermeldingen werkt (voor als dit een commercieel gebruikersaccount is bijv.)	Ja
friends_count	Het aantal gebruikers die deze gebruiker zelf volgt (gebruikers waarvan de Twitterberichten in zijn timeline verschijnen).	Ja
protected	Boolean of de gebruiker toestemming moet geven voordat gebruikers zijn berichten mogen inzien.	Ja
name	De volledige naam van de gebruiker (door de gebruiker zelf in te voeren).	Nee
profile_image_url	Adres van de profielfoto van de gebruiker.	Nee

Tabel 2: De attributen van een van het gebruikerobject behorende bij een Twitterbericht

hun Twitterberichten ook daadwerkelijk in het Engels geschreven zijn. We zullen dus ook naar een manier van taaldetectie moeten zoeken.

8 Gegevensset

Voor bepaalde mogelijkheden die ik hieronder presenteer toets ik vermoedens aan een set Twitterberichten van verschillende gebruikers. Ik zal daarom eerst een toelichting geven hoe deze gegevensset eruit ziet en hoe deze tot stand is gekomen.

8.1 Verzamelen van gebruikers

De Twittergebruikers waar ik de Twitterberichten van heb gedownload heb ik verkregen met behulp van een simple scrape methode. Je begint met een gebruiker en van deze gebruiker download je alle Twitterberichten. Tijdens het opslaan van de Twitterberichten scan je deze berichten ook nog op zoek naar andere gebruikersnamen. Deze zijn te herkennen aan het teken voor de gebruikersnaam. Deze nieuwe gebruikersnamen worden tijdens het zoeken in een queue gezet. Wanneer alle berichten van de eerste gebruiker zijn opgeslagen in de database, wordt de volgende gebruikersnaam uit de queue opgehaald. Deze gebruikersnaam is de eerste die gevonden is tijdens het doorzoeken van de Twitterberichten van de eerste gebruiker. Voor deze volgende gebruiker wordt dezelfde procedure uitgevoerd. Op deze manier wordt het Twitternetwerk op een breath-first search manier uitgediept. Dit heeft echter twee nadelen. Ten eerste wordt hier gezocht naar gebruikers waarnaar gerefereerd is door andere gebruikers. Gebruikers waar nooit naar gerefereerd wordt, worden op deze manier nooit in beschouwing genomen. Dat is in dit geval geen probleem; Gebruikers waar nooit naar gerefereerd wordt, zijn vaak geen actieve gebruikers en daar willen we ons toch niet op richten tijdens dit onderzoek. Daarnaast kent deze methode nog een nadeel dat in het begin de gedownloade Twitterberichten sterk afhankelijk is van de keuze van de eerste gebruiker. Aangezien na de eerste gebruiker gebruikers gescand worden waar de eerste gebruiker iets tegen gezegd heeft, is dit zeker niet willekeurig gekozen. Deze gebruikers hebben namelijk vaak iets gemeen met de eerste gebruiker (zelfde studie of collega's). Bij selectie worden echter zeer veel gebruikers gefilterd (zodat absoluut zeker was dat de gebruikers Nederlands zijn) dus we mogen wel aannemen dat onze verzameling Twitterberichten divers genoeg is.

Van de Gebruikers heb ik de volgende gegevens opgeslagen en zijn dus beschikbaar voor tests:

- userID
- screen_name
- full_name
- location
- bio
- created_at (profiel)
- statusses_count

Voor elk Twitterbericht heb ik het volgende opgeslagen:

- tweetID
- userID
- text
- created_at (Twitterbericht)
- source

8.2 Selectie

Voordat ik alle Twitterberichten heb gedownload heb ik eerst een deel van de Twitterberichten geanalyseerd om te bepalen of ze wel voldeden aan de belangrijkste eisen. Om deze test uit te voeren heb ik van een gebruiker de eerste 200 berichten gedownload (het downloaden van de Twitterberichten kan maximaal in batches van 200 berichten, dus ik heb de condities getest na het uitvoeren van de eerste request voor deze gebruiker).
Ten eerste moest de gebruiker wel bestaan. Het komt namelijk ook regelmatig voor dat er naar een gebruiker wordt gerefereerd die nooit bestaan heeft of op dit moment niet meer bestaat. Je kunt namelijk ook je Twitter gebruikersnaam wijzigen (zolang de nieuwe naam nog maar niet bestaat) en het kan ook voorkomen dat mensen hun Twitteraccount verwijderd hebben. Allereerst moeten we dus een request doen om berichten op te halen van een bepaalde gebruikersnaam en controleren of we daadwerkelijk wel een resultaat terug krijgen van de Twitter servers.
Ten tweede moeten we kijken of de gebruiker zijn Twitterberichten wel openbaar heeft gemaakt. Twittergebruikers kunnen namelijk hun berichten ook beveiligen zodat alleen andere Twittergebruikers waar zij handmatig toestemming voor hebben moeten geven de berichten van deze gebruiker kunnen inzien. Het is niet mogelijk om deze gebruikers automatisch te analyseren omdat we dan eerst een verzoek zouden moeten doen om de berichten te mogen bekijken en dit is implementatie technisch lastig. Het duurt namelijk waarschijnlijk een redelijke tijd voordat de gebruiker het verzoek geaccepteerd heeft, en dan zou het scrape-programma om de zoveel tijd moeten controleren of er al toegang is verschaft. Daarnaast is het ook niet de bedoeling van dit onderzoek om gebruikers te analyseren die hun Twitterberichten beveiligd hebben. Voor hun is er namelijk minder risico dat er misbruik gemaakt wordt van de gegevens die zij publiceren aangezien zij zelf controle hebben over met wie ze de informatie delen en met wie niet.
Ten derde moeten we op een of andere manier zorgen dat de gebruikers waarvan we de Twitterberichten downloaden voornamelijk Nederlandse berichten publiceren. Dit omdat we een bepaalde taal moeten kiezen om ons op te richten (anders zijn de resultaten niet te combineren) en ik heb hier voor Nederlands gekozen. Zoals al eerder te lezen was bij de uitleg van de attributen van Twitterberichten, konden we de taal die als attribuut is meegegeven aan Twitterberichten niet gebruiken. De mogelijke waarden die daar in te stellen zijn, zijn te beperkt voor dit doel en Nederlands komt niet voor tussen de mogelijke keuzes. Dit attribuut kunnen we hier dus niet voor gebruiken. Ook het attribuut van de tijdzone kunnen we helaas niet gebruiken. Bij dit attribuut was het wel mogelijk om Amsterdam als waarde in te stellen, echter het was geen verplicht veld bij het aanmaken van een Twitteraccount. Vrijwel niemand heeft dit dus ingevuld en ook dit attribuut kunnen we daarom niet gebruiken. Als oplossing heb ik hier een service van Google gebruikt. Google Language Detector. Deze service kun je via een API een tekst geven en dan geeft de service terug van welke taal de text afkomstig is en met welke zekerheid dit zo is. Van de 200 berichten die ik van de gebruiker al heb gedownload (mocht de gebruiker er zoveel hebben gepubliceerd), stuur ik er steeds 30 willekeurig gekozen berichten naar de Google Language Detector. Wanneer deze van minimaal 15 berichten met minimaal 50% zekerheid uitwijst dat het om een Nederlandse tekst gaat, ga ik verder met het downloaden en opslaan van de berichten van deze gebruiker. Anders worden de berichten niet opgeslagen en de gebruiker uit de database verwijderd.

8.3 Details gegevensset

Mijn doel was om in eerste instantie een dataset te verkrijgen van Twitterberichten van 1000 gebruikers. Vanwege verschillende limieten (van Twitter en Google) heeft dit een zeer lange tijd ingenomen.

Aantal gebruikers: 947.
Totaal aantal Twitterberichten: 1.066.082.
Gemiddeld aantal berichten per gebruiker: 1126.
Datum downloaden: Maart-April 2011.

Van de 947 Twittergebruikers die we bekeken hebben, zijn er maar 16 inactief. Deze gebruikers hebben gedurende het bestaan van hun Twitteraccount minder dan 10 berichten gepubliceerd en zijn daarom wel aan te merken als niet actieve Twittergebruikers. Dit is een kleine aanwijzing dat er waarschijnlijk niet zo heel veel Twittergebruikers zijn die alleen een gebruikersaccount hebben aangemaakt om berichten van andere gebruikers te lezen en niet om zelf berichten te publiceren. Het is natuurlijk niet helemaal duidelijk vast te stellen welke grens van het totale aantal berichten die geplaatst moeten zijn gebruikt moet worden

om gebruikers tot de groep van actieve gebruikers te rekenen of niet. Echter voor onze dataset geldt dat als je de grens bij 20 berichten legt, er maar 33 gebruikers als inactief gezien kunnen worden. Bij de grens van 50 is dit 64 en bij de grens van 100 berichten is dit 116. We kunnen dus wel stellen dat de gebruikers in onze dataset voor het grootste deel actief zijn met het plaatsen van berichten op Twitter en niet alleen passieve gebruikers zijn die enkel een Twitteraccount hebben aangemaakt om de Twitterberichten van andere mensen te kunnen lezen.

9 Leefpatroon Informatie in Twitterberichten

9.1 Publicatiefrequentie

Volgens de support pagina's van Twitter [16] kun je van een gebruiker geen publicaties meer ophalen die ouder zijn dan 3200 publicaties. We kunnen van elke gebruiker dus maximaal 3200 berichten in de geschiedenis terug zien. We zijn bij het formuleren van ons vermoeden uitgegaan van het ultieme geval dat gebruikers gemiddeld elk half uur berichten waar ze op dat moment mee bezig zijn. Wanneer zij dit elke dag 16 uur doen (wanneer men slaapt kan men geen bericht publiceren), dan kom je uit op een gemiddelde van 32 berichten per dag. Aangezien wij maximaal 3200 berichten kunnen terugzien, zou dit betekenen dat we ongeveer 100 dagen in de geschiedenis kunnen analyseren. (Hierbij gaan we er dan wel vanuit dat alle berichten gebruikt worden om te vertellen wat er gaande is, en niet om een gesprek te voeren met een andere Twittergebruiker of nieuwsfeiten te publiceren enz.) Dit zou dus betekenen dat we logboeken hebben van de gebeurtenissen van een gebruiker gedurende 14 weken (wat ongeveer gelijk is aan 3,5 maand). Om onze eerste deelvraag te kunnen beantwoorden moeten we ook gaan kijken of het een realistische schatting is dat (actieve) gebruikers gemiddeld elk halfuur een bericht plaatsen. Wat is de daadwerkelijk Twitterfrequentie van de Twittergebruikers? Is dit voldoende voor ons beoogde systeem om te kunnen functioneren (aangenomen dat alle berichten qua semantiek bruikbaar zijn)?
Tijdens het downloaden van de Twitterberichten is ook het attribuut statussus_count opgeslagen. Dit attribuut geeft voor elke gebruiker aan hoeveel Twitterberichten de gebruiker heeft gepubliceerd sinds het aanmaken van zijn of haar Twitteraccount. Daarnaast wordt bij elk request voor een Twitterbericht aan de servers van Twitter ook de datum en tijd van het aanmaken van de desbetreffende gebruiker meegegeven, dus deze hebben we ook opgeslagen. Als laatst kunnen we voor elke gebruiker bepalen wanneer het moment was dat de Twittergebruiker het laatste bericht plaatste dat wij in onze verzameling Twitterberichten hebben. Tesamen met de datum van aanmaken van de Twittergebruiker kunnen we nu het tijdsbestek uitrekenen waarin de door ons verzamelde berichten zijn geplaatst. Op basis van dit tijdsbestek en het totale aantal Twitterberichten dat iemand gepubliceerd heeft, kunnen we van elke gebruiker uit gaan rekenen wat het gemiddelde aantal berichten is dat hij of zij per dag publiceert. Vervolgens kunnen we dan concluderen voor hoeveel gebruikers dit boven het gestelde minimum van 32 berichten per dag ligt. Twitter heeft overigens ook hier limieten gesteld. Je kunt maximaal 100 berichten per uur plaatsen en maximaal 1000 berichten per dag. Deze limieten hebben de ontwikkelaars van Twitter ingevoerd om misbruik van het medium (reclame verspreiden e.d.) te voorkomen.
Onze eerste vinding is dat wanneer we alle gebruikers meerekenen, de gemiddelde publicatie frequentie van alle gebruikers op 5.8 berichten per dag uitkomt. Wanneer we echter de inactieve gebruikers niet meerekenen, zitten we op een gemiddelde van 6.6. We hebben hier de grens van minimaal 100 berichten gekozen.
Op de volgende pagina is een tabel (Tabel 3 op pagina 20) geplaatst met voor de 90 Twittergebruikers met het hoogste aantal gemiddelde publicaties per dag. Hierin zijn een paar dingen opvallend: Een gebruiker heeft veruit het hoogste gemiddelde, namelijk een gemiddelde van 160 berichten per dag. Dit zou uitkomen op een publicatiefrequentie van een bericht per 6 minuten (uitgaande van een dag van 16 uur). Dit is natuurlijk verdacht hoog en wanneer we gaan bekijken welke berichten deze gebruiker publiceert, blijkt ook dat het niet om een mens gaat maar om een bot die geautomatiseerd berichten plaatst. Deze bot zoekt naar verkeerde spellingen van het woord "sowieso" en corrigeert de gebruikers die hier een spelfout in maken door middel van het refereren naar deze gebruiker. Vanaf de tweede gebruiker in de ranglijst lijken het echter wel "echte" gebruikers te zijn.
Kort samengevat komen de gegevens over de publicatiefrequentie neer op het volgende:

- 34 van de 947 personen halen het door ons gestelde vereiste van 30 publicaties per dag.
- 83 gebruikers hebben een gemiddelde publicatiefrequentie van boven de 16 per dag. Dit zou betekenen dat 83 mensen ongeveer elk uur een bericht plaatsen. Ook dit zou nog genoeg kunnen zijn voor onze

tool. Echter, dit is nog maar 9 procent van onze dataset.

- 177 mensen hebben een publicatiefrequentie van boven de 8. Dit is ongeveer elke twee uur een bericht.
- 437 mensen plaatsen gemiddeld 2 berichten of minder op een dag. Dit is 46% van de door ons bekeken Twittergebruikers.
- 181 mensen plaatst gemiddeld minder dan een bericht per dag op Twitter. Deze mensen kunnen we dus waarschijnlijk niet analyseren met de tool aangezien je voor patronen meer datapunten per dag nodig hebt.

We kunnen uit deze gegevens halen dat er wel degelijk Twittergebruikers zijn die een gemiddeld aantal berichten per dag publiceert dat boven de 30 ligt. Echter, dit is niet veel. Slechts 34 gebruikers van de 947 bekeken gebruikers halen dit gemiddelde en dat is maar 3,5%. Wanneer dit percentage een goede afspiegeling is van alle Nederlandse gebruikers, zou dit kunnen betekenen dat er maar een zeer selecte groep Nederlandse Twittergebruikers genoeg berichten publiceren om te kunnen analyseren met de beoogde tool. Nu wordt de werking van de tool niet uitgesloten wanneer er gebruikers geanalyseerd worden met een lagere publicatiefrequentie dan 30 berichten per dag, aangezien we berichten van verschillende weken over elkaar heen zullen schuiven tot een weekrooster. Echter, hoe minder bruikbare datapunten de tool heeft, des te onzeker is het resultaat van de analyse door de beoogde tool. Dit geldt zeker ook omdat we bij het bekijken van het publicatiegemiddelde nog geen berichten hebben uitgesloten die niets vertellen over wat de gebruiker op een gegeven moment aan het doen is. De berichten die hier geteld zijn, moeten dus nog worden gefilterd op relevantie en daardoor houdt je minder bruikbare berichten over.

9.2 Meest gebruikte onderwerpen

Om te kijken of de informatie die we via Twitterberichten van gebruikers krijgen op het eerste gezicht wel betrekking hebben op hedendaagse gebeurtenissen, heb ik een woordfrequentie analyse gedaan. Aangezien dit computationeel erg zwaar is, heb ik niet de gehele dataset van 1.066.082 berichten bekeken maar de helft [2]. Ik heb hiermee de berichten bekeken van 576 Twittergebruikers. Met deze methode abstraheer je natuurlijk wel van de volledige context waarin een bepaald woord gebruikt wordt, maar wellicht kunnen we wel een beeld krijgen van de meest gebruikte woorden op Twitter en kijken of deze te maken kunnen hebben met leefpatronen of niet. Algemene bevindingen zijn:

- In totaal zijn er 4.461.637 woorden geteld.
- Het meest gevonden woord is het woord "niet", dit woord kwam 81961 keer voor.
- De meest gevonden woorden zijn vooral hulpwerkwoorden, bijvoeglijk naamwoorden, voegwoorden en bijwoorden.
- Tijdsverwijzingen zijn ook erg veelgebruikt. Het woord "vandaag" is 15666 keer gevonden het woord "morgen" 13979 keer en het woord "vanavond" 9884 keer.
- Van het totale aantal gevonden woorden (359.879) zijn er 257.974 die maar een keer zijn gevonden. Ongeveer 72% van de woorden die gebruikt zijn, zijn dus uniek.

Wanneer we enkel kijken naar werkwoorden en zelfstandignaamwoorden, krijgen we de tabel zoals deze te zien is op 21 (Tabel 9.2).
Omdat we ook geïnteresseerd zijn in de tijdsverwijzingen (zie sectie tijdsverwijzingen op pagina 25) hebben we ook woorden die hier betrekking op hebben getoond in de tabel. Uit de case studies (in de volgende secties) bleek namelijk dat gebruikers vaak ook tijdsverwijzingen in hun berichten plaatsten. Tijdens de woordfreqentie analyse blijkt nu dus ook dat de woorden die globale tijdsverwijzingen aangeven vaak worden gebruikt. Dit maakt het bepalen van het tijdstip waarop een activiteit plaatsvindt een stuk lastiger aangezien we de tijd waarop het bericht is gepubliceerd niet meer direct kunnen gebruiken. We moeten eerst een correctie van de tijd doen op basis van deze tijdsverwijzingen in het bericht. De vaak gevonden woorden die duiden op een tijdsverwijzing zijn: vandaag, vanavond, week, weekend, dagen, gisteren, avond, vanmiddag, vrijdag, zaterdag, zondag, weken, maandag, minuten, geleden, donderdag, gister, woensdag, binnenkort en april. Hier verderop meer over.

[2] De woordfrequentie analyse duurde een maand op een standaard kantoormachine, nu is mijn implementatie vast niet het meest efficiënt, maar dit soort analyses nemen over het algemeen erg veel tijd in beslag.

UserID	Totaal Berichten	Dagen	Gem. Per Dag	UserID	Totaal Berichten	Dagen	Gem. Per Dag
106558253	70187	440	159.51590909091	27855707	18362	743	24.7133243607
26556456	60378	736	82.035326086957	73211879	13999	575	24.346086956522
55050797	42779	635	67.368503937008	26889458	18015	740	24.344594594595
56061424	42621	633	67.331753554502	18018548	20628	854	24.154566744731
80375986	36633	547	66.970749542962	24906715	18173	756	24.03835978836
713333	95084	1540	61.742857142857	176725583	5689	241	23.605809128631
75090873	34890	568	61.426056338028	18727316	19140	813	23.542435424354
28081432	43508	734	59.275204359673	189099917	4719	201	23.477611940299
36447473	40528	700	57.897142857143	39739684	16205	695	23.31654676259
20541596	39794	777	51.214929214929	134910374	8303	359	23.128133704735
16822936	43972	902	48.749445676275	24675888	16732	759	22.044795783926
51496252	30507	656	46.504573170732	18586556	17979	821	21.898903775883
103049036	20106	440	45.695454545455	19703385	17550	803	21.855541718555
22903315	33904	769	44.088426527958	191425333	4148	195	21.271794871795
189616887	8135	195	41.717948717949	39814273	14779	700	21.112857142857
22132751	31767	767	41.417209908735	80296574	11158	536	20.817164179104
56559272	25290	625	40.464	19386691	15901	791	20.102402022756
228350371	4135	108	38.287037037037	23472198	15368	768	20.010416666667
148672853	10112	266	38.015037593985	14706328	21111	1065	19.822535211268
132897455	13463	363	37.088154269972	87716359	10227	516	19.81976744186
90043396	18762	506	37.079051383399	135403981	6868	350	19.622857142857
43966590	24367	673	36.206537890045	22630828	14610	758	19.274406332454
18571417	29090	811	35.869297163995	134751920	6866	359	19.125348189415
20326729	27909	781	35.734955185659	105236243	8544	447	19.114093959732
20683641	27663	790	35.016455696203	59147427	11741	617	19.029173419773
14300696	37679	1091	34.536205316224	125372812	7242	381	19.007874015748
19445747	27826	810	34.353086419753	29494978	13670	723	18.907330567082
56608891	20959	626	33.480830670927	72307215	10514	576	18.253472222222
62338709	20432	611	33.440261865794	16254349	17058	943	18.089077412513
41670599	21948	683	32.134699853587	45818902	11789	655	17.998473282443
20274637	25489	795	32.061635220126	233852982	1520	86	17.674418604651
24241313	24330	761	31.971090670171	43858241	11518	673	17.11441307578
107535868	13726	432	31.773148148148	201428093	3025	177	17.090395480226
180319496	6760	225	30.044444444444	92038459	8289	487	17.020533880903
26019356	21110	739	28.565629228687	104446679	7587	455	16.674725274725
115471993	11588	406	28.541871921182	204333757	2851	172	16.575581395349
23525048	21512	757	28.417437252312	96975923	7723	476	16.224789915966
98902765	13254	468	28.320512820513	196929186	3067	190	16.142105263158
118141975	11309	401	28.201995012469	50730748	10171	637	15.967032967033
92571858	13795	491	28.095723014257	27319057	11633	738	15.762872628726
42953965	17828	680	26.217647058824	77061950	8513	546	15.591575091575
18266124	21672	837	25.89247311828	42611098	10522	679	15.496318114875
1605791	37874	1465	25.852559726962	20593919	12153	785	15.48152866242
270085001	256	10	25.6	25314632	11516	746	15.436997319035
103582393	11426	453	25.222958057395	15872575	14727	955	15.420942408377

Tabel 3: Publicatiefrequentie van de eerste 90 Twittergebruikers wanneer je ze sorteert op freqentie van hoog naar laag.

woord	frequentie	woord	frequentie	woord	frequentie
vandaag	15666	vrijdag	2675	album	1627
vanavond	9884	kans	2656	gemist	1620
mensen	9028	film	2542	gister	1616
twitter	8052	lezen	2524	geworden	1611
week	8022	lees	2488	media	1608
maken	7853	site	2422	liggen	1602
komen	6773	leven	2421	proberen	1600
zien	6692	kamer	2384	mannen	1573
moeten	5732	auto	2340	programma	1571
thuis	5222	utrecht	2332	richting	1568
eten	5098	amsterdam	2292	hoofd	1562
huis	4860	zaterdag	2265	rijden	1529
slapen	4695	muziek	2259	woensdag	1527
werk	4688	radio	2250	optreden	1520
werken	4668	houden	2242	ipad	1517
foto	4541	blijven	2242	gevonden	1507
hoop	4506	vakantie	2233	college	1502
gebruiken	4431	denken	2214	vorige	1502
weekend	4212	nummer	2209	moeder	1499
gefeliciteerd	4192	zondag	2204	probleem	1497
gedaan	4154	weken	2142	verhaal	1497
zeggen	4079	tweets	2094	zoeken	1488
bedankt	4062	spelen	2071	genieten	1464
succes	4035	maandag	2046	dinsdag	1454
gezien	4021	beginnen	2044	sneeuw	1434
klaar	3992	internet	2042	schrijven	1434
geweest	3944	boek	2023	binnenkort	1432
zitten	3913	geld	2013	vrienden	1418
weten	3809	uitzending	1986	verjaardag	1413
nijmegen	3673	slaap	1984	bericht	1405
horen	3578	onderweg	1982	actie	1394
bezig	3430	mail	1965	filmpje	1394
dagen	3396	zoek	1958	bellen	1375
gisteren	3273	geloof	1927	geniet	1372
nederland	3242	leren	1895	volgers	1366
tweet	3199	minuten	1889	facebook	1364
wachten	3139	geleden	1888	winnen	1358
avond	3082	vrouw	1877	kopen	1357
trein	3046	euro	1822	stad	1350
vanmiddag	2945	website	1800	vrouwen	1342
wakker	2934	begin	1747	luisteren	1341
vroeg	2926	video	1743	lachen	1333
gemaakt	2923	debat	1740	gekregen	1330
fotos	2842	lopen	1726	google	1326
studenten	2829	wereld	1724	april	1322
school	2807	dagje	1720	twitteren	1319
werkt	2783	kinderen	1712	single	1317
vragen	2751	onderwijs	1670	drinken	1296
koffie	2731	donderdag	1659	begonnen	1286
krijgen	2688	vergeten	1653	gesprek	1281

Tabel 4: Meest gevonden werkwoorden en zelfstandig naamwoorden met hun frequenties

Leefpatroon	geval 1	geval 2	geval 3	geval 4	geval 5
School	32	0	0	30	0
Reizen	16	10	3	43	7
Slapen	60	95	18	60	17
Feest	5	0	4	43	3
Eten	25	52	10	66	5
Vrije Tijd	21	27	16	39	19
Tv	46	9	8	140	44
Werken	27	76	19	75	7
Sporten	1	11	0	6	8
Verzorging	4	8	4	13	0

Tabel 5: Aantal gevonden berichten die te maken hadden met leefpatronen per geanalyseerde gebruiker.

Uit de tabel blijkt dat een deel van de meest gebruikte woorden een relatie hebben met leefpatronen. Dit ondersteund dus het vermoeden dat er op Twitter veel dingen gepubliceerd worden over dagelijkse gebeurtenissen. Voor het leefpatroon "werken" waar we naar opzoek zijn, zie je dat het woord "werken" en het woord "werk" redelijk hoog in de lijst staat. Daarnaast komen de woorden "eten" en "drinken" wat lijkt te behoren tot de leefpatroon "eten" ook veel voor. Naast deze leefpatronen, zijn ook de leefpatronen "slapen", "school" en "vrije tijd" zichtbaar in de woordfrequentie analyse. We kunnen dus voorzichtig stellen dat deze onderwerpen vaak voorkomen in Twitterpublicaties en dat een mens met behulp van context wellicht patronen zou moeten kunnen ontdekken in de Twitterberichten van een gebruiker. In de volgende sectie ga ik daarom voor een aantal gevallen berichten handmatig categoriseren.

9.3 Handmatige toewijzing van categoriën

Omdat Hashtags in voorgaande secties niet geheel geschikt bleken om van een bericht te bepalen wat het onderwerp van dit bericht is en bij welke leefpatroon dit bericht hoort, zal ik voor 5 mensen met een hoge publicatiefrequentie de berichten handmatig gaan indelen in categoriën. Hiermee wil ik dan vervolgens uitsluitsel doen of hashtags op zich niet bruikbaar zijn voor het analyseren van leefpatronen of dat Twitter over het algemeen te weinig gebruikt wordt om berichten over het dagelijks leven en het dagelijks handelen te publiceren. Ik kies hiervoor de mensen met het hoogste gemiddeld aantal publicaties per dag omdat deze naar alle waarschijnlijkheid het meeste publiceren over dagelijkse gebeurtenissen en bij deze mensen zou hierdoor met de hoogste zekerheid patronen vast te stellen zijn.

9.3.1 Case studies

Ik heb voor de vijf case studies de vijf gebruikers gekozen die bij de analyse naar de publicatiefrequentie het hoogste gemiddelde aantal berichten per dag behaalden. We hebben hierbij de bots uitgesloten (wanneer er duidelijk te zien was dat de berichten geautomatiseerd werden gepubliceerd). Bij dit onderdeel probeer ik de berichten handmatig in te delen in groepen op basis van de relatie met een leefpatroon. Wanneer uit een bericht dus (door een mens) handmatig geconcludeerd kan worden dat het bericht een indicatie is voor het uitvoeren van een bepaalde gebeurtenis die te maken heeft met een bepaald leefpatroon, wordt deze opgenomen in de bijbehorende groep. De groepen gebeurtenissen die bij leefpatronen horen die wij bekijken per case zijn: School, Reizen (of vervoer), Slapen, Feest, Eten, Vrije Tijd, Tv, Werken, Sport en Verzorging. Deze indeling is gebaseerd op het vermoeden dat de mensen die het meest berichten plaatsen op Twitter jonge studenten of scholieren zijn.

9.3.2 Bijzonderheden geval 1

De eerste gebruiker waarvan we de berichten hebben bekeken is de gebruiker met userID 26556456. Deze gebruiker had volgens de vorige sectie een publicatiegemiddelde van ongeveer 82%. Wat bij deze gebruiker direct opvalt, is dat ze Twitter voornamelijk gebruikt als middel om gesprekken te voeren met mensen. 78,4% (2491 berichten van de 3176) van de berichten beginnen met een '@'-teken en zijn dus direct gericht aan een andere persoon. Veel van deze stukjes van gesprekken bleken niet bruikbaar, maar soms kon je, door het gebruik van context, wel bepaalde gebeurtenissen afleiden die te maken hadden met een leefpatroon.

Voor deze gebruiker hebben we 3176 berichten bekeken. Deze berichten waren geplaatst in een tijdsbestek van 21 dagen. (Dit is een gemiddelde van 151 berichten per dag, dus blijkbaar is de publicatiefrequentie van deze persoon toegenomen de laatste tijd, aangezien eerder bleek dat het gemiddelde 82 berichten per dag was.) Je kunt voor elk bekeken leefpatroon zien hoeveel deze gebruiker over dit patroon bericht heeft in de kolom "geval 1" van Tabel 5 op Pagina 22.

Op het eerste gezicht zijn er voor deze gebruiker dus genoeg berichten om een globale schets te krijgen van de momenten dat de gebruiker slaapt, eet, werkt en tv kijkt. Opvallend hieraan is dat deze gebruiker redelijk consequent voordat ze gaat slapen vermeld dat ze gaat slapen en dat ze 's morgens vroeg ook meteen publiceert dat ze is opgestaan. Dit patroon kan bij deze persoon dan ook het beste worden afgeleid. Helaas is dit patroon eigenlijk als een van de enige patronen makkelijk af te leiden met de aanname dat mensen die slapen geen berichten zullen publiceren. Om een slaappatroon te achterhalen kun je dus ook kijken naar wanneer de gebruiker (vrijwel) nooit berichten op Twitter plaatst. Hier hoef je eigenlijk geen semantische analyse voor te doen.
Een ander opvallend gegeven voor deze persoon is dat het in eerste instantie leek alsof de persoon nog op school zat. Ze had het namelijk over "naar de les gaan" en dergelijke. Uit een bericht bleek echter dat het om een docente ging, en dat we dus de groep "school" eigenlijk moeten combineren met de groep "werken". Op deze manier krijgen we voor het leefpatroon "werken 59 datapunten.

> Ik ga maar weer eens naar de les. Laters.
> @rmk1972 geef zelf les,Nederlands

Op basis van bovenstaande gegevens (en de leeftijd van de gebruiker), kon ik afleiden dat de persoon niet naar school ging om les te volgen, maar om les te geven. Hierdoor komt de groep gebeurtenissen die te maken hebben met het leefpatroon "school" eigenlijk te vervallen en behoren alle berichten die te maken leken te hebben met "school" tot de groep van gebeurtenissen die met "werken" te maken hebben. Dit is voor een algoritme denk ik zeer lastig om af te leiden.

9.3.3 Bijzonderheden geval 2

Voor deze gebruiker hebben we 2917 berichten bekeken. Deze berichten waren geplaatst in een tijdsbestek van 65 dagen. (Dit is een gemiddelde van 44 berichten per dag, dus blijkbaar is de publicatiefrequentie van deze persoon afgenomen de laatste tijd, aangezien eerder bleek dat het gemiddelde van deze persoon 67 berichten per dag was.) Je kunt voor elk bekeken leefpatroon zien hoeveel deze gebruiker over dit patroon bericht heeft in de kolom "geval 2" van Tabel 5 op Pagina 22.

Deze persoon doet naast werk ook het huishouden, vandaar dat we huishoudelijke taken ook tot de categorie "werken" hebben gerekend. Deze gebruiker publiceert het meeste over de tijdstippen dat ze aan het werk is en de tijdstippen dat ze met eten bezig is of aan het eten is. Deze patronen lijken dan ook het beste afgeleid te kunnen worden uit haar berichten. Opvallend aan deze gebruiker was wel dat werktijden redelijk nauwkeurig werden vermeld in het bericht. Wanneer je deze dus goed zou kunnen parseren, zou je veel informatie kunnen verkrijgen over wanneer deze persoon aan het werk is. De berichten zien er bijvoorbeeld zo uit:

> Vanmiddag werken van 1 tot half 9 Nog 5 uren te gaan. #Nachtdienst

9.3.4 Bijzonderheden geval 3

Voor deze gebruiker hebben we 3082 berichten bekeken. Deze berichten waren geplaatst in een tijdsbestek van 16 dagen. (Dit is een gemiddelde van 192 berichten per dag, dus blijkbaar is de publicatiefrequentie van deze persoon toegenomen de laatste tijd, aangezien eerder bleek dat het gemiddelde 67 berichten per dag was.) Je kunt voor elk bekeken leefpatroon zien hoeveel deze gebruiker over dit patroon bericht heeft in de kolom "geval 3" van Tabel 5 op Pagina 22.

Deze gebruiker heeft een erg hoog publicatiegemiddelde, maar erg weinig berichten van de berichten die ze publiceert, heeft ze zelf geschreven. Deze gebruiker lijkt graag mee te doen aan loterijen die gespeeld worden via Twitter. Bedrijven kunnen, om hun naamsbekendheid te vergroten, reclameberichten plaatsen

op Twitter en vervolgens een prijs verloten onder de mensen die deze berichten herhaalt. Dit zodat mensen die geabonneerd zijn op de berichten van deze persoon (en deze berichten dus ontvangen in hun timeline) dit bericht ook zien. Deze gebruiker lijkt systematisch mee te doen met dit soort acties en herhaald dus erg veel van dit soort berichten. Hierdoor heeft deze gebruiker dus een erg hoog publicatiegemiddelde, maar eigenlijk zijn er maar weinig berichten die over haar eigen situatie gaan. Slechts 1224 van de 3082 berichten waren niet direct gekopieerd. Voor analyse hebben we dus vrij weinig informatie over deze gebruiker. Dit is de reden dat de aantallen van berichten die te maken hebben met de leefpatronen genoemd in Tabel 5 zo laag zijn.

Verder bleek uit de handmatige analyse van de Twitterberichten van deze persoon dat het een huisvrouw betreft met twee kinderen. Er kon niet uit de berichten opgemaakt worden dat deze persoon nog studeerde of dat ze een baan had. Vandaar dat we berichten met betrekking tot huishoudelijke taken en taken die te maken hadden met de opvoeding van de kinderen onder het leefpatroon "werken" hebben geplaatst. Toch geeft dit aan dat de statische indeling van de mogelijke leefpatronen niet voldoende is en eigenlijk zouden we een oplossing moeten bedenken die verschillende soorten patronen kan ontdekken in de berichten zonder dat de soort patronen vooraf zijn vastgesteld.

9.3.5 Bijzonderheden geval 4

Voor deze gebruiker hebben we 2558 berichten bekeken. Deze berichten waren geplaatst in een tijdsbestek van 21 dagen. (Dit is een gemiddelde van 121 berichten per dag, dus blijkbaar is de publicatiefrequentie van deze persoon toegenomen de laatste tijd, aangezien eerder bleek dat het gemiddelde 67 berichten per dag was.) Je kunt voor elk bekeken leefpatroon zien hoeveel deze gebruiker over dit patroon bericht heeft in de kolom "geval 4" van Tabel 5 op Pagina 22.

Uit de analyse blijkt dat het hier een huisvrouw betreft met een kind. Deze gebruiker rapporteert ook veel over de momenten dat haar zoon naar school moet, dus de berichten die onder het leefpatroon "school" vallen hebben meer betrekking op de zoon van de gebruiker dan op de gebruiker zelf. Dat neemt niet weg dat dit belangrijke informatie is die invloed heeft op de dagelijkse activiteiten van de gebruiker. Als je namelijk kleine kinderen hebt, kun je stellen dat wanneer ze naar school zijn je meer tijd hebt voor het huishouden of andere activiteiten. Aangezien deze persoon geen baan lijkt te hebben hebben we ook bij dit geval indicatoren van activiteiten die met het huishouden te maken hebben en met het opvoeden van de kinderen gerekend tot de categorie "werken". Dat we steeds op basis van context moeten vaststellen dat de categoriën zoals ze in het begin van de handmatige analyse zijn vastgesteld niet passend zijn, geeft aan dat we eigenlijk bij geautomatiseerde analyse ook niet kunnen gaan zoeken naar vaste categoriën van leefpatroon-activiteiten. Hier meer over bij de sectie algemene bevindingen.

9.3.6 Bijzonderheden geval 5

Voor deze gebruiker hebben we 3041 berichten bekeken. Deze berichten waren geplaatst in een tijdsbestek van 75 dagen. (Dit is een gemiddelde van 40 berichten per dag, dus blijkbaar is de publicatiefrequentie van deze persoon de laatste tijd afgenomen, aangezien eerder bleek dat het gemiddelde 61 berichten per dag was.) Je kunt voor elk bekeken leefpatroon zien hoeveel deze gebruiker over dit patroon bericht heeft in de kolom "geval 1" van Tabel 5 op Pagina 22.

Deze persoon bleek een journalist te zijn en gebruikte hierdoor zijn Twitteraccount op een wat professioneler niveau. Vandaar dat hij veel reageert op andere Twittergebruikers en publiceert hij minder over zijn privé leven en dus zijn dagelijkse werkzaamheden. 2404 berichten waren direct aan anderen gericht (79%) en deze reacties waren vaak aan werk gerelateerd. We kwamen dus op basis van zijn Twitterberichten niet veel te weten over zijn leefpatronen, zoals in de tabel te zien is. Echter, het meeste kwamen we te weten over de tijdstippen dat de persoon televisie keek en naar welke programma's hij dan keek.

Ook bij het analyseren van deze Twittergebruiker liepen we tegen het feit aan dat de categoriën van leefpatronen niet goed aansloten bij de leefsituatie van de gebruiker. De gebruiker noemde namelijk vaak concerten en evenementen die hij bezocht, en dit zouden wij normaal gesproken tot de categorie "feest" of "vrije tijd" rekenen. Echter, tijdens de handmatige analyse kwamen wij erachter dat de persoon in kwestie een journalist was die verslag moest doen van deze evenementen. De evenementen moeten we dan eigenlijk rekenen tot gebeurtenissen die te maken hebben met zijn werk.

9.3.7 Algemene bevindingen

Uit bovenstaande case studies zijn een paar nieuwe bevindingen aan het licht gekomen, maar sommige eerder gestelde vermoedens moeten we ook bijstellen.

Tijdsverwijzingen

Een probleem dat ik ontdekte tijdens het handmatig analyseren van de berichten van de proefpersonen was dat de proefpersonen vaak wel vertelde wat ze wanneer doen, maar dat dit vaak met tijdsverwijzingen werd gepubliceerd. Hierdoor kun je dus niet meer de tijd van het plaatsen van het bericht zomaar gebruiken, maar zou je dit moeten bijstellen met de semantische gegevens uit het bericht. We moeten dus ook rekening proberen te houden met tijdsverwijzingen en moeten eigenlijk een manier zien te vinden die de tijdsverwijzingen die in de tekst van het bericht staan parseert. Dit gegeven maakt ons model vele malen moeilijker. Om een agenda te herconstrueren kan namelijk niet meer de plaatsingstijd en datum gebruikt worden, maar deze tijd en datum moet eerst nog gecorrigeerd worden op basis van de gevonden tijdsverwijzingen in een bericht. Een paar voorbeelden:

> Ik ga morgen lekker naar de Efteling:).

Dit bericht geeft vrij veel informatie over het leefpatroon "vrije tijd". Echter, als we al een manier hebben om dit bericht geautomatiseerd te interpreteren als een indicator voor het patroon "vrije tijd", dan zal dit een foutief datapunt opleveren. Er wordt namelijk gewoon het tijdstip van plaatsing gebruikt, terwijl "morgen" als tijdsverwijzing in het bericht aangeeft dat de activiteit de dag erna pas plaatsvindt.

> Klaar met vergadering 1. Van 16.30u tot 18.00u vergadering 2. Pfff.

In dit bericht zitten drie belangrijke gegevens verstopt. Namelijk dat de gebruiker net klaar is met vergaderen (we kunnen dus stellen dat de gebruiker ongeveer een uur hiervoor aan het vergaderen was), maar ook dat er binnenkort nog een vergadering bijgewoond zal worden. Om dit te parseren naar gebeurtenissen op bepaalde tijdstippen moet dit bericht dus gekoppeld worden aan het patroon "werken". Dit zal niet zo moeilijk zijn door het woord "vergadering" wat redelijk makkelijk geautomatiseerd gevonden kan worden. Maar daarnaast moet de tijdsverwijzing dat er een vergadering is van 16.30u tot 18.00u ook geparseerd worden en moet op basis hiervan en de tijd van plaatsing van het bericht de tijden van de gebeurtenis vastgesteld worden.

> Ik moet er ff niet aan denken dat ik tot 21.30u moet werken vandaag... de zon wint t op deze lange dag:)!

Voor dit voorbeeld geldt hetzelfde. Dit bericht bevat veel informatie over de werktijden op de desbetreffende dag. Echter, dit staat redelijk los van de tijd waarop dit bericht is geplaatst. Er wordt door de gebruiker vooruit gekeken in haar planning, en dit heeft weinig te maken met de tijd waarop het bericht is geplaatst.

Context

In sommige situaties bleek dat je uit een bericht op zich niet zo veel conclusies kon trekken over wat de gebruiker aan het doen was, maar dat je, doordat je meerdere berichten van de gebruiker in chronologische volgorde bekeek, wel een conclusie kon trekken over de huidige gebeurtenis. De samenhang tussen verschillende berichten kunnen we dus blijkbaar niet altijd verwaarlozen. Dit verschijnsel is ook te verklaren door het feit dat mensen soms direct na het plaatsen van een bericht nog een tweede bericht plaatsen die het een en ander nog eens toelichten. Ook de limitatie van 140 tekens per Twitterpublicatie draagt bij aan dit effect. Wellicht moeten we bij het analyseren van leefpatronen in berichten niet elk bericht afzonderlijk bekijken, maar alle berichten die in die tijd dicht bij elkaar zijn geplaatst als een bericht zien.
Een voorbeeld[3]:

> Wat eten we vanaaf? Suggesties? Graag makkelijke dingen #geenkeukenprinses #geenzinomtekoken 2011-03-15 14:50:00
> Dankzij @fred050grunn weet ik wat ik ga eten: uitsmijter!! 2011-03-15 14:59:00
> @GJHahn klinkt ook goed, maar ik ga voor de uitsmijter :) 2011-03-15 15:00:00

[3]We hebben hier de publicatietijden ook vermeld zodat het duidelijk is dat de berichten snel na elkaar zijn gepubliceerd, deze tijden zijn overigens in GTM +0000

Dit voorbeeld bestaat uit drie berichten. In het eerste bericht staat een vraag wat de gebruiker zal gaan eten die avond. Vervolgens wordt hier een antwoord op gevonden. Wanneer deze drie berichten niet gecombineerd zouden worden, zou de analyse kunnen uitwijzen dat er om 3 uur 's middags een uitsmijter gegeten is. Uit de context (die ook een tijdsverwijzing bevat) blijkt dat dit niet juist is, het gaat namelijk over het avondmaal. Wanneer de context in dit geval dus niet in beschouwing zou worden genomen, zou dit resulteren in twee foutieve datapunten.

Woordverbasteringen

De berichten op Twitter worden voornamelijk geplaatst door mensen. De berichten zijn vaak van informeel karakter en daardoor is de gebruiker vaak erg creatief met de berichtgeving. Tijdens de case studies ben ik dan ook tegengekomen dat Twittergebruikers regelmatig toch vaak afwijken van de Nederlandse spelling. Je ziet dat woorden fonetisch geschreven worden, eigen afkortingen worden gebruikt en daarnaast veel stijlfiguren worden gehanteerd. Ook het gebruik van woorden uit andere talen komt regelmatig voor. Juist omdat mensen dagelijkse dingen relatief vaak verwoorden op Twitter, wordt de behoefte van de gebruikers steeds groter om de informatie op een creatieve manier te publiceren. Dit zal de geautomatiseerde semantische analyse van Twitterberichten er niet makkelijker op maken. Dit is dus een extra moeilijkheid. Wanneer we geen rekening houden met woordverbasteringen en stijlfiguren, zullen veel berichten genegeerd worden. Voorbeelden hiervan zijn:

> Zeau. De ochtendles zit erop. Zometeen thuis lunsju. En dan vanaaf weer terug :S.
> Off 2 class. Laterrrrrrrrrrrrrrr.
> Sterkte nog, ik ben over duh helluf :). Bijna wiekent!
> Staaaatsiejon Den Bosch. Here we go.

Hashtags

Hashtags worden vaak aan berichten toegevoegd die te maken hebben met leefpatronen. Het meest lijken ze gebruikt te worden bij berichten waaraan je kunt afleiden dat de gebruiker TV kijkt. Voorbeelden hiervan zijn:

> Jah,dit is leuk:)! #thesingoff
> Gentle Voices.. hm..ik ben nu al fan :)! #thesingoff
> Ik zet mezelf voor de tv en baal dat ik er niet bij ben. COME ON FEYENOORD!!!!!!!!! #FEYnac

Het lijkt erop dat je op basis van deze hashtags ook patronen kunt ontdekken. Hashtags zijn er ook voor bedoeld om te gebruiken wanneer er berichten worden geplaatst die betrekking hebben op publieke evenementen of media gebeurtenissen. We zullen bij het gedeelte van de semantische analyse van Twitterberichten in dit document dan ook nog verder onderzoek naar het gebruik van Hashtags doen.

Soort patronen

We hebben ook gezien dat de soort leefpatronen niet vast moeten staan. Aangezien we uitgingen van het vermoeden dat juist jongere studenten of scholieren veel gebruik maakte van Twitter, hadden we hier de indeling van leefpatronen op gebaseerd. Echter, niet elke gebruiker heeft te maken met een leefpatroon "School", met een leefpatroon "Sporten" enz. Daarnaast blijken gebruikers leefpatronen te hebben die wij niet eens hebben meegenomen in onze indeling van leefpatronen (opvoeding kinderen, huishouden etc.). Dit kunnen wij als mens bij een handmatige analyse opmaken uit de totale verzameling van berichten, achtergrondinformatie en interpretatie. Maar een computer kan geen begrip krijgen van de gezinssituatie, woonsituatie of leefsituatie. Hieruit blijkt dus dat een oplossing waarbij we zoeken naar een vaste indeling van leefpatronen eigenlijk geen volledige oplossing is. Het beste kunnen we zoeken naar een oplossing die geautomatiseerd patronen zoekt in een verzameling Twitterberichten en dan op basis van het meest voorkomende woord in de berichten die tot dit gevonden patroon behoren gebruiken als naam voor het leefpatroon.

9.4 Afweging voldoende informatie

Nu we gekeken hebben naar de gemiddelde publicatiefrequentie van Twittergebruikers, de gebruikte onderwerpen en de verzameling van berichten van de gebruikers met de hoogste publicatiefrequenties, hebben we een beter beeld van hoe Twitter gebruikt wordt en welke informatie betreffende leefpatronen uit de Twitterpublicaties te halen zijn. We hebben gezien dat de gemiddelde publicatiefrequentie rond de 5.8 lag bij de gebruikers die in onze dataset waren opgenomen. Dit is lager dan we als uitgangspunt hadden genomen en dit

houdt in dat een betrouwbare analyse naar leefpatronen zeker niet bij elke Twittergebruiker toegepast kan worden. Maar 3.5% van de gebruikers uit onze dataset heeft een publicatiegemiddelde dat hoger ligt dan 30 berichten per dag. De publicatiefrequentie die wij graag gezien hadden met als uitgangspunt dat de gebruiker elk half uur een bericht plaatst, wordt dus maar door een zeer kleine groep behaald. De tool die leefpatronen kan analyseren zou dan dus maar op een zeer kleine groep gebruikers betrouwbaar kunnen werken. Echter, een lagere publicatiefreqentie wil niet meteen zeggen dat er geen informatie over leefpatronen gevonden kan worden, wanneer deze frequentie lager ligt, kunnen er berichten over een langer tijdsbestek bekeken worden (gezien het limiet van 3200 berichten die nog te downloaden zijn). Aangezien we deze berichten toch willen projecteren op een week, zou dit nog steeds informatie over leefpatronen kunnen bevatten.

Ook hebben we gekeken naar welke woorden het meest gebruikt werden in Twitterberichten. Hierbij waren de resultaten redelijk positief. De meest gebruikte woorden leken namelijk direct een verband te hebben met leefpatronen en zouden we (wanneer we abstraheren van de context) kunnen stellen dat onderwerpen die te maken hebben met dagelijkse dingen redelijk populair zijn. Dit spreekt dus voor ons vermoeden dat leefpatronen af te leiden zouden zijn uit de berichten. Echter, of deze woorden ook gebruikt worden om aan te geven wat voor een handelingen de Twittergebruikers op dat moment aan het uitvoeren waren, is helaas niet af te leiden uit deze woordfrequentie analyse. Wel hebben we ook gezien dat woorden die tijdsverwijzingen aangeven veel gebruikt worden. Hieruit blijkt dat we dus niet direct de tijd van het plaatsen van het bericht als de tijd waarop de activiteit plaats vond kunnen gebruiken.

Daarnaast hebben we voor vijf personen (met het hoogste publicatiegemiddelde) handmatig voor een paar leefpatronen de Twitterberichten ingedeeld. Hieruit kwam naar voren dat vooral voor de patronen "tv", "slapen" en "werken" de meeste berichten gevonden waren. Toch bleek ook dat het vooraf vaststellen van de categoriën niet de juiste keuze was. Voor twee van de geanalyseerde personen bleek dat hun leefsituatie anders was dan in eerste instantie vanuit gegaan was. Dit waren namelijk geen schoolgaande mensen, maar vrouwen die zorgdraagden voor het huishouden en de opvoeding van de kinderen.

Concluderend kunnen we stellen dat de Twitterberichten van mensen met een hoge publicatiefrequentie inderdaad redelijk wat informatie bevatten over dagelijkse dingen die herhaaldelijk terug lijken te komen. Ook de meest gebruikte woorden op Twitter lijken een relatie te hebben met leefpatronen en een mens kan bij gebruikers redelijk goed achterhalen wat een persoon de hele week door doet. Echter, niet iedere Twittergebruiker plaatst zoveel berichten zodat de gewenste nauwkeurigheid behaald wordt. Ook zijn we tegengekomen dat er veel tijdsverwijzingen gebruikt worden in de berichten en de context waarin de berichten geplaatst worden is ook niet volledig te verwaarlozen. Dit zal het geautomatiseerd extraheren van leefpatronen aanzienlijk moeilijker maken en leefpatronen zullen zeker niet bij elke gebruiker evengoed zichtbaar worden. Ook zal het geautomatiseerd filteren van berichten op basis van leefpatronen erg nauwkeurig moeten plaatsvinden want de berichten die duidelijke aanwijzingen geven zijn schaars. In de volgende sectie zullen wij dan ook gaan behandelen wat hier de beste methode voor is.

10 Semantische Analyse

We moeten uit alle Twitterberichten van een Twittergebruiker af gaan leiden wat hij op welk moment aan het doen is. Hiervoor moeten we de kern van elk Twitterbericht kunnen achterhalen. We moeten namelijk voor elk bericht proberen te bepalen of een bericht een gebeurtenis beschrijft die te maken heeft met een bepaald leefpatroon. In onze formele theorie, die we eerder hebben gepresenteerd, is deze stap aangeduid met de functie $describes(b : B, e : E)$. We moeten voor de tool die we willen ontwerpen een manier vinden om te bepalen of berichten wel of niet bij gebeurtenissen van een bepaalde leefpatroon horen. Hieronder zullen we een aantal mogelijkheden uiteenzetten en bekijken of ze geschikt zijn voor ons probleem of ze rekening kunnen houden met de in de vorige sectie gedentificeerde moeilijkheden. Deze moeilijkheden waren:

- Mensen maken in hun berichten veelvuldig gebruik van tijdsverwijzingen waardoor de tijd van het plaatsen van het bericht niet direct meer te gebruiken is. Dit is in het algemeen op te lossen met een tijdscorrectie van de Twitterberichten zoals we die verderop presenteren. Zie sectie 10.8 op pagina 36.
- De context van berichten is niet altijd te verwaarlozen, soms is het enkel mogelijk de situatie in te schatten wanneer je meerdere (snel achter elkaar geplaatste) berichten combineert. Dit behandelen we in sectie 10.7 op pagina 35.

- Woordverbasteringen komen ook regelmatig voor. De Twittergebruiker wil graag op een creatieve en informele manier dingen vertellen aan de buitenwereld en dit volgt niet altijd de gangbare spelling- en taalregels. Een (algemene) oplossing hiervoor presenteren we in sectie 10.9 op pagina 37.

- We kunnen de soorten patronen waar we naar zoeken niet statisch definiëren. De toepasbaarheid van leefpatronen verschillen tussen gebruikers erg sterk en het is niet uit te sluiten dat door de programmeur een soort leefpatroon niet opgemerkt wordt en over het hoofd wordt gezien. We zullen hieronder bij het bekijken van de verschillende methoden van het geautomatiseerd zoeken naar leefpatronen in de verzameling van Twitterberichten ook bekijken of deze overweg kunnen met niet vooraf gedefiniëerde set van leefpatronen of niet. De noodzakelijkheid van de mogelijkheid om variabele soorten leefpatronen te kunnen analyseren moet echter nog nader bekeken worden in mogelijk vervolg onderzoek. Zie sectie 12 op pagina 40

10.1 Zoeken naar primaire aanduiding leefpatroon

We moeten eerst eens gaan bekijken of het niet al een voldoende resultaat geeft als we enkel naar woorden zoeken die primair aanduiden wat de activiteit is waar de gebruikers op dat moment mee bezig zijn. Voor het leefpatroon dat wij "'eten" hebben genoemd, is dat bijvoorbeeld het woord "eten" en "drinken". Wellicht krijgen we voor dit patroon al genoeg berichten als we enkel op deze twee woorden zoeken. Om dit te bekijken, gaan we even terug naar de gevonden berichten tijdens onze handmatige analyse van Twitterberichten wat we behandeld hebben in sectie 9.3. Wanneer we enkel naar de berichten kijken van het leefpatroon "eten" zien we dat als we een spellingscontrole uitvoeren, we een deel van de berichten al kunnen terugvinden als we gewoon naar het woord "eten" op zoek gaan. Voor geval 1 krijgen we dan 4 van de eerder gevonden 25 resultaten, voor geval 2 21 van de 52 resultaten, voor geval 3 2 van de 10 resultaten, voor geval 4 47 van de 66 en voor geval 5 1 van de 5 resultaten. Voor dit leefpatroon lijkt het dus nog een redelijk resultaat op te leveren om enkel naar het woord "eten" te zoeken. In onze eerdere analyse van woordfrequentie 9.2 bleek dat het woord "eten" over het algemeen ook vaak gebruikt werd. Echter, eten is ook iets dat elke Twittergebruiker moet doen om te overleven en buiten het feit dat je het anders kunt omschrijven, is de kans groot dat Twittergebruikers wel eens berichten "dat ze gaan eten". Er is dan ook weinig variatie mogelijk om maaltijden aan te duiden zonder in te gaan op details van het gerecht (iets wat mensen niet vaak doen in een Twitterbericht). Dit geldt niet zo sterk voor leefpatronen als werken en al helemaal niet voor het leefpatroon dat wij "vrije tijd" hebben genoemd. Vrijwel geen enkele gebruiker zal berichten dat hij öp dit moment vrije tijd heeft". Voor geen enkele van de eerder bekeken gevallen krijg je dan ook een resultaat als je naar "'vrije tijd" zoekt. We kunnen dan ook wel stellen dat naar mate er meer variatie is in de soort activiteiten die bij een leefpatroon horen, hoe lastiger het is om bijbehorende berichten met enkele steekwoorden op te zoeken. Enkel zoeken naar deze primaire aanduiding van leefpatronen is dus niet voldoende voor een goede analyse van leefpatronen.

10.2 Zoeken naar meest gebruikte woorden

We hebben net gezien dat enkel zoeken naar de primaire aanduiding niet voldoende zal zijn om alle leefpatronen uit een verzameling Twitterberichten te extraheren. Een andere simpele methode is om te gaan kijken naar de meest gebruikte woorden in de verzameling Twitterberichten van een Twittergebruiker en op basis van deze woorden de berichten die daar bij horen te verzamelen. Op deze manier kun je ook patronen herkennen in momenten die met dezelfde verwoording zijn omschreven. Deze methode heeft als voordeel dat je niet vast zit aan een bepaalde voorselectie van leefpatronen waarnaar gezocht kan worden. Hiermee lossen we dus het eerder genoemde probleem dat de soorten leefpatronen waar naar gezocht wordt vast zouden staan op. Zo'n beetje elke soort leefpatroon zou zichtbaar kunnen worden wanneer je gaat zoeken naar berichten die de meest voorkomende woorden bevatten. Deze methode heeft echter ook een groot nadeel. Omdat mensen graag op verschillende manieren dingen willen berichten met een andere bewoording (omdat mensen zoals eerder al besproken vaak creatief zijn), zullen voor gebeurtenissen die bij eenzelfde leefpatroon horen meerdere patronen gevonden worden wanneer deze methode wordt gebruikt. Wellicht komt het patroon niet eens uit de analyse als resultaat naar voren als er op zeer veel verschillende manieren een gebeurtenis verwoord kan worden. Wanneer deze methode toch wordt toegepast, en de meest voorkomende woorden geselecteerd zouden worden, krijg je als het waren allemaal subpatronen die bij eenzelfde leefpatroon zouden kunnen horen (omdat de gebeurtenissen van dit leefpatroon met meerdere woorden kan worden vastgelegd). De link tussen de verschillende subpatronen die gevonden zijn en het uiteindelijke leefpatroon waar ze bij horen wordt door

deze methode nooit gelegd. Dit zou dan nog open blijven voor interpretatie van de gebruiker van de tool. Deze zou dan zelf nog moeten bepalen welke subpatronen bij elkaar horen en een leefpatroon definiëren. Daarnaast zal deze oplossing in verhouding behoorlijk veel processorkracht vereisen om steeds bij elke analyse van leefpatronen een volledige woordfrequentie analyse te doen op de beschikbare Twitterberichten van de geselecteerde gebruiker.

10.3 Hashtags

Omdat Twitter ook een mogelijkheid heeft om berichten een tag (een soort kenmerk) te geven, is het ook een idee om daar naar te kijken. Deze tags worden Hashtags genoemd en zijn steekwoorden in het bericht waar een # voor is gezet. Op het moment dat mensen vervolgens naar berichten met deze hashtag zoeken (met de zoek mogelijkheid van Twitter) komt ook dit bericht voor in de resultaten (mits je Twitterberichten niet beveiligd zijn). Wellicht hoeft er geen uitgebreide semantische analyse gedaan te worden op de Twitterberichten wanneer blijkt dat gebruikers met behulp van hashtags al aangeven waar het bericht over gaat. Wanneer het hashtag symbool (#) consequent gebruikt wordt om de gebeurtenis waarover het bericht gaat aan te geven, kunnen we op basis van enkel deze hashtags de semantiek van de berichten bepalen en kijken bij welke leefpatronen het bericht hoort. We zullen dus moeten bekijken of gebruikers hashtags consequent gebruiken maar ook of ze daadwerkelijk hashtags gebruiken om dagelijkse gebeurtenissen aan te geven (waar wij in eerste instantie in geïnteresseerd zijn).

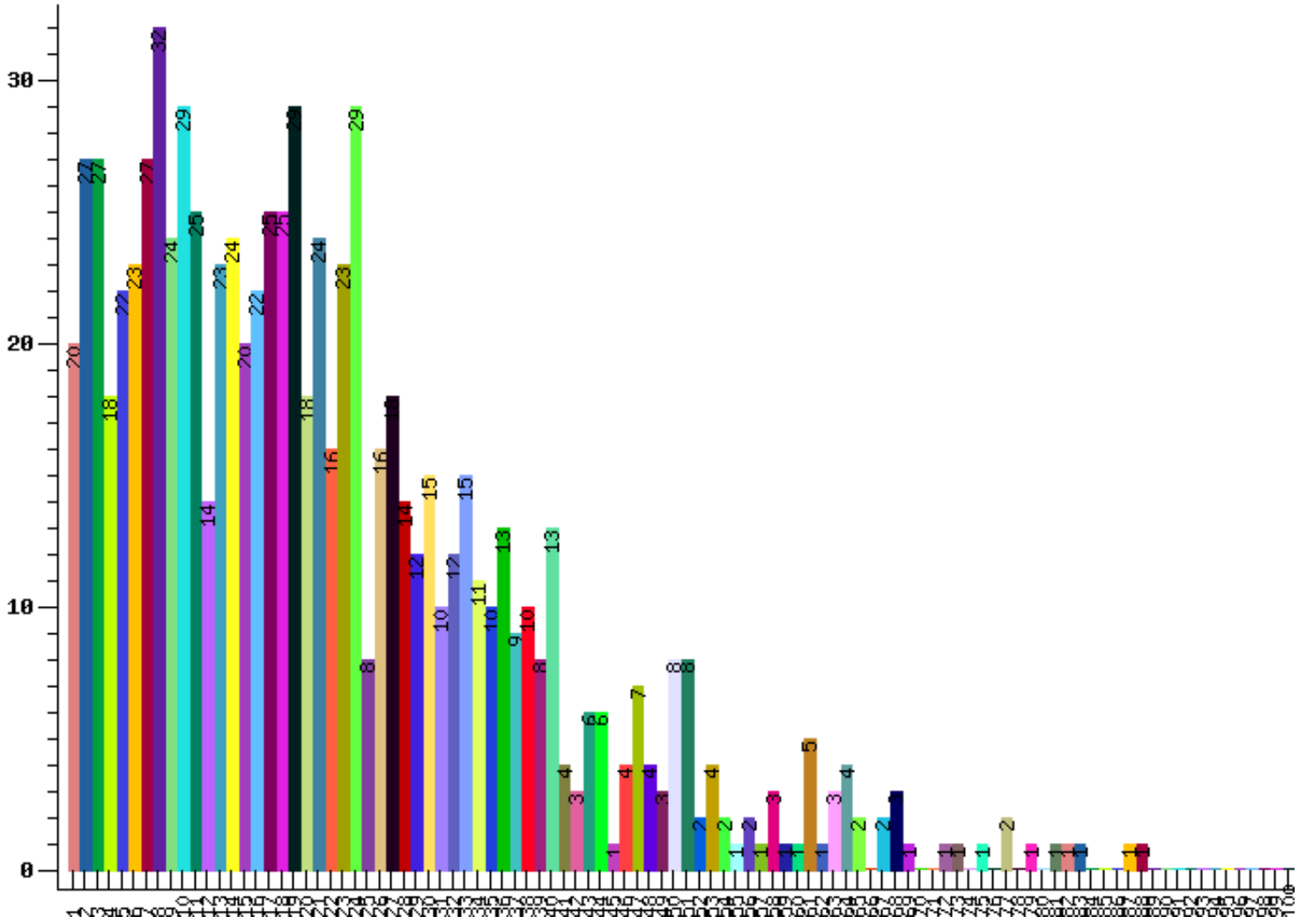
10.3.1 Gebruik Hashtags

Allereerst heb ik gekeken of Hashtags wel voldoende gebruikt worden door de gebruikers. Hieruit bleek dat er gemiddeld 248 berichten voorzien waren van een of meer hashtags. Echter waren er gemiddeld ook 947 Twitterberichten geplaatst zonder een enkele hashtag. Met een totaal van 1195 Twitterberichten gemiddeld, komt het aantal Twitterberichten dat gemiddeld voorzien is van een hashtag uit op een 21%, ruim een-vijfde deel. Een andere opmerkelijke vondst is dat 65 van de 1195 Twittergebruikers die ik bij deze test heb bekeken helemaal nooit gebruik maakt van Hashtags. Hieronder volgt een grafiek met een histogram van het gebruik van hashtags (Zie Figuur 2 op pagina 30). Per percentage (x-as) is de frequentie van gebruikers met een dergelijke percentage van Twitterberichten met hashtag uitgezet (op de y-as).

10.3.2 Meest gebruikte hashtags

Voor de verzameling Twitterberichten die we hebben verworven tijdens het systematisch downloaden, heb ik ook bekeken wat de meest gebruikte hashtags zijn. Voor elke hashtag heb ik bijgehouden hoe vaak de tag is gebruikt door de verzameling gebruikers. In totaal is er door de 947 gebruikers gezamenlijk 304.870 keer een hashtag gebruikt in hun Twitterberichten. We hebben in alle Twitterberichten 304.870 verschillende Hashtags gevonden. De hashtag "#durftevragen" is het meeste gebruikt en is 6750 keer gevonden in de Twitterberichten. Dit lijkt veel, maar eigenlijk is dit maar 2,2% van de hashtags die gevonden zijn. Hashtags worden dus over het algemeen niet consequent gebruikt. Dit is te verklaren door het feit dat er nooit regels, conventies of richtlijnen zijn opgesteld voor het gebruik van hashtags. Iedereen kan zelf bepalen welke woorden hij voorziet van een # teken en welke niet. Van de 304.870 verschillende hashtags zijn er 60.141 uniek (19,7%). Deze zijn in alle Twitterberichten slechts een keer gevonden.
Op een van de volgende bladzijden staat een tabel (Tabel 6 op pagina 31) met de honderd meest gebruikte hashtags en het patroon waar de tag waarschijnlijk betrekking op heeft.

Zoals je uit tabel 6 makkelijk kun opmaken, worden de meeste hashtags gebruikt om aan te geven waar verschillende informatie betrekking op heeft. Hashtags worden meer gebruikt in combinatie met nieuwsfeiten of media gerelateerde opmerkingen dan om aan te geven welke activiteit een bepaald bericht betrekking op heeft. Hashtags die verwijzen naar tv programma's, muziek, politieke partijen, of nieuwsfeiten die in de media veel aandacht hebben gekregen, komen dan ook zeer vaak voor in de verzameling Twitterberichten die ik bekeken heb. Dit komt goed overeen met de vondsten in het onderzoek van Kathryn Corrick [3]. Hierin wordt namelijk ook gesteld dat de meeste trends over Entertainment gaan(28%). Volgens de supportpagina's van Twitter is een hashtag bedoeld om het onderwerp of kernwoord van je bericht aan te geven [17]. Dit met achterliggende gedachten dat wanneer mensen geïnteresseerd zijn wat anderen over een bepaald onderwerp denken, kunnen zoeken naar alle berichten met als hashtag dit onderwerp. Voor televisieprogramma's en politiek lijkt dit systeem dus goed te werken. Per politieke partij kun je zoeken wat er door anderen gezegd



Figuur 2: Aantal Twittergebruikers (y-as) dat een bepaald percentage (y-as) van zijn Twitterberichten van Hashtag voorzag.

Hashtag	Aantal	Patroon	Hashtag	Aantal	Patroon
durftevragen	6750	<geen >	tk2010	350	politiek
iPhone	4110	technologie	tmobile	347	technologie
fail	3126	<geen >	cda	336	politiek
3fm	2393	muziek	nuiphone	335	technologie
tvoh	2336	tv	Groenlinks	333	politiek
Nijmegen	2110	reizen	hunt	333	e-commerce
dwdd	2005	tv	zucht	321	emotie
FF	1888	<geen >	F1	317	tv / sport
fb	1545	<geen >	fe11	316	<geen >
penw	1434	tv	D66	311	politiek
twexit	1420	slaap	utrecht	306	reizen
twvhj	1315	<geen >	ochtend	306	slaap
sr10	1236	muziek	nieuwsuur	303	tv / nieuws
h	1208	<geen >	omroepgld	297	tv
nieuws	1098	nieuws	leuk	295	emotie
NS	1077	reizen	RT	294	<geen >
widm	968	tv	carnaval	285	uitgaan
in	916	<geen >	erecode	285	<geen >
Androidworld	874	technologie	vacature	281	werk zoeken
BZV	802	tv	Apple	277	technologie
wk2010	741	tv / sport	spiritinthesky	275	<geen >
android	735	technologie	lente	272	<geen >
ajax	731	tv / sport	pvda	270	politiek
lastfm	702	muziek	GTST	268	tv
wordpress	679	technologie	SPANNEND	268	emotie
ps2011	644	politiek	Tempoteam	268	werk zoeken
koffie	614	consumptie	SxSW	264	muziek
ibood	613	e-commerce	zandhazendurp	264	<geen >
kenniscrisis	552	<geen >	top2000	263	muziek
lol	544	emotie	DeJaap	261	<geen >
ipad	515	technologie	RunKeeper	260	<geen >
rtvnh	497	tv	trots	258	emotie
pvv	473	politiek	japan	256	reizen
zinin	462	emotie	rutweetup	254	<geen >
	461	<geen >	goedemorgen	251	slaap
dtv	459	tv	oznz	249	tv
sstnl	450	tv	hardlopen	248	sport
PowNews	445	tv / nieuws	effeekdom	245	muziek
rosmalen	439	reizen	ohohcherso	245	tv
nos	425	tv / nieuws	KPN	241	technologie
ekstraweekend	416	muziek	gezellig	240	emotie
xfactor	411	tv	VVD	238	politiek
twitter	401	<geen >	FFF	238	<geen >
nosdebat	393	tv / politiek	sport28	238	tv / sport
HvA	383	tv	Amsterdam	237	reizen
sneeuw	380	<geen >	VI	236	tv / sport
nowplaying	375	muziek	ned	233	<geen >
egypte	373	reizen	moerdijk	233	reizen
ochtendopradioNH	366	muziek	VK	230	nieuws
feyenoord	354	tv / sport	Radio1	229	muziek

Tabel 6: Meest gebruikte hashtags met hun voorkomens en het patroon waar ze toe zouden kunnen behoren.

wordt over politieke gebeurtenissen. Ook voor muziek en nieuwsfeiten worden de hashtags redelijk consequent gebruikt. Onze vindingen bij dit onderdeel bevestigen dan ook het vermoeden dat we eerder geformuleerd hebben dat hashtags het meest gebruikt worden voor TV programma's en andere media gebeurtenissen. Shoko Wakamiya, Ryong Lee en Kazutoshi Sumiya hebben in hun onderzoek [19] dan ook laten zien dat de populariteit van televisieprogramma's ook goed te meten is op basis van het gebruik van de hashtags die bij een televisieprogramma horen. Op basis van de populariteit van de hashtags kan dus bepaald worden wat de kijkcijfers voor een bepaald programma zijn. Echter, voor dagelijkse activiteiten die behoren tot een leefpatroon wordt dit blijkbaar minder consequent gedaan. De lijst met hashtags is dan ook helemaal niet representatief voor de lijst met de meest gebruikte woorden die we gezien hebben in de Sectie "Meest gebruikte onderwerpen" op pagina 19. Wanneer we gaan zoeken naar hashtags die gebruikt zijn om activiteiten aan te duiden die wel met een leefpatroon te maken zouden kunnen hebben, zien we dat deze veel minder vaak gebruikt worden. Hieronder 70 hashtags die gevonden zijn in de berichten en die mij wel relevant lijken bij het bepalen van leefpatronen (Tabel 7 op pagina 33).

Het aantal voorkomens van deze hashtags valt eigenlijk sterk tegen. Voor veel hashtags geldt ook nog eens dat het toewijzen van een leefpatroon aan de tag speculatief is. Bijvoorbeeld degene op de laatste plek "ah1596". Deze hashtag verwijst naar een filiaal van de winkelketen Albert Hein. 40 keer komt deze hashtag voor in de verzameling. Nu kan het zijn dat mensen die bij dit filiaal werken in hun publicaties aangeven wanneer ze bij de AH aan het werk zijn. Echter, het kan ook toevallig zo zijn dat veel verschillende mensen bericht hebben dat ze boodschappen gaan doen bij dit filiaal van de Albert Hein. Een ander voorbeeld is de hashtag "Heineken". Deze tag kan ook gebruikt worden om activiteiten van verschillende leefpatronen aan te geven. Het kan zijn dat er op dat moment Heineken gedronken wordt op een terrasje (wanneer het bij het patroon vrije tijd zou horen) maar het zou ook goed kunnen dat tijdens het uitgaan Heineken wordt gedronken. Als laatste is het ook nog goed mogelijk dat een Twitterbericht enkel over de kwaliteit, reclame, of andere eigenschappen van het product gaat. In dat geval behoort het helemaal niet tot een indicator van een leefpatroon. Aangezien we niet de context van het Twitterbericht gebruiken bij het bepalen van de leefpatronen, kunnen er makkelijk foutieve classificaties voorkomen. Dit tezamen met het gegeven dat hashtags meer gebruikt worden om te verwijzen naar nieuws, politiek, tv programma's en muziek te verwijzen denk ik niet dat hashtags voldoende zijn om te gebruiken voor ons systeem bij het bepalen welke Twitterberichten bij welk leefpatroon horen.

Voor het bepalen tot welk patroon of gegevenssoort een bepaalde hashtag behoort heb ik in sommige gevallen gebruik gemaakt van een internet woordenboek voor hashtags. Deze is te vinden op `http://tagdef.com/`. Op deze website kunnen mensen zelf betekenissen toekennen aan hashtags en kunnen ze een beoordeling geven aan ingezonden betekenissen van anderen. De betekenis is echter dynamisch aangezien iedereen zelf een hashtag mag verzinnen. Verschillende Twittergebruikers kunnen dan ook dezelfde hashtag gebruiken voor totaal verschillende dingen. Ik heb steeds de betekenis met de hoogste beoordeling gebruikt.

10.3.3 Hashtag gebruik per Twittergebruiker

Als we kijken naar het gebruik van hashtags per gebruiker om te zien hoeveel gebruikers hashtags wel consistent en consequent gebruiken, zien we eigenlijk meteen al dat een bepaalde hashtag door een bepaalde gebruiker niet vaker dan een of twee keer gebruikt wordt. 345 gebruikers gebruikten een bepaalde tag gemiddeld een keer. 399 gebruikers gebruikten een bepaalde tag 2 keer. Daarnaast waren er nog 101 gebruikers die een tag 3 of 4 keer gebruikten en een paar uitzonderingen die een bepaalde tag vaker gebruikten. Dit laatste is te verklaren door het feit dat mensen soms hun eigen hashtag verzinnen en die consequent te gebruiken in al hun berichten. Dit zodat ze makkelijk kunnen zoeken naar berichten van zichzelf en alle herhalingen van publicaties (mensen kunnen namelijk ook berichten van anderen herhalen). Dit is dus helaas ook niet genoeg om duidelijke patronen te vinden.

10.4 Associatiegraaf

We hebben hierboven een paar simpele methoden gezien om berichten die te maken hebben met leefpatronen geautomatiseerd te extraheren uit een verzameling Twitterberichten. Deze methoden bleken toch als voornaamste tekortkomingen te hebben dat ze niet voldoende gebeurtenissen konden categoriseren op basis van welk leefpatroon de berichten toebehoorden. We moeten dus voor betere methodes meer ingaan op de daadwerkelijke semantiek van berichten.
Een mogelijkheid om de semantiek van de berichten te analyseren is door een associatiegraaf te gebruiken.

Hashtag	Aantal	Patroon	Hashtag	Aantal	Patroon
twexit	1420	slaap	werkspot	76	werk
koffie	614	consumptie	tentamen	73	studie
ochtend	306	slaap	nespresso	71	consumptie
carnaval	285	uitgaan	luiezondag	71	vrije tijd
goedemorgen	251	slaap	gaap	69	slaap
hardlopen	248	sporten	zzz	69	slaap
Studeren	203	studie	school	69	studie
onderwijs	176	studie	werken	67	werk / studie
psychologie	174	studie	proost	67	consumptie
schaatsen	166	hobby / vrije tijd	connexxionfail	66	reizen
vakantie	154	vakantie	bus	65	reizen
opweg	154	reizen	deadlineavond	64	studie
klm	150	reizen / werk	saxion	63	studie
studiemarathon	148	studie	vroeg	62	slaap
twuste	144	slaap	lunch	61	consumptie
trusten	144	slaap	duinrell	61	vrije tijd
stickykoffie	132	consumptie	ovchip	58	reizen
efteling	121	vrije tijd	webdesign	57	hobby / werk
ru	116	studie	onderwijsavond	57	studie / werk
bier	112	consumptie	tudelft	56	studie
goodnight	111	slaap	tentamens	55	studie
file	105	reizen	wintersport	55	vakantie
honger	103	consumptie	heineken	52	consumptie
stage	101	studie	vertraging	51	reizen
werk	100	werk / studie	brak	49	uitgaan
trein	97	reizen	eten	48	consumptie
watetenwevandaag	93	consumptie	connexxion	48	reizen
welterusten	91	slaap	scriptie	45	studie
nomnomnom	90	consumptie	prorail	45	reizen
slaaplekker	88	slaap	spits	43	reizen
schiphol	85	reizen	studenten	43	studie
slaap	81	slaap	dinnertime	42	consumptie
nsfail	80	reizen	studie	40	studie
truste	79	slaap	stageplaatsen	40	studie
weltrusten	77	slaap	ah1596	40	werken

Tabel 7: Hashtags met hun aantal voorkomens die waarschijnlijk wel gebruikt kunnen worden om een leefpatroon te extraheren

Wanneer je een associatiegraaf hebt met gewichten tussen alle woorden, kun je om bijvoorbeeld het leefpatroon "werken" te verkrijgen alle Twitterberichten uit de verzameling te halen die het woord "werken" bevatten of een van de woorden bevatten die een voldoende hoog relatiegewicht heeft met het woord "werken". Wanneer een relatiegewicht genoeg hoog is wordt vooraf ingesteld door de programmeur, gebaseerd op welke waarde de betrouwbaarste resultaten gaven (uit een serie eerder geteste waarden). De Twitterberichten die je op deze manier verkrijgt zouden aan moeten geven dat de Twittergebruiker op de momenten van plaatsen van het bericht bezig was met het leefpatroon "werken".

10.5 Associatiegraaf maken

Voor bovenstaande methoden moeten we wel eerst een associatiegraaf hebben waarin de relaties tussen bepaalde woorden worden gedefinieerd. Wanneer twee woorden in eenzelfde Twitterbericht voorkomen, kun je het gewicht dat de sterkte van de relatie tussen twee woorden aangeeft verhogen. Dit kun je voor een sample van Twitterberichten doen om je associatiegraaf te crieëren. Voor het analyseren van nieuwe Twitterberichten gebruikt het systeem dan de vooraf geleerde associatiegraaf.
Daarnaast zou je voordat je patronen gaat zoeken ook per gebruiker een associatiegraaf kunnen maken met enkel de berichten van deze Twittergebruiker. Op deze manier kun je ook gebruiker-specifieke namen (die betrekking hebben op de gebeurtenissen behorende tot een bepaald leefpatroon, bijvoorbeeld de naam van het bedrijf waar de persoon werkt) en persoonlijke woordassociaties (veelvuldig gebruikte metaforen) gebruiken voor de analyse. Je combineert de relatiegewichten uit de algemene associatiegraaf en de gebruikers associatiegraaf vervolgens en op basis hiervan ga je kijken welke woorden te maken hebben met een door de programmeur ingevoerd woord dat sterk gerelateerd is aan een leefpatroon (in het voorbeeld in de vorige sectie het woord "werken").

Echter, het probleem van het uitrekenen van een associatiegraaf op basis van een set Twitterberichten is computationeel veel te groot. Wanneer je alleen al een verzameling van 1000 Twittergebruikers zou bekijken en van al deze gebruikers 3200 Twitterberichten zou ophalen waar het gemiddelde aantal woorden per bericht 15 is, zou je ongeveer $2^{48miljoen}$ relatiegewichten tussen woorden moeten uitrekenen. Zelf een associatiegraaf afleiden uit het gebruik van woorden in Twitterberichten is dus niet haalbaar.

10.5.1 Nederlandse Woord Associatie Project

Voor de Nederlandse taal is in 2003 een grootschalig onderzoek opgestart door Simon De Deyne en Gert Storms [6], het zogenaamde Nederlandse Woord Associatie Project. Dit project is er op gericht een associatie graaf samen te stellen op basis van gegevens die vrijwillige deelnemers invoeren. Als deelnemer wordt je gevraagd voor 18 woorden de belangrijkste 3 woorden op te schrijven die de persoon met het gegeven woord associeert [7]. Helaas zijn de volledige gegevens van de graaf die tot nu toe is ontwikkeld niet gepubliceerd, maar deze gegevens zouden we als basis kunnen gebruiken voor onze tool wanneer deze gebruik zou maken van een associatiegraaf. De auteurs van dit onderzoek zijn eerder meer geïnteresseerd in het taalkundige en psychische aspect van woordassociaties. Hun eerste vindingen zijn te lezen in het stuk "Word associations: Network and semantic properties" [**?**]. Wel geven de auteurs aan dat hun toekomstige doel is om via een website het publiek de mogelijkheid te bieden een grafische weergave van de associaties van een bepaald woord te bekijken. Of de gehele dataset ooit publiek gemaakt wordt is niet bekend.

10.5.2 Enkel vooraf gedefinieerde patronen

Wanneer we deze techniek zouden gaan toepassen, hebben we voor een probleem dat we eerder geïdentificeerd hebben geen oplossing: Er wordt geen rekening gehouden met de soorten van leefpatronen die bij verschillende leefsituaties horen. De programmeur moet bij deze oplossing namelijk nog steeds zelf aangeven dat er gezocht moet worden naar berichten die te maken hebben met bijvoorbeeld "werken" of "school". We hebben eerder al gezien, dat de toepasselijkheid van een leefpatroon per gebruiker verschilt. Dit probleem is wel op te lossen door voor een uitgebreide set van leefpatronen te zoeken naar berichten die daar betrekking op hebben en enkel patronen weer te geven waarvan een minimaal aantal berichten gevonden is, maar met deze methoden zullen nooit alle patronen ontdekt worden.

10.6 Support Vector Machine

Om te bepalen of een bepaald Twitterbericht een indicator is voor een bepaald soort natuurgeweld, hebben Sakaki en zijn team een Support Vector Machine [12] gebruikt. Elk bericht dat uit een bepaald geografisch gebied afkomstig is wordt door de Support Vector Machine bekeken. Het wordt dan geclassificeerd of het te maken zou kunnen hebben met bijvoorbeeld een aardbeving of een tyfoon of niet. Ze hebben voor het geval van het identificeren van aardbevingen en tyfonen de SVM 597 voorbeelden van Twitterberichten die een aardbeving of tyfoon rapporteren geleerd en ook een hoeveelheid negatieve voorbeelden. Ze gebruiken de SVM om berichten te filteren die wél het woord "earthquake" of "typhoon" bevatten, maar niet daadwerkelijk een rapportering van een dergelijke natuurverschijnsel zijn. Echter is dit een iets makkelijker probleem dan het probleem dat onze tool moet oplossen. De tool van Sakaki en zijn team zoekt namelijk naar Twitterberichten met het woord "earthquake" of "thyfoon" erin. Ze krijgen hierdoor dus een set Twitterberichten die sowieso te maken hebben met aardbevingen of tyfonen. Ze hoeven nu enkel nog maar de false positives eruit te halen die bestaan uit Twitterberichten die wel over aardbevingen of tyfonen gaan maar niet echt een gebeurtenis van natuurgeweld rapporteren (iemand kan bijvoorbeeld berichten dat hij naar een conferentie over aardbevingen gaat). Om deze te filteren volstaat een SVM. Echter, bij onze aanpak hebben we een grote set van berichten van een bepaalde gebruiker, en moeten we deze in gaan delen in verschillende categoriën van leefpatronen. Het voldoet niet om enkel bepaalde woorden die bij leefpatronen te horen te zoeken in de verzameling van berichten van de gebruiker aangezien er op heel veel verschillende manieren kan worden aangeduid met welke activiteit ze momenteel bezig zijn. Eventueel zouden we wel voor elk leefpatroon waar we naar willen zoeken in de verzameling van Twitterberichten van een bepaald persoon een aparte SVN kunnen trainen die voor elk bericht uitsluitsel kan geven of het bericht bij een bepaald leefpatroon hoort of niet, maar onze dataset is per tijdseenheid ook veel kleiner. Sakaki maakt gebruik van real-time informatie, maar heeft wel veel meer bronnen op een bepaald moment. Wanneer een bericht dat over natuurgeweld gaat onterecht wordt afgekeurd door de SVM, heeft dit nog geen grote gevolgen. Hoogstwaarschijnlijk zullen er namelijk binnen korte termijn meer berichten volgen over hetzelfde natuurgeweld dat wel als positief resultaat wordt gezien door de SVM. Als er bij onze tool echter een bericht wordt afgekeurd dat wel bij een bepaald leefpatroon hoorde, missen we gelijk een belangrijk tijdstip in onze data. Dit kan net het tijdstip zijn geweest om een bepaald patroon net wel te vinden. Voor een bepaalde gebeurtenis op een bepaalde tijd hebben we vaak maar een rapportage en false negatives zijn in ons geval dus veel problematischer.

10.6.1 Aanpak

Een mogelijke aanpak voor het verkrijgen van berichten die te maken hebben met leefpatronen is:

- Voor alle mogelijke patronen waarin we geïnteresseerd zijn een SVM trainen. Dit kan door handmatig een voor elk patroon een set positieve en negatieve voorbeelden van berichten samen te stellen en deze aan de SVM te leren.
- Alle SVM (voor elk leefpatroon een), voor elk Twitterbericht een uitspraak laten doen.
- Voor alle berichten de tijd opzoeken en corrigeren zodat de tijdstippen allemaal in dezelfde week vallen.
- Een correctie toepassen voor semantische tijdsverwijzingen[4].

Deze oplossing is computationeel goed haalbaar. Echter, voor een vast aantal soorten leefpatronen moet je in deze situatie een SVM gaan trainen. Ook hierbij is de tool dus niet erg flexibel en zit de tool vast aan vooraf bepaalde leefpatronen waar hij naar kan zoeken. We hebben eerder bij onze handmatige analyse van Twitterberichten al gezien dat dit problemen op kan leveren wanneer mensen in leefsituaties zitten waar de programmeur bij het ontwikkelen van de tool geen rekening mee gehouden heeft.

10.7 Oplossing context probleem

Omdat we in het voorgaande hoofdstuk gezien hebben dat het vaak voorkomt dat mensen berichten die met elkaar te maken hebben snel achter elkaar publiceren kunnen we de Twitterberichten niet altijd als een eenheid beschouwen. De berichten zijn op Twitter maximaal 140 tekens lang, maar we kunnen bij onze analyse van de semantiek van het bericht in principe oneindig lange teksten gebruiken. Hoe langer de tekst,

[4] Zie de sectie "Tijdscorrectie op basis van semantiek" op pagina 36

hoe meer mogelijke woorden die zouden kunnen refereren aan een gebeurtenis die te maken heeft met een leefpatroon. De beste oplossing om rekening te houden met de context van bepaalde berichten is dan ook het combineren van berichten die snel na elkaar zijn geplaatst. Op basis van een bepaalde grenswaarde die de maximale tijd bepaald waarin berichten elkaar moeten opvolgen om met elkaar gecombineerd te worden moet vooraf (door de programmeur) bepaald worden. Waar we precies deze grenswaarde het beste kunnen stellen moet experimenteel worden vastgesteld. Dit experiment zou in een mogelijk vervolgonderzoek uitgevoerd kunnen worden.

10.8 Tijdscorrectie op basis van semantiek

We hebben zowel bij de woordfrequentie analyse 9.2 op pagina 19 als bij de handmatige analyse van Twitterberichten 9.3 op pagina 22 van een aantal personen ontdekt dat er toch veelvuldig gebruik wordt gemaakt van tijdverwijzingen. Hierdoor kun je er niet meer vanuit gaan dat de persoon de activiteit die het bericht omschrijft ook daadwerkelijk op dat moment aan het uitvoeren is. Om dit probleem op te lossen zullen we de tijd van het bericht moeten aanpassen op basis van de geparseerde tijdsaanduiding in het bericht. Ook tijdsverwijzingen kunnen namelijk op oneindig aantal manieren genoteerd worden en deze zijn ook onderheven aan verbasteringen en taalfouten.

10.8.1 Onduidelijke tijdsverwijzingen

Uit de woordfreqentie analyse kwam naar voren dat de volgende tijdsverwijzingen bij de 150 meest voorkomende woorden zaten: vanavond, week, weekend, dagen, gisteren, avond, vanmiddag, vrijdag, zaterdag, zondag, weken, maandag, minuten, geleden, donderdag, gister, woensdag, binnenkort en april. Voor deze verwijzingen kunnen in principe de tijden van de berichten gemuteerd worden. Bijvoorbeeld voor de weekdagen geldt dat de berichten verschoven kunnen worden naar de desbetreffende weekdag in het weekrooster dat we willen reconstrueren. Voor woorden als "vanavond", "gisteren" en "gister" kan de tijd van het bericht ook een dag vooruit of achteruit verschoven worden. Maar hiermee wordt een probleem zichtbaar: Wanneer wij verwijzen naar een andere dag wordt hier doorgaans geabstraheerd van het tijdstip waarop de genoemde activiteit plaats heeft gevonden of plaats zal vinden. Hierdoor wordt de tijd die aan het bericht is gekoppeld dus eigenlijk onbetrouwbaar. De tijd waarop het bericht geplaatst is, heeft namelijk totaal geen betrekking meer op het tijdstip die bij de genoemde activiteit hoort. We kunnen wanneer deze tijdsverwijzingen geïntroduceerd worden in een bericht dus niets meer zeggen over het tijdstip waarop de activiteit wordt uitgevoerd, enkel nog de dag.

Hetzelfde geldt eigenlijk voor de tijdsverwijzingen "vanavond", "vanmiddag", en andere tijdsverwijzingen wel naar een moment op dezelfde dag verwijzen, maar geen absolute betekenis hebben van een tijdstip aanduiding. Ook hiervoor geldt dat je enkel een bereik aan kunt geven van tijdstippen waarvoor de activiteit die in het bericht wordt omschreven kan gelden. Om berichten dus in een weekrooster te kunnen zetten, moeten we waarschijnlijk de mogelijkheid bieden om berichten een bereik van tijd te geven in plaats van een exact tijdstip. Op deze manier kun je bij bijvoorbeeld de aanduiding "vanavond" het bericht een bereik geven het tijdstip van het bericht tot aan het einde van de desbetreffende dag wanneer het bericht na 18:00 is geplaatst of een bereik van 18:00 tot 24:00 wanneer het bericht eerder dan 18:00 is geplaatst. Het is nog wel haalbaar om voor dit soort tijdsverwijzingen die bestaan uit een woord een dergelijke lijst met vertalingen te maken.

10.8.2 Letterlijke tijdsaanduidingen in bericht

Naast de onduidelijke tijdsverwijzingen kwamen we in berichten ook letterlijke tijdsverwijzingen tegen die bijvoorbeeld precies de tijdstippen aangaven van wanneer tot wanneer de persoon in kwestie moest werken. Deze berichten waren bijvoorbeeld in de vorm:

> Van 16.30u tot 18.00u vergadering
> Ik moet nog 4.5 uren werken.
> Morgen al om 7 uur werken!!
> Moe. Nog 2.5 uur werken

Het geeft zeer waardevolle inzichten als we deze berichten goed zouden kunnen parseren. In deze berichten worden namelijk letterlijk de werktijden verteld en dit is voor het leefpatroon "werken" hetgeen we naar op zoek zijn. Ook hiervoor geldt dat deze informatie een tijdsbereik aan het bericht geeft. Wanneer er gesteld

wordt dat de persoon nog 4.5 uur moet werken, kun je hier niet alleen uithalen dat de persoon op het moment van het plaatsen van het bericht aan het werk is, maar dat dit ook het geval zal zijn voor de komende 4.5 uur.
Deze tijdsaanduidingen zijn ook te herkennen omdat de tijden vaak worden aangegeven met cijfers (dit heeft ook te maken met het feit dat je in een Twitterbericht maar 140 tekens kunt plaatsen, dus uitschrijven van tijden neemt dan te veel ruimte in.) Er zijn dan ook regels op te stellen voor het parseren van tijdsaanduidingen. Het valt buiten de scope om deze parseerfunctie tot in de detail te gaan bekijken, maar een paar voorbeelden van de regels die gebruikt zouden kunnen worden zijn te vinden in tabel 8 op pagina 38. Hierbij worden de getallen genummerd aangegeven met een G. Berichttijden worden veranderd met de functies $addTime(berichttijd, uu : mm)$ en $substrTime(berichttijd, uu : mm)$ die respectievelijk de tijd van argument 2 toevoegen aan de tijd van het bericht in argument 1 dan wel de tijd van argument 2 aftrekken van de tijd van het bericht van argument 1. Daarnaast hebben we nog de functies $setTime(berichttijd, uu : mm)$ en $setTimeRange(berichttijd, uu : mm, uu : mm)$ nodig om de berichttijd op een bepaalde tijdswaarde te zetten of een tijdsberijk te geven. Deze functies houden er rekening mee dat de datum van het bericht mogelijk ook verandert wanneer de tijden worden aangepast. De tijden die aan de functies meegegeven worden zijn in het formaat uu:mm waar uu staat voor de uren en mm staat voor de minuten. S wordt bij de aanduidingen gebruikt als aanduiding van een scheidingsteken, er kan bijvoorbeeld gebruikt zijn: ':', '.' of '-'. Dat een dergelijke set van regels niet in alle gevallen zal werken blijkt al uit het volgende voorbeeld (dat gevonden is in de berichtenset van geval 1 van onze case studie):

> Naar de les. Nog 'maar' 7,5 uur.

Hieruit blijkt dat de parsering van tijdsaanduiding ook rekening zal moeten houden met de mogelijkheid dat er halve uren worden aangeduid (met een komma) en dat er ook nog woorden tussen "nog" en de cijfers van de tijdsaanduiding kunnen staan. Hetzelfde geldt eigenlijk ook voor de positie van de cijfers van de tijdsaanduiding en "uur", mensen kunnen namelijk ook zoiets publiceren als:

> Naar de les. Nog 7,5 zware uren.

Deze regels om tijdsaanduidingen te parseren zouden dus nog verder uitgebreid moeten kunnen worden om alle mogelijke tijdsverwijzingen in berichten te kunnen interpreteren. Voor dergelijke tijdsverwijzingen kunnen regels in de vorm van bijvoorbeeld reguliere expressies worden opgesteld en wanneer een bericht overeenkomt met een reguliere expressie kunnen bepaalde mutaties op de tijdsaanduiding die bij het bericht hoort worden gedaan. Dit is een mogelijk aspect dat bekeken kan worden in eventueel vervolgonderzoek. Ook omdat het parseren van tijden uit natuurlijke taal een universeel probleem is waar nog geen wetenschappelijk onderzoek naar lijkt te zijn gedaan (voor de Nederlandse taal).

10.9 Verbastering van de taal

10.9.1 140 Tekens

Twitterberichten hebben een maximale grootte van 140 tekens. Dit limiet is gesteld omdat Twitter een microblog dienst is. Het is niet de bedoeling dat de gebruikers via deze dienst hele verhalen gaan delen want hier zijn normale blog diensten voor. De essentie van Twitter is juist om kleine berichten te delen zodat mensen die jou volgen ook snel kunnen lezen wat je aan het doen bent of wat er aan de hand is. Toch zie je regelmatig dat de lengtegrens van 140 woorden een lastige beperking is en dat mensen eigenlijk meer ruimte nodig hebben om hun bericht te publiceren. Dit is helemaal vaak het geval als mensen hun bericht aan iemand anders richten (want het vernoemen van een andere gebruiker kost ook ruimte) of wanneer gebruikers een hyperlink naar een website willen delen (wat vaak ook veel ruimte inneemt). Soms wordt het bericht bij een tekort aan ruimte in meerdere Twitterberichten opgedeeld, of wordt het bericht door de applicatie waarmee berichten worden gepubliceerd automatisch op een andere website geplaatst en wordt enkel het begin van het bericht met een link naar het gehele bericht op Twitter gepubliceerd. Maar vaak zie je dat mensen dingen gaan afkorten. Juist omdat mensen vaak net een aantal karakters te veel in hun bericht hebben. Hierbij worden niet alleen maar gangbare afkortingen gebruikt, maar worden ook eigen afkortingen bedacht. Bijvoorbeeld: "van" wordt "vn" en "het" wordt "t". Dit gebeurd helaas niet alleen met voegwoorden en lidwoorden maar ook met zelfstandig naamwoorden. Hierdoor wordt het extra lastig om geautomatiseerd te analyseren waar een Twitterbericht over gaat en bij welk leefpatroon het bericht behoord.

Aanduiding	Vertaling
Nog G1G2 uur	$setTimeRange(berichttijd, berichttijd, addTime(berichttijd, G1G2 : 00))$
Nog G1G2 uren	$setTimeRange(berichttijd, berichttijd, addTime(berichttijd, G1G2 : 00))$
Tot G1G2 uur	$setTimeRange(berichttijd, berichttijd, addTime(berichttijd, G1G2 : 00))$
Over G1G2 uur	$addTime(berichttijd, G1G2 : 00)$
Over G1G2 minuten	$addTime(berichttijd, 00 : G1G2)$
Over G1G2 min	$addTime(berichttijd, 00 : G1G2)$
Van G1G2 tot G3G4	$setTimeRange(berichttijd, G1G2 : 00, G3G4 : 00)$
Van G1G2SG3G4 tot G5G6SG7G8	S is een scheidingsteken, bijvoorbeeld (':', '.','-'). $setTimeRange(berichttijd, G1G2 : 00, G3G4 : 00)$
Van G1G2u tot G3G4u	$setTimeRange(berichttijd, G1G2 : 00, G3G4 : 00)$
Van G1G2SG3G4u tot G5G6SG7G8u	S is een scheidingsteken, bijvoorbeeld (':', '.','-'). $setTimeRange(berichttijd, G1G2 : 00, G3G4 : 00)$
Om G1G2 uur	$setTime(berichttijd, G1G2 : 00)$
Om G1G2SG3G4 uur	$setTime(berichttijd, G1G2 : G3G4)$
Half G1G2	$setTime(berichttijd, G1G2 : 30)$
Kwart voor G1G2	$setTime(berichttijd, G1G2 - 1 : 45)$
Kwart over G1G2	$setTime(berichttijd, G1G2 : 15)$

Tabel 8: Voorbeelden van interpretaties van tijdsaanduidingen.

10.9.2 Spellingsfouten

Tijdens de handmatige analyse van de Twitterberichten viel ons ook op dat de berichten vaak vol met spellingsfouten zitten, dit kan verschillende redenen hebben:

- De gebruiker besteed minder aandacht aan hetgeen de gebruiker schrijft omdat het door de gebruiker niet gezien wordt als officiële publicatie. Twitter wordt vooral gebruikt in de informele sfeer en de algemene opvatting is dat teksten die gepubliceerd worden dan niet erg verzorgd hoeven te zijn. Het gaat om de gedachte die telt en die komt ook over wanneer er mogelijk een spellingsfout in het bericht zit.

- Het toetsenbord dat gebruikt wordt is in veel gevallen een toetsenbord op een mobiel apparaat. Deze kent vaak een aantal beperkingen in bijvoorbeeld afmeting en haptische waarneming. Bij een mobiel apparaat kunnen deze eigenschappen meestal niet volledig toereikend doorgevoerd worden omdat het apparaat dan al snel te groot wordt. Omdat de invoer niet optimaal is, worden er ook meer fouten gemaakt tijdens het typen. Dit zorgt voor spellingsfouten en tesamen met eerder genoemde punt, worden deze spellingsfouten ook sneller gepubliceerd.

- Mensen willen vaak op een creatieve en originele manier hun bericht overbrengen. Hiervoor wordt er doorgaans bewust afgeweken van de gangbare taalregels en spelling. Wat je vaak tegenkomt is dat mensen aanduidingen fonetisch in hun bericht plaatsen. Mensen kunnen het dan wel begrijpen wanneer ze het bericht lezen, maar voor een geautomatiseerde analyse is dit onbruikbaar.

- Zoals eerder besproken: Het limiet van 140 tekens zorgt ervoor dat mensen op een onnatuurlijke manier woorden gaan inkorten of spaties verwijderen.

Wanneer een tekst spellingsfouten bevat maakt dit de geautomatiseerde analyse van de berichten een stuk lastiger. We zullen dus bij het ontwikkelen van de beoogde tool om levenspatronen van mensen te achterhalen rekening moeten houden met mogelijke spellingsfouten. We kunnen dus het beste een automatische spellingscontrole uitvoeren op de berichten voordat we deze verder gaan analyseren. Een voorbeeld van een dergelijk algoritme is het Soundex algoritme dat ontwikkeld is door Russel en O'Dell dat al gepatenteerd is in 1918[13]. David Holmes en M. Catherine McCabe hebben in hun onderzoek [9] gekeken naar hoe ze de nauwkeurigheid van Soundex kunnen verbeteren [9] en een aantal verbeteringen voorgesteld. Voor de tool zouden deze algoritmen ook gebruikt kunnen worden om spellingsfouten uit de berichten te halen. Het

Soundex algoritme is gebaseerd op het principe dat woorden waarbij de klanken van het woord het beste overeenkomen met het woord dat niet in een woordenboek is gevonden, vaak de beste spellingssuggesties zijn. Een onbekend woord wordt dan eerst naar een klanken codering vertaald en vervolgens vergeleken met de klankencodering van elk woord in het woordenboek. De beste overeenkomst wordt vervolgens als suggestie gegeven. Dit principe lost ook meteen het probleem op dat mensen regelmatig woorden fonetisch schrijven. Mensen begrijpen woorden namelijk alleen als we de fonetische klanken herkennen. Dit moet dus overeen komen met de klanken van het woord wanneer het correct gespeld is. Wanneer mensen de fonetische schrijfwijze dus kunnen interpreteren (en hier mogen we vanuit gaan, anders heeft het publiceren van een dergelijk bericht weinig betekenis), zou deze manier van spellingscorrectie ook het goede woord moeten kunnen vinden.

Ook haalt dit algoritme van Holmes volgens hun onderzoeken ook overbodige herhalingen van letters eruit. Op deze manier kunnen dus ook de volgende uitroepen toch geparseerd worden (gevonden bij geval 3 van de handmatige analyse op pagina 22):

goeiemorgennnnnnnnnnnnnnnnnn
goeiemorgennnnnnnnnn
goeieeeeeeeeeeeeeeeemorgen ben nog moe
goeiemorgeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee

Met de spellingscontrole gebaseerd op Soundex kunnen bovenstaande uitroepen van "goeiemorgen" waarschijnlijk allemaal gecorrigeerd worden. Op deze manier heb je 4 berichten waar het woord "goeiemorgen" correct instaat en kan dit door de tool beter geïnterpreteerd worden. Het woord "goeiemorgen" kan namelijk gezien worden als indicator voor het leefpatroon slapen en kan het begin van de dag aangeven.

11 Conclusie

We hebben eerder in dit document al geconcludeerd dat er naar alle waarschijnlijkheid bij mensen met een relatief hoge publicatiefrequentie genoeg informatie over leefpatronen te vinden is in de verzameling van leefpatronen. We hebben hierbij gezien dat wanneer we deze informatie geautomatiseerd uit de informatie van Twitterberichten willen halen moeten letten op een aantal aspecten. Deze aspecten waren dat we naar context moeten kijken, rekening moeten houden met tijdsverwijzingen en woordverbasteringen.

Om onze tweede deelvraag te beantwoorden moesten we gaan bekijken hoe we de moeilijkheden die we eerder tijdens het behandelen van onze eerste deelvraag zijn tegengekomen gaan oplossen om een geautomatiseerde analyse mogelijk te maken.
We hebben nu gezien dat we rekening kunnen houden met de context door berichten die dicht na elkaar geplaatst zijn te zien als een bericht en we hebben gezien dat voor tijdsverwijzingen gemakkelijk regels op te stellen zijn die de tijdsverwijzing interpreteren en de berichttijd bijstellen. Ook hebben we voor de tijd waarop de informatie uit een bericht geldt een bereik geïntroduceerd. Wanneer een tijdsverwijzing gecontroleerde vaagheid bevat (de gebruiker verwijst niet duidelijk genoeg door bijvoorbeeld woorden als "vanavondöf "morgen"konden we namelijk niet meer duidelijk een punt aangeven wanneer de gebeurtenis plaatsvindt of plaatsvond. Dit maakt het ook mogelijk om berichten te interpreteren waarin de gebruiker letterlijk aangeeft van en tot welke tijd ze een activiteit gaan uitvoeren. Ook hebben we gezien dat de spellingsfouten, taalfouten en het fonetisch taalgebruik geïnterpreteerd kunnen worden door voor de analyse een spellingscontrole, gebaseerd op het Soundex algoritme toe te passen dat werkt op basis van fonetische klanken van woorden. Hiermee zouden eventuele ongebruikelijke fouten die voorheen niet te interpreteren waren door de computer toch interpreteerbaar zijn.

Enkel voor de basismethode om berichten te vinden bij bepaalde leefpatronen hebben we nog geen goede afweging gemaakt welke waarschijnlijk het beste resultaat zal opleveren. Om patronen te zoeken kan toch beter gezocht worden naar bepaalde woorden die betrekking hebben op gebeurtenissen die behoren tot een bepaald soort leefpatroon. Het is zoals hierboven besproken niet handig om enkel naar de meest voorkomende woorden te zoeken en de berichten die dit woord bevatten te zien als berichten die bij een leefpatroon horen aangezien je voor verschillende activiteiten die wel bij hetzelfde leefpatroon horen vaak verschillende aanduidingen hebt. We kunnen dus het beste een methode kiezen die woorden categoriseert bij bepaalde vaststaande leefpatronen waarna gezocht wordt. Hierbij heb je dus wel een vaste set leefpatronen waar je naar

zoekt en is het dus zaak om rekening te houden met zo veel mogelijk leefpatronen die in verschillende leefsituaties voor kunnen komen. Of het verschil tussen vaststaande leefpatronen waar je naar zoekt en variabele soorten leefpatronen veel verschil maakt in de resultaten zou in eventueel vervolgonderzoek nog bekeken moeten worden. Zie Sectie 12.
Om woorden bij elkaar te zoeken die met hetzelfde leefpatroon te maken hebben kunnen we ons denk ik het beste richten in de associatiegraaf. Deze methode is computationeel wel erg groot, maar de resultaten van het Nederlandse Woordassociatie Project[6] kan hier uitkomst bieden. Het voordeel van deze aanpak is dat de associaties van dit project echt gebaseerd zijn op menselijke input. Wanneer wij associaties willen gaan gebruiken om te analyseren welke berichten betrekking hebben op een bepaalde leefpatroon-activiteit hebben we ook te maken met tekst die door mensen is opgesteld. Wanneer mensen zoeken naar woorden om de gebeurtenis te berichten zullen ze met eenzelfde associatie mechanisme (dat onbewust in hun hoofd wordt gebruikt) woorden afwegen om in hun bericht te gebruiken. Ditzelfde mechanisme is bij de Nederlandse Woordassociatie Project test[7] die proefpersonen uitvoerden ook op de proef gesteld.

Echter, wanneer de auteurs van het Nederlandse Woordassociatie Project[6] de door hun verkregen associatiegraaf nog niet openbaar hebben gemaakt of niet van plan zijn openbaar te maken, kan de aandacht toch het beste gericht worden naar de Support Vector Machine aanpak die besproken is in sectie 10.6 op pagina 35. Het maken van een representatieve associatiegraaf op basis van verzamelingen Twitterberichten wordt al snel een computationeel veel te groot probleem en er zal een zeer grote set berichten bekeken moeten worden om een betrouwbare associatiegraaf te kunnen genereren. De methode die gebruik maakt van een Support Vector Machine werkte ook bij het onderzoek van Sakaki[14] naar het gebruik van Twitterberichten als indicatoren voor natuurgeweld. We hebben dus een redelijke kans van slagen als we voor alle soorten leefpatronen een verzameling SVM trainen (voor elke bepaalde basisactiviteit die bij deze leefpatroon hoort). SVM kon wel redelijk goed overweg met verschillende benoemingen van dezelfde gebeurtenissen maar kon niet overweg met verschillende gebeurtenissen voor een bepaald leefpatroon. Voor bijvoorbeeld het leefpatroon "eten" en "werken" zou de SVM goed kunnen presteren, maar met het leefpatroon "vrije tijd" zou deze aanpak nog moeilijkheden hebben. Er kan namelijk een breed scala van gebeurtenissen onder de categorie "vrije tijd" vallen en waarschijnlijk is deze categorie te breed voor een goed getrainde SVM. Wel was het zaak de SVM uitgebreid te trainen. Dit zal in het begin veel werk zijn aangezien alle voorbeelden die aan de SVM worden gegeven met de hand zullen moeten worden geselecteerd. Hiervoor zullen dus meerdere handmatige analyses van leefpatronen moeten plaatsvinden.

Al met al hebben we gezien dat er bij mensen met een hoge publicatiefrequentie redelijk wat informatie over leefpatronen zichtbaar is in hun Twitterberichten. Dit is een redelijke indicatie dat het implementeren van de beoogde tool daadwerkelijk resultaat op zou kunnen leveren. Er zijn tijdens dit vooronderzoek ook geen vondsten gedaan die volledig uitsluiten dat een geautomatiseerde methode van het analyseren van leefpatronen te implementeren is. Enkel de complexiteit van het probleem lijkt aan de hoge kant. Zeker wanneer de genoemde oplossing waarbij de resultaten van het Nederlandse Woordassociatie Project gebruikt worden niet mogelijk blijkt te zijn (bijvoorbeeld omdat de auteurs niet hun volledige resultaten publiceren op een manier dat andere onderzoekers het kunnen gebruiken in onderzoek). De implementatie wordt dan een redelijk grote opgave maar lijkt nog steeds mogelijk met behulp van Suport Vector Machines. Toch lijkt het mij nog steeds zinvol dat er verder gekeken wordt naar een implementatie van de tool die leefpatronen zoekt in Twitterberichten, aangezien de informatie over leefpatronen wel aanwezig bleek te zijn in Twitterpublicaties en de in de sectie relevantie (sectie 3.2 op pagina 6) gepresenteerde redenen om hier onderzoek naar te doen nog steeds van toepassing zijn. Naast verdere implementatie van de tool die leefpatronen kan ontdekken in Twitterberichten, zijn er ook nog andere aspecten die meegenomen kunnen worden in een eventueel vervolgonderzoek. Deze zijn uiteengezet in de volgende sectie.

12 Mogelijk Vervolgonderzoek

12.1 Uitsluiten gespreksberichten

Zoals we eerder al besproken hebben, kan Twitter ook gebruikt worden om berichten direct te richten aan een bepaalde andere Twittergebruiker. Op deze manier kun je Twitter ook gebruiken om gesprekken te voeren,

hoewel dit medium daar niet voor bedoeld is[5]. Tijdens onze handmatige analyse van Twitterberichten ontstond ook het vermoeden dat de informatie die gedeeld wordt in Twitterberichten meestal niet relevant is voor het analyseren van leefpatronen. Wanneer uit een bericht dat aan een persoon gericht is wel informatie over deze patronen gehaald konden worden, was dit vaak een gesprek als reactie op een algemene aankondiging van een gebeurtenis aan iedereen. De informatie over de leefpatronen konden in die gevallen meestal ook uit dat algemene bericht gehaald worden. Het vermoeden is dus dat we weinig informatie over leefpatronen verliezen als we besluiten de berichten die direct aan anderen gericht zijn uit te sluiten (beginnende met een teken). Dit is interessant om in een mogelijk vervolgonderzoek nog te bekijken omdat we hiermee vaak een groot deel van de verzameling Twitterberichten kunnen uitsluiten waardoor we minder berichten echt semantisch hoeven te analyseren. Dit levert naar alle waarschijnlijkheid een snelheidswinst op.

12.2 Verschil vaste en variabele soorten leefpatronen

We hebben eerder al gezien bij de handmatige analyse van Twitterberichten tijdens onze vijf case studies 9.3 dat de soorten leefpatronen waar wij naar keken tijdens de analyse niet voor elke leefsituatie geschikt was. Zo was het soort leefpatroon "School" in vrijwel alle gevallen niet (echt) van toepassing en misten we in een aantal gevallen een leefpatroon "Huishouden" of "Opvoeding van kinderen". Nu kun je bij de handmatige analyse stellen dat het soort leefpatroon enkel een benaming is voor een patroon waar je naar zoekt. Echter, voor geautomatiseerde analyse betekent een ander soort patroon in veel van de mogelijke methodes van analyse dat er naar andere aspecten van de Twitterberichten gezocht moet worden (bijvoorbeeld bij het gebruik van een associatiegraaf een ander beginwoord in de graaf, of bij de hashtag methode een andere hashtag waarna gezocht wordt. Een aspect dat in mogelijk verder onderzoek nog bekeken kan worden, is of het zoeken naar patronen het beste geheel variabel mogelijk moet zijn (zodat er geen enkele aanname gemaakt wordt over het soort leefpatroon waarnaar gezocht wordt, en je dus naar algemene patronen in berichten zoekt) of dat een grote set vooraf gedefinieerde leefpatronen voldoende zijn om voor een ruime groep Twittergebruikers leefgewoonten af te leiden aan de hand van hun Twitterpublicaties.

12.3 Populariteit uitgaansactiveiten

We zagen al eerder dat er op basis van Twitterberichten van mensen redelijk goed te analyseren valt welke televisieprogramma's populair zijn en welke niet. Tijdens onze handmatige analyse van Twitterberichten voor een vijftal Twittergebruikers viel mij ook op dat gebruikers vooral een Twitterbericht publiceren wanneer ze een uitstapje maken dat voor hun speciaal is en naar voor hun ongebruikelijke publieke activiteiten gaan. Over deze activiteiten geven de Twittergebruikers vaak informatie over hoe kun beleving is en dit is wellicht waardevolle feedback voor de organisatoren van de bezochte activiteiten. In een mogelijk vervolgonderzoek is het wellicht ook interessant om de mogelijkheid te analyseren om uit de Twitterberichten te extraheren op welke momenten van het seizoen welke activiteiten veel in trek zijn en of het mogelijk is om feedback voor activiteiten te verzamelen die op Twitter gepubliceerd wordt. Voor theatervoorstellingen is het bijvoorbeeld interessant om te weten welk genre van voorstellingen op bepaalde momenten populair zijn op welke momenten van het jaar. Maar ook voor pretparken is het bijvoorbeeld handig om te weten wat er over bezoek aan het park gezegd wordt. Vaak worden enquêteformulieren van het pretpark weinig ingevuld aangezien mensen hier niet altijd toe bereid zijn terwijl mensen vaak wel een oordeel hebben over bepaalde aspecten. Dit geldt vooral voor negatief commentaar. Wanneer een bezoeker van een pretpark namelijk uren in de rij heeft gestaan voor een bepaalde attractie, zal de bezoeker die toevallig ook een Twittergebruiker is, eerder zijn ongenoegen uiten op Twitter dan dat hij of zij een enquêteformulier invult.

Referenties

[1] Martin Bryant *6 Reasons why Twitter Geolocation is a really, really bad idea*, augustus 2009, `http://thenextweb.com/2009/08/21/5-reasons-twitter-geolocation-bad-idea/`, Gezien: Maart 2011.

[2] Martin Bryant *Twitter Geo-fail? Only 0.23% of tweets geotagged*, januari 2010, `http://thenextweb.com/2010/01/15/twitter-geofail-023-tweets-geotagged/`, Gezien: Maart 2011.

[5] Twitter heeft hier zelfs extra beperkingen voor ingesteld om gesprekken te voorkomen die meer weg hebben van Instant Messaging, zo hebben ze bijvoorbeeld een maximaal aantal berichten per uur ingesteld en wanneer je dit aantal overschrijdt, wordt je tijdelijk de toegang tot de Twitter website ontzegt.

[3] Kathryn Corrick *The State of the Twittersphere, Februari 2011*, Februari 2011, http://www.scribd.com/doc/49019436/The-state-of-the-Twittersphere-February-2011, Gezien: Mei 2011.

[4] Kenneth Ward Church, Patric Hank, *Word Association Norm, Mutual Information, and Lexicography*, 1990, Computational linguistics Volume 16 Issue 1, maart 1990.

[5] Gobinda G. Chowdhury, *Natural Language Processing*, 2003, Annual Review of Information Science and Technology Volume 37, Issue 1, pages 5189, 2003.

[6] Simon De Deyne, Gert Storms, *Beschrijving van het Nederlandse Woordassociatie Project*, 2008, http://ppw.kuleuven.be/concat/simon/info_asso.html, Gezien: Mei 2011.

[7] Simon De Deyne, Gert Storms, *Deelnamemogelijkheid studie over woordassociaties*, 2008, http://www.kuleuven.be/lsa/associations/, gezien: Mei 2011.

[8] De Deyne S., Storms G., *Word associations: Network and semantic properties*, 2008, Behavior Research Methods, 40, 213-231.

[9] Holmes, D. and McCabe, M.C., *Improving precision and recall for soundex retrieval*, 2002, Information Technology: Coding and Computing, 2002. Proceedings. International Conference on, pages 22-26.

[10] Bernardo A. Huberman, Daniel M. Romero and Fang Wu *Social networks that matter: Twitter under the microscope*, december 2008, arxiv.org

[11] Akshay Java, Xiaodan Song, Tim Finin, Belle Tseng, *Why we twitter: understanding microblogging usage and community*, 2007, ACM New York, NY, USA 2007, ISBN: 978-1-59593-848-0.

[12] T. Joachims *Text categorization with support vector machines*, 1998, In Proc ECMĹ98, pages 137-142.

[13] RC Russel *The soundex coding system Patent*, 1918, US patents.

[14] Takeshi Sakaki, Makoto Okazaki, Yutaka Matsuo *Eathquake Shakes Twitter Users: Real-time Event Detection by Social Sensors*, April 2010, WWẂ10 Proceedings of the 19th international conference on World wide web, ACM New York, NY, USA 2010, ISBN: 978-1-60558-799-8.

[15] The Twitter Homepage www.twitter.com.

[16] The Twitter API support pages about rate limiting http://support.twitter.com/entries/15364-about-twitter-limits-update-api-dm-and-following and http://dev.twitter.com/pages/every_developer, Mei 2011.

[17] Twitter support page: *What Are Hashtags ("#" Symbols)?*, Gezien: April 2011.

[18] H.M.W. Verbeek, J.C.A.M. Buijs, B.F. van Dongen, W.M.P. van der Aalst, *ProM 6: The Process Mining Toolkit*, 2010, Department of Mathematics and Computer Science, Eindhoven University of Technolog, Netherlands.

[19] Shoko Wakamiya, Ryong Lee, Kazutoshi Sumiya, *Towards better TV viewing rates: exploiting crowd's media life logs over Twitter for TV rating*, 2011, ICUIMC '11 Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication, ACM New York, USA, ISBN: 978-1-4503-0571-6.