Kennisontsluiting ten behoeve van management uit de blogosfeer

Antoon KEUNEN

promotor :
Prof. Jeanne SCHREURS

Eindverhandeling voorgedragen tot het bekomen van de graad
Handelsingenieur in de beleidsinformatica

universiteit hasselt

Inhoudsopgave

Samenvatting

Door de opkomst van breedbandverbindingen en de toegenomen populariteit van onder andere Web 2.0 kan iedereen op het internet tegenwoordig over zowat alles zijn bijdrage en mening kwijt. Men kan dit doen door gebruik te maken van verschillende technologieën waaronder weblogs. Momenteel zijn er al meer dan 70 miljoen weblogs. Samen vormen ze een soort sociaal netwerk: de blogosfeer.

Er zijn al verschillende onderzoeken gebeurd naar de invloed die de blogosfeer heeft op de publieke opinie. De resultaten lopen erg uiteen zodat het op dit moment nog moeilijk is om hier een definitieve uitspraak over te doen. Weblogs hebben echter al voldoende bewezen dat ze een bepalende factor kunnen zijn.

De blogosfeer bevat een enorme hoeveelheid informatie over een erg breed aantal zaken. De blogosfeer onder het licht houden kan voor bedrijven interessante informatie aan het licht brengen.

Bedrijven kunnen door het oprichten van een corporate blog zelf deel gaan uitmaken van de blogosfeer. Slechts weinig bedrijven hebben op dit moment deze stap gezet. Naast heel wat potentiële voordelen stelt het bedrijf zich ook bloot aan mogelijke gevaren die moeilijk zijn in te schatten.

Indien een bedrijf geconfronteerd wordt met een groot aantal berichten, kan het nodig zijn om te kiezen voor een gerichte aanpak. Een eerste mogelijkheid is om relevante invloedrijke weblogs te identificeren en te volgen.

Een tweede optie is om met behulp van tekstanalysesoftware kennis te ontsluiten uit weblogs. Om te verduidelijken hoe men hierbij te werk kan gaan, worden zowel text mining als web mining besproken. Text mining is het automatisch ontsluiten van kennis uit tekstuele data. Web mining focust zich op het ontsluiten van kennis uit data op het internet.

Voor het praktijkgedeelte werd samengewerkt met i.Know[1]. Als case werd gekozen voor de heisa die afgelopen zomer ontstond toen enkele Dell laptops in brand schoten. De blogosfeer speelde een bepalende rol om dit probleem in de aandacht te krijgen. Dell kondigde hierdoor een van de grootste terugroepingen uit de recente geschiedenis aan.

Met behulp van PHP-code werd een lijst bekomen met 6.997 URL's van blogposts die tussen 22 juni en 1 oktober deze zaak aanhaalden. Bij 3.490 hiervan kon de inhoud van de blogpost worden weggeschreven naar een txt-bestand.

Zonder de teksten zelf te analyseren kon reeds een en ander geconcludeerd worden uit onder andere het aantal posts per dag en de pagina's waarnaar werd gelinkt.

Een sentiment classification werd uitgevoerd door gebruik te maken van de 'General Inquirer'[2] aan de hand van de categorieën 'positiv' en 'negativ' in de 'Harvard-IV-4 dictionary'. Deze methode heeft een aantal zwakke punten. Eén daarvan is de behandeling van woordgroepen.

Aan de hand van de methode die i.Know hanteert om zinnen in te delen in clusters werd gekeken in hoever het mogelijk is dit te verbeteren. i.Know identificeerde een groot aantal clusters in de blogposts die waarschijnlijk een positieve of negatieve connotatie hebben. Een deel hiervan werd handmatig ingedeeld als 'positief', 'neutraal' of 'negatief'. Aan de hand hiervan werd een sentiment classification uitgevoerd. De behaalde resultaten werden vervolgens vergeleken met de resultaten behaald door de 'General Inquirer'.

[1] www.iknow.be

[2] http://www.wjh.harvard.edu/~inquirer

Woord vooraf

Deze eindverhandeling werd geschreven met het oog op het behalen van het diploma Master Handelsingenieur in de Beleidsinformatica. Bij de realisatie van deze eindverhandeling kon ik rekenen op de hulp en steun van verschillende personen. Hen zou ik via deze weg willen bedanken.

In de eerste plaats wil ik mijn promotor, Prof. Jeanne Schreurs, bedanken voor haar deskundige hulp en advies.

Verder gaat mijn dank uit naar de mensen van i.Know die wensten mee te werken aan het tot stand komen van deze eindverhandeling. In het bijzonder dank ik Saskia Debergh voor haar begeleiding en advies.

Tenslotte wil ik mijn ouders bedanken voor hun steun tijdens mijn studies.

Hoofdstuk 1: Inleiding

1.1 Probleemsituering

Het Amerikaanse tijdschrift Time riep 'You' uit tot Persoon van het Jaar 2006, oftewel 'u' als internetgebruikende 'Citizen of the New Digital Democracy'. De motivatie hierachter was de sterk toegenomen rol die user-generated content (door de gebruiker gegenereerde inhoud) begon te spelen.

Door de opkomst van breedbandverbindingen en de toegenomen populariteit van onder andere Web 2.0 kan de internetgebruiker zijn bijdrage en mening over zowat alles op het internet plaatsen. Hij kan dit doen door gebruik te maken van verschillende technologieën zoals blogs, podcasts, webvideo's en wiki's.

De invloed van deze user-generated content heeft zich de laatste jaren al vaak laten gelden. Verschillende nieuwsfeiten werden door de 'nieuwe media' op de kaart gezet en werden pas later opgepikt door de reguliere media.

Weblogs maken deel uit van deze mix. Momenteel zijn er al meer dan 70 miljoen. Samen vormen ze een soort sociaal netwerk: de blogosfeer. De blogosfeer onder het licht houden kan voor bedrijven interessante informatie aan het licht brengen. De blogosfeer bevat een enorme hoeveelheid informatie over een erg breed aantal zaken.

Bedrijven kunnen de blogosfeer gebruiken om een beeld te krijgen van wat er online over hen gezegd wordt. Aan de hand van de blogosfeer kan men bijvoorbeeld ook ontdekken welke nieuwe trends er opkomen.

1.2 Probleemstelling

Het opzet van deze eindverhandeling is na te gaan op welke manier bedrijven gebruik kunnen maken van de informatie die vervat zit in weblogs.

Het omzetten van deze informatie in kennis is echter niet steeds even eenvoudig. Een bedrijf kan uiteraard beslissen om louter een bepaalde tijd te besteden aan het lezen van blogposts. Een meer gerichte aanpak zoals het lezen van potentieel invloedrijke berichten vraagt enig inzicht in de blogosfeer. Men kan nog een stap verder gaan door grote hoeveelheden blogposts te analyseren.

Een andere manier om als bedrijf meer inzicht te krijgen in de blogosfeer is door er zelf deel van uit te maken. Dit kan door het oprichten van een corporate blog. Een corporate blog biedt opportuniteiten op verschillende terreinen maar houdt ook enkele gevaren in.

Het praktijkgedeelte van deze thesis zal een specifiek onderdeel van het gekozen onderwerp nader bekijken. Aan de hand van een gevalsstudie zal een deel van de theorie aan de praktijk worden getoetst om meer inzicht te verwerven in zowel de opportuniteiten als de moeilijkheden van een sentiment classification.

Er werd besloten om een sentiment classification uit te voeren op blogposts door gebruik te maken van de 'General Inquirer'[3]. Omdat de 'General Inquirer' werkt aan de hand van unigrammen en hier nadelen aan verbonden zijn, zal verkennend gekeken worden of deze methode mogelijk verbeterd kan worden door gebruik te maken van clusters in plaats van unigrammen.

[3] http://www.wjh.harvard.edu/~inquirer/

Deze eindverhandeling zal proberen een overzicht te geven van de huidige mogelijkheden die bedrijven hebben om in te spelen op de informatie binnen de blogosfeer. De centrale onderzoeksvraag luidt als volgt:

"Op welke manieren kunnen bedrijven kennis vervat in de blogosfeer ontsluiten?"

Omdat het niet gemakkelijk is om op deze vraag dadelijk een antwoord te geven, wordt er gewerkt met een aantal subvragen:

- Wat is een weblog?
- Welke opportuniteiten en gevaren zijn er verbonden met het opstarten van een corporate blog?
- Op welke manier kan informatie in teksten ontsloten worden?
- Hoe kan informatie uit data op het internet gehaald worden?
- Op welke wijze kan dit toegepast worden op de blogosfeer?
- Welke mogelijke motieven hebben bedrijven om dit te doen?
- Wat zijn de sterke en minder sterke punten van een sentiment classification uitgevoerd door de 'General Inquirer'?
- In welke mate kan een sentiment classification op basis van clusters tot betere resultaten leiden dan de resultaten behaald door de 'General Inquirer'?

1.3 Methode van werken

Deze eindverhandeling bestaat uit twee grote delen. In het eerste deel wordt er aan de hand van een literatuurstudie gekeken naar enkele gebieden die verband houden met het ontsluiten van kennis uit de blogosfeer. In het tweede deel wordt er een gevalsstudie uitgewerkt.

<u>Literatuurstudie</u>

In de universiteitsbibliotheken van Diepenbeek (UHasselt) en Leuven (Faculteit ETEW) werd gezocht naar relevante publicaties. Vermits het een recent onderwerp betreft, zijn er weinig of geen specifieke boeken over te vinden. Vandaar dat er vooral een beroep werd gedaan op artikels uit recente tijdschriften. Er werd ook op

het internet gezocht naar relevante artikels over het onderwerp. Hier kon vrij veel relevante informatie gevonden worden. Dit is logisch. Het onderwerp heeft immers rechtstreeks te maken met het internet en de blogosfeer is een nieuw medium dat erg populair is. Omdat het vaak white papers betrof, was het aangewezen er met een gezonde dosis scepticisme mee om te gaan.

Gevalsstudie

Voor het praktijkgedeelte werd samengewerkt met i.Know. Er werd gekozen om de berichten in de blogosfeer in verband met de ophef rond enkele in brand geschoten Dell laptops nader te bekijken.

Allereerst diende een corpus met de inhoud van de berichten samengesteld te worden. Vervolgens werd gekeken welke informatie men kan halen uit het aantal berichten per dag. Ook de links werden geanalyseerd. Vervolgens werd een sentiment classification uitgevoerd. Dit gebeurde met behulp van de 'General Inquirer'. Dit is een systeem voor automatische inhoudsanalyse van tekstuele data. De blogposts uit het corpus werden ingedeeld in drie categorieën: 'positief', 'neutraal' en 'negatief'. Deze methode heeft enkele tekortkomingen, wat in bepaalde gevallen tot verkeerde resultaten kan leiden. Met behulp van een methode op basis van clusters werd verkennend gekeken of verbeteringen mogelijk zijn. De resultaten van beide methodes werden vervolgens vergeleken.

1.4 Overzicht van de thesis

De eigenlijke literatuurstudie begint in hoofdstuk twee. In dit hoofdstuk wordt nader ingegaan op enkele manieren om kennis te ontsluiten uit ongestructureerde en semi-gestructureerde data. Zowel text mining als web mining worden besproken. Beide zijn relevant. Omdat de meeste webposts uit tekst bestaan, is het interessant uit te zoeken hoe kennis automatisch uit tekstuele data ontsloten kan worden. Web mining focust op het ontsluiten van kennis uit het internet.

In hoofdstuk drie wordt besproken hoe bedrijven gebruik kunnen maken van de informatie die vervat zit in de blogosfeer. Allereerst wordt er gekeken wat de blogosfeer nu juist inhoudt. De voor- en nadelen die verbonden zijn aan het opstarten van een corporate weblog worden tegen elkaar afgewogen. Bedrijven met een corporate weblog kunnen gemakkelijker inzicht verwerven in wat er leeft in de blogosfeer. Ook kunnen ze een en ander opsteken door het gedrag van de bezoekers te analyseren. Verder wordt gekeken naar de mogelijke motieven die bedrijven kunnen hebben om de blogosfeer in het oog te houden. Als laatste wordt ingegaan op de verschillende methoden die men kan gebruiken om dit te realiseren.

In hoofdstuk vier wordt een gevalsstudie uitgewerkt. Allereerst wordt beschreven hoe het corpus werd samengesteld. Vervolgens worden, aan de hand van dit corpus, enkele zaken, zoals het aantal posts per dag en de links, besproken. De belangrijkste analyse die in hoofdstuk vier wordt aangehaald is een sentiment classification door de 'General Inquirer'. De gehanteerde werkwijze, de resultaten en enkele mogelijke verbeteringen worden besproken.

Hoofdstuk vijf diept één mogelijke verbetering uit. De werkwijze om te werken op basis van clusters wordt besproken. Ook hier worden de gehanteerde werkwijze, de resultaten en mogelijke verbeteringen besproken. De resultaten van de twee verschillende methodes worden tevens vergeleken.

In hoofdstuk zes tenslotte wordt er een overzicht gegeven van de verschillende conclusies die uit dit alles getrokken kunnen worden.

Hoofdstuk 2: Ontsluiten van kennis uit semi-gestructureerde en ongestructureerde bronnen

2.1 Text Mining

Text mining is het automatisch ontsluiten van kennis uit tekstuele data. Het heeft vooral als doel om nieuwe kennis op te doen. Aangezien geschat wordt dat ongeveer 85% van alle data een ongestructureerde vorm heeft[4], zijn de mogelijkheden die text mining biedt erg uitgebreid.

Tabel 1 situeert text mining. Het vereist doorgaans kennis van verschillende vakgebieden zoals natural language processing, machine learning, information retrieval, data mining, ...

Tabel 1: Situering text mining (Text mining: tools, techniques, and applications, AvaQuest)[5]

	Search (goal oriented)	Discover (opportunistic)
Structured data	Data retrieval	Data mining
Unstructured data	Information retrieval	Text mining

Computers zijn niet in staat om natuurlijke talen te begrijpen zoals mensen dat kunnen. Daarentegen kunnen ze wel grote hoeveelheden tekst op een relatief korte tijd verwerken. Bij text mining probeert men de menselijke linguïstische capaciteiten zo goed als mogelijk over te dragen op computers.

[4] Merrill Lynch (Bron: http://www.dmreview.com/article_sub.cfm?articleId=6287)

[5] http://www.knowledgetechnologies.net/proceedings/presentations/treloar/nathantreloar.ppt

Bondig zal ik hieronder enkele vaak voorkomende toepassingen binnen text mining behandelen.[6]

Information extraction

Het doel van information extraction is op een automatische manier bepaalde informatie binnen ongestructureerde data om te zetten naar gestructureerde data. Information extraction wordt vaak ook door andere text mining toepassingen gebruikt.

Topic tracking

Een topic tracking systeem heeft als doel een gebruiker automatisch te voorzien van informatie over een aantal onderwerpen. Eventueel kan de selectie zelfs automatisch worden samengesteld aan de hand van het gedrag van de gebruiker. De meer geavanceerde systemen kunnen de selectie bijschaven aan de hand van wat de gebruiker doet met de door het systeem aangedragen items. Eventueel kan zijn surfgedrag meegenomen worden om te bepalen welke items hij waarschijnlijk wenst te lezen.

Summarization

Text summarization is het automatisch samenvatten van teksten. De lezer kan hierdoor een overzicht krijgen wat er zich in een tekst bevindt. De gebruiker kan meestal aangeven hoe uitgebreid hij de samenvatting wil hebben. Het is ook mogelijk om text summarization toe te passen op meerdere teksten.

Categorization

Bij categorization worden documenten ingedeeld in een bepaalde categorie. Categorization kan bijvoorbeeld gebruikt worden om aan topic tracking te doen.

Clustering

Bij clustering worden gelijkaardige documenten gegroepeerd. Het verschil met categorization is dat bij clustering de indeling niet op voorhand bepaald is.

---

[6] Tapping the power of text mining, Communications of the ACM, Sept. 2006

Concept linkage

Concept linkage linkt documenten met gemeenschappelijke begrippen. Het wordt vooral gebruikt als onderzoekstool. Zo is het bijvoorbeeld handig in de medische wereld waar zoveel onderzoek gedaan wordt dat het voor iedereen moeilijk wordt om de verbanden te zien. Aan de hand van concept linkage kunnen nieuwe onderzoekshypothesen gevonden worden door te zoeken naar links die nog niet onderzocht zijn.

Information visualization

Information visualization of visual text mining geeft als resultaat een visuele weergave van grote hoeveelheden tekst. De output kan bijvoorbeeld in de vorm van een mindmap zijn.

Question answering

Een andere toepassing van text mining is question answering. Aan de hand van natural language processing zullen question answering systemen een antwoord zoeken aan de hand van een gegeven vraag van de gebruiker.

2.2 Web Mining

2.2.1 Inleiding

Het internet bevat een enorme hoeveelheid data. Door web mining toe te passen probeert men een gericht deel van deze data te ontsluiten en aan de hand van analyse om te zetten naar bruikbare kennis. Dit gebeurt met behulp van data mining technieken. Net als text mining vereist web mining doorgaans kennis van verschillende vakgebieden zoals natural language processing, machine learning, information retrieval, data mining, ...

De informatie op het internet is echter even overvloedig als heterogeen. Door het gebrek aan een eenduidige structuur in het internet is web mining geen gemakkelijke opgave. Het internet is dynamisch, de inhoud en structuur verandert

voortdurend. Verder bevat het internet een enorme hoeveelheid redundante of niet bruikbare informatie.

Ondanks deze niet makkelijk te overkomen moeilijkheden zijn er een aantal zaken die web mining aantrekkelijk maken. Een groot deel van de informatie op het internet is gemakkelijk toegankelijk. Het aanbod van informatie is erg ruim. Men kan over praktisch alles wel informatie vinden. Het grootste deel van de data op het web is semi-gestructureerd. De meeste pagina's bevatten links die naar andere pagina's binnen of buiten de site verwijzen. Deze links in kaart brengen kan bruikbare informatie opleveren.

Web mining wordt doorgaans in drie subtopics onderverdeeld:

- Web Usage Mining
- Web Structure Mining
- Web Content Mining

2.2.2 Web usage mining

Web servers slaan vaak data over surfgedrag van bezoekers op in web logs. Deze web logs bevatten data zoals de manier waarop de bezoeker op de website terechtkwam, welke pagina's hij bezocht, in welke volgorde hij de verschillende pagina's bezocht, ... Op basis hiervan tracht men uit te zoeken hoe men de website best kan aanpassen om beter aan de wensen van de bezoeker te voldoen.

2.2.3 Web structure mining

De structuur van webpagina's kan bruikbare informatie opleveren. Web structure mining focust vooral op het gebruik van hyperlinks in webpagina's. Zo geeft het aantal links naar een webpagina een indicatie van de populariteit van die webpagina.

2.2.4 Web content mining

Web content mining is de verzamelnaam voor de verschillende methodes die uit de inhoud van webpagina's informatie ontsluiten. Het kan gaan over het halen van

informatie uit tekst, audio of video. Het ontsluiten van tekstuele data krijgt momenteel het meeste aandacht waardoor web content mining ook wel eens web text mining genoemd wordt. Toch verschilt web content mining van text mining omdat data op het web veelal semi-gestructureerd is, terwijl text mining steeds wordt toegepast op ongestructureerde data.

De text mining toepassingen zoals bondig besproken onder hoofdstuk 2.1, kunnen ook toegepast worden op tekstuele data op het internet. Zo kan topic tracking toegepast worden op online nieuwsberichten die met behulp van categorization worden geselecteerd. Andere specifieke toepassingen van web content mining zijn onder andere structured data extraction, information integration en schema matching, sentiment classification en opinion analysis.

Structured data extraction

Structured data extraction focust op het uitselecteren van gestructureerde data vervat in websites. Het gaat over data die uit databases gehaald worden, zoals bijvoorbeeld data over producten in een online winkel. Vermits het internet over het algemeen semi-gestructureerd is, is het mogelijk om al dan niet manueel aan te geven welke data er uitgehaald dienen te worden.

Bedrijven zouden bijvoorbeeld automatisch kunnen bijhouden hoe hun producten het doen op veilingsites zoals eBay. Dit kan vergeleken worden met de producten van de concurrentie.

Information integration en schema matching

Het internet bevat veel websites die gelijkaardige data bevatten. Deze informatie samenbrengen kan grote voordelen met zich meebrengen. Het is echter niet gemakkelijk om semantisch gelijkaardige data automatisch te identificeren en te matchen.

Er zijn verschillende mogelijke toepassingen. Zo kan men bijvoorbeeld data op verschillende veilingsites ontsluiten en deze integreren. Een andere mogelijke toepassing is het integreren van web query interfaces. Met kan op deze manier

bijvoorbeeld een website creëren die de zoekmachines van de verschillende luchtvaartmaatschappijen combineert.

Sentiment classification

Sentiment classification of sentiment analysis is een text mining toepassing. Documenten worden geclassificeerd naargelang de stemming die ervan uitgaat.

Het internet is een interessant domein om sentiment classification op toe te passen. De internetgebruiker kan tegenwoordig op het internet over zowat alles zijn bijdrage en mening kwijt. Men kan berichten plaatsen op internetforums, weblogs en allerhande sites. Met behulp van sentiment classification kan men een beeld krijgen van de toon in deze berichten.

Onderzoek is gedaan naar het classificeren van reviews op websites zoals epionions.com, imdb.com, amazon.com, ... De semantische oriëntatie van de gehele review wordt berekend. De reviews worden meestal ingedeeld in twee categorieën: positief en negatief.

De resultaten die hierbij gehaald worden zijn reeds redelijk accuraat. Onderzoek van Turney uit 2002 behaalde een gemiddelde nauwkeurigheid van 74%. De juistheid van de indeling liep uiteen per categorie van 65,83% voor filmrecensies tot 84% voor reviews over auto's. Ander onderzoek van Pang, Lee en Vaithyanathan uit 2002 testte verschillende sentiment classification technieken uit. Het beste resultaat werd behaald aan de hand van de support vector machine methode op basis van unigrammen. Men bereikte hiermee een nauwkeurigheid van 82,9%. In later onderzoek (Pang en Lee, 2004) werd door het wegfilteren van objectieve zinnen de nauwkeurigheid verhoogd tot 86,4%. Zo heeft de zin 'I love this movie' een duidelijke positieve connotatie, in tegenstelling tot de zin 'This is a love movie'. Ander onderzoek van Whitelaw, Grag en Argamon uit 2005 behaalde een nauwkeurigheid van 90,2%.

Sentiment classification kan ook toegepast worden op individuele zinnen. De nauwkeurigheid die hier gehaald wordt is momenteel nog een stuk lager.

Het toepassingsgebied van sentiment classification is erg ruim. Zo kunnen bedrijven het toepassen om een beeld te krijgen over de houding van de internetgebruikers ten opzichte van hun bedrijf. Men kan dit uitzetten over de tijd heen en vergelijken met de score die behaald wordt door de rechtstreekse concurrenten. Het kan ook gebruikt worden om erg ontevreden klanten op te sporen.

De Verenigde Staten hebben bijvoorbeeld een project opgestart om met behulp van sentiment classification een beeld te krijgen over wat er over hen gezegd wordt in de media van andere landen.[7] De bedoeling is vooral om erg negatieve stukken in kranten en tijdschriften te identificeren.

Sentiment classification kan ook gebruikt worden om de reacties van beleggers op een bepaalde gebeurtenis te onderzoeken. Zo bleek uit onderzoek van Tetlock uit 2006 dat de aandelenkoersen correleerden met het resultaat van een sentiment classification van het dagelijkse commentaarstuk 'Abreast of the Market' in de Wall Street Journal dat de dag daarvoor verscheen.

Opinion analysis

Opinion analysis, ook wel opinion mining genoemd, gaat verder dan sentiment classification. Bij opinion analysis worden eerst de zinnen geïdentificeerd die een bepaalde mening uitdragen. Terwijl sentiment classification enkel kijkt naar de semantische oriëntatie wordt bij opinion analysis ook gekeken waar de mening over gaat. Eventueel wordt ook gekeken wie de mening uit. Dit zal in de meeste gevallen de schrijver zelf zijn, maar het kan ook dat hij de mening van een andere persoon vermeldt.

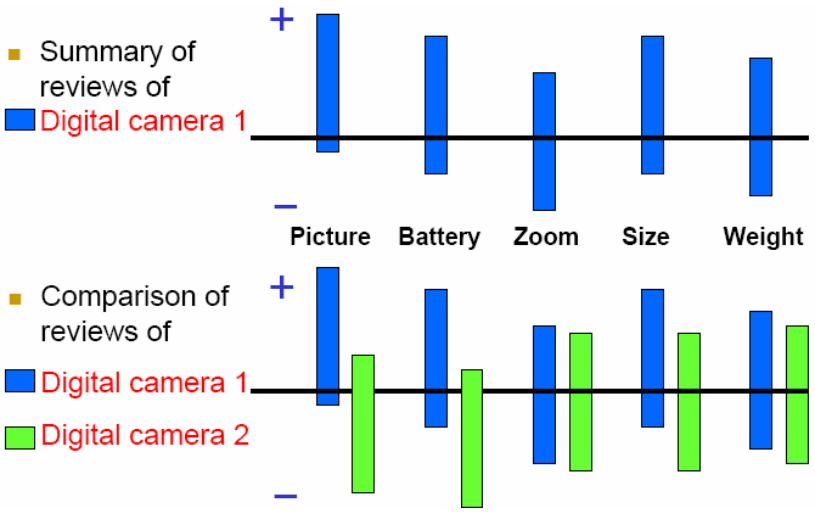
Opgemerkt dient te worden dat sommigen sentiment classification als een relatief eenvoudige vorm van opinion analysis zien. Anderen maken het onderscheid zoals hier gemaakt wordt.

---

[7] http://www.nytimes.com/2006/10/04/us/04monitor.html?ex=1180584000&en=f0dfbf939224120f&ei=5070

Het meeste onderzoek naar opinion analysis is net als bij sentiment classification gedaan op reviews. Doorgaans worden er een aantal features ingegeven waarover mogelijk in de review een mening over gegeven zal worden. De nauwkeurigheid van opinion mining is lager dan die van sentiment classification. Verschillende methodes zijn met wisselend succes getest. Bij reviews waar men de voordelen en nadelen apart bespreekt wordt logischerwijze een hogere nauwkeurigheid behaald.

De resultaten kunnen grafisch weergegeven worden. Het is interessant om deze resultaten te vergelijken met bijvoorbeeld de resultaten voor een gelijkaardig product van een naaste concurrent. Een voorbeeld daarvan wordt gegeven in Figuur *1*.



Figuur 1: Visuele samenvatting en vergelijking (Opinion Mining, Bing Liu, 2006)[8]

Een mogelijke aanpak is om specifiek te zoeken naar zinnen waar een waardeoordeel wordt uitgesproken aan de hand van een vergelijking. Men kan zoeken naar zinnen waar iets beter wordt ingeschat dan iets anders aan de hand van woorden zoals 'better', 'ahead', 'beats', ... Tevens kan men zinnen identificeren waarvan een gelijkheid uitgaat op basis van woorden zoals 'equal to', 'same as', ... Tenslotte kan men zinnen met een superlatief selecteren die woorden als 'best' en 'better than all' bevatten.

[8] http://www.cs.uic.edu/~liub/teach/cs583-spring-07/opinion-mining.pdf

Hoofdstuk 3: De blogosfeer als kennisbron

Dit hoofdstuk zal focussen op de verschillende manieren waarop bedrijven op de blogosfeer kunnen inspelen. Eerst zal ingegaan worden op de vraag wat de blogosfeer nu juist is. Vervolgens zullen de voor- en nadelen van het opstarten van een corporate blog besproken worden. Door zelf deel uit te maken kunnen bedrijven de blogosfeer beter aanvoelen. Het opstarten van een corporate weblog omvat echter veel meer aspecten. Deze zullen bondig aangehaald worden. Vervolgens worden enkele mogelijke motieven besproken waarom bedrijven zouden kunnen beslissen om rekening te houden met de blogosfeer. Tenslotte wordt ingegaan op enkele methoden die gebruikt kunnen worden om deze motieven te verwezenlijken.

3.1 Weblogs

3.1.1 Definitie

Een weblog, of kortweg blog, is een website waarop één of meer mensen op regelmatige basis nieuwe berichten plaatsen. Op een weblog verschijnen de berichten in antichronologische volgorde, de nieuwste berichten staan dus bovenaan. Vaak wordt er aan de bezoeker de mogelijkheid gegeven om te reageren. De inhoud kan tekst, foto's, tekeningen, audio en video's omvatten.

Weblogs maken deel uit van de zogenaamde sociale media. Hiermee wordt een verzameling online communicatievormen en -technieken bedoeld waarmee user-generated content verspreid kan worden. Hieronder vallen bijvoorbeeld Wikipedia, YouTube, MySpace, Second Life en Gather.com.

Interactie met anderen is voor de meeste bloggers erg belangrijk. Ze lezen andere blogs, plaatsen commentaar, refereren op hun eigen blog naar andere blogposts en citeren uit posts. Men kan alle blogs samen dus als een soort sociaal netwerk zien. Dit wordt de blogosfeer genoemd.

Verschillende webloggers maken gebruik van Trackback. Dit is een systeem om de interactie tussen webloggers te verduidelijken. Indien blogger A een bericht schrijft over een bericht van blogger B, zal, indien beide weblogs Trackback ondersteunen, er bij het bericht van blogger A een link verschijnen naar dit bericht van blogger B.

3.1.2 Enkele cijfers

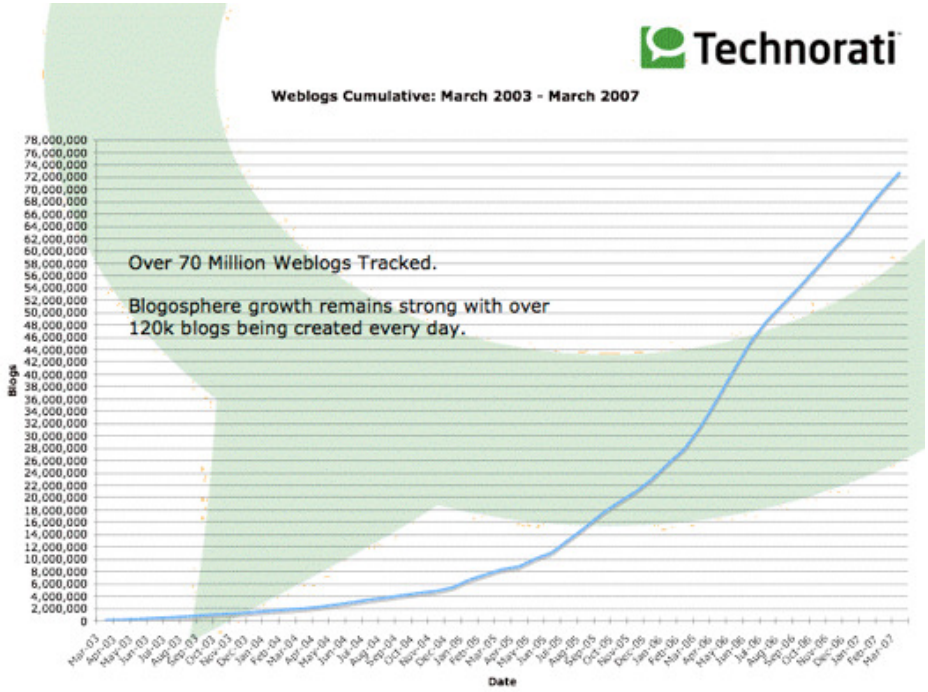
Technorati[9], een van de populairste zoekmachines voor weblogs, brengt elk kwartaal een overzicht uit van de huidige stand van zaken in de blogosfeer. In 'The State of the Live Web, April 2007'[10] staan volgende cijfers. Momenteel volgt Technorati meer dan 70 miljoen blogs. Elke dag worden wereldwijd zo'n 120.000 nieuwe weblogs aangemaakt. Dit is gelijk aan 1,4 nieuwe weblogs per seconde. Deze cijfers dienen echter enigszins gerelativeerd worden. Een groot aantal weblogs is niet meer actief. Een groot probleem binnen de blogosfeer zijn splogs. Dit zijn spam blogs zonder betekenisvolle inhoud die automatisch worden aangemaakt. Door het plaatsen van links naar websites, trachten ze deze websites bij zoekmachines hoger in de resultaten te laten komen.

Er worden dagelijks zo'n 1,4 miljoen berichten op weblogs geplaatst. Dit komt overeen met 16 posts per seconde. Uit Figuur 3 valt af te leiden dat het aantal berichten per dag het laatste jaar niet meer lijkt toe te nemen. Een mogelijke verklaring hiervoor is de toenemende populariteit van andere sociale media zoals bijvoorbeeld YouTube. Onderzoeksbureau Gartner voorspelde eind 2006 dat bloggen in 2007 zou pieken en daarna aan een terugval zou beginnen.

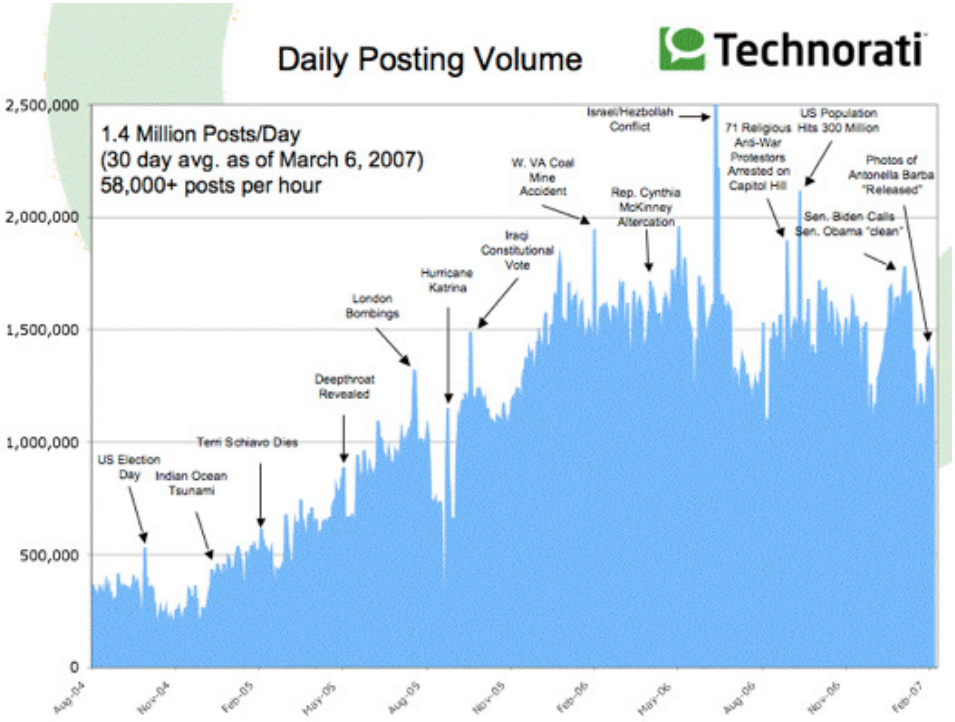
---

[9] http://www.technorati.com/

[10] http://www.sifry.com/alerts/archives/000493.html



Figuur 2: aantal weblogs (Technorati, The State of the Live Web, April 2007)



Figuur 3: posts per dag (Technorati, The State of the Live Web, April 2007)

Japans is de meest voorkomende taal in de blogosfeer. Zo'n 37% van de berichten wordt in het Japans geschreven. Ongeveer evenveel posts (36%) worden in het Engels gemaakt. Andere vaak voorkomende talen zijn: Chinees (8%), Italiaans (3%), Russisch (2%), Frans (2%), Portugees (2%), Duits (1%) en Perzisch (1%). Nederlands valt met ongeveer 1% net buiten de top 10.

Dat Japans meer voorkomt dan Engels is te verklaren door de grote regionale verschillen in populariteit. In Figuur 4 zijn de resultaten terug te vinden van een bevraging door onderzoeksbureau Edelman. Zo blijkt 74% van de Japanners wel eens een blog te lezen. In België en Duitsland ligt dit percentage met ongeveer 15% een stuk lager. Ook de frequentie dat er gelezen wordt verschilt sterk. In Japan leest men ongeveer 4,5 dagen per week een blog. In België en Duitsland is dat slechts 0,4 dagen per week.

	READ BLOGS	DO NOT READ BLOGS	% OF "INFLUENCERS"* READING BLOGS
JAPAN	74%	26%	91%
SOUTH KOREA	43%	57%	63%
CHINA	39%	53%	53%
UNITED STATES	27%	69%	34%
UNITED KINGDOM	23%	75%	35%
FRANCE	22%	68%	37%
ITALY	16%	73%	30%
POLAND	16%	85%	45%
GERMANY	15%	84%	27%
BELGIUM	14%	79%	41%

Figuur 4: populariteit blogs (Edelman, A Corporate Guide to the Global Blogosphere)[11]

[11] http://www.edelman.com/image/insights/content/WhitePaper011107sm.pdf

3.1.3 Invloed

Dat de blogosfeer bepaalde zaken op de agenda kan zetten en een invloed kan uitoefenen op de publieke opinie is al langer duidelijk. Zo verscheen op 16 mei 2007 op technologieblog Engadget een bericht waarin stond dat de introductie van de Apple iPhone en het besturingssysteem Leopard was uitgesteld.[12] Alhoewel al kort later zou blijken dat het bericht fout was, daalden de aandelen van Apple binnen de 10 minuten na het verschijnen van het bericht van $107,89 naar $103,42. Dit komt overeen met een verlies aan marktkapitalisatie van 4 miljard dollar.

De meningen zijn sterk verdeeld over de mate waarin de blogosfeer invloed heeft. De resultaten van de onderzoeken hiernaar lopen erg uiteen zodat het op dit moment volgens mij moeilijk is om hierover een definitieve uitspraak over te doen.

Uit onderzoeken van Jupiter Research en Millward Brown blijkt dat men erg weinig vertrouwen heeft in de inhoud van weblogs. Slechts in een klein aantal gevallen gaat men te rade op het internet alvorens een aankoop te doen. Dit is vooral het geval bij onder andere het beslissen van een vakantiebestemming of de aankoop van producten zoals mobiele telefoons, camera's en software.

Andere onderzoeken komen tot heel andere conclusies. Een onderzoek van Ipsos Moris[13] concludeert dat weblogs na kranten als het meest geloofwaardig worden ingeschat. Vooral mensen uit hogere inkomstenklassen zouden veel vertrouwen hebben in weblogs. Meer dan de helft van de ondervraagde Europeanen gaf aan dat ze meer geneigd waren een bepaald product te kopen nadat ze positieve commentaar erover in user-generated content hadden gelezen. Zo'n 34% gaf aan dat ze in het verleden reeds een keer besloten een bepaald product niet aan te kopen omdat ze er in user-generated content negatieve commentaar over hadden gelezen.

---

[12] http://www.engadget.com/2007/05/16/iphone-delayed-until-october-leopard-delayed-again-until-januar

[13] http://www.ipsos-mori.com/polls/2006/blogging.shtml

Dit verschil is volgens mij grotendeels te wijten aan de diversheid van de blogosfeer. Lezers zullen bepaalde weblogs meer vertrouwen dan andere en zullen elk bericht afzonderlijk nog eens beoordelen naar betrouwbaarheid. Een negatieve review over een technologisch product op een van de toonaangevende technologieblogs zal waarschijnlijk als betrouwbaarder worden gepercipieerd dan negatieve commentaar over datzelfde product op een reisweblog.

De meeste weblogs hebben erg weinig lezers. Daartegenover staat een klein aantal weblogs die erg veel lezers hebben. Zo heeft technologieblog Engadget ongeveer 8 miljoen unieke bezoekers per maand. Er zijn steeds meer weblogs met hoge autoriteit. In de reguliere media wordt er tegenwoordig meer dan ooit tevoren verwezen naar weblogs.

In Figuur 4 op pagina 24 is te zien dat volgens de enquête uitgevoerd door Edelman 'influencers' relatief gezien vaker weblogs lezen dan niet-'influencers'. Edelman noemt mensen 'influencers' indien ze relatief veel actie nemen zoals bijvoorbeeld het sturen van een brief naar een krant, het tekenen van een petitie of indien ze lid zijn van een lobbygroep.

3.2 Corporate Blogs

3.2.1 Definitie

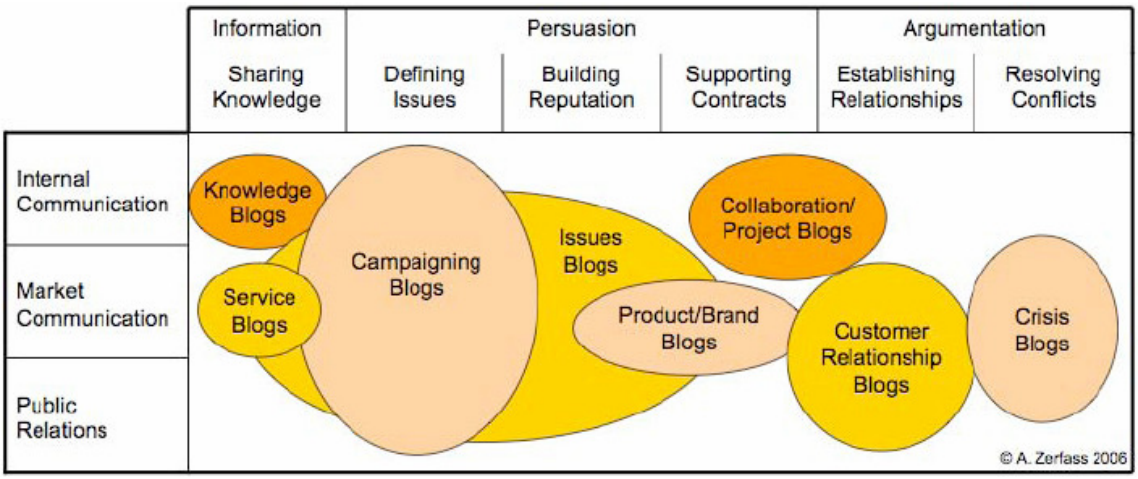
Een corporate weblog is een aan een specifiek bedrijf verbonden weblog. Een dergelijke blog kan zowel intern als extern zijn en dient vooral als communicatie- en marketingkanaal. Een bedrijf kan uiteraard meerdere corporate weblogs hebben.

Vermits deze eindverhandeling gaat over het ontsluiten van kennis uit de blogosfeer, wordt vooral gefocust op externe corporate weblogs die door iedereen te lezen zijn.

3.2.2 Indeling

Er bestaan verschillende soorten corporate blogs. Een veel gebruikte indeling is de classificatie van Ansgar Zerfaß in Figuur 5. Op basis van bedrijfsdoelstelling en

functioneel gebied onderscheidt Zerfaß knowledge blogs, service blogs, campaigning blogs, issues blogs, product/brand blogs, CRM blogs, collaboration/project blogs en crisis blogs.



Figuur 5: Classificatie corporate blogs (Ansgar Zerfaß, EuroBlog 2006)[14]

Een andere indeling (zie Tabel 2) is gebaseerd op basis van auteur en aandachtsgebied. De indeling houdt enkel rekening met permanente weblogs met een externe focus. Men maakt onderscheid tussen een CEO-blog, bedrijfsblog, expert blog en topic blog.

Een CEO-blog wordt geschreven door de algemeen directeur van een bedrijf. De CEO schrijft meestal voornamelijk over zijn visie op het bedrijf. Soms vermeldt hij ook persoonlijke, niet-werk gerelateerde gebeurtenissen in zijn leven.

[14] http://www.euroblog2006.org/symposium/program/assets/EuroBlog2006_Zerfass.pdf

Tabel 2: Types corporate blogs (gebaseerd op: The Business Value of Blogging, Lewis, maart 2007)[15]

Door : / Focus :	Eén persoon	Meerdere personen
Bedrijf	CEO blog	bedrijfsblog
Onderwerp	expert blog	topic blog

Bij een bedrijfsblog worden de bijdragen door verschillende werknemers geplaatst. Doorgaans zijn deze personen elk gespecialiseerd in een bepaald aspect van het bedrijf en plaatsen ze berichten over voor het bedrijf relevante gebeurtenissen op de bedrijfsblog.

Een expert blog wordt onderhouden door één werknemer. Doorgaans is deze persoon erg bedreven binnen een bepaald vakgebied.

Ook bij een topic blog wordt er binnen een bepaald vakgebied gebleven. Het verschil met de expert blog is dat er meerdere personen berichten plaatsen. Een ander verschil is dat er op een topic blog meestal meer berichten verschijnen. Men tracht een volledige bron van informatie te zijn binnen het onderwerp dat de weblog behandelt. Bij een expert blog ligt de nadruk meer op de kwaliteit van de analyse en minder op de kwantiteit.

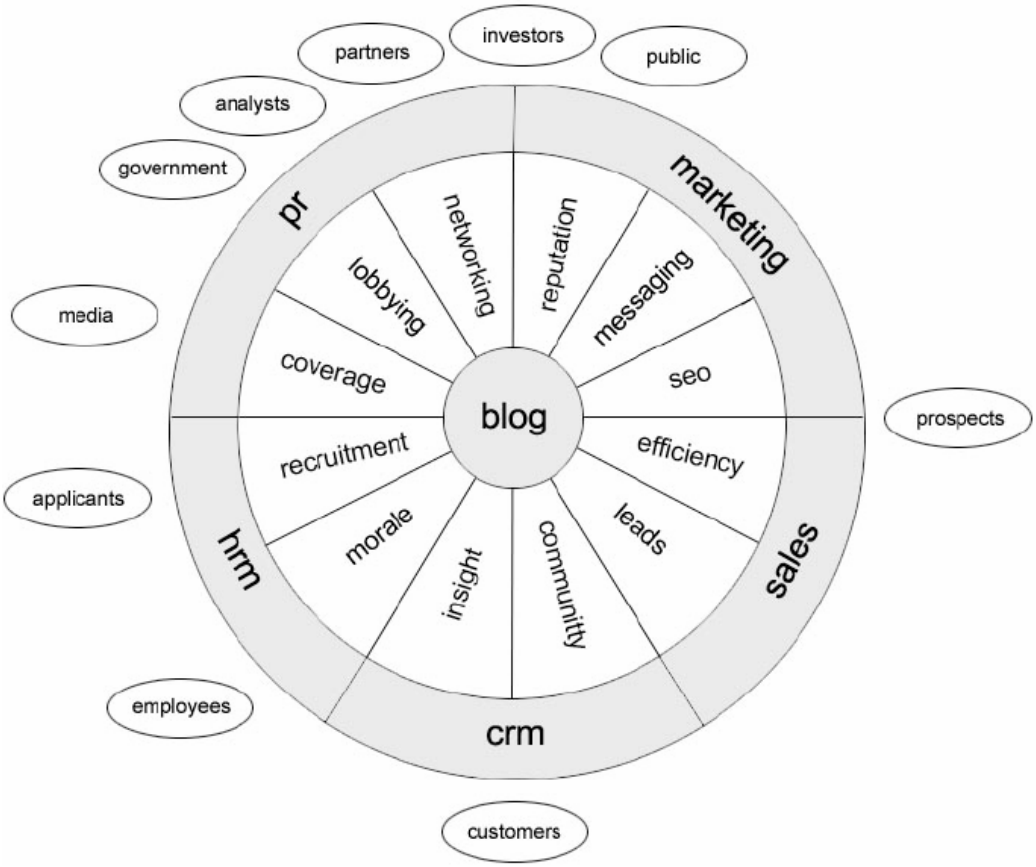
Onderzoeksbureau Lewis, waar deze indeling op gebaseerd is, onderscheidt verder ook nog de guru blog en de industry blog. Aangezien deze types blog minder gelinkt zijn aan een bepaald bedrijf neem ik deze niet mee op.

3.2.3 Voordelen en opportuniteiten

Het opzetten van een blog is erg eenvoudig. Er is bijvoorbeeld open-source software beschikbaar die gemakkelijk aan te passen is.

[15] www.lewis360.com/downloads/Business_value_of_blogging.pdf

De mogelijke voordelen bij het opstarten van een corporate weblog zijn divers. Een vrij compleet overzicht wordt volgens mij gegeven in het 'Blogging value wheel & stakeholders' van onderzoeksbureau Lewis, terug te vinden in Figuur 6. Zoals te zien is, hebben de mogelijke voordelen van een corporate weblog betrekking op verschillende afdelingen binnen het bedrijf.



Figuur 6: 'Blogging value wheel & stakeholders' (The Business Value of Blogging, Lewis, 2007)[16]

Wat betreft personeelsbeleid kan een corporate blog een positieve invloed hebben op de moraal van de werknemers. Via een corporate blog zullen tevens werkzoekenden

[16] http://www.euroblog2006.org/symposium/program/assets/EuroBlog2006_Zerfass.pdf

een beeld kunnen krijgen over de bedrijfscultuur wat mogelijk tot meer sollicitaties leidt.

Door het oprichten van een corporate blog kan men via de interactie met de lezers en andere bloggers inzicht krijgen in de mening van de klant. Het is mogelijk dat er een soort community van vaste lezers ontstaat, wat de band met het bedrijf versterkt.

Een corporate blog kan de reputatie van een bedrijf verbeteren. Een goede corporate blog zal er ook toe leiden dat andere bloggers naar de corporate blog en naar de website van het bedrijf linken. Dit zorgt ervoor dat het bedrijf hoger komt te zijn in de resultaten van internet zoekmachines.

In bepaalde gevallen zal het oprichten van een corporate blog op een directe wijze leiden tot klantenwerving onder lezers van de blog. Verkopers kunnen via de corporate blog op de hoogte blijven van de recente ontwikkelingen en kunnen de blog gebruiken om de visie van het bedrijf uiteen te zetten.

Andere blogs en andere media zullen gemakkelijker linken naar blogposts op de corporate blog wat zal leiden tot meer aandacht. Deze extra aandacht kan aangewend worden om te lobbyen. Vermits de blogosfeer een soort sociaal netwerk is, kunnen via een corporate blog mogelijk nieuwe contacten gelegd worden of kunnen bestaande contacten makkelijker onderhouden worden.

Deze voordelen wegen bij verschillende types blogs verschillend door. De invulling hiervan door Lewis is terug te vinden in Bijlage 1.

3.2.4 Nadelen en gevaren

Het onderhouden van een succesvolle corporate weblog is niet zo eenvoudig. Om voldoende interesse te verkrijgen, dient er regelmatig gepost te worden. Dit vraagt uiteraard tijd. Een corporate blog met weinig berichten zal veeleer een negatieve indruk nalaten. Vooraleer een bedrijf de stap zet naar een corporate blog, dient men volgens mij dan ook te onderzoeken of er voldoende materiaal is om over te

schrijven. Een bedrijf in een innovatieve sector zal bijvoorbeeld veel berichten kunnen plaatsen. De kans op een succesvolle corporate blog is hier dan ook groter.

Omdat een corporate blogger het bedrijf en het onderwerp waarover hij schrijft erg goed moet kennen, wordt een corporate blog over het algemeen onderhouden door ervaren werknemers. Deze werknemers dienen ook goed te kunnen schrijven.

Een bedrijf stelt zich open aan een aantal gevaren. Het begeeft zich op onbekend terrein waar andere regels gelden. Vermits er vrijwel steeds aan de bezoeker de mogelijkheid wordt gegeven om te reageren op een bericht, moet men kunnen omgaan met kritiek. Het censureren van dergelijke kritiek is meestal geen optie omdat dit door de blogosfeer niet gesmaakt wordt.

De juiste toon vinden is ook niet zo gemakkelijk. Het is uiteraard de bedoeling om het bedrijf in een positief daglicht te stellen, maar bloggers zullen weinig geloof hechten aan een corporate blog met enkel uitgesproken positieve berichten.

3.2.5 Enkele bedenkingen

Het gebruik van corporate blogs is op dit ogenblik vrij beperkt. Een onderzoek van Lewis[17] ondervond dat in januari 2007 slechts zo'n 5% van de bedrijven binnen de IT-, media-, telecom- en professionele dienstensector een permanente, naar buiten gerichte corporate blog hadden. Het aantal corporate blogs buiten deze sectoren ligt waarschijnlijk nog lager. Ongeveer 14% van de Amerikaanse bedrijven hebben een weblog. In West-Europa bedraagt dit slechts 2,5%.

Deze cijfers geven aan dat vele bedrijven waarschijnlijk de nadelen en gevaren verbonden met een corporate blog hoger inschatten dan de voordelen en opportuniteiten. Een andere mogelijkheid is dat vele bedrijven een afwachtende houding aannemen en de resultaten van de corporate weblogs van andere bedrijven afwachten.

---

[17] www.lewis360.com/downloads/Business_value_of_blogging.pdf

Een mogelijkheid om vertrouwd te geraken met de blogosfeer is het oprichten van een tijdelijke blog. Dit kan bijvoorbeeld bij het aantreden van een nieuwe CEO, waarbij deze op een blog zijn eerste ervaringen met het bedrijf bespreekt binnen een vooraf aangegeven periode.

Van de bedrijven die een corporate blog hebben opgestart, waren volgens een studie van Porter Novelli[18] meer dan drie kwart van mening dat het opstarten van een corporate blog had geleid tot meer media-aandacht of website verkeer. Hetzelfde aantal vond dat de gewenste resultaten bij het opstarten van de blog bereikt waren. Nochtans waren 70% van de ondervraagde bedrijven ontevreden over de mate van interactie met bezoekers en andere bloggers.

3.3 Motieven

3.3.1 'Word of mouth' inschatten

Een voor de hand liggend motief voor het monitoren van weblogs is om een beeld te krijgen van wat er over het bedrijf gezegd wordt. In vele gevallen is het interessant om dit te vergelijken met hetgeen er over de directe concurrenten gezegd wordt. Men kan meer inzicht verkrijgen in de meningen op de markt waar het bedrijf actief is.

3.3.2 Blogs van werknemers monitoren

Indien een bedrijf zijn personeelsleden toelaat om corporate blogs op te richten kan het nodig zijn deze blogs in het oog te houden. De kans bestaat dat werknemers te veel interne informatie naar buiten brengen of overgaan tot het plaatsen van illegale informatie, zoals bijvoorbeeld berichten die het auteursrecht schenden.

Verschillende webloggers zijn reeds ontslaan, in het blogjargon ook wel 'dooced' genoemd, omdat ze zaken schreven die volgens hun werkgever niet door de beugel konden. Het Amerikaanse leger houdt reeds enige tijd de warblogs van soldaten in oorlogsgebied in het oog.

---

[18] www.porternovelli.com/Corporate_Blog_Learnings.pdf

Een onderzoek van het Britse human-resourcebedrijf Croner concludeerde dat 39% van de bloggers toegaf wel eens informatie op hun blog te plaatsen die tot hun ontslag zou kunnen leiden.[19]

3.3.3 Trends detecteren

Weblogs bieden een mogelijkheid om trends te detecteren en hier tijdig op te reageren. Bedrijven kunnen mede aan de hand hiervan beter voorspellen welke producten het goed gaan doen. Bedrijven kunnen door weblogs in het oog te houden waarschijnlijk beter inschatten welke dingen er onder de bevolking leven en ze kunnen hierop inspelen.

3.3.4 Actie ondernemen

Aangezien de blogosfeer een soort sociaal netwerk is, bestaat de mogelijkheid om hier als bedrijf aan deel te nemen. Een corporate blog is hier een voorbeeld van.

Een andere mogelijkheid is rechtstreeks contact op te nemen met bepaalde invloedrijke bloggers. Dit kan ertoe leiden dat ze positiever over het bedrijf gaan schrijven of het bedrijf op hun weblog meer aandacht zullen geven.

Een derde mogelijkheid is om klanten met klachten te vinden en hen te helpen. Men kan hen een mail sturen met de gepaste procedure om iets aan het probleem te doen. UPC, een Nederlands telecommunicatiebedrijf, startte in juli 2006 een webcare-programma. Medewerkers screenden onder andere weblogs en internetforums op discussies over UPC. Soms mengden ze zich in de discussie. In januari 2006 werkten er 3 mensen die er samen ongeveer één fulltimebaan aan besteedden.[20] Soms reageerden ze ook op mensen met specifieke klachten. Zo stuurden ze mails met verontschuldigingen, werd er raad gegeven of werd er om

---

[19] http://www.croner.co.uk/croner/jsp/Editorial.do?contentId=714662

[20] http://www.nrc.nl/media/article601866.ece/Eerste_hulp_bij_webdiscussie

contactgegevens verzocht om de zaak op te lossen. Indien er bij de weblog geen mailadres werd gegeven, werd er mogelijk op de weblog zelf gereageerd.

3.3.5 Snel reageren

De blogosfeer kan gebruikt worden om relevante gebeurtenissen snel op te sporen en in te schatten. Indien er buitensporig veel geschreven wordt over een bedrijf, kan dit duiden op een mogelijke crisissituatie. Door de blogosfeer in het oog te houden kan men ook snel reageren op foutieve informatie.

Een voorbeeld waar een bedrijf door snel te reageren waarschijnlijk heel wat negatieve publiciteit vermeed is UPC. Mogelijk mede dankzij hun webcare-programma kwamen zij er al snel achter dat in augustus 2006 GeenStijl, een van de populairste Nederlandse weblogs, volgend fragment op hun weblog plaatsten naar aanleiding van een medewerker die al anderhalve maand wachtte op een UPC-aansluiting:

"*We gaan een maand lang UPC kapot maken. Joris met de digicam naar kantoor, speciaal UPC-forum. 1,6 miljoen unieke bezoeker per maand weet je! Wij hebben meer lezers dan UPC klanten heeft. Registreren we www.overstappenisweleng.nl en vragen we sales of ze er een KPN, Xs4all, Zonnet of Sonera bannert bij knallen, is iedereen blij. Vervolgens geven we onze bezoekers een maand lang de tijd om over te stappen, daarna bannen we alle UPC-ip's. Ik heb het nu nog maar over de eerste 24 uur van ons sloopwerk...*"[21]

Hoewel het niet duidelijk was of GeenStijl de dreiging werkelijk meende, bracht UPC de eerstvolgende werkdag om 8 uur 's morgens de kabelaansluiting in orde.

Vermits berichten van bloggers onmiddellijk door iedereen te lezen zijn, kunnen ze ook gebruikt worden om snel te peilen naar de meningen over een recente gebeurtenis. Dit kan bijvoorbeeld een aankondiging van een bedrijf zijn, een uitlating van een politicus of een gebeurtenis die mogelijk effect heeft op de aandelenmarkt.

---

[21] http://www.geenstijl.nl/mt/archieven/015603.html

3.3.6 Onderzoek

De enorme hoeveelheid informatie over een groot aantal zaken maakt de blogosfeer een interessant onderzoeksdomein. Men kan bijvoorbeeld zoeken naar nieuwe onderzoekshypothesen met behulp van concept linkage.

3.4 Methoden

3.4.1 Relevante posts lezen

Voor bepaalde bedrijven zal het voldoende zijn om regelmatig gericht weblogs te lezen om een idee te krijgen wat er over hen in de blogosfeer gezegd wordt. Indien er weinig over het bedrijf, de concurrentie en de markt waarop ze actief zijn wordt geschreven, heeft het weinig nut om geavanceerde software in te zetten.

Een bijkomend voordeel van het monitoren van weblogs is dat de reacties van bloggers op een bepaalde gebeurtenis onmiddellijk opgevraagd kunnen worden. In tegenstelling tot een opiniepeiling krijgt men dus al snel een beeld van hoe bloggers reageren.

Monitoren kan eenvoudigweg door in een zoekmachine voor blogs een zoekterm in te geven. Een betere aanpak zou zijn om zich met behulp van RSS-feeds automatisch op de hoogte te houden. Er kan gratis ingeschreven worden op een RSS-feed bij een zoekmachine die aan de hand van een ingegeven zoekterm, telkens snel weergeeft welke posts er zijn bijgekomen die aan de zoekterm beantwoorden. Eventueel kunnen deze posts in een database worden opgeslagen, zodat ze niet verloren gaan wanneer de gebruiker ze verwijdert of wanneer de blog verdwijnt. Ook dit kan automatisch gebeuren. Een probleem hierbij is dat verscheidene blogs niet de gehele post in de RSS-feed beschikbaar stellen. Zo sporen ze mensen die de hele post willen bekijken aan om ook de weblog zelf te bezoeken.

3.4.2 Invloedrijke bloggers identificeren

Er zijn twee redenen voor het speciaal behandelen van invloedrijke bloggers.

Indien er over een bedrijf behoorlijk veel geschreven wordt, is het niet rendabel om al deze posts te gaan lezen. De meeste weblogs worden slechts door erg weinig mensen gelezen. De invloed die dergelijke blogs hebben, is dan ook vrij beperkt. Het is veel belangrijker om de relevante, invloedrijke bloggers te vinden. Zij onderbouwen hun mening vaak beter, worden door velen als betrouwbaar binnen hun specialiteit gezien en hun mening wordt vaker gelezen. De invloed die zij op de opinie hebben, is dan ook een stuk groter.

Een tweede reden om invloedrijke bloggers eruit te pikken gaat een stap verder. Verscheidene bedrijven communiceren rechtstreeks met deze bloggers. Ze sturen hen exclusieve informatie. Dit levert in vele gevallen een win-winsituatie op. De blogger kan exclusief nieuws publiceren, het bedrijf krijgt extra aandacht. Het is bovendien ook waarschijnlijk dat de blogger iets minder geneigd zal zijn om op een negatieve manier over het bedrijf te schrijven. Verscheidene bedrijven gaan nog verder en sturen de bloggers prototypes op van hun producten of nodigen hen uit om al dan niet op hun kosten bepaalde voorstellingen of conferenties bij te wonen.

Het kan ook nodig zijn de invloed van weblogs te onderzoeken om na te gaan op welke weblogs men best kan adverteren.

Het vinden van invloedrijke bloggers gebeurt doorgaans aan de hand van Web structure mining. Het aantal links naar elke blogger wordt geteld. Weblogs waarnaar vaak verwezen wordt, zijn hoogstwaarschijnlijk blogs met invloed. Verschillende zoekmachines voor blogs, zoals bijvoorbeeld technorati[22] en blogpulse[23], publiceren een ranking van invloedrijke weblogs. Sommige zoekmachines laten het toe om deze ranking automatisch op te vragen.

Men kan het meten van de invloed verfijnen door andere factoren te laten meetellen zoals uitgaande links, het gemiddelde aantal commentaren per artikel, de frequentie van posten, het aantal bezoekers, het aantal unieke bezoekers en het profiel van deze bezoekers.

---

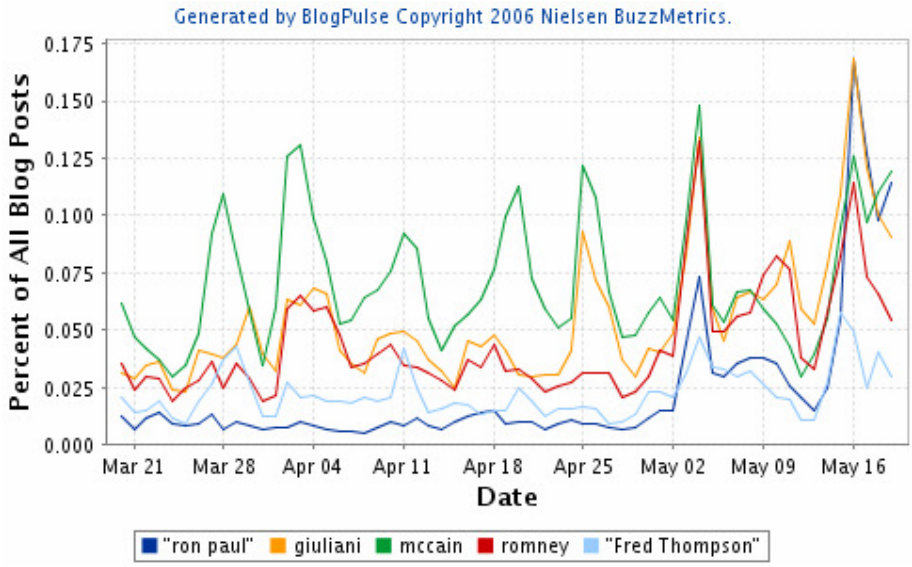
[22] http://technorati.com

[23] http://www.blogpulse.com

3.4.3 Trends zoeken

Onderzoeken hoeveel er over een bepaald onderwerp geschreven wordt kan tot zinvolle inzichten leiden. Bij een bepaalde gebeurtenis kan men dadelijk zien hoeveel berichten hierover verschijnen in de blogosfeer. Dit kan mogelijk een indicatie zijn van het belang dat bloggers aan deze gebeurtenis hechten.

Door simpelweg het aantal keren te tellen dat een woord of woordencombinatie voorkomt, kan men een beeld krijgen over de mate waarin over een bepaald onderwerp gepraat wordt.



Figuur 7: Voorbeeld trend search (Blogpulse)[24]

In Figuur 7 kan men aflezen in hoeveel procent van de posts enkele Republikeinse presidentskandidaten vermeld worden in de periode tussen half maart en half mei 2007. Het is duidelijk te zien dat er rond de tijd van de debatten op 3 mei en 15 mei over alle kandidaten meer gepost werd.

[24] http://www.blogpulse.com

Men kan categorization (zie hoofdstuk 2.2.4) toepassen op berichten en aan de hand hiervan bijhouden hoe vaak er over een bepaald thema geschreven wordt.

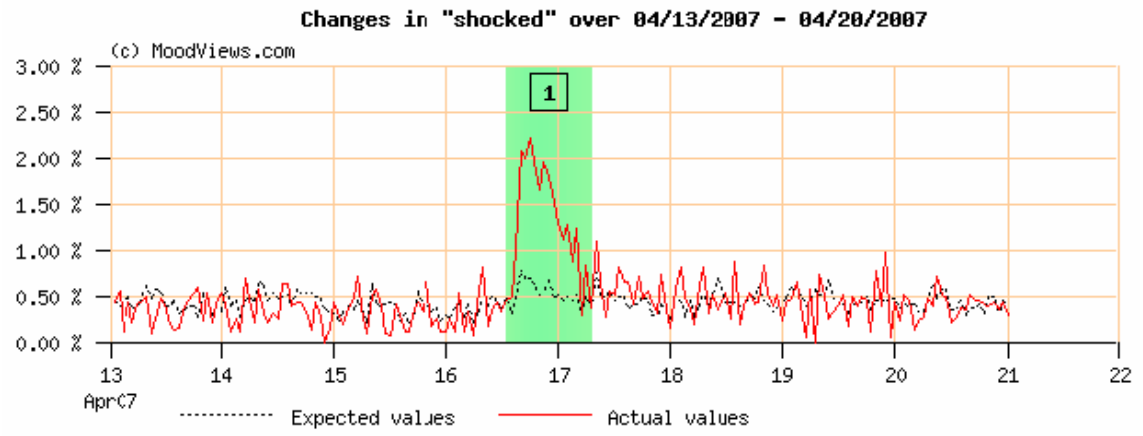
Onderzoek heeft reeds aangetoond dat de mate waarin over een bepaald product geschreven wordt alvorens het op de markt is, een voorspellende waarde heeft om in te schatten hoe succesvol het product zal zijn. Zo toonde onderzoek van Gruhl e.a. (2005) dit aan voor boeken. Tong (2001) ondervond dat er een correlatie bestaat tussen het aantal keren dat er in nieuwsgroepen naar films gerefereerd wordt en het behaalde kassucces.

3.4.4 Sentiment classification

Sentiment classification zoals beschreven in hoofdstuk 2.2.4, kan toegepast worden op weblogs.

Blogposts over een bepaald onderwerp kunnen verzameld worden. Met behulp van sentiment classification kan een beeld gevormd worden over de stemming van de schrijver ten tijde van het schrijven.

Figuur 8 illustreert dit. Moodviews archiveert alle blogposts gemaakt op livejournal.com en puurt hieruit de gemoedstoestand van de betrokken bloggers. Het aantal posts dat als 'shocked' wordt ingedeeld, wordt in procenten in beeld gebracht. Er is een duidelijke stijging te zien op 16 april. Daarenboven zoekt het programma bij grote pieken automatisch naar woorden die buitensporig veel voorkomen. De resultaten hiervan waren:
'*Peak (1) 04/16/2007 14h - 04/17/2007 08h, Overused terms during the peak period: virginia, techs, shoot, campuses, gunman, students, blacksburg, police, universal, kill, tragedies, deadliest, classroom, guns*.'



Figuur 8: Voorbeeld sentiment classification (MoodViews, Moodsignals)[25]

Sentiment classification kan ook voor andere toepassingen gebruikt worden. Zo blijkt uit onderzoek van Mishne en Glance uit 2006 dat er betere resultaten behaald worden bij het voorspellen van het succes van films uit user-generated content indien er een sentiment classification mee in rekening wordt genomen.

Door het selecteren van specifieke blogposts kan er een beeld gevormd worden van de algemene stemming over een bepaald onderwerp. Eventueel kan eerst enkel het relevant gedeelte uit deze blogposts geselecteerd worden.

3.4.5 Opinion analysis

De grootste potentiële waarde om kennis te ontsluiten uit weblogs ligt waarschijnlijk bij opinion analysis (zie 2.2.4). Er zijn reeds bedrijven die dit soort analyse aanbieden zoals bijvoorbeeld Cymfony[26], Nielsen BuzzMetrics[27], IBM[28], en Umbria[29]. Opinion analysis toepassen is niet eenvoudig. Het is op dit moment meestal nog erg

---

[25] http://moodviews.com/Moodsignals

[26] http://www.cymfony.com

[27] http://www.nielsenbuzzmetrics.com

[28] http://new.marketwire.com/2.0/rel.jsp?id=669667

[29] http://www.umbriacom.com

arbeidsintensief en de nauwkeurigheid van de behaalde resultaten is niet steeds erg hoog.

Hoofdstuk 4: Classificatie met behulp van de General Inquirer

4.1 Inleiding

Om een deel van de theorie aan de praktijk te toetsen en meer inzicht te verwerven in zowel de opportuniteiten als de moeilijkheden van een sentiment classification, wordt in dit hoofdstuk een gevalsstudie uitgewerkt.

Hiervoor werd samengewerkt met i.Know[30]. Van hen kwam ook het idee voor het onderwerp van de case: de heisa in de zomer van 2006 rond de Dell laptops die in brand schoten. Er werd besloten een sentiment classification uit te voeren op blogposts door gebruik te maken van de 'General Inquirer'[31].

Als begindatum van de onderzochte periode werd gekozen voor 21 juni. Op 21 juni verscheen namelijk het eerste bericht over een ontplofte Dell laptop. Als einddatum werd geopteerd voor 1 oktober. Omdat verschillende producenten een terugroeping aankondigden, bleef het thema redelijk lang in de aandacht. Daarbij schoot er op 20 september nog een Dell laptop in brand.

In hoofdstuk 4.2 wordt kort het bedrijf i.Know geschetst. Een overzicht van de gebeurtenissen wordt gegeven in hoofdstuk 4.3. In hoofdstukken 4.4 en 4.5 wordt de manier besproken waarop het corpus, dat gebruikt werd voor het uitvoeren van de sentiment classification, werd bekomen. Respectievelijk in hoofdstuk 4.6 en 4.7 wordt gekeken naar de kennis die ontsloten kan worden door te kijken naar het aantal posts per dag en de pagina's waarnaar gelinkt werd. In hoofdstuk 4.8 worden de resultaten van de sentiment classification van de blogposts door de 'General Inquirer' beschreven en geïnterpreteerd. Hoofdstuk 4.9 geeft bondig een overzicht van enkele mogelijke verbeteringen.

---

[30] www.iknow.be

[31] http://www.wjh.harvard.edu/~inquirer/

4.2 i.Know

i.Know[32] is een softwarebedrijf uit Diepenbeek. Ze bieden informatieoplossingen om beter te kunnen omgaan met de grote hoeveelheden ongestructureerde informatie waarmee bedrijven geconfronteerd worden. Ze focussen zich op 4 sectoren waar de hoeveelheid te verwerken informatie traditioneel erg groot is: de medische sector, de juridische sector, de mediasector en de financiële sector.

De voornaamste aangeboden toepassingen zijn int.for® Categorization, int.for® Summarization, int.for® Searching (information retrieval met clustering) en int.for® Distribution (topic tracking). De toepassingen van i.Know kunnen toegepast worden op zowel Engelstalige als Nederlandstalige teksten. Momenteel is men bezig dit uit te breiden naar Franstalige teksten.

4.3 Chronologie

Op 14 augustus 2006 kondigde Dell aan 4,1 miljoen lithium-ion batterijen van draagbare computers terug te roepen omdat er een risico bestond dat deze onder bepaalde omstandigheden te heet konden worden. Dit kon leiden tot brand en zelfs explosiegevaar. De Amerikaanse US Consumer Product Safety Commission (CPSC) had op dat moment weet van 6 incidenten met oververhitte batterijen van de teruggeroepen types in Dell laptops. Bij geen ervan vielen gewonden. De kosten van deze actie werden geschat rond de 250 miljoen dollar.

In de maanden daarop besloten ook andere laptopproducenten tot een terugroeping van potentieel gevaarlijke batterijen. In totaal werden zo'n 10 miljoen batterijen teruggeroepen. Al deze batterijen waren geproduceerd door Sony. Door fouten in het productieproces zouden in sommige batterijen onzuiverheden zoals minuscuul kleine metalen deeltjes gekomen zijn. Deze deeltjes kunnen de polymeerlaag tussen de

[32] http://www.iknow.be

kathode en de anode van de batterij beschadigen, hetgeen een kans tot kortsluiting veroorzaakt.

De beslissing om al deze batterijen terug te roepen wordt vaak aangehaald als voorbeeld van de toegenomen invloed van het internet en de nieuwe media die daarmee gepaard gaan. Een artikel in BusinessWeek van 30 augustus 2006 had als titel: "The Battery Recall: A Win for the Web".[33] Het internet speelde dan ook een centrale rol om het probleem in de aandacht te brengen. Op 21 juni verscheen in de Inquirer, een Britse tabloid nieuws website over IT, een reeks foto's met een in brand staande Dell laptop op een conferentie in Japan.[34] Het nieuws verspreidde zich razendsnel op het web. Talloze blogs pikten het op, waaronder invloedrijke zoals Gizmodo, Flashdot, Engadget, ... Vele bloggers drukten hun bezorgdheid uit over de mogelijkheid dat een laptop in brand zou schieten terwijl iemand hem op de schoot aan het gebruiken was, of terwijl men zich in een vliegtuig bevond.

Het imagoprobleem voor Dell werd er niet kleiner op toen op korte tijd foto's van twee andere in brand gevlogen Dell laptops op het web verschenen. In de blogosfeer verschenen hier weer verschillende postings over. Het leek erop alsof Dell geen andere keus meer had dan te reageren. Zo schreef Engadget, een erg invloedrijke blog, op 31 juli: "We'll keep posting these until we see a recall or a solution, so please Dell, treat 'em right."[35]

Dell besloot midden augustus tot een van de grootste terugroepingen uit de recente geschiedenis. Ook hier lieten de nieuwe media zich gelden. Het nieuws van deze beslissing lekte uit en verscheen vroegtijdig, alweer, in de Inquirer.[36] Dit dwong Dell om de aankondiging te vervroegen en een persconferentie te houden terwijl hun nieuwe website[37] hierover nog niet gereed was.

---

[33] http://www.businessweek.com/technology/content/aug2006/tc20060830_642667.htm

[34] www.theinquirer.net/?article=32550

[35] http://www.engadget.com/2006/07/31/dell-laptop-number-3-explodes

[36] http://www.theinquirer.net/?article=33642

[37] http://www.dellbatteryprogram.com

De vraag kan gesteld worden of er zonder het internet ook dergelijke drastische maatregelen zouden genomen zijn. Van de ongeveer 4,5 miljoen batterijen die in de Verenigde Staten teruggeroepen werden, waren er in januari 2007 16 gevallen van oververhitte batterijen bekend, waarvan er 2 leidden tot brandwonden. Dit terwijl een groot deel van deze batterijen reeds jaren in gebruik was.

Dell bedankte de bloggers later voor de snelheid en de vastberadenheid waarmee ze het probleem op de agenda plaatsten. Het was niet de eerste keer dat Dell geconfronteerd werd met de blogosfeer. Midden 2005 schreef Jeff Jarvis op zijn blog 'BuzzMachine'[38] een reeks negatieve posts[39] over zijn ervaring met de klantenservice van Dell. De 'Dell Hell' posts werden door verschillende bloggers opgepikt. Het was de aanleiding tot een stroom van negatieve kritiek op de dienstverlening. Dell sloot zijn online forums zelfs tijdelijk vanwege de grote hoeveelheid negatieve posts. In juli 2006 stelden ze een 'customer service representative' aan, die continu de internetreputatie van Dell in de gaten dient te houden.

4.4 URL's selecteren

De eerste stap was om een lijst URL's met potentieel relevante blogingaves te verkrijgen. Hiervoor maakte ik gebruik van de zoekmachine op de site www.blogger.com: http://search.blogger.com. Google, eigenaar van Blogger besliste recentelijk enkel de vrijwel identieke zoekmachine http://blogsearch.google.com te blijven ondersteunen. De gebruikte methode om tot een lijst met URL's te komen blijft mits enkele kleine aanpassingen mogelijk.

De zoekmachine biedt de mogelijkheid de eerste 100 resultaten van een zoekopdracht op te vragen met behulp van een RSS-feed. Omdat RSS-feeds XML-bestanden zijn, is het vrij eenvoudig hier de nodige informatie uit te halen.

---

[38] http://www.buzzmachine.com

[39] http://www.buzzmachine.com/archives/cat_dell.html

Met behulp van zelfgeschreven PHP-code (zie Bijlage 2) konden zoektermen worden ingegeven tezamen met een periode waarbinnen gezocht moest worden. Voor elke dag binnen deze periode werd de RSS-feed opgevraagd. Hieruit werden de URL, de titel en de datum gehaald en opgeslagen in een database. Opgemerkt dient te worden dat aangezien de RSS-feed maximaal 100 resultaten kon bevatten en er voor één dag meer dan 100 resultaten konden zijn, niet steeds elk resultaat van de zoektermen opgeslagen kon worden.

Er werd gebruik gemaakt van 8 zoektermen die verband houden met deze case. De zoektermen werden gekozen zodat er een grote kans was dat de URL's die bekomen werden ook daadwerkelijk over de gekozen case gingen. Initieel werden de zoektermen zonder het woord 'laptop' ingegeven. Dit leidde echter tot een significant aantal posts die weinig tot niets met deze zaak te maken hadden. Daarom werd besloten de zoektermen strenger te maken. Zoals reeds werd vermeld in hoofdstuk 4.1, werd als begindag 21 juni 2006 en als einddag 1 oktober 2006 ingegeven. De zoektermen zijn in Tabel 3 terug te vinden, samen met het aantal resultaten dat ze opleverden.

Tabel 3: Zoektermen

Zoektermen	Aantal
Dell laptop fire	3.284
Dell laptop fires	3.399
Dell laptop explosion	552
Dell laptop explodes	1.065
Dell laptop exploded	587
Dell laptop flames	1.673
Dell laptop flame	439
Dell laptop recall	3.302
Totaal:	14.301
Unieke:	6.997

Het ingeven van deze zoektermen leidde tot een lijst met 14.301 URL's die mogelijk relevant waren. Na het verwijderen van duplicaten bleven 6.997 unieke URL's over.

4.5 Blogposts uithalen

De volgende stap was de inhoud uit de blogposts achter de URL's te verkrijgen. Omdat de software van i.Know gebruik kan maken van txt-files als input voor tekstanalyse, diende de inhoud opgeslagen te worden in txt-files. Het leek mij ook interessant om na te gaan naar welke URL's er gelinkt werd.

In Bijlage 3 is de php-code te vinden die geschreven werd om deze taak tot een goed einde te brengen. De URL's die zich in de database bevonden, werden opgehaald. Met behulp van een lus werd er voor elk van deze 6.997 URL's de volgende stappen doorlopen:

- Broncode opvragen
- Relevant deel van de broncode selecteren
- Links uithalen
- Geselecteerd gedeelte bewerken
- Opslaan in txt-file

Om deze stappen te verduidelijken zal ik een korte blogpost, zie Figuur 9, als voorbeeld aanhalen.


THURSDAY, JUNE 22, 2006

When laptops go bad

Dell laptop explodes at Japanese conference

"AN INQUIRER READER attending a conference in Japan was sat just feet away from a laptop computer that suddenly exploded into flames, in what could have been a deadly accident."

Oooh nasteh.

Ads by Google
View ads about:
»

POSTED BY GRID212 AT 12:30 PM

0 COMMENTS | TRACKBACK

LINKS TO THIS POST:

CREATE A LINK

Figuur 9: Voorbeeld blogpost[40]

4.5.1 Broncode opvragen

De broncode werd opgevraagd met behulp van de PHP-functie file_get_contents(). Deze functie retourneert het resultaat als een string. Een aantal weblogs bleek echter niet meer te bestaan. Bij 1.984 URL's werd een HTTP-foutmelding 'Error 404' verkregen. Daarenboven werd er geen bruikbaar resultaat behaald bij 100 blogposts van het type ???.spaces.live.com. Bij de overige 4.913 URL's werd er wel een string met daarin de broncode verkregen.

Een klein gedeelte van de broncode van de blogpost uit Figuur 9 wordt weergegeven in Figuur 10.

---

[40] http://grid212.blogspot.com/2006/06/when-laptops-go-bad-dell-laptop.html


```
<!-- Begin .post -->
<div class="post"><a name="115097942316825365"></a>

        <div class="post-body">
      <div>
    <div style="clear:both;"></div><h3 class="post-title">When laptops go bad</h3><br /><a
href="http://www.theinquirer.net/?article=32550">Dell laptop explodes at Japanese conference</a><br /><br />"AN INQUIRER
READER attending a conference in Japan was sat just feet away from a laptop computer that suddenly exploded into flames, ir
what could have been a deadly accident."<br /><br />Oooh nasteh.<div style="clear:both; padding-bottom:0.25em"></div>
    </div>
    </div>

        <p id="linkunit">

<script type="text/javascript"><!--
google_ad_client = "pub-9081959924745729";
google_ad_width = 200;
google_ad_height = 90;
google_ad_format = "200x90_0ads_al_s";
//2007-04-15: Link Unit
google_ad_channel = "1247607021";
//-->
</script>
<script type="text/javascript"
  src="http://pagead2.googlesyndication.com/pagead/show_ads.js">
</script>

        </p>

    <p class="post-footer">
      <em>posted by Grid212 at <a href="http://grid212.blogspot.com/2006/06/when-laptops-go-bad-dell-laptop.html"
title="permanent link">12:30 PM</a></em>
        <span class="item-action"><a href="http://www2.blogger.com/email-post.g?blogID=24533790&postID=115097942316825365"
title="Email Post"><span class="email-post-icon"> </span></a></span><span class="item-control blog-admin pid-
260477452"><a style="border:none;" href="http://www2.blogger.com/post-edit.g?blogID=24533790&postID=115097942316825365"
title="Edit Post"><span class="quick-edit-icon"> </span></a></span>
    </p>

  </div>
  <!-- End .post -->


  <!-- Begin #comments -->

  <div id="comments">
```

Figuur 10: Deel broncode blogpost uit Figuur 9

4.5.2 Relevant deel van de broncode selecteren

De volgende stap, uit deze string het relevante deel selecteren, was een stuk ingewikkelder. Omdat html-pagina's semi-gestructureerd zijn, opteerde ik ervoor om hiervan gebruik te maken. De meeste weblogs gebruiken betekenisvolle html-tags om het begin en einde van een post af te bakenen. Voorbeelden zijn het gebruik van ' ' of '<!-- start entry of your post -->'.

Gezien de grote hoeveelheid aan gebruikte weblog software, is het niet zo eenvoudig om de juiste tags te selecteren. Mijn eerste idee was dan ook om mij te beperken tot een paar populaire blogformaten zoals ???.blogspot.com en ???.livejournal.com. Maar ook hier verschilden de gebruikte html-tags aanzienlijk. Dit komt omdat de meeste aanbieders aan de blogger een grote mate van vrijheid bieden om hun blog zelf in te richten. Zo is livejournal open source software en kan het dus gemakkelijk aangepast worden. Ook stelde ik vast dat verschillende weblogs op standaardsoftware draaien maar dat dit niet aan hun naam te zien is.

Daarom besloot ik toch te proberen om voor alle 4.913 broncodes het relevante gedeelte te proberen selecteren. In de php-code (zie Bijlage 3) werd een array gevuld met mogelijke begin- en eindtags. Vervolgens werd in een lus per array-element gekeken of beide tags in de tekst voorkwamen. Indien dit het geval was, werd het gedeelte van de broncode dat werd ingesloten door deze tags, opgeslagen in een string en werd de lus beëindigd. De volgorde waarin de tags in de array zitten is dus van belang. Erg specifieke tags, zoals bijvoorbeeld ' ', bevinden zich vooraan in de array terwijl naar het einde van de array vooral meer algemene tags, zoals bijvoorbeeld ' ', terug te vinden zijn. De logica hierachter is dat de meer specifieke tags een grotere kans hebben om het relevante gedeelte correct af te bakenen.

Voor het voorbeeld uit Figuur 10 zijn de gewenste begin- en eindtags ' ' en '<p class="post-footer">' aangegeven in de broncode. Het eindresultaat van deze stap zal in dit geval dus een string zijn met alles wat zich tussen deze twee tags bevindt. Andere mogelijke begin- en eindtags zijn '<!-- Begin .post -->' en '<!-- End .post -->'. Deze zijn minder nauwkeurig aangezien ze te veel zouden ophalen.

Bij 3.490 van de 4.913 broncodes werd er op deze manier succesvol een deel geselecteerd. In Bijlage 4 is een lijst terug te vinden die weergeeft welke begin- en eindtags meer dan 30 keer gebruikt werden. Bij 1.423 broncodes werd er geen resultaat gehaald. Hier zijn diverse mogelijke redenen voor. Een eerste mogelijkheid is dat de broncode simpelweg geen betekenisvolle html-tags bevat die gebruikt kunnen worden. De door mij gebruikte methode is dan niet geschikt om tot een resultaat te komen. Een andere mogelijke reden is dat de broncode wel betekenisvolle html-tags bevat, maar dat deze tags niet in de array zitten. Het is ook mogelijk dat de weblog op het moment van het opvragen van de broncode niet meer bestond en vervangen werd door een andere inhoud, zoals bijvoorbeeld een reclamesite.

De behaalde nauwkeurigheid viel redelijk mee. In een aantal gevallen werden te veel data uitgeselecteerd. Het ging hier meestal over metadata die net voor of na de inhoud van de blogingave stonden. Deze metadata hebben over het algemeen echter weinig invloed op een sentiment classification. Bij een aantal posts verwees de URL niet langer naar de relevante blogpost maar naar een meer recente blogpost. In dit geval werd dus de verkeerde inhoud geselecteerd. Een ander probleem was dat sommige blogposts verwijderd waren. Meestal werd in dit geval een HTTP-foutmelding verkregen, maar in een aantal gevallen werd de inhoud van een foutmelding zoals bijvoorbeeld 'No such entry.' geselecteerd.

4.5.3 Links uithalen

De derde stap bestond erin te kijken waar de bloggers naartoe linkten. Dit gebeurde met behulp van de functie haalurls(). In totaal werden er 59.152 links opgeslagen in de database.

In Figuur 10 is te zien dat er één link staat in het gedeelte tussen de begin- en eindtag: 'http://www.theinquirer.net/?article=32550'.

4.5.4 Geselecteerd gedeelte bewerken

Vooraleer de inhoud van de posts naar de txt-files geschreven kon worden, moest eerst het geselecteerde gedeelte van de broncode opgeschoond worden. Javascripts en HTML-en PHP-codes dienden verwijderd te worden. Het gebruik van de PHP-functie strip_tags() bracht een ander probleem aan het licht. Door de HTML-tags te verwijderen, werd de tekst vaak minder leesbaar.

Indien bijvoorbeeld alle codes verwijderd worden uit de broncode van Figuur 10 zou de opgehaalde tekst als volgt beginnen:
'When laptops go badDell laptop explodes at Japanese conference"AN INQUIRER...'

Het is duidelijk dat dit niet het gewenste resultaat is. Er bestaan verschillende open source converters om HTML naar tekst om te zetten. Toch besloot ik om deze niet te gebruiken omdat ze niet direct de flexibiliteit boden die ik zocht.

Tekstanalysesoftware hecht grote waarde aan leestekens om zinnen af te bakenen. Vermits bloggers vaak niet al te nauwkeurig omspringen met de gebruikte taal, koos ik ervoor om dit enigszins te compenseren door enkele punten toe te voegen. Alvorens de code te verwijderen werden enkele HTML-tags, zoals '</p>' en '</h1>', die normaal enkel aan het einde van een zin voorkomen vervangen door een punt en een spatie. Andere HTML-tags, zoals '</br>' en ' ', die meestal enkel aan het einde van een woord staan, werden vervangen door een spatie. Na het verwijderen van de code werd dan gezocht naar overbodige spaties en punten om eventueel te veel toegevoegde tekens weer te verwijderen.

Voor het voorbeeld in Figuur 10 werd de opgehaalde tekst hierdoor als volgt:
'*When laptops go bad. Dell laptop explodes at Japanese conference "AN INQUIRER READER attending a conference in Japan was sat just feet away from a laptop computer that suddenly exploded into flames, in what could have been a deadly accident." Oooh nasteh.*'
Eigenlijk zou er nog een punt moeten toegevoegd worden tussen 'conference' en '"AN', maar dit is naar mijn mening moeilijker te realiseren.

4.5.5 Opslaan in txt-file

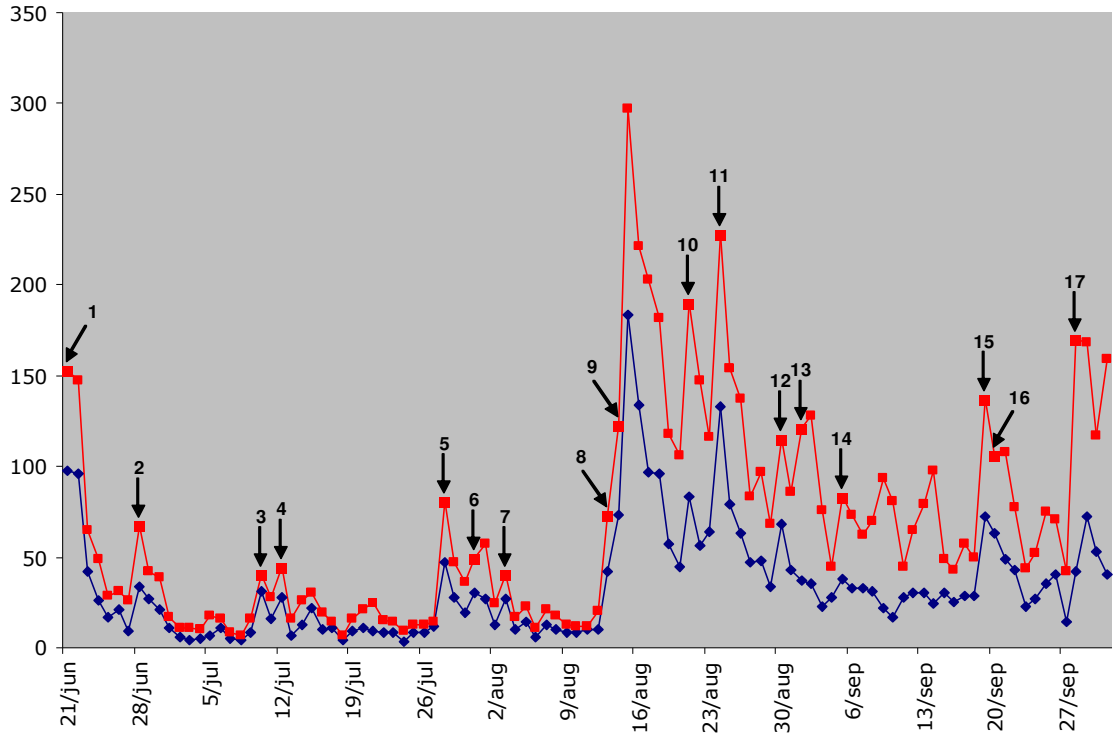
De bewerkte tekst werd weggeschreven naar een txt-bestand. Als naam voor het bestand werd het id van de overeenkomstige URL in de database gegeven.

4.6 Datum

Zonder echt naar de inhoud van de blogposts te kijken, kan er al bruikbare informatie uit geselecteerd worden. Door te kijken naar het aantal posts per dag, kan er een beeld gevormd worden over het belang dat bloggers aan een bepaald nieuwsfeit hechten. In Figuur 11 zijn er twee lijnen te zien. De rode lijn geeft het aantal posts per dag weer van de 6.997 URL's die onder hoofdstuk 4.4 bekomen werden. De blauwe lijn laat enkel de posts zien waarvan de inhoud succesvol werd omgezet en die gebruikt zullen worden voor tekstanalyse. Te zien is dat de verdeling

van teksten in het bekomen corpus ongeveer overeenstemt met de verdeling van de URL's.

Niet onverwacht ligt de dag met de meeste berichten (15 augustus) rond het tijdstip dat Dell de terugroeping aankondigde. Dit gebeurde op 14 augustus. Ook valt het op dat er meer gepost werd over de eerste in brand geschoten laptop (22 juni) dan over de twee volgende (25 juli en 31 juli). Het is ook te zien dat invloedrijke weblogs het aantal berichten over een bepaalde zaak sterk kunnen beïnvloeden. Zo wordt er op 25, 26 en 27 juli nauwelijks geschreven over de tweede in brand geschoten laptop. Pas toen het bericht op 28 juli op enkele invloedrijke weblogs, waaronder Engadget, verscheen werd er plots door verschillende bloggers over geschreven.



Figuur 11: aantal posts per dag (rood) en aantal posts waarvan de inhoud kon geselecteerd worden (blauw)

1. 21 juni: Dell laptop vliegt in brand op een Japans congres. Artikel in de 'Inquirer': http://www.theinquirer.net/?article=32550

2. 28 juni: Dell kondigt aan dat het een onderzoek naar de zaak gestart heeft.
3. 10 juli: Enkele berichten over geruchten dat een andere Dell laptop in brand geschoten is.
4. 12 juli: Verschillende posts over Parodie "My PC Is On Fire": http://www.youtube.com/watch?v=FPejYdBM11I
5. 28 juli: Nieuws over een andere uitgebrande Dell laptop op enkele invloedrijke blogs: http://community.tomshardware.com/dellpost.html?t=192887 (25 juli) http://www.engadget.com/2006/07/28/another-dell-laptop-ignites/
6. 31 juli: Berichten over een derde uitgebrande Dell laptop: http://blogs.smh.com.au/mashup/archives/005359.html
7. 3 augustus: Man beweert dat zijn truck uitbrandde nadat zijn Dell laptop in brand schoot: http://www.consumeraffairs.com/news04/2006/08/dell_fire.html
8. 13 augustus: Nieuws over de geplande terugroeping van 4.1 miljoen batterijen lekt uit. http://www.theinquirer.net/default.aspx?article=33642
9. 14 augustus: Dell kondigt terugroeping aan. http://www.dell.com/content/topics/global.aspx/corp/pressoffice/en/2006/2006_08_14_rr_002
10. 21 augustus: Verschillende posts over een man die beweert dat zijn huis uitbrandde door een in brand geschoten Dell laptop. http://www.heraldtribune.com/apps/pbcs.dll/article?AID=/20060818/NEWS/608180446/
11. 24 augustus: Apple kondigt aan 1.8 miljoen batterijen terug te roepen. http://www.cpsc.gov/cpscpub/prerel/prhtml06/06245.html
12. 30 augustus: Verschillende posts verwijzen naar een artikel in Business Week: "The Battery Recall: A Win for the Web" http://www.businessweek.com/technology/content/aug2006/tc20060830_642667.htm
13. 1 september: Berichten over een ontplofte laptop in het Verenigd Koninkrijk.
14. 5 september: Panasonic kondigt aan 6.000 batterijen terug te roepen.
15. 19 september: Toshiba roept 340.000 batterijen terug. http://www.csd.toshiba.com/cgi-in/tais/su/su_sc_dtlView.jsp?soid=1482876

16. 20 september: Gebouw van Yahoo! ontruimd nadat een Dell laptop in brand schiet.
    http://www.engadget.com/2006/09/20/dell-battery-explodes-at-yahoo-hq-hundreds-evacuat/
17. 28 september: Lenovo/IBM kondigen aan 526.000 batterijen terug te roepen.

4.7 Links

Het leek mij interessant om te kijken naar welke pagina's de bloggers verwezen. In 4.5.3 werden 59.152 links uit de blogposts gehaald en opgeslagen in een database. Van de 3.490 posts waarvan de inhoud werd opgehaald, bevatten 951 – dit is ongeveer 27% - geen enkele link. Gemiddeld bevatte een post ongeveer 17 links. Dit aantal wordt zwaar vervormd door een klein aantal posts die erg veel links bevatten. Zo werden er bij 12 blogposts meer dan 1.000 links opgehaald. De mediaan ligt met de waarde 2 dan ook logischerwijze een stuk lager. Het werkelijke aantal links ligt waarschijnlijk nog lager omdat er bij een aantal posts buiten de inhoud van de posts ook metadata werden opgehaald. Deze metadata bevatten vaak links zoals permalink, tags, datum, digg it, trackback, ... Bijlage 5 geeft een idee over de verdeling van het aantal links per post.

Met behulp van een SQL-query werd een lijst bekomen die aangaf welke links het meeste voorkwamen. De meest voorkomende links zijn in Bijlage 6 terug te vinden. Wat opvalt is dat het overgrote deel van deze links geen verband houden met de case. De vier meest voorkomende links, 'http://www.cashbazar.com', 'http://www.organicgreens.us', 'http://www.home-based-business-team.com' en 'http://www.end-your-debt.com' zijn duidelijk niet relevant. Pas op de vijfde plaats is er een verwijzing terug te vinden naar een relevante pagina: 'http://www.theinquirer.net/?article=32550'.

De verklaring hiervoor is volgens mij splogs. Splogs zijn spam blogs die gebruikt worden om de ranking van bepaalde sites op zoekmachines te beïnvloeden. Ze bevatten geen bruikbare inhoud en bestaan over het algemeen louter uit een erg groot aantal links. Om een groot deel van deze splogs eruit te filteren kon van dit

laatste gebruik gemaakt worden. Door enkel de links uit blogposts met minder dan 100 links mee te tellen, werd de tabel bekomen die in Bijlage 7 is terug te vinden. Wat opvalt is dat de links nu meestal wel relevant zijn en dat het aantal verwijzingen naar de pagina's vrijwel gelijk blijft. Het uitsluiten van de posts met meer dan 100 links heeft dus waarschijnlijk niet geleid tot het wegfilteren van veel relevante links.

Sommige links verwijzen naar dezelfde pagina. Zo is het duidelijk dat 'http://www.theinquirer.net/default.aspx?article=32550' eigenlijk identiek is aan 'http://www.theinquirer.net/?article=32550'. Gelijkaardige links werden bij elkaar gevoegd. Tevens werden niet relevante links, zoals bijvoorbeeld '#comment' verwijderd. De einduitkomst kan teruggevonden worden in Tabel 4.
Niet onverwacht wordt er het meest gelinkt naar het artikel van de 'Inquirer' dat de hele zaak inluidde. Er wordt vaak gelinkt naar de webpagina's van de producenten waar terug te vinden is welke batterijen in aanmerking komen. Ook het ontbreken van verwijzingen naar traditionele media springt in het oog. Verder valt op dat er naar een groot aantal artikels gelinkt wordt op de website www.engadget.com.

In totaal verschenen er op www.engadget.com meer dan 15 artikels die specifiek over de problemen met oververhitte batterijen in Dell laptops gingen. In totaal werd er 412 keer verwezen naar Engadget. In 'The State of the Live Web, April 2007'[41] rangschikt Technorati engadget.com als de meest invloedrijke blog ter wereld. Van alle online informatiebronnen staat het op de 19de plaats. Ook Blogpulse.com rangschikt www.engadget.com op basis van het grootste aantal verwijzingen als eerste.[42]

[41] http://technorati.com/weblog/2007/04/328.html

[42] http://www.blogpulse.com/profile?url=www.engadget.com (nr.1 op 30 mei 2007)

Tabel 4: Links

Link:	Aantal:
http://www.theinquirer.net/?article=32550	330
https://www.dellbatteryprogram.com/	253
http://miluping.com/dellnews/category/uncategorized/	68
https://support.apple.com/ibook_powerbook/batteryexchange/	54
http://www.engadget.com/2006/07/28/another-dell-laptop-ignites/	41
http://www.engadget.com/2006/06/22/dude-your-dell-is-on-fire/	35
http://bl-technology.desertscapeit.com/?cat=1	32
http://www.dell.com	36
http://www.gizmodo.com/gadgets/laptops/dell-laptop-explodes-in-flames-182257.php	35
http://www.engadget.com/2006/07/31/dell-laptop-number-3-explodes/	30
http://www.engadget.com/2006/08/14/dell-recalls-4-1-million-batteries/	27
http://forumz.tomshardware.com/hardware/Dude-Dell-freaking-blew-ftopict192887.html	21
http://www.engadget.com/2006/09/20/dell-battery-explodes-at-yahoo-hq-hundreds-evacuat/	20
http://www.consumeraffairs.com/news04/2006/08/dell_fire.html	19
http://www.heraldtribune.com/apps/pbcs.dll/article?AID=/20060818/NEWS/608180446/-1/Help0530	16
http://www.cpsc.gov/cpscpub/prerel/prhtml06/06231.html	15
http://www.engadget.com/2006/08/03/dell-laptop-ammo-no-go/	12
http://www.cpsc.gov/cpscpub/prerel/prhtml06/06245.html	12
http://www.engadget.com/2006/06/29/dell-looking-into-flaming-laptop-incident/	12
http://www.engadget.com/2006/07/20/dell-knew-about-dozens-of-burned-laptops-two-years-before-reca/	12
http://www.direct2dell.com/one2one/archive/2006/08/14/1803.aspx	12

http://www.smh.com.au/news/biztech/safety-first-for-carryon-dells/2006/08/23/1156012601607.html	11
http://www.engadget.com/2006/09/16/thinkpad-explodes-at-lax-ignites-bomb-scare/	11
http://techfreep.com/virgin-atlantic-bans-dell-apple-laptops.htm	11
http://www.theinquirer.net/?article=33321	9

4.8 General Inquirer

Alvorens over te gaan tot het uitvoeren van een sentiment analysis werden eerst enkele txt-files geschrapt. Het betreft hier bestanden die groter waren dan 100 kilobyte. Dit is erg veel voor een blogpost. De kans is groot dat het om een niet-relevante post of een spamblog gaat. Op deze manier werden 50 blogposts geschrapt waardoor het corpus nog uit 3.430 blogposts bestond.

4.8.1 General Inquirer

De 'General Inquirer'[43] is een methode voor automatische inhoudsanalyse van tekstuele data. Het systeem kan Engelstalige teksten analyseren op basis van lexicons. Zo'n lexicon bevat verschillende woordenlijsten. Elke woordenlijst brengt woorden samen die tot een gemeenschappelijk semantisch veld behoren.

Ik zal gebruik maken van twee categorieën in de 'Harvard-IV-4 dictionary': 'Positiv' en 'Negativ'. 'Positiv' is een woordenlijst met 1.915 woorden met een positieve uitstraling zoals bijvoorbeeld 'ability', 'accomplish' en 'accept'. De categorie 'Negativ' bevat 2.291 woorden zoals 'abandon', 'aggression' en 'angry'.

De 'General Inquirer' gaat na hoeveel woorden er per categorie in een tekst voorkomen. Het systeem telt niet enkel het aantal exacte overeenkomsten. Indien een woord meerdere betekenissen kan hebben, probeert het na te gaan welke

[43] http://www.wjh.harvard.edu/~inquirer/

betekenis in de gebruikte context relevant is. Zo heeft het woord 'fun' meestal een positieve bijklank, maar heeft het in de woordencombinatie 'making fun (of)' een negatieve connotatie. In de woordenlijst is enkel de grondvorm van woorden opgenomen, de infinitief voor werkwoorden, het enkelvoud voor substantieven en de onverbogen vorm voor adjectieven. Teksten bevatten echter vaak vervoegde of verbogen vormen. Het programma herkent deze vormen en koppelt ze aan hun grondvorm. In de woordenlijst vindt men enkel 'fail' terug, maar het woord 'failing' zal automatisch gematcht worden met de infinitief 'fail'.
Woorden in een bepaalde categorie wegen steeds even sterk door. Alhoewel het woord 'disgusting' bij de meeste mensen een sterkere negatieve connotatie zal oproepen dan het woord 'bad', zullen ze beide even zwaar doorwegen voor de categorie 'Negativ'.

Als output geeft deze methode voor elk van de verschillende categorieën een getal, dat weergeeft hoeveel woorden uit deze categorie in de tekst voorkomen.

4.8.2 Werkwijze

Een demoversie van de 'General Inquirer' is online vrij te gebruiken.[44] Ik probeerde deze versie uit en stelde vast dat het automatisch ophalen van de resultaten vrij vlot verliep. De demoversie is vooral bedoeld voor kleine teksthoeveelheden. Omdat de blogposts over het algemeen vrij kort waren, oordeelde ik dat deze site volstond om tot bruikbare resultaten te komen. De PHP-code die hiervoor gebruikt werd is terug te vinden in Bijlage 8.

Er waren echter enkele problemen. Allereerst werkte het systeem niet wanneer bepaalde tekens, zoals ö of ô, werden ingevoerd. Dit werd opgelost door enkel reguliere tekens toe te laten en alle andere tekens eruit te filteren. Een tweede probleem was dat de inhoud van sommige blogposts te groot was om in één keer verwerkt te worden. Dit was het geval indien de inhoud meer dan 2.000 tekens bedroeg. De tekst moest dan gesplitst worden. Omdat splitsen middenin een zin de resultaten kan beïnvloeden, werd getracht te splitsen aan het einde van een zin. Ik

[44] http://www.webuse.umd.edu:9090

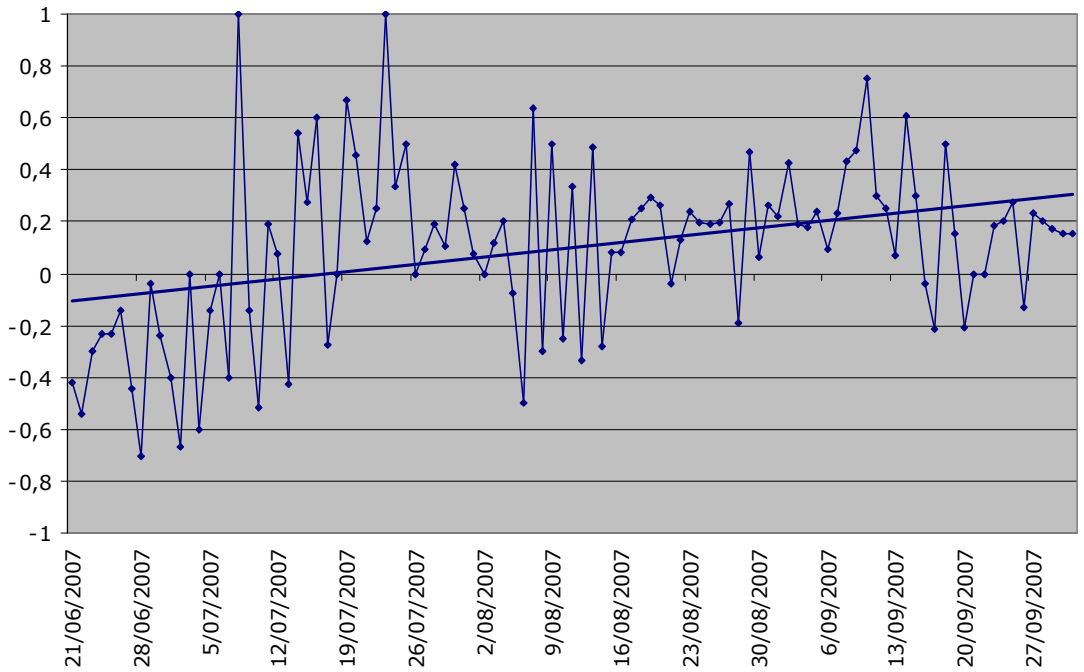
stelde in te splitsen vanaf de eerst voorkomende punt na 1.700 tekens. Het laatste probleem dat ik tegenkwam, was dat bepaalde woorden het systeem deden crashen waardoor er geen geldig resultaat verkregen werd. Dit gebeurde bij woorden zoals bijvoorbeeld 'zune' en 'zulu'. Omdat ik hier geen enkele logica achter zag, kon ik dit probleem niet oplossen. Voor de blogposts die woorden zoals deze bevatten, werd er geen resultaat bekomen. Bij 26 van de 3.430 txt-bestanden waarvan de inhoud werd doorgestuurd, was dit het geval. Voor de overige 3.404 blogposts volgt hier een bespreking van de behaalde resultaten.

4.8.3 Resultaten

Zoals reeds vermeld, werd als output voor elk van de categorieën een getal bekomen dat weergeeft hoe vaak woorden uit deze categorie in de tekst voorkwamen. De 3.404 blogposts bevatten gezamenlijk 1.562.010 woorden. Van deze woorden labelde de 'General Inquirer' 43.179 woorden (2,76%) als negatief en 61.381 als positief (3,39%). De meest voorkomende positieve woorden waren: 'FREE#1' (1.858 keer), 'HOME' (1.689 keer) en 'GOOD#1' (1.406 keer). Bij de negatieve woorden kwamen 'EXPLODE' (1.920 keer), 'PROBLEM' (1.606 keer) en 'NEED#1' (821 keer) het vaakst voor. Een meer uitgebreide lijst is terug te vinden in Bijlage 9.

De blogposts werden in drie categorieën ingedeeld: 'positief', 'negatief' en 'neutraal'. 'Positief' indien er meer positieve woorden dan negatieve woorden in voorkwamen. 'Negatief' indien ze minder positieve woorden dan negatieve woorden bevatten. Bij een gelijk aantal werden ze als 'neutraal' gezien. Op deze manier waren 1.662 posts 'positief' (48,82%), 1.377 'negatief' (40,45%) en 365 'neutraal' (10,72%).

Vervolgens werd voor elke dag een score berekend die een indicatie geeft van de toon in de posts van die dag. Posts uit de categorieën 'positief', 'negatief' en 'neutraal' kregen respectievelijk de waardes 1, -1 en 0. De score weergegeven in Figuur 12 is het gemiddelde van deze waardes. Indien bijvoorbeeld voor een bepaalde dag 2 posts 'positief' zijn, 1 'negatief' en 1 'neutraal' zal de score voor deze dag 0,25 bedragen.



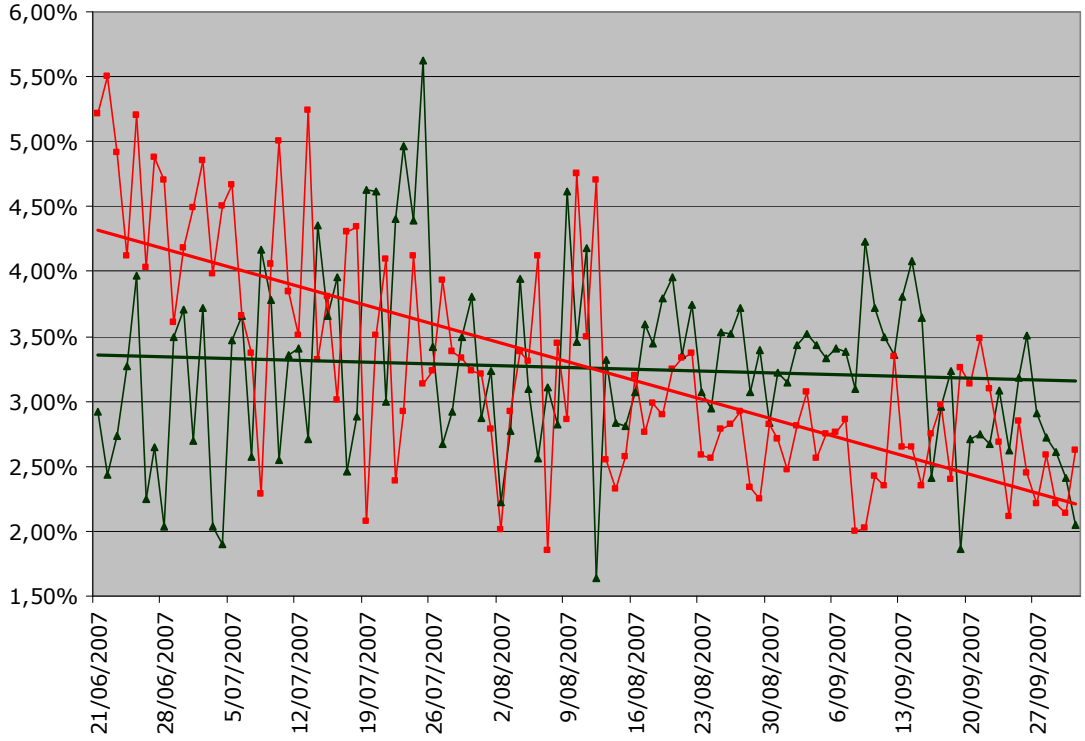
Figuur 12: gemiddelde score per dag

Te zien is dat de posts in het begin van de onderzochte periode gemiddeld negatiever worden ingeschat dan diegenen naar het einde van de periode. Vooral in de eerste weken lijkt er een positieve trend te zien te zijn in de scores. Later blijven deze scores relatief stabiel. De posts die in de twee weken na het bericht over de eerste ontplofte laptop geschreven zijn, bevatten over het algemeen meer 'negatieve' woorden dan 'positieve' woorden. Opvallend is wel dat de posts die geschreven worden na de berichten over de twee andere laptops die in brand schoten (28 juli en 31 juli) gemiddeld gezien gematigder zijn. Er is ook geen duidelijke kentering te zien rond het tijdstip (14 augustus) dat Dell de terugroeping aankondigde.

Een bekommernis was of het resultaat niet sterk vervormd werd door opname van scores van dagen met weinig posts. Dit lijkt echter niet het geval te zijn. In Bijlage 10 is op Figuur 24 te zien dat de scores over het algemeen inderdaad gematigder

worden naargelang het aantal posts toeneemt, maar de globale trend blijft, zoals te zien is op Figuur 25 in Bijlage 10, ongeveer ongewijzigd.

In Figuur 13 wordt de evolutie van het gemiddeld percentage 'negatieve' en 'positieve' woorden weergegeven. Voor elke post werd het percentage van deze twee categorieën woorden berekend. De score voor een bepaalde dag was dan het gemiddelde van de percentages bij alle posts van die dag.



Figuur 13: Gemiddeld % positieve woorden (groen) en gemiddeld % negatieve woorden (rood)

Ook hier werd nagegaan of het behaalde resultaat niet beïnvloed werd door de resultaten behaald op dagen met weinig posts. Uit Figuur 26 en Figuur 27 in Bijlage 11 blijkt dat wanneer enkel dagen worden meegeteld met meer dan 25 posts, de trend relatief ongewijzigd blijft.

Uit Figuur 13 blijkt dat het gemiddeld percentage 'positieve' woorden per dag ruwweg stabiel blijft en rond de 3,25% schommelt. Er zijn wel enkele schommelingen te zien zoals gemiddeld genomen een lager percentage in de eerste weken. Daarna volgt een periode met enkele dagen met hoge percentages 'positieve' woorden. Deze dagen zijn echter veelal dagen met erg weinig posts. Ook hier is er geen merkbare trend te bespeuren vanaf 14 juli, de dag dat Dell besloot tot een terugroeping. Het percentage 'positieve' woorden stijgt niet merkbaar.

Bij het percentage 'negatieve' woorden is een duidelijke neerwaartse trend op te merken. In de eerste weken is het gemiddelde percentage (rond de 4,5%) duidelijk hoger dan in de rest van de periode. Een mogelijke verklaring hiervoor is het artikel in 'The Inquirer'[45] dat de zaak aan het licht bracht. Dit artikel is ook duidelijk negatief. Het artikel door de 'General Inquirer' laten analyseren geeft als resultaat: 4 'positieve' woorden, 8 'negatieve' woorden, percentage 'positieve' woorden: 2,34% en percentage 'negatieve' woorden 4,68%. Verschillende bloggers namen grote stukken van dit artikel letterlijk over.

Ook hier zien we dat het aantal 'negatieve' woorden niet sterk stijgt na de verschijning van foto's op het internet van de twee andere laptops die in brand geschoten. De dag dat Dell de terugroeping aankondigde, 14 augustus, lijkt evenmin een kantelpunt te zijn in het gemiddeld percentage 'negatieve' woorden. Al overschrijdt deze waarde na deze datum de 3,5% niet meer, wat in de periode daarvoor wel frequent gebeurde.

Wat betekenen deze waarden nu exact? Vermits er enkel gekeken wordt naar het aantal 'positieve' en 'negatieve' woorden is het antwoord op deze vraag niet eenduidig. Het lijkt voor de hand te liggen dat Dell posts die duidelijk als 'negatief' bestempeld worden, minder graag zal zien voorbijkomen dan posts die als 'positief' bestempeld worden. Bij het overgrote gedeelte van de posts leek mij dit ook te kloppen maar omdat er niet gekeken wordt waarover deze 'negatieve' of 'positieve'

[45] http://www.theinquirer.net/default.aspx?article=32550

woorden gaan, is dit niet steeds het geval. Een post met veel positieve woorden, spreekt niet altijd positief over Dell. Evenmin is een post met overwegend negatieve woorden steeds antireclame voor Dell. Hieronder zal ik kort 2 voorbeelden aanhalen. Bij beide voorbeelden wordt volgens mij de toon correct ingeschat door de 'General Inquirer'. Maar de conclusies die op basis hiervan voor Dell getrokken dienen te worden, zijn tegenstrijdig met deze toon.

De tekst in de txt-file van het eerste voorbeeld luidt als volgt:
'*How mass media craps in your brain. 22 June 2006. Here's just one example of how mass media craps in your brain - an article in The Inquirer about Dell laptop explosion. I picked this one as it is number one in Bloglines' Top Links for June 21, 2006. Read it. You'll notice a whole bunch of crap that was intended for your brain. Need any hints? Here are a few:. "Dell laptop" (why on earth the brand of the laptop is important in this single case?). "exploded into flames, in what could have been a deadly accident" (I'm sorry… deadly?). "his advice is …stay away, away, away" (Stupid Silly… Unplug the damn thing! Or cut the power in the room. It's all about electricity.). "For the record, this is a Dell machine" (for which record? And why are you so sure and concerned about it anyway?). "It is only a matter of time until such an incident breaks out on a plane" (Oh, p'lease! It's just a laptop. Even if something like this will ever happen on the plane, a trained stuard will take of the fire with fire extinguisher. You don't have to scary everyone just yet.). There's more. Either it's all is a joke, or I don't get it. Damn! I don't get it even if it's a joke….*'[46]

In de tekst worden 14 'negatieve' woorden (de vol onderstreepte woorden) en 4 'positieve' woorden (de met stippellijn onderstreepte woorden) door de 'General Inquirer' herkend. Het percentage 'negatieve' woorden bedraagt 6,31%, het percentage 'positieve' woorden is gelijk aan 1,80%. Het is echter duidelijk dat de schrijver van de post het artikel van de 'Inquirer' als erg overdreven beoordeelt en vindt dat er geen reden tot paniek is. Ondanks de erg negatieve score lijkt dit dus eerder goed nieuws voor Dell in de gegeven situatie. Het is met deze methode dus

[46] http://mamchenkov.net/wordpress/2006/06/22/how-mass-media-craps-in-your-brain

erg moeilijk om onderscheid te maken tussen negatieve reacties op Dell naar aanleiding van dit bericht of negatieve reacties op het bericht zelf.

Een ander voorbeeld:
'*And Now a Message From our Sponsor. 10 July 2006. I'd just like to take a moment to bash Dell computers for making shitty products. The motherboard in the GF's Dell Inspiron 1000 failed after about 18 months and Dell wanted nearly $500 to fix it. Fortunately there is some small bit of justice in the world: this account of a Dell laptop exploding into flames - complete with pictures - is helping to kill Dell's already-suffering reputation, and Wired just downgraded Dell OFF of its list of the "Top 40" most Wired companies. Leave a comment. Add to Memories. Link*."[47]

Deze tekst bevat 3 'negatieve' woorden (de vol onderstreepte woorden) en 4 'positieve' woorden (de met stippellijn onderstreepte woorden). Respectievelijk 2.97% en 3.96% van de woorden zijn 'negatief' en 'positief'. Alhoewel de 'General Inquirer' de toon van deze post naar mijn mening correct inschat, - de schrijver is blijkbaar verheugd om het nieuws over de in brand geschoten laptop te horen - is het duidelijk dat deze post niet positief is over Dell.

Het is opvallend dat de 'General Inquirer' in de twee voorbeelden een aantal woorden met een duidelijke positieve of negatieve connotatie niet opmerkt. Het gaat om woorden zoals 'crap', 'bash' en 'shitty'. Informele woorden die op weblogs en internetforums nochtans vrij frequent gebruikt worden. In de laatste tekst wordt ook 'suffering' niet meegerekend omdat het niet correct in de tekst voorkomt: 'already-suffering'. Ook 'downgraded' wordt niet als negatief bestempeld. Alhoewel 'downgrade' een duidelijke negatieve connotatie heeft, maakt het geen deel uit van de categorie 'negativ' uit de 'Harvard-IV-4 dictionary'.

[47] http://sethgecko13.livejournal.com/211645.html

4.9 Mogelijke verbeteringen

4.9.1 Betrouwbaarheid input

Om tot betrouwbare resultaten te komen zouden meer inspanningen gedaan moeten worden om een betrouwbaar corpus van txt-files samen te stellen.

Een eerste punt waar aandacht aan besteed dient te worden is vermijden dat niet-relevante blogposts in het corpus belanden. Het kan zijn dat de URL die een zoekmachine retourneert niet meer naar de originele blogpost verwijst. In de plaats kan bijvoorbeeld een meer recente post, een foutmelding of een reclamesite staan. Men dient er dus voor te zorgen dat deze niet in het corpus komen. Dit kan bijvoorbeeld eenvoudigweg door te controleren of de titel, die men bij de resultaten van de zoekmachine bekomt, nog voorkomt op de site.

Een andere mogelijkheid is dat de URL nog steeds naar de juiste blogpost verwijst, maar dat deze blogpost niet relevant is. Het kan een spamblog zijn. Dergelijke spamblogs dienen verwijderd te worden. Dit kan door het verwijderen van grote bestanden en/of door het verwijderen van posts met erg veel links. Een andere mogelijkheid is dat het niet om een spamblog gaat, maar om een niet-relevante blogpost die toevallig de zoektermen bevat. Een mogelijke verbetering zou dan ook zijn om de bekomen teksten met behulp van categorization te filteren.

Een tweede aandachtspunt is om de inhoud van de relevante blogpost op een zo goed mogelijke manier te kunnen wegschrijven naar het corpus.

De methode om het relevante gedeelte uit de broncode te selecteren zou verder verfijnd kunnen worden. Bij de huidige code wordt soms nog te veel materiaal uit de broncode geselecteerd waardoor vooral metadata voor of achter het bericht worden opgehaald.

Op het internet wordt er vaak niet veel aandacht besteed aan het verzorgen van het taalgebruik. Zo worden leestekens frequent weggelaten. Nochtans wordt door tekstanalysesoftware veel waarde gehecht aan leestekens. Ik heb getracht dit

enigszins op te vangen door punten toe te voegen waar dit mogelijk was. Het is zeker mogelijk dit verder te verbeteren.

Er wordt op het internet gebruik gemaakt van verschillende tekencoderingsstandaarden zoals bijvoorbeeld ISO 8859-1 en UTF-8. Vooraleer men de inhoud van een post kan wegschrijven naar een txt-file dient men de inhoud te decoderen. Het decoderen liep in de meeste gevallen zoals het moest. Toch kan hier nog verbetering geboekt worden. Een logische verbetering zou zijn om aan de hand van de broncode de gebruikte tekencoderingsstandaard op te vragen en op basis hiervan te decoderen.

4.9.2 Volledigheid input

In hoofdstuk 4.5.2 werden de relevante gedeeltes van de broncodes geselecteerd aan de hand van begin- en eindtags. Indien er geen begin- en eindtags gevonden werden, kon de inhoud niet worden geselecteerd. Door de lijst met begin- en eindtags uit te breiden zouden er meer blogposts kunnen uitgeselecteerd worden.

Bij het selecteren van de lijst URL's in hoofdstuk 4.4 konden er per dag maximaal 100 URL's opgehaald worden. Het is mogelijk dit aantal te verhogen en alle URL's voor een bepaalde dag op te halen.

De beste manier om ervoor te zorgen dat de meeste relevante blogposts gevonden worden, is om continu te zoeken naar nieuwe bijgekomen blogposts. Dit kan door het inschrijven op een RSS-feed. De meeste tools voor de blogosfeer zijn gefocust op nieuwe blogposts.

4.9.3 General Inquirer

Een eerste mogelijkheid om de resultaten van de 'General Inquirer' te verbeteren, is het toevoegen van woorden aan de categorieën 'positiv' en 'negativ' in de 'Harvard-IV-4 dictionary'. De voorbeelden in hoofdstuk 4.8.3 illustreren al duidelijk dat vele informele woorden met een duidelijke connotatie zoals bijvoorbeeld 'crap', 'bash' en 'shitty' door de 'General Inquirer' niet herkend worden. Omdat dit soort woorden op

het internet, en dus ook op weblogs, relatief frequent gebruikt wordt, dienen deze toegevoegd te worden. Ook smileys zouden kunnen worden toegevoegd, alhoewel ze op weblogs minder vaak gebruikt worden dan bijvoorbeeld op internetforums.

Een tweede mogelijke verbetering is het schrappen van objectieve zinnen. Op dit moment worden woorden in objectieve zinnen evenzeer meegeteld. Zoals in hoofdstuk 2.2.4 reeds werd aangehaald, stijgt de nauwkeurigheid van sentiment classification indien er geen rekening wordt gehouden met objectieve zinnen.

Een derde mogelijkheid is het inbouwen van verschillende gradaties. Momenteel tellen alle woorden even sterk door. Toch heeft het woord 'excellent' een sterkere positieve connotatie dan het woord 'good'. Er bestaan reeds lexicons waar woorden op een dergelijke manier worden ingedeeld. Een voorbeeld hiervan is SentiWordNet[48].

Weblogposts gaan in tegenstelling tot reviews niet steeds over slechts één onderwerp. Indien men dus een sentiment classification wil van hetgeen er over één bepaald thema gezegd wordt, zal het nodig zijn om de blogposts te doorlopen en de relevante zinnen te selecteren.

Er wordt momenteel geen rekening gehouden met negaties. Zo zal in de woordgroep 'not safe' één positief woord geteld worden. 'Unsafe' daarentegen zal als negatief woord geteld worden. Het is echter duidelijk dat beide identiek dezelfde betekenis hebben.

Een andere mogelijke verbetering is het uitbreiden naar woordgroepen. Hierin zal in het volgende hoofdstuk verder op ingegaan worden.

[48] http://sentiwordnet.isti.cnr.it

Hoofdstuk 5: Classificatie op basis van clusters

5.1 Inleiding

De 'General Inquirer' maakt met betrekking tot woordgroepen in bepaalde gevallen de verkeerde analyse. Dit is omdat de 'General Inquirer' enkel unigrammen telt en geen rekening houdt met bijvoorbeeld een adjectief dat bij een substantief hoort en de connotatie beïnvloedt. Zo heeft 'risk' alleenstaand veeleer een negatieve connotatie maar hebben woordencombinaties als 'low risk' en 'zero risk' veeleer een positieve connotatie. Men zou de analyse kunnen uitbreiden naar bigrammen. Ook dit is niet feilloos. Zo heeft 'zero risk bias' dan weer een negatieve connotatie.

Een mogelijk betere aanpak zou zijn om eerst zinnen te analyseren en woorden en woordgroepen te selecteren die een soort geheel vormen. Dit wordt ook wel text segmentation genoemd. i.Know deelt zinnen in in clusters.

Clusters kunnen uit één of meerdere woorden bestaan. Een voorbeeld:
'*The Consumer Product Safety Commission, which announced the voluntary recall along with Apple, said the company had reported nine incidents of batteries overheating, including two that resulted in minor burns and others that caused some property damage*.'
In deze zin werden de volgende clusters geïdentificeerd: 'consumer product safety commission', 'which', 'announced', 'voluntary recall', 'along with', 'apple', 'said', 'company', 'had reported', 'nine incidents', 'of', 'batteries overheating', 'including', 'two', 'that resulted in', 'minor burns', 'and', 'others', 'that caused' en 'property damage'.

Het opzet van dit gedeelte is om een eerste verkenning te doen om te achterhalen in welke mate een sentiment classification op basis van clusters in plaats van unigrammen de resultaten positief kan beïnvloeden. In hoofdstuk 5.2 wordt de gevolgde werkwijze toegelicht. Hoofdstuk 5.3 bespreekt de resultaten. Een

vergelijking met de 'General Inquirer' wordt gemaakt in hoofdstuk 5.4. Tenslotte worden in hoofdstuk 5.5 enkele mogelijke verbeteringen besproken.

5.2 Werkwijze

Omdat het handmatig labelen van de clusters enorm tijdrovend zou zijn, werd getracht clusters te selecteren die een grote kans hebben om een positieve of negatieve connotatie te hebben. Hiervoor werd gebruik gemaakt van de reeds eerder gebruikte categorieën in de 'Harvard-IV-4 dictionary': 'Positiv' en 'Negativ'.

i.Know bezorgde mij 2 lijsten. De eerste lijst bevatte woorden en woordencombinaties die vermoedelijk een positieve connotatie hadden. Allereerst werden de txt-bestanden doorlopen om de zinnen op te delen in clusters. Vervolgens werd gezocht naar clusters die een woord bevatten uit de categorie 'positiv' in de 'Harvard-IV-4 dictionary'. Deze clusters werden dan weggeschreven naar een excel-bestand tezamen met het aantal keren dat ze voorkwamen in alle teksten. Dit resulteerde in een lijst met 27.463 woorden en woordencombinaties. Op dezelfde manier werd op basis van de categorie 'negativ' in de 'Harvard-IV-4 dictionary' een lijst bekomen met 19.873 woorden en woordencombinaties.

Een volgende stap was om de meest voorkomende clusters in te delen in drie categorieën: 'positief', 'neutraal' en 'negatief'. Zoals reeds eerder werd vermeld, kunnen clusters uit één of meerdere woorden bestaan.

De clusters die uit één woord bestonden moesten niet manueel ingedeeld worden omdat de indeling van de 'Harvard-IV-4 dictionary' overgenomen kon worden. Het excel-bestand dat bekomen werd op basis van de Harvard-categorie 'positiv' bevatte 1.199 woorden. Er waren 1.443 woorden in het bestand op basis van de Harvard-categorie 'negativ'.

Er werd gekozen om alle woordencombinaties die meer dan 3 keer voorkwamen manueel in te delen in een van de drie categorieën. Voor het bestand samengesteld aan de hand van de Harvard-categorie 'positiv', kwam dit neer op 2.537

woordencombinaties. Hiervan werden er 142 als 'negatief' ingevuld, 1.207 als 'neutraal' en 1.188 als 'positief'. Bij de lijst samengesteld op basis van de Harvard-categorie 'negativ' werden 1.414 woordencombinaties geselecteerd. 690 werden ingedeeld als 'negatief', 579 als 'neutraal' en 145 als 'positief'.

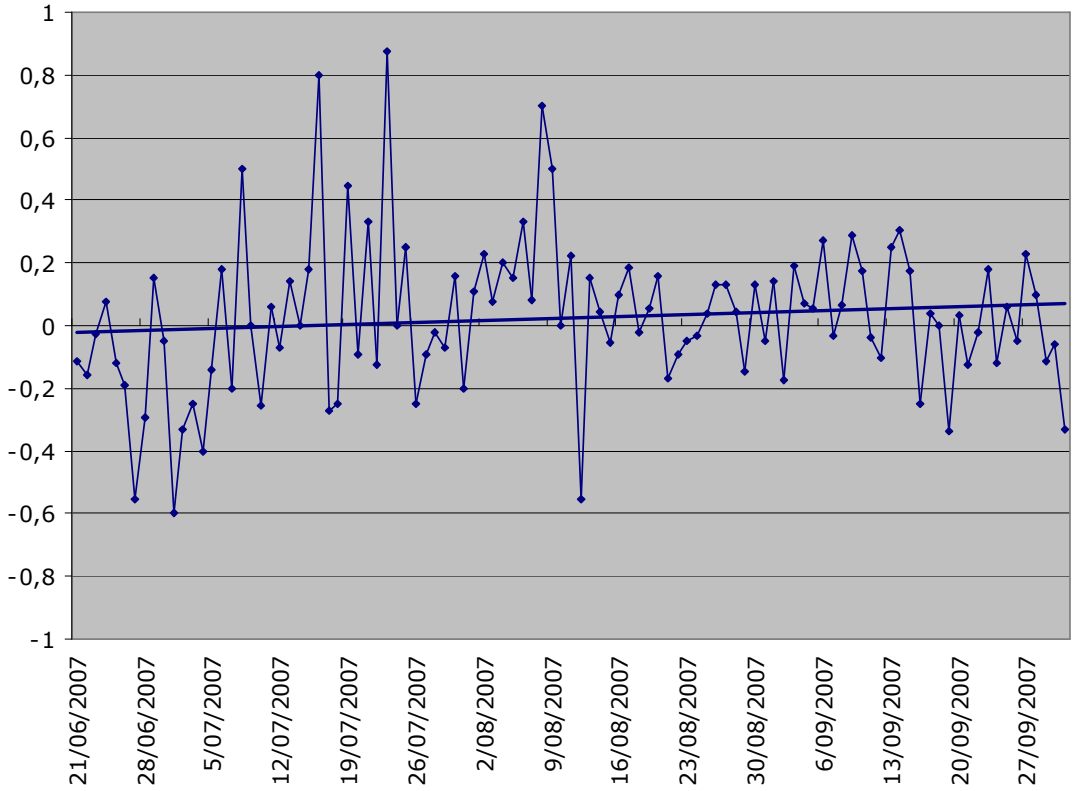
De twee bestanden werden samengevoegd en de duplicaten werden verwijderd. Dit resulteerde in een lijst met 6.309 clusters waarbij een connotatie werd aangegeven. 2.401 clusters waren gelabeld als 'positief', 1.810 als neutraal en 2.098 als 'negatief'.

5.3 Resultaten

In totaal bevatten de 3.440 blogposts 858.188 clusters. Van deze clusters waren 12.323 clusters (1,44%) 'positief', 10.205 'negatief' (1,19%) en 14.235 'neutraal' (1,66%). De meest voorkomende clusters waaraan een connotatie was gegeven waren enkelvoudige woorden. Bij de positieve clusters waren dit: 'well' (731 keer), 'home' (430 keer) en 'kind' (270 keer). Bij de negatieve clusters kwamen 'problem' (536 keer), 'cost' (193 keer) en 'ill' (150 keer) het vaakst voor. De meest voorkomende neutrale clusters waren: 'fire' (1.632 keer), 'company' (863 keer) en 'right' (435 keer). Een meer uitgebreide is terug te vinden in Bijlage 13.

Net als bij de werkwijze bij de 'General Inquirer' werden de blogposts in drie categorieën ingedeeld: 'positief', 'negatief' en 'neutraal'. Blogposts met meer positieve clusters dan negatieve clusters werden als 'positief' gezien. Posts met minder positieve clusters dan negatieve clusters werden 'negatief' bestempeld. Bij een gelijk aantal werden ze als 'neutraal' gezien. Er waren evenveel posts 'positief' als 'negatief', namelijk 1.199 (34,96%). De overige 1.032 posts (30,09%) waren 'neutraal'.

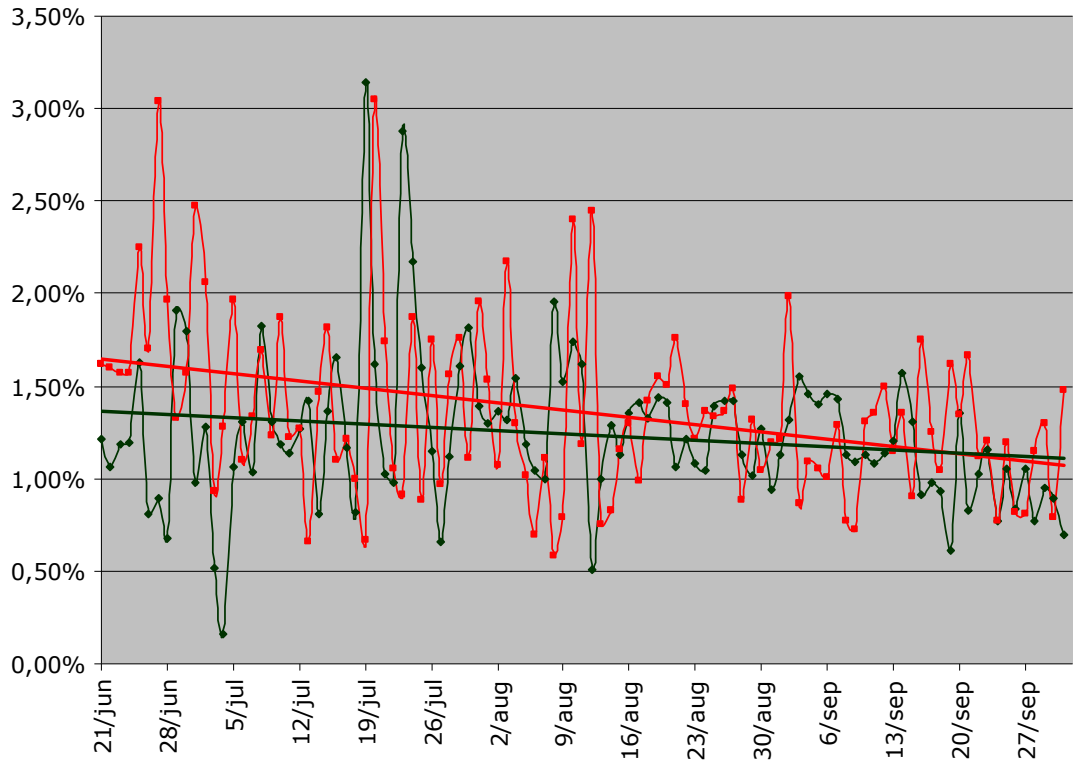
Ook hier werden aan de categorieën 'positief', 'negatief' en 'neutraal' respectievelijk de waardes 1, -1 en 0 gegeven. Voor elke dag werd vervolgens het gemiddelde van deze waardes berekend. Het resultaat hiervan wordt weergegeven in Figuur 14.



Figuur 14: Gemiddelde waarde per dag

Er is een lichte opwaartse trend te zien. Deze trend is echter een stuk minder sterk dan degene die bij de analyse van de 'General Inquirer' behaald werd. In de beginperiode zijn de posts gemiddeld negatiever. Later blijft de toon relatief stabiel. Na de aankondiging van de terugroeping op 16 augustus lijken er minder uitschieters te zijn.

In Figuur 15 wordt het gemiddeld percentage positieve clusters en het gemiddeld percentage negatieve clusters weergegeven. De gemiddelde score per dag wordt berekend als het gemiddelde van het percentage clusters per post zodat lange posts niet te zwaar doorwegen.



Figuur 15: Gemiddeld % positieve clusters (groen) en gemiddeld % negatieve clusters (rood)

Zowel het gemiddeld percentage positieve clusters als het gemiddeld percentage negatieve clusters daalt. Dit kan er op duiden dat de posts gemiddeld genomen gematigder worden in toon.

Wat opviel was dat er weinig clusters waren die als 'positief' of 'negatief' werden herkend. Zo waren er 663 berichten waarin geen enkele positieve of negatieve cluster werd gevonden. Bij 567 posts was er slechts één cluster met een niet-neutrale indeling. Voor 476 blogposts bedroeg dit aantal 2. Er waren slechts 437 blogposts – of 7,85% - waarin meer dan 10 positieve en negatieve clusters werden geïdentificeerd.

De belangrijkste reden hiervoor is dat woorden enkel in hun grondvorm werden uitgeselecteerd. Meervoudsvormen of vervoegde werkwoorden werden niet herkend.

Clusters met een woord uit de 'Harvard-IV-4 dictionary' categorieën 'negativ' en 'positiv' dat zich niet in de grondvorm bevond, werden dus niet opgenomen. Zo werd de cluster 'explosion' wel als 'negatief' meegeteld maar de cluster 'explosions' niet.

Een andere reden is dat woordgroepen die minder dan twee keer voorkwamen niet handmatig werden ingedeeld.

5.4 Vergelijking resultaten General Inquirer en i.Know

5.4.1 Vergelijking woordencombinaties

De werkwijze om woordgroepen te behandelen als 'positief', 'negatief' of 'neutraal' overeenkomstig de som van de losse woorden waaruit de woordgroep is samengesteld, heeft tekortkomingen die moeilijk opgelost kunnen worden. Woordgroepen zijn in semantisch opzicht namelijk niet zomaar eenvoudigweg de som van losse woorden. Zo zal de 'General Inquirer' in de woordgroep 'bad experience' één negatief woord en één positief woord vinden. Deze woordgroep zal in zijn geheel als 'neutraal' gezien worden bij de gebruikte scoringsmethode. Maar de score 'negatief' zou hier beter op zijn plaats zijn.

Een betere indeling van woordgroepen zal logischerwijze leiden tot een hogere accuraatheid. Daarom wordt er een overzicht gegeven van de verschillen tussen de manuele indeling van de woordgroepen en de indeling door de 'General Inquirer'. Bij de eerste werkwijze worden de woordgroepen op basis van de som van de losse woorden als 'positief', 'negatief' of 'neutraal' gecategoriseerd. Bij de tweede werkwijze worden de woordgroepen handmatig als 'positief', 'negatief' of 'neutraal' gecategoriseerd. De mate waarin deze resultaten verschillen geeft tevens een indicatie van de mate waarin de methode op basis van woordclusters tot betere resultaten zou kunnen leiden.

Opgemerkt dient te worden dat een verschil in score behaald door de twee methodes niet automatisch betekent dat de handmatige indeling beter is. Het manueel indelen in drie categorieën was niet steeds even eenduidig. Indien verschillende mensen de lijsten zouden invullen, zouden er ongetwijfeld ook verschillen zijn.

Zoals reeds vermeld, werden van de 2.537 clusters met minstens één woord uit de 'Harvard-IV-4 dictionary' categorie 'positiv' er 142 als 'negatief' ingevuld, 1.207 als 'neutraal' en 1.188 als 'positief'. Van de op 1.414 woordencombinaties op basis van de Harvard-categorie 'negativ' werden er 690 ingedeeld als 'negatief', 579 als 'neutraal' en 145 als 'positief'.

Deze getallen zijn op het eerste gezicht erg frappant. Nochtans is het niet correct te stellen dat er bij deze lijsten slechts 1.878 (1.188 en 690) van de 3.951 woordencombinaties – dit is 47,5% - bij de twee werkwijzen hetzelfde worden ingedeeld. Dat een woordencombinatie voorkomt in de lijst met mogelijk negatieve of positieve woordencombinaties wil enkel zeggen dat minstens één van de woorden in de woordencombinatie voorkomt in de categorie 'positiv' of de categorie 'negativ' van de 'Harvard-IV-4 dictionary'. De 'General Inquirer' houdt rekening met de context waarin de woorden voorkomen, zodat het mogelijk is dat de woordencombinatie door de 'General Inquirer' anders wordt ingedeeld. Een voorbeeld zal dit verduidelijken. 'Left hand' komt voor in de lijst met mogelijke negatieve woordencombinaties, omdat 'hand' in de categorie 'negativ' van de 'Harvard-IV-4 dictionary' voorkomt. In de context van 'out of hand' heeft 'hand' namelijk een negatieve connotatie. In de woordencombinatie 'left hand' wordt 'hand' door de 'General Inquirer' als neutraal gezien en wordt de gehele woordencombinatie dus als neutraal bestempeld. Vermits de indeling van de woordencombinatie door de 'General Inquirer' afhankelijk kan zijn van de context, is het niet zo makkelijk om voor alle woordencombinaties volledig juist na te gaan hoe ze zouden worden ingedeeld door de 'General Inquirer'.

Een vergelijking maken wordt dan ook al een stuk zinvoller indien de handmatige indeling wordt vergeleken met de indeling van de volledige woordencombinaties door de 'General Inquirer'. De woordencombinaties uit de twee lijsten werden hiervoor ingelezen in een database en met behulp van de code in Bijlage 12 werden ze door de 'General Inquirer' gehaald. De resultaten hiervan zijn terug te vinden in Tabel 6 en Tabel 7 in Bijlage 14 onder 'General Inquirer'. In totaal werden 47.336 woorden en woordencombinaties geanalyseerd. De door mij toegepaste scoringsmethode zou

er toe geleid hebben dat van de woordgroepen er 21.605 als 'positief' zouden worden ingedeeld, 12.924 als 'neutraal' en 12.807 als 'negatief'.

Van de 1.333 woordencombinaties die handmatig ingedeeld werden als 'positief' beoordeelde de 'General Inquirer' er 1.100 (82,52%) als 'positief'. 169 woordencombinaties (12,68%) analyseerden ze als 'neutraal'. Het betrof bijvoorbeeld woordencombinaties waarvan alle woorden als neutraal werden ingeschat zoals 'right choice' en 'quality products' ofwel woordencombinaties die evenveel woorden bevatten die 'negatief' en 'positief' werden ingeschat zoals bijvoorbeeld 'cost efficiency'. De overige 64 woordencombinaties (4,80%) werden door de 'General Inquirer' als 'negatief' ingedeeld. De woordencombinaties 'low risk' en 'nothing bad' zijn hiervan voorbeelden.

Bij de 1.786 woordencombinaties die manueel werden ingedeeld als 'neutraal' waren de verschillen met de indeling van de 'General Inquirer' groter. Slechts 764 woordencombinaties (42,78%) werden ook door de 'General Inquirer' als 'neutraal' beschouwd. Van de overige combinaties werden 809 (45,30%) ingedeeld als 'positief'. Het ging daarbij om woorden die volgens mij geen duidelijke connotatie hadden zoals bijvoorbeeld 'consumer product safety commission' en 'basic facts'. 213 woordencombinaties (11,93%) werden door de 'General Inquirer' als 'negatief' ingedeeld. Het betrof woorden zoals bijvoorbeeld 'average cost'.

In totaal werden 832 woordencombinaties handmatig als 'negatief' bestempeld. Het grootste deel van deze woordencombinaties, namelijk 512 combinaties (61,54%) werden gelijkaardig door de 'General Inquirer' ingedeeld. 58 woordencombinaties (6,97%) waaronder bijvoorbeeld 'less responsive' en 'major setback' werden door de 'General Inquirer' als 'positief' ingedeeld. De overige 262 (31,49%) als 'neutraal'. Een voorbeeld hier is 'serious injuries' waar 'serious' door de 'General Inquirer' als 'positief' wordt gezien en 'injuries' als 'negatief'.

In totaal werd zo'n 60,14% van de 3.951 handmatig ingevulde woordencombinaties hetzelfde ingedeeld door de 'General Inquirer'. In Tabel 6 en Tabel 7 in Bijlage 14 zijn meer uitgebreide cijfers terug te vinden.

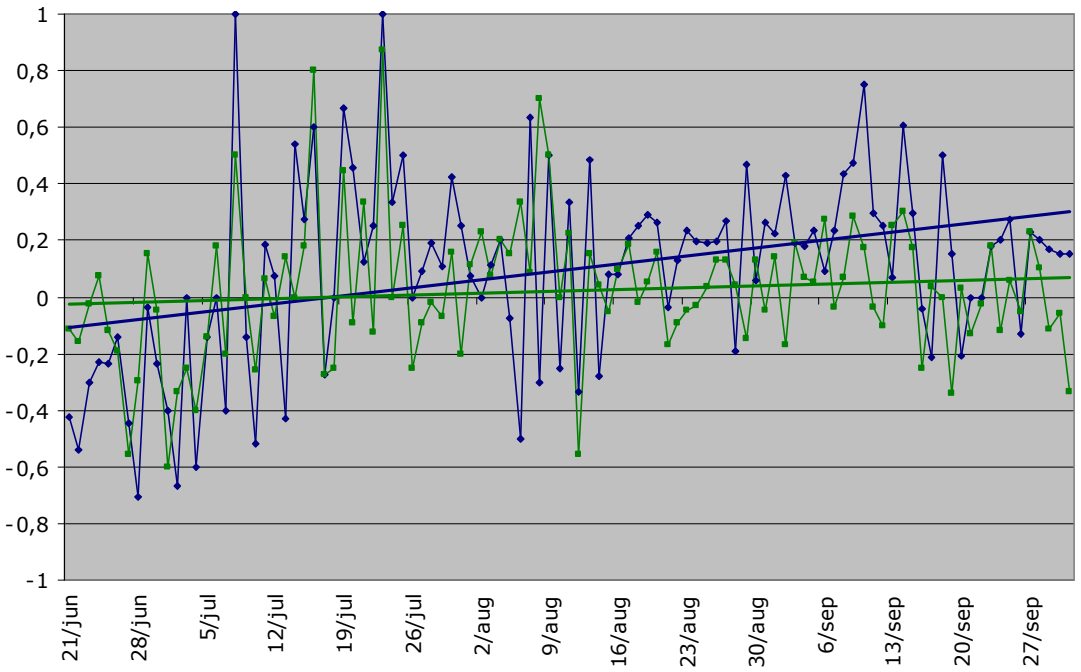
5.4.2 Vergelijking resultaten

Tabel 5 vergelijkt de indeling van de blogposts op basis van clusters met de indeling van de 'General Inquirer'. Hieruit wordt duidelijk dat slechts 55,64% hetzelfde wordt ingedeeld. Het grootste verschil is dat er bij de indeling op basis van clusters een groter aantal posts als 'neutraal' wordt ingedeeld. Dit is waarschijnlijk te wijten aan het relatief klein aantal clusters dat in de teksten herkend wordt. Hierdoor is de kans dat een post als 'neutraal' wordt ingedeeld groter. Ook is te zien dat zo'n 15,60% van de posts volledig tegengesteld wordt ingedeeld.

Tabel 5: Vergelijking indeling posts

Indeling mbv clusters	Indeling General Inquirer	Aantal
Negatief	Negatief	753
Negatief	Neutraal	102
Negatief	Positief	337
Neutraal	Negatief	430
Neutraal	Neutraal	203
Neutraal	Positief	387
Positief	Negatief	194
Positief	Neutraal	60
Positief	Positief	938
	Totaal:	3.404

In Figuur 16 worden de gemiddelde scores van de classificatie door de 'General Inquirer' en de classificatie aan de hand van clusters vergeleken. Wat opvalt is dat de classificatie op basis van clusters over het algemeen gematigder is. De stijging is een stuk minder vergeleken met de stijging van de gemiddelde score van de 'General Inquirer'. De verklaring hiervoor is mogelijk het groot aantal posts met erg weinig 'positieve' en 'negatieve' clusters. Hierdoor worden 1.032 blogposts als 'neutraal' bestempeld. Bij de indeling op basis van de 'General Inquirer' waren er slechts 365 blogposts 'neutraal'.

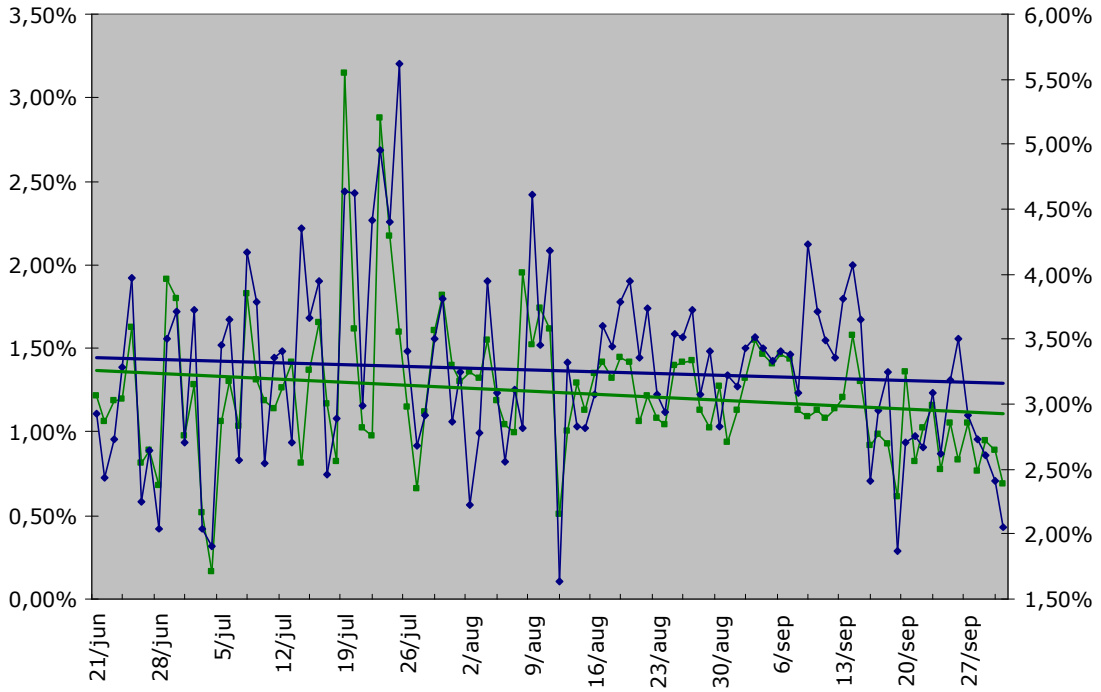


Figuur 16: gemiddelde score 'General Inquirer' (blauw) en op basis van clusters(groen)

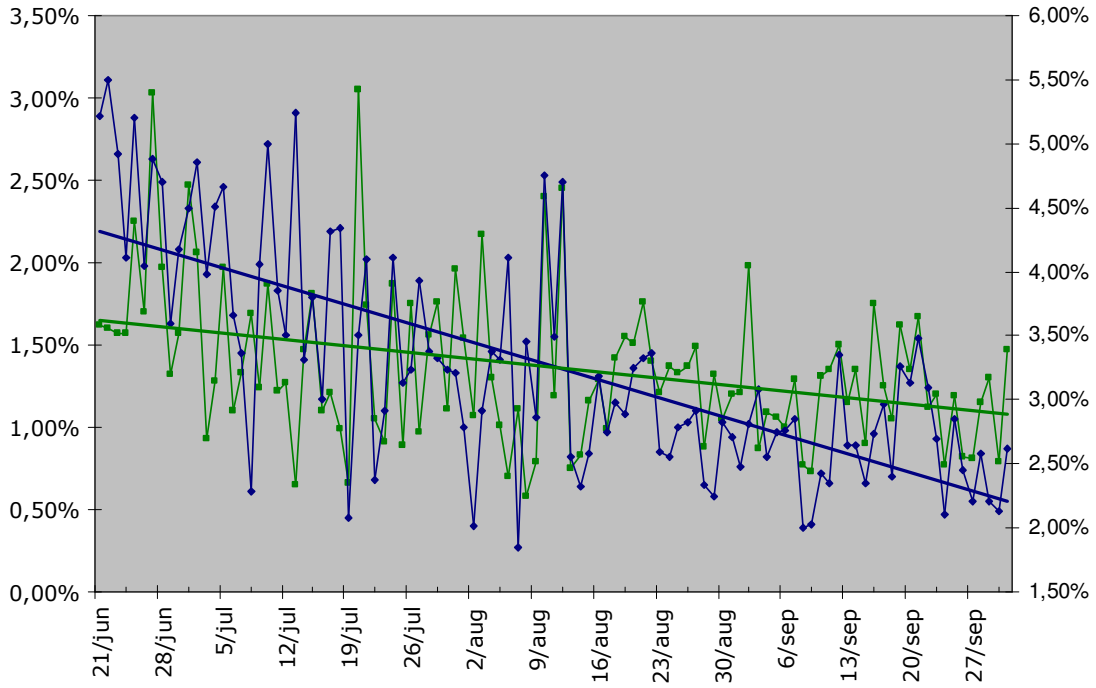
In Figuur 17 en Figuur 18 wordt respectievelijk het percentage positieve woorden/clusters en het percentage negatieve woorden/clusters voor beide methoden vergeleken. Het percentage positieve woorden en clusters komt redelijk goed overeen, al dient er rekening gehouden te worden met de verschillende schaal van de assen. Het percentage negatieve woorden en clusters komt minder overeen. Het is te zien dat het percentage negatieve woorden een stuk sterker daalt dan het percentage negatieve clusters.

In de eerste dagen is er bijvoorbeeld een redelijk groot verschil te merken. Bij de indeling volgens de 'General Inquirer' is het percentage negatieve woorden tijdens deze periode erg hoog. Zoals reeds eerder aangehaald onder hoofdstuk 4.8.3, is de verklaring hiervoor volgens mij te vinden in het artikel van de 'Inquirer' dat de zaak aan het licht bracht. Het artikel wordt in de eerste paar dagen na het verschijnen door verschillende bloggers in zijn geheel of gedeeltelijk overgenomen. Het artikel

bevat volgens de indeling van de 'General Inquirer' 4,68% negatieve woorden. De methode op basis van clusters pikt in dit artikel verscheidene negatieve woorden niet op. Verschillende negatieve woorden zijn niet in hun grondvorm. Het gaat om woorden zoals 'exploded', 'explosions' en 'explodes'. Dit kan een mogelijke verklaring voor het verschil zijn.



Figuur 17: % positieve woorden 'General Inquirer' (blauw – rechtse as) en op basis van clusters (groen – linkse as)



Figuur 18: % negatieve woorden 'General Inquirer' (blauw – rechtse as) en op basis van clusters (groen – linkse as)

Om de verschillen tussen de twee methodes verder te verduidelijken zullen hieronder aan de hand van een voorbeeld[49] enkele verschillen tussen beide systemen bondig besproken worden.

Indeling door de 'General Inquirer' (positief, negatief):
'Apple recalls laptop batteries. 24 August 2006. Sony has just had a bad month. After the Dell fiasco where thousands of laptop batteries were recalled, the company is dealing with a similar experience with Apple who is recalling 1.8 million batteries. Users of the iBook G4 and PowerBook G4 laptop models sold between October 2003 and August 2006 are being asked to return their laptop batteries as they may overheat and catch on fire. According to the Consumer Product Safety Commission: "These lithium ion batteries can overheat, posing a fire hazard to consumers." Apple

[49] http://computerconsultingblog.blogspot.com/2006/08/apple-recalls-laptop-batteries.html

spokesman, Steve Dowling, said, "We discovered that some Sony batteries in previous models of Power PC-based PowerBooks and iBooks do not meet Apple's standards for safety and performance. None of Apple's Intel-based laptops are affected." Learn more in the full article (link above). Added by Computer Consulting Kit.'

Indeling op basis van clusters (positief, neutraal, negatief):
'*Apple recalls laptop batteries. 24 August 2006. Sony has just had a bad month. After the Dell fiasco where thousands of laptop batteries were recalled, the company is dealing with a similar experience with Apple who is recalling 1.8 million batteries. Users of the iBook G4 and PowerBook G4 laptop models sold between October 2003 and August 2006 are being asked to return their laptop batteries as they may overheat and catch on fire. According to the Consumer Product Safety Commission: "These lithium ion batteries can overheat, posing a fire hazard to consumers." Apple spokesman, Steve Dowling, said, "We discovered that some Sony batteries in previous models of Power PC-based PowerBooks and iBooks do not meet Apple's standards for safety and performance. None of Apple's Intel-based laptops are affected." Learn more in the full article (link above). Added by Computer Consulting Kit.'*

Verschillende zaken vallen op. De indeling op basis van clusters pikt een aantal woorden zoals 'bad' en 'fiasco' met een duidelijke connotatie niet op. Dit komt omdat deze woorden deel uitmaken van een cluster die niet is ingedeeld in een categorie. De clusters 'dell fiasco' en 'bad month' komen enkel in deze blogpost voor en werden dus niet benoemd.

De cluster 'Consumer Product Safety Commission' werd manueel ingedeeld als 'neutraal'. Bij de 'General Inquirer' worden echter twee woorden als 'positief' geïdentificeerd: 'safety' en 'commission'. Het is volgens mij duidelijk dat dit niet correct is.

Het valt ook op dat 'return' door beide methodes verschillend wordt ingedeeld. Dit komt doordat de 'General Inquirer' rekening houdt met de context. Zo wordt bij het

werkwoord ‘return’ onderscheid gemaakt tussen enkele betekenissen. Indien het ‘teruggaan’ betekent, wordt het niet als positief gezien. In de betekenis van ‘iets teruggeven’ wordt het echter wel als positief ingedeeld. Bij de methode waar ingedeeld wordt op basis van clusters wordt slechts gewerkt met één betekenis en wordt er niet gekeken naar de context.

Het lijkt mij duidelijk dat van de twee gehanteerde methodes de sentiment classification op basis van de ‘General Inquirer’ het meest betrouwbaar is. Bij de methode op basis van clusters worden in de teksten te weinig positieve en/of negatieve woorden of woordencombinaties herkend.

5.5 Mogelijke verbeteringen

De hier gehanteerde sentiment classification op basis van clusters heeft nog ernstige tekortkomingen.

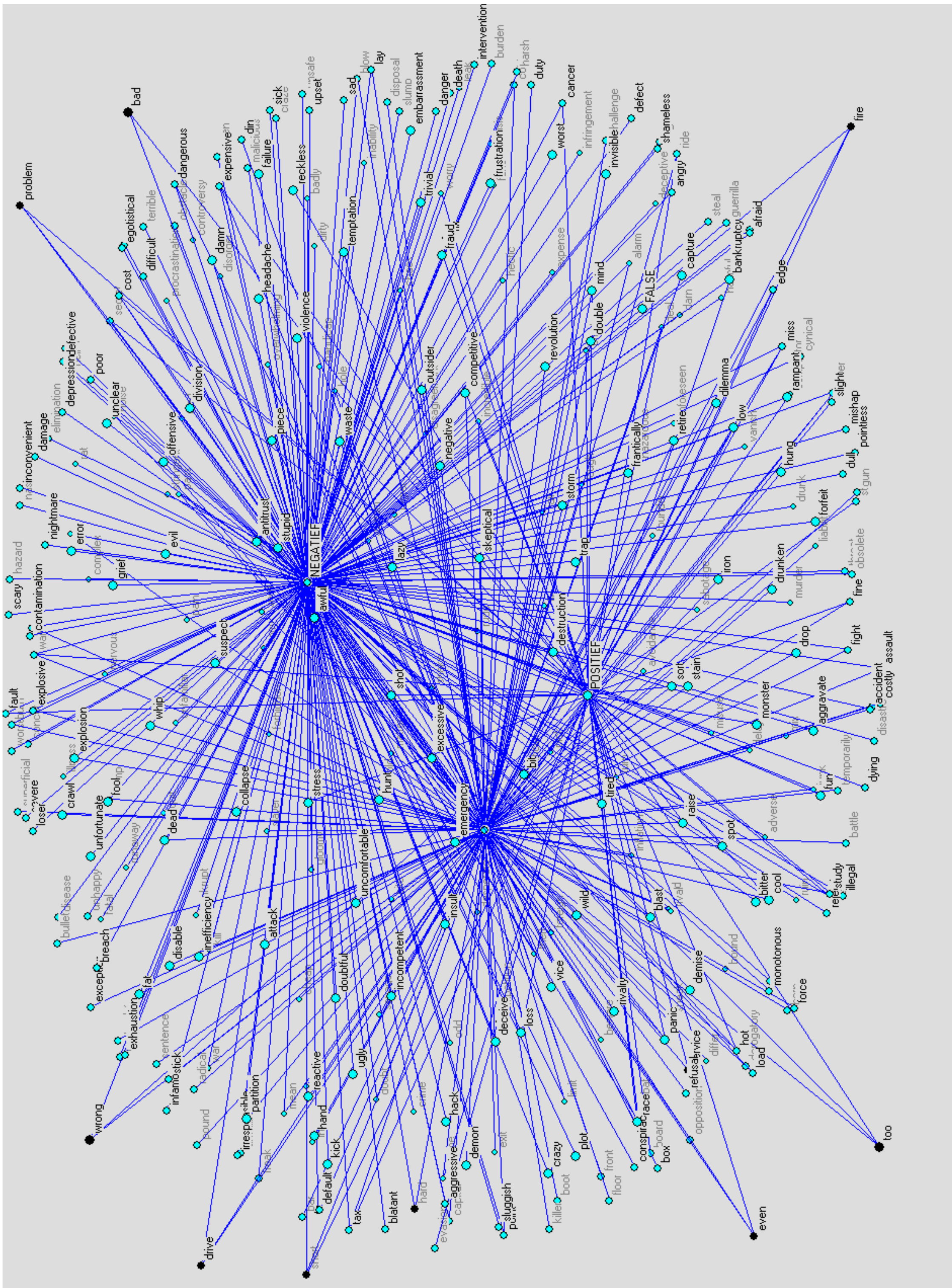
De belangrijkste tekortkoming op dit moment is het feit dat er te weinig positieve en/of negatieve clusters gevonden worden in de teksten. De vermoedelijke hoofdreden hiervoor is dat de lijsten met potentiële positieve of negatieve clusters enkel worden samengesteld op basis van de grondvorm van de woorden in de categorieën ‘positiv’ en ‘negativ’ in de ‘Harvard-IV-4 dictionary’.

Een ander zwak punt is dat er geen rekening wordt gehouden met de context waarin de clusters voorkomen. Indien men met deze methode betere resultaten wil halen dan de resultaten behaald door de ‘General Inquirer’, zal dit volgens mij ingebouwd moeten worden. Dit omdat de frequentst voorkomende clusters enkelvoudige woorden zijn. Bij deze enkelvoudige woorden houdt de ‘General Inquirer’ wél rekening met de context.

Een ander nadeel van een methode die werkt op basis van clusters is de tijd die gespendeerd moet worden aan het manueel labelen van de woordgroepen. De werkbaarheid van de methode hangt volgens mij dan ook af van de mate waarin dit geautomatiseerd kan worden. Indien een woordgroep bijvoorbeeld bestaat uit een

adjectief en een substantief kunnen bepaalde woordgroepen automatisch ingedeeld worden. Sommige woorden zijn namelijk contextonafhankelijk. Het gaat hier bijvoorbeeld over adjectieven zoals 'good' en 'excellent'. Wanneer er reeds een groot aantal woordgroepen zijn ingedeeld, kunnen er automatisch verbanden gelegd worden tussen woorden die op eenzelfde of net omgekeerde manier ingedeeld worden.

Bij de gehanteerde werkwijze kan bijvoorbeeld gekeken worden bij welke unigrammen de connotatie steeds hetzelfde blijft. In Figuur 19 wordt dit visueel weergegeven. Een unigram wordt verbonden met een categorie, 'positief', 'neutraal' of 'negatief', indien er een cluster met het unigram in die categorie werd ingedeeld. Op deze manier kan men contextonafhankelijke woorden automatisch herkennen. Deze woorden zullen slechts met één categorie verbonden worden. Hiervoor zou wel onder andere rekening gehouden moeten worden met zowel negatie als de woordsoort. Zo wordt 'bad' in Figuur 19 verbonden met zowel 'positief' als 'negatief' alhoewel het een contextonafhankelijk woord is. De woordgroep 'nothing bad' werd echter ingedeeld als 'positief. '



Figuur 19: Visuele weergave van labeling unigrammen in clusters (bron: i.Know)

Hoofdstuk 6: Conclusies

Uit de cijfers gepresenteerd in hoofdstuk 3.1.2 blijkt dat het aantal weblogs de laatste jaren explosief is toegenomen. Ook het aantal posts per dag is fel gestegen, al lijkt dit aantal het afgelopen jaar te stagneren. Het is op dit ogenblik moeilijk te voorspellen of deze trends zich in de toekomst gaan verder zetten. Maar onafgezien daarvan, lijkt het mij duidelijk dat de blogosfeer groot genoeg is om ermee rekening te houden.

Het aantal bedrijven dat gebruik maakt van een corporate weblog is op dit ogenblik nogal beperkt. Buiten de verscheidene opportuniteiten die het oprichten van een corporate weblog met zich meebrengt, zijn er ook verschillende onzekerheden. De belangrijkste reden dat corporate blogs niet doorbreken is waarschijnlijk dan ook het feit dat bedrijven niet graag de controle over de bedrijfscommunicatie verliezen. Een corporate weblog kan namelijk pas een succes worden, wanneer men interactie met de bezoekers aanmoedigt door te antwoorden op reacties en door in te spelen op relevante posts van andere bloggers.

De invloed die de blogosfeer op de publieke opinie heeft, is vooralsnog moeilijk in te schatten. Verschillende onderzoeken kwamen tot tegengestelde conclusies. Dat het internet de aankoopbeslissingen kan beïnvloeden is echter al langer geweten.

De blogosfeer als informatiebron gebruiken heeft een aantal voordelen ten opzichte van andere informatiebronnen. Een bericht dat op een weblog geplaatst wordt, is onmiddellijk door iedereen op te vragen. In tegenstelling tot bijvoorbeeld peilingen, kunnen in de blogosfeer de reacties op een bepaalde gebeurtenis erg snel worden ingeschat. De blogosfeer is semi-gestructureerd, waardoor discussies met behulp van links, trackbacks en pings gevolgd of gereconstrueerd kunnen worden. De waarde die aan een gebeurtenis gehecht wordt, kan worden ingeschat door het aantal malen dat erover geschreven wordt. Op dezelfde wijze kunnen trends in een vroeg stadium opgemerkt worden.

De invloed van individuele weblogs verschilt sterk. De meeste weblogs worden slechts door weinig personen gelezen. Enkele websites hebben een erg groot aantal lezers. Het kan daarom nuttig zijn om de invloed van bloggers op de markt waar men actief is in kaart te brengen. Dit kan door web structure mining toe te passen en aan de hand van het aantal links naar een bepaalde weblog de invloed in te schatten. Aan de hand hiervan kan men gerichter monitoren, communiceren met bloggers en adverteren.

Bedrijven kunnen de blogosfeer gebruiken om de 'word of mouth' over henzelf en de concurrentie in te schatten. Zo kan er gekeken worden waarover het meeste geklaagd wordt. Men dient er wel steeds rekening mee te houden dat de blogosfeer geen representatieve doorsnede is van de bevolking.

Het automatisch ontsluiten van kennis uit de blogosfeer heeft raakvlakken bij zowel text mining als web mining. Er kan van verschillende text mining technieken gebruik gemaakt worden. De belangrijkste zijn categorizaton, topic tracking, sentiment classification en opinion analysis. Vermits user-generated content vaak meningen bevat, is het erg geschikt om sentiment classification en opinion analysis op toe te passen.

Bij sentiment classification worden de blogposts ingedeeld naargelang de stemming. De methode heeft echter een aantal nadelen. Omdat blogposts vaak erg kort zijn, is het bekomen van een correcte indeling moeilijker dan bij langere teksten. Zelfs indien de stemming in blogposts correct kan ingeschat worden, is het niet steeds makkelijk om dit correct te interpreteren. Indien er bijvoorbeeld een sterk negatieve stemming wordt gevonden in een bepaalde post, zegt dit niet waarover deze negatieve stemming gaat. De betrouwbaarheid van sentiment classification is op dit ogenblik reeds behoorlijk.

Bij opinion analysis wordt ook gekeken waarover de uitgesproken mening gaat. Eventueel wordt ook in kaart gebracht wie deze mening uit. De mogelijkheden die opinion analysis biedt zijn dan ook een stuk groter dan die bij sentiment

classification. Aangezien opinion analysis een stuk moeilijker te realiseren is, ligt de bereikte nauwkeurigheid lager.

Uit de gevalsstudie werd duidelijk dat sentiment classification niet zo gemakkelijk uit te voeren is. Er zijn verschillende problemen waarmee men geconfronteerd wordt.

Om een betrouwbaar corpus samen te stellen, moeten enkele moeilijkheden overwonnen worden. Allereerst moet men de relevante blogposts identificeren. Enkel werken op basis van zoektermen leidt ertoe dat ook niet-relevante posts worden opgenomen. Vervolgens moet de inhoud correct worden geselecteerd. Verder moet er bij het toepassen van sentiment classification rekening gehouden worden met de kenmerken van de blogosfeer. De manier van schrijven is, in vergelijking met zakelijke teksten, vaak onverzorgd en er wordt meer gebruikt gemaakt van informele woorden.

Op het bekomen corpus werd een sentiment classification uitgevoerd aan de hand van de 'General Inquirer'. Alhoewel ik vond dat de indeling over het algemeen redelijk accuraat was, waren er toch enkele punten waarop verbetering mogelijk is. Zo werd geen rekening gehouden met negatie. Verder werd de sterkte van de semantische oriëntering niet meegerekend. Subjectieve zinnen werden ook niet geschrapt. Omdat een blogpost over meerdere onderwerpen kan gaan, zouden eigenlijk ook enkel relevante zinnen mogen worden meegerekend. Als laatste zou ook de globale indeling van woordgroepen kunnen verbeterd worden.

Om te kijken op welke manier dit laatste punt verbeterd zou kunnen worden, werd een sentiment classification op basis van clusters uitgevoerd. De gehanteerde methode had echter teveel tekortkomingen om hier duidelijke conclusies uit te trekken. Een vergelijking van de woordgroepen bracht wel aan het licht dat er een relatief groot verschil was tussen de beide indelingen. Omdat dit een aanwijzing is dat de methode op basis van clusters potentieel tot betere resultaten kan leiden, loont het volgens mij de moeite om deze methode verder te onderzoeken.

Lijst van de geraadpleegde werken

Aeserud, K., 'Bonding by blogging', *Profit*, 25 (2006) nr.4, p119-119

Armstrong, S., 'Bloggers for hire', *New Statesman* 135 (2006) nr. 4807, p26-27

Aschenbrenner, A. en S. Miksch, 'blog mining in a corporate environment', Smart Agent Technologies, september 2005, http://ieg.ifs.tuwien.ac.at/techreports/Asgaard-TR-2005-11.pdf

Attardi, G. en M. Simi, 'Blog Mining Through Opinionated Words', In *Proceedings of The Fifteenth Text Retrieval Conference* (TREC 2006)
Ook online beschikbaar: http://trec.nist.gov/pubs/trec15/papers/upisa.blog.final.pdf

Balog, K., G. Mishne en M. de Rijke, 'Why Are They Excited? Identifying and Explaining Spikes in Blog Mood Levels', In: *11th Meeting of the European Chapter of the Association for Computational Linguistics* (EACL 2006), April 2006
Ook online beschikbaar: http://staff.science.uva.nl/~mdr/Publications/Files/eacl2006-moodsignals.pdf

Bartlett, M., 'The New Word Of Mouth', *Credit Union Journal* 10 (2006) nr. 45, p24-24

Borkard, V., K. Deshmukh en S. Sarawagi, 'Automatic segmentation of text into structured records', *Proceedings of the 2001 ACM SIGMOD international conference on Management of data* (2001), Santa Barbara, California, United States, p175-186
Ook online beschikbaar: www.it.iitb.ac.in/~sunita/papers/sigmod01.pdf

Carr, N., 'Lessons in Corporate Blogging', *Business Week Online*, 18 juli 2006, p9-9

Chau, M. en J. Xu, 'Mining communities and their relationships in blogs: A study of online hate groups', *International Journal of Human Computer Studies* 65 (2007) nr. 1, p57-70

Columbus, L., 'Blog Mining gets real', *CRM Buyer.com* (2005), http://www.crmbuyer .com/story/43483.html

Communications of the ACM, 'Blog-Free CEOs' 49 (2006) nr. 10, p10-10

Communications of the ACM, 'Blog Profiles' 48 (2005) nr. 11, p9-10

Cooke, M., 'The importance of blogging', *International Journal of Market Research*, 48 (2006) nr. 6, p645-646

Croner, 'Bloggy Hell', 25 mei 2007, http://www.croner.co.uk/croner/jsp/Editorial.do ?cache=true&contentId=714662

Dave, K., S. Lawrence en M. Pennock, 'Mining the peanut gallery: opinion extraction and semantic classification of product reviews', *In Proceedings of the Twelfth International World Wide Web Conference* (WWW2003 2003)
Ook online beschikbaar: www.kushaldave.com/p451-dave.pdf

De Standaard, 'Belgische bedrijven sceptisch over bloggen', 29 december 2006

De Standaard, 'Corporate blogging komt niet van de grond', 16 maart 2007

DM Review, 'The Problem with Unstructured Data', februari 2003, http://www. dmreview.com /article_sub.cfm?articleId=6287

Donato, M., Blogs: 'Marketing Evolves Online', *Sales & Marketing Management* 158 (2006) nr. 7, p23-23

Du, H. en C. Wagner, 'Weblog success: Exploring the role of technology', *International Journal of Human Computer Studies* 64 (2006) nr. 9, p789-798

Dwyer, P., 'Measuring the value of electronic word of mouth and its impact in consumer communities', *Journal of Interactive Marketing* 21 (2007) nr. 2, p63-79

Edelman, 'A Corporate Guide to the Blogosphere: The new model of peer-to-peer communications', januari 2007, http://www.edelman.com/image/insights/content/ WhitePaper011107sm.pdf

Enright, A., 'Dell learns power of the blog', *Marketing News* 40 (2006) nr. 20, p17-18

Esuli, A., 'Opinion Mining', *Language and Intelligence Reading Group*, Pisa, Italy, 14juni 2006, http://medialab.di.unipi.it/web/Language+Intelligence/OpinionMining06 -06.pdf

Esuli, A. en F. Sebastiani, 'Determining term subjectivity and term orientation for opinion mining', *Proceedings of EACL-06, 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, IT, 2006, p193-200
Ook online beschikbaar: http://nmis.isti.cnr.it/sebastiani/ Publications/EACL06.pdf

Fan, W. e.a., 'Tapping the power of text mining', *Communications of the ACM* 49 (2006) nr. 9, p77-82

Furukawa, T. e.a., 'Analysis of user relations and reading activity in weblogs', *Electronics and Communications in Japan (Part I: Communications)* 89 (2006) nr. 89, p88-96

Galitsky, B. en B. Kovalerchuk, 'Mining the blogosphere for contributors' sentiments', In *Proceedings of AAAI 2006 Spring Symposium on Computational Approaches to Analyzing Weblogs* (AAAI-CAAW 2006), maart 2006

Ghahremani, Y., 'The joy of text', *CFO* 22 (2006) nr. 1, p79-81

Goldie, L., 'Blogs influence customer shopping', *New Media Age*, 23 november 2006, p11-11

Gordon, S., 'Rise of the blog', *IEE Review* 52 (2006) nr. 3, p32-35

Gregg, D. En S. Walczak, 'Adaptive Web: Information Extraction', *Communications of the ACM* 49 (2006) nr. 5, p. 78-84

Gruhl, D. e.a., 'The predictive power of online chatter', *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, New York, NY, USA, 2005, p78–87

Hamm, S., 'The Battery Recall: A Win for the Web', 30 augustus 2006, http://www.businessweek.com/technology/content/aug2006/tc20060830_642667.htm

Hu, M. en B. Liu, 'Mining and summarizing customer reviews', *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2004), Seattle, Washington, USA, augustus 22-25, 2004
Ook online beschikbaar: http://www.cs.uic.edu/~liub/publications/kdd04-revSummary.pdf

Hu, M. en B. Liu., 'Mining Opinion Features in Customer Reviews', *Proceedings of Nineteeth National Conference on Artificial Intellgience* (2004), San Jose, USA, juli 2004
Ook online beschikbaar: http://www.cs.uic.edu/~liub/publications/aaai04-featureExtract.pdf

i.Know NV, 'Technical White Paper', http://www.iknow.be/viewdocument.php?id=4

Ipsos Mori, 'The Business Impact Of Blogs', 21 november 2006, http://www.ipsos-mori.com/polls/2006/blogging.shtml

Java, A. e.a., 'Modeling the Spread of Influence on the Blogosphere', technical report, maart 2006, http://ebiquity.umbc.edu/_file_directory_/papers/262.pdf

Jindal, N. en B. Liu, 'Mining Comparative Sentences and Relations', *Proceedings of 21st National Conference on Artificial* (AAAI-2006), Boston, Massachusetts, USA, juli 16-20, 2006
Ook online beschikbaar: http://www.cs.uic.edu/~liub/publications/aaai06-comp-relation.pdf

Karger, D. en Q. Dennis, 'What would it mean to blog on the semantic web?', Web Semantics: Science, Services and Agents on the World Wide Web, 3 (2005) nr. 2-3, p147-157

Klosek, J., 'Corporate Blogs: Handle With Care', *Business Week Online*, 14 december 2006, p6-6

Ku, L.-W., Y.-T. Liang en H.-H. Chen, 'Opinion extraction, summarization and tracking in news and blog Corpora'. *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs* (AAAI-CAAW 2006), maart 2006
Ook online beschikbaar: http://nlg18.csie.ntu.edu.tw:8080/opinion/SS0603KuLW.pdf

Lamont, J., 'Business intelligence: The text analysis strategy', *KMWorld* 15 (2006) nr. 10, p8-9 en 30

Lewis, 'the business value of blogging', maart 2007, www.lewis360.com/downloads/Business_value_of_ blogging.pdf

Lipton, E., 'Software Being Developed to Monitor Opinions of U.S.', *New York Times*, 4 oktober 2006

Liu, B., 'Web Content Mining', *The 14th International World Wide Web Conference* (WWW-2005), Chiba, Japan, mei 10-14 2005, http://www.cs.uic.edu/~liub/Web-Content-Mining-2.pdf

Liu, B., 'Chapter 11: Opinion Mining', slides van hoofdstuk 11 uit het boek: *Web Data Mining - Exploring Hyperlinks, Contents and Usage Data*, december 2006, http://www.cs.uic.edu/~liub/teach/cs583-spring-07/opinion-mining.pdf

Liu, B. en K. Chen-Chuan-Chang, 'Editorial: special issue on web content mining', *ACM SIGKDD Explorations Newsletter* 6 (2004) nr. 2, p1-4

Marketingfacts, 'Blogs & Word-of-Mouth Marketing', http://www.slideshare.net/marketingfacts/blogs-wordofmouth-marketing

Marketingfacts, 'Dell vs Blogosphere', http://www.slideshare.net/marketingfacts/dell-vs-blogosphere

Marketingfacts, 'UPC: Digitale televisie en nu?', http://www.slideshare.net/marketingfacts/upc-digitale-televisie-en-nu/

Marketing News, 'Dell learns power of the blog', 40 (2006) nr. 20, p17-18

Mercado-Kierkegaard, S., 'Blogs, lies and the doocing: The next hotbed of litigation?', *Computer Law and Security Report* 22 (2006) nr. 2, p127-136

Mishne, G., en N. Glance, 'Predicting Movie Sales from Blogger Sentiment', *Spring Symposium on Computational Approaches to Analysing Weblogs* (AAAI-CAAW 2006), 2006

Mishne, G. en N. Glance, 'Predicting Movie Sales from Blogger Sentiment', *In: AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs* (AAAI-CAAW 2006), Maart 2006, http://staff.science.uva.nl/~gilad/pubs/aaai06-linkpolarity.pdf

Mishne, G. en M. de Rijke, 'Capturing Global Mood Levels using Blog Posts', *In*: *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs* (AAAI-

CAAW 2006), maart 2006, http://staff.science.uva.nl/~gilad/pubs/aaai06-blogmoods.pdf

Mishne, G. en M. de Rijke, 'MoodViews: Tools for Blog Mood Analysis', *In: AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs* (AAAI-CAAW 2006), maart 2006, http://staff.science.uva.nl/~gilad/pubs/aaai06-mooddemo.pdf

Mitchell, R., 'Drowning in Unstructured Data', *Computerworld* 39 (2005) nr. 12, p26-26

Mitra, M., 'Web Mining: an Overview', http://www.ewh.ieee.org/r10/calcutta/comsoc/MNGN06/talk2.pdf

NRC Handelsblad, 'Eerste hulp bij webdiscussie', 17 januari 2007, http://www.nrc.nl/media/article601866.ece/Eerste_hulp_bij_webdiscussie

Pang, B. en L. Lee, 'A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts', *Proceedings of ACL* (2004), p271-278
Ook online beschikbaar: http://www.cs.cornell.edu/home/llee/papers/cutsent.pdf

Pang, B., L. Lee en S. Vaithyanathan, 'Thumbs up? Sentiment classification Using Machine Learning Techniques', In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP 2002), p79-86
Ook online beschikbaar: http://www.cs.cornell.edu/home/llee/papers/sentiment.pdf

Pfeiffer, E., 'Taming the Beast: The battle to tame Unstructured Data', *CIOInsight,* Supplement Issue 68, mei 2006, p46-56

Porter Novelli, 'Corporate Blog Learnings: The Discovery Age', juli 2006, http://www.porternovelli.com/Corporate_Blog_Learnings.pdf

Redondo, J., 'Estudio sobre uso, interés, conocimiento y percepción de la blogosfera española', *Zed Digital*, 12 april 2007, http://www.zeddigital.es/Presentaci% F3n_estudio_Blogs_ZedDigital.zip

Sanjuan, E. en F. Ibekwe-SanJuan, 'Text mining without document context', Information Processing & Management 42 (2006) nr. 6, p1532-1552

Schmerken, I., 'Analyzing Web Talk', *Wall Street & Technology*, Februari 2007 Advanced Trading, p10-11

Shane, L., 'Military issues content warning to combat-zone bloggers', *Stars and Stripes*, 1 oktober 2005, http://stripes.com/article.asp?section=104&article=31111 &archive=true

Sifry, D., 'The State of the Live Web', *Technorati*, 5 april 2007, http://www.sifry. com/alerts/archives/000493.html

Sprague, R., 'Business Blogs and Commercial Speech: A New Analytical Framework for the 21st Century', *American Business Law Journal* 44 (2007) nr. 1, p127-159

Stone, P., 'Inquirer Home Page', http://www.wjh.harvard.edu/~inquirer/

Stumme, G., A.Hotho en B. Berendt, 'Semantic Web Mining', *Web Semantics: Science, Services and Agents on the World Wide Web* 4 (2006) nr. 4, p124-143

Tetlock, P., 'Giving Content to Investor Sentiment: The Role of Media in the Stock Market', *Journal of Finance* 62 (2007) nr. 3, p1139-1168

The Economist, 'Blogging bosses', 23 Jan. 2007, web-only, http://www.economist.com/business/displaystory.cfm?story_id=8580521

Tong, M. 'An operational system for detecting and tracking opinions in on-line discussions', In *Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification*, 2001

Treloar, N., 'Text mining: Tools, Techniques, And Applications', *Avaquest*, 2002, http://www.knowledgetechnologies.net/proceedings/presentations/treloar/nathantreloar.ppt

Trumbach, C., 'Addressing the information needs of technology managers: Making derived information usable', *Technology Analysis & Strategic Management* 18 (2006) nr. 2 , p221-243

Turney, P., 'Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews', *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (ACL 2002), Philadelphia, Pennsylvania, USA, 8-10 juli 2002, pp 417-424
Ook online beschikbaar: http://www.iit-iti.nrc-cnrc.gc.ca/iit-publications-iti/docs/NRC-44946.pdf

Twibell, D., 'Getting Sentimental', *Financial Planning* 34 (2004) nr. 9, p141-144

Voight, J., 'How Consumers Help Build a Brand's DNA', *Adweek* 48 (2007) nr. 5, p16-18

Wasserman, T., 'Consumers Don't Trust Blogs', *Brandweek* 47 (2006) nr. 32, p10-10

Weglarz, G., 'Two worlds of data: Unstructured and Structured', *DM Review* 14 (2004) nr. 14, p19-21

Whitelaw, C., N. Garg en S. Argamon, 'Using Appraisal Taxonomies for Sentiment Analysis', *Proceedings of the 14th ACM international conference on Information and knowledge management* (MCLC 2005), p625-631

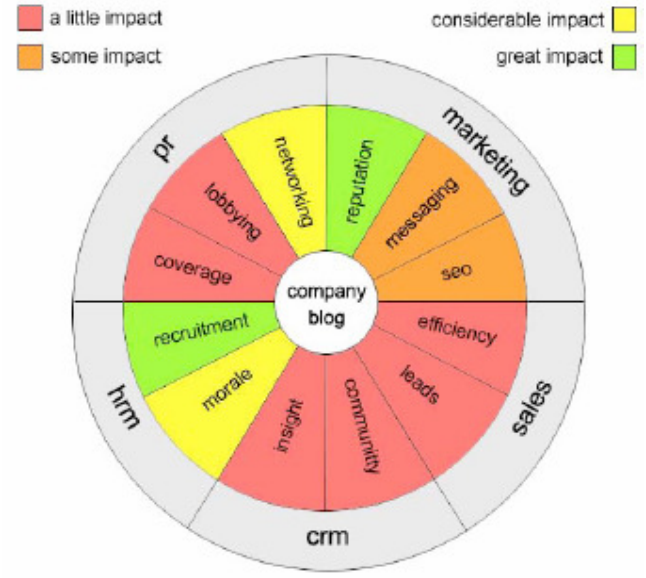
Ook online beschikbaar: www.cs.rhul.ac.uk/home/alexc/year3/appraisal_sentiment .pdf

Zerfaß, A., 'Social Software, Business Excellence and Communication Strategies: A framework for theorizing about weblogs, podcasts, wikis and RSS', *EuroBlog 2006 - International Research Symposium „Public Relations and Social Software*", Stuttgart, Duitsland, 18 maart 2006, http://www.euroblog2006.org/symposium/program/assets/EuroBlog2006_Zerfass.pdf

Bijlagen

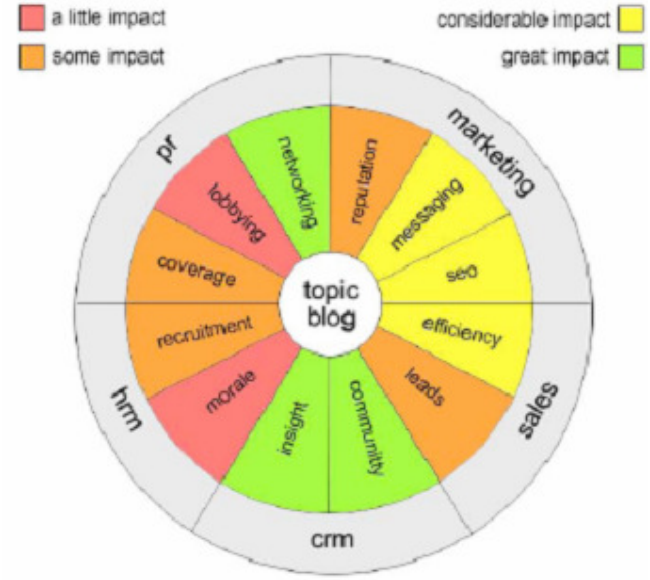
Bijlage 1 Blogging value wheel

Figuur 20: Blogging value wheel: CEO blog



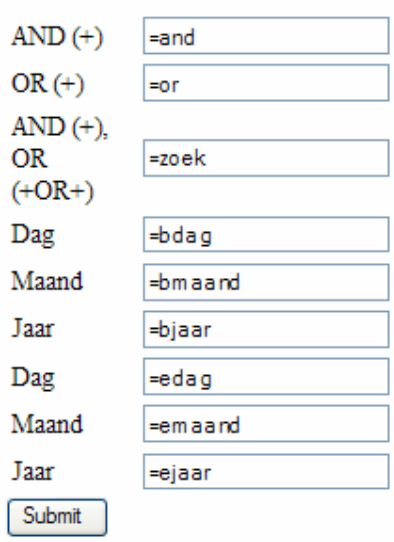
Figuur 21: Blogging value wheel: company blog

Figuur 22: Blogging value wheel: expert blog



Figuur 23: Blogging value wheel: topic blog

Bijlage 2 Selectie Inhoud




```
include('connectie.php');
session_start();
$db = $_SESSION['database'];

set_time_limit(0);

$and=$_POST['and'];
$or=$_POST['or'];
$zoek=$_POST['zoek'];
$bdag=$_POST['bdag'];
$bmaand=$_POST['bmaand'];
$bjaar=$_POST['bjaar'];
$edag=$_POST['edag'];
$emaand=$_POST['emaand'];
$ejaar=$_POST['ejaar'];

$dag = mktime(0,0,0,$bmaand,$bdag,$bjaar);
$einddag = mktime(0,0,0,$emaand,$edag,$ejaar);

$i= 0;
while ($dag <= $einddag) {
$url = "http://search.blogger.com/blogsearch_feeds?as_lq=&hl=en&as_q=" . $and .
"&as_epq=&as_oq=" . $or . "&as_eq=&as_qdr=a&as_drrb=b&as_mind=" . date(j,
$dag) . "&as_minm=" . date(n, $dag) . "&as_miny=" . date(Y, $dag) . "&as_maxd="
. date(j, $dag) . "&as_maxm=" . date(n, $dag) . "&as_maxy=" . date(Y, $dag) .
```

```
"&amp;amp;lang=some&lr=lang_en&safe=images&q=" . $zoek . "&ui=blg&ie=utf-
8&num=100&output=rss";

	$dag = mktime(0,0,0,date(m, $dag), date(d, $dag) + 1, date(Y, $dag));

	$sFile = file_get_contents($url);
	$filename = 'test.xml';
	file_put_contents($filename, $sFile);

	if (file_exists('test.xml')) {
		$xml = simplexml_load_file('test.xml');
	} else {
		exit('Failed to open test.xml.');
	}
	$a = 1;
	$query = "INSERT INTO " . $db . " ( id , url , datum , titel ) VALUES ";
	foreach ($xml->channel->item as $item) {
		$ns_dc = $item->children('http://purl.org/dc/elements/1.1/');
		$i++;
		foreach($xml->channel as $channel);
		$aantal = count($channel) - 3;
		if ($a < $aantal) {
		$query= $query . "(NULL,'" . $item->link . "','" . $ns_dc->date . "','" .
		htmlentities($item->title) . "'), ";
		} else {
		$query= $query . "(NULL,'" . $item->link . "','" . $ns_dc->date . "','" .
		htmlentities($item->title) . "');";
		}
		$a++;
	}
	if ($aantal > 0){
		mysql_query($query);
	}
}
```

Bijlage 3 Code: inhoud URL naar txt-file

```
include('connectie.php');

session_start();
$db = $_SESSION['database'];

set_time_limit(0);

$search = array ("\t",
                                "\n",
                                "\r",
                                "\0",
                                "\x0B",
                                '<br>',
                                '<BR>',
                                '<br />',
                                '<BR />',
                                '<br/>',
                                '<BR/>',
                                '</li>',
                                '</LI>',
                                '</p>',
                                '</P>',
                                '<p>',
                                '<P>',
                                '</tr>',
                                '</TR>',
                                '</div>',
                                '</DIV>',
                                '</span>',
                                '</SPAN>',
                                '...',
                                '</h1>',
                                '</h2>',
                                '</h3>',
                                '</h4>',
                                '</H1>',
                                '</H2>',
                                '</H3>',
                                '</H4>');

$replace = array ('',
                                 '',
                                 '',
                                 '',
                                 '',
                                 ' ',
```

```
' ',
' ',
' ',
' ',
' ',
'. ',
'. ',
'. ',
'. ',
'. ',
'. ',
' ',
' ',
' ',
' ',
' ',
' ',
'xxddffeerreeffddxx',
'. ',
'. ',
'. ',
'. ',
'. ',
'. ',
'. ',
'. ');
```

```
$array[] = array("<div class=\"blog-post-content\">", "</div>", 1);
$array[] = array("<div class=\"post-content\">", "</div>", 1);
$array[] = array("<p class=\"blogContent\">", "</p>", 1);
$array[] = array("<div class=\"itemtext\">", "<small class=\"metadata\">", 0);
$array[] = array("<div class=\"itemtext\">", "</div>", 1);
$array[] = array("<p class=\"blogContent\">", "<p class=\"blogContentInfo\">", 0);
$array[] = array("<div class=\"post-body\">", "<p class=\"post-footer\">", 0);
$array[] = array("<div class='post-body'>", "<div class='post-footer'>", 0);
$array[] = array("<div class='post-body'>", "<p class=\"postmetadata\"", 0);
$array[] = array("<div class=\"contentbody\">", "<div class=\"tags\">", 0);
$array[] = array("<!-- start entry of your post -->", "<!-- end entry of your post --
>", 0);
$array[] = array("<div class=\"storycontent\">", "</div> <!-- storycontent -->",
0);
$array[] = array("<div class=\"storycontent\">", "<div class=\"post-footer\"", 0);
$array[] = array("<div class=\"storycontent\">", "</div>", 1);
$array[] = array("<div class=\"entry-content\">", "<!-- .entry-content -->", 0);
$array[] = array("<div class=\"entry-content\">", "</div>", 1);
$array[] = array("<div class=\"postentry\">", "<div class=\"sociable\">", 0);
$array[] = array("<div class=\"postentry\">", "</div>", 1);
$array[] = array("<div class=\"entry-body\">", "<p class=\"entry-footer\">", 0);
```

```
$array[] = array("<div class=\"entry-body\">", "</div>", 1);
$array[] = array("<div class=\"asset-content\">", "<div class=\"asset-footer\">",
0);
$array[] = array("<!-- content -->", "<div id='Comments'>", 0);
$array[] = array("<!-- content -->", "<div id=\"MiniSiteMap\">", 0);
$array[] = array("<!-- content -->", "<!-- end minimap -->", 0);
$array[] = array("<!-- content -->", "<!-- end content -->", 0);
$array[] = array("<div class=\"entrytext\">", "* Social Bookmark Script", 0);  //
??????
$array[] = array("<div class=\"entrytext\">", "<p class=\"postmetadata", 0);
$array[] = array("<div class=\"entrytext\">", "</div>", 1);
$array[] = array("<div class=\"blogbody\">", "<div
class=\"itemCategoryLinksStyle\">", 0);
$array[] = array("<div class=\"blogbody\">", "<div class=\"itemFooterStyle\">",
0);
$array[] = array("<div class=\"blogbody\">", "</div>", 1);
$array[] = array("<div class=\"post\"", "<p class=\"post-footer\">", 0);
$array[] = array("<div class=\"postText\">", "</div>", 1);
$array[] = array("<div class=\"entrybody\">", "<div class=\"postinfo\">", 0);
$array[] = array("<div class=\"entrybody\">", "</div>", 1);
$array[] = array("<div class=\"blogPost\">", "</div>", 1);
$array[] = array("<div class=\"post-body\">", "<p class=\"post-meta-ind\">", 0);
$array[] = array("<div class=\"post-body\">", "<div class=\"post-meta-ind\">", 0);
$array[] = array("<div class=\"post-body\">", "<div id=\"comments\">", 0);
$array[] = array("<div class=\"post-body\">", "<div class=\"comments\">", 0);
$array[] = array("<div class=\"postbody\">", "<!-- technorati tags start -->", 0);
$array[] = array("<div class=\"post-body\">", "</div>", 1);
$array[] = array("<div class=\"articleBody\">", "</div>", 1);
$array[] = array("<div class=\"entry\">", "<!-- technorati tags begin -->", 0);
$array[] = array("<div class=\"entry\">", "<p class=\"postmetadata", 0);
$array[] = array("<div class=\"content\">", "<p class=\"postmetadata", 0);
$array[] = array("<div class=\"content\">", "<ul class=\"metalinks\">", 0);
$array[] = array("<div id=\"mainContent\"", "<div id=\"bottomSponsors\">", 0);
$array[] = array("<div id=\"mainContent\"", "</div>", 1);
$array[] = array("<div id=\"main-content\"", "</div>", 1);
$array[] = array("<div class=\"entry\">", "<p class=\"meta\"", 0);
$array[] = array("<div class=\"entry\">", "</div>", 1);
$array[] = array("<span class=\"entry\">", "</span>", 1);
$array[] = array("<span class='story'>", "</span>", 1);
$array[] = array("<div class=\"post\">", "<!-- you can start editing here. -->", 0);
$array[] = array("<div class=\"content\">", "<!-- you can start editing here. -->",
0);
$array[] = array("<div class=\"post\"", "</div>", 1);
$array[] = array("<div class=\"posting\"", "</div>", 1);
$array[] = array("<div class=\"posts\"", "</div>", 1);
$array[] = array("<div class=\"itemBodyStyle\">", "</div>", 1);
$array[] = array("<div id=\"content\">", "<div class=\"comments\">", 0);
$array[] = array("<div class=\"content\">", "<p class=\"posted\"", 0);
$array[] = array("<div class=\"item\">", "<div class=\"references\">", 0);
$array[] = array("<span id=\"intelliTXT\">", "</span>", 1);
```

```
$array[] = array("<div class=\"content\">", "</div>", 1);
$array[] = array("<div class=\"body\">", "</div>", 1);
$array[] = array("<div class=\"item\">", "</div>", 1);


function haalinhoud($sFilef, $begin, $einde, $div) {

	$arraytitel[] = array("<h1", "</h1");
	$arraytitel[] = array("<h2", "</h2");
	$arraytitel[] = array("<h3", "</h3");
	$arraytitel[] = array("<h4", "</h4");
	$arraytitel[] = array("<title", "</title");

	$bpos = stripos(strtolower($sFilef), strtolower($begin));
	$sFilef = substr($sFilef, $bpos);

	if ($div == 1) {
		switch ($einde) {
		  case "</div>":
			$bdiv = "<div";
			$ediv = "</div";
			break;
		  case "</span>":
			$bdiv = "<span";
			$ediv = "</span";
			break;
		  case "</p>":
			$bdiv = "<p";
			$ediv = "</p";
			break;
		}
		$hulp = substr($sFilef, 1);
		$hulp2 = substr($sFilef, 1);
		$rest = 0;
		$rest2 = 0;
		$laatste = false;
		while ($laatste == false AND $posdiv !== false AND $posdiv2 !==
		false) {
			$posdiv = stripos(strtolower($hulp), $bdiv);
			$posdiv2 = stripos(strtolower($hulp2), $ediv);
			if (($posdiv == false AND $posdiv !== 0) OR ($posdiv + $rest)
			> ($posdiv2 + $rest2)){
				$inhoud = substr ($sFilef, 0, $posdiv2 + $rest2);
				$laatste = true;
			} else {
				$hulp = substr ($hulp, $posdiv + 1);
				$hulp2 = substr ($hulp2, $posdiv2 + 2);
				$rest = $rest + $posdiv + 1;
				$rest2 = $rest2 + $posdiv2 + 2;
			}
```

```
		}
	} else {
		$epos = stripos(strtolower($sFilef), strtolower($einde));
		$inhoud = substr ($sFilef, 0, $epos);
	}

	//titel verwijderen
	$laatste2 = false;
	$ii = 0;
	foreach ($arraytitel as $result2) {
		if ((strpos(strtolower($inhoud), strtolower($arraytitel[$ii][0])) !==
		false) AND (strpos(strtolower($inhoud), strtolower($arraytitel[$ii][0]))
		< 50) AND (strpos(strtolower($inhoud), strtolower($arraytitel[$ii][1]))
		!== false) AND ($laatste2 == false)) {

		$hulp1 = substr ($inhoud, stripos(strtolower($inhoud),
		$arraytitel[$ii][1]));
		$hulp2 = substr ($inhoud, 0, stripos(strtolower($inhoud),
		$arraytitel[$ii][0]));

			$inhoud = $hulp2 . ". " . $hulp1;
			$laaste2 = true;
		}
		$ii++;
	}

	return $inhoud;
}

function haalurls($inhoudf, $id) {
	$laatste = 0;
	$rest = $inhoudf;

	$i = 0;
	while ($laatste == 0) {
		$bpos = stripos(strtolower($rest), " href=\"");

		$rest = substr ($rest, $bpos+7);
		$epos = stripos($rest, "\"");
		$link = substr ($rest, 0, $epos);

		if (stripos(strtolower($rest), " href=\"") == false) {
			$laatste = 1;
		}

$query = "INSERT INTO links (id, idurl, link) VALUES (NULL, $id, '$link'); ";
		mysql_query($query);
		$i++;
	}
```

```
        return $i;
}

function spatiesetc($inhoudf) {
        while (strpos($inhoudf, "  ") !==false) {
                $inhoudf = str_replace("  ", " ",$inhoudf);
        }
        while (strpos($inhoudf, ". .") !==false) {
                $inhoudf = str_replace(". .", ".",$inhoudf);
        }
        while (strpos($inhoudf, " .") !==false) {
                $inhoudf = str_replace(" .", ".",$inhoudf);
        }
        while (strpos($inhoudf, "..") !==false) {
                $inhoudf = str_replace("..", ".",$inhoudf);
        }

        $inhoudf = str_replace("?.", "?",$inhoudf);
        $inhoudf = str_replace("!.", "!",$inhoudf);
        return $inhoudf;
}


$query = "SELECT url, id, titel, date_format(datum, '%e %M %Y') as datumf FROM
$db WHERE url NOT LIKE '%spaces.live%'";
$resultqry = mysql_query($query);
$aantal = mysql_num_rows ($resultqry);

if($aantal == NULL){
        echo "Geen rijen aangetroffen.";
} else {
        while ($row = mysql_fetch_assoc($resultqry)){
                $url= $row["url"];
                $id = $row["id"];
                $titel = $row["titel"];
                $datum = $row["datumf"];

                $sFile = @file_get_contents($url);
                if (empty($sFile))
                        {
                                // Web page empty/access failure
                                echo "Url niet gevonden.";
                } else {
                        $inhoud = NULL;
                        $i = 0;
                        $laatste = false;
                        foreach ($array as $result) {
                                if ((strpos(strtolower($sFile), strtolower($array[$i][0]))
                                !== false) AND (strpos(strtolower($sFile),
```

```
        strtolower($array[$i][1])) !== false) AND
        (empty($inhoud))) {
                $inhoud = haalinhoud($sFile, $array[$i][0],
                $array[$i][1], $array[$i][2]);
                $laatste = true;
        }
        $i++;
}

$inhoud = str_replace($search, $replace, $inhoud);
$inhoud = trim($inhoud);
$inhoud = spatiesetc($inhoud);

if (empty($inhoud)) {
        echo "Kan gegevens niet uithalen.";
} else {
        //Haal eventuele URL's op
        if (stripos(strtolower($inhoud), strtolower(" HREF=\""))
        !== false) {
                $aantal = 0;
                $aantal = haalurls($inhoud, $id);
        } else {
                $aantal = 0;
        }

        $inhoud = strip_tags($inhoud);

        $filename = "$id.txt";

        $inhoud =
        html_entity_decode($inhoud,ENT_QUOTES,'UTF-8');
        $inhoud =
        html_entity_decode($inhoud,ENT_QUOTES,'ISO-8859-
        15');
        $inhoud = html_entity_decode($inhoud,ENT_QUOTES);

        $inhoud = strip_tags($inhoud);
        $inhoud = spatiesetc($inhoud);

        $inhoud = str_replace("xxddffeerreeffddxx",
        "...",$inhoud);

        $titel = strip_tags($titel);
        $titel = html_entity_decode($titel,ENT_QUOTES,'UTF-8');
        $titel = html_entity_decode($titel,ENT_QUOTES,'ISO-
        8859-15');
        $titel = html_entity_decode($titel);
        $titel = strip_tags($titel);
        $titel = $titel . ". ";
        $datum = $datum . ". ";
```

```
                        if ((strpos($inhoud, ". ") == 0) AND (strpos($inhoud, ". ")
                        !== FALSE)) {
                                $inhoud = substr($inhoud, 2);
                        }

                        $file = fopen("txt/" . $filename, "w");
                        fputs($file, $titel);
                        fputs($file, $datum);
                        fputs($file, $inhoud);
                        fclose($file);
                }
        }
    }
}
```

Bijlage 4 Lijst begin- en eindtags

Begintag	Eindtag	Aantal
		750
	<p class="postmetadata	424
		294
		179
<div class="post"		172
<!-- content -->		170
		163
		130
	<p class="postmetadata	123
		73
		73
		72
		72
		71
<!-- content -->		68
	<small class="metadata">	56
		49
		47
		46
		42
		36
		36
	<p class="posted"	35
<div class="post"		35
	<!-- you can start editing here. -->	32

<div id="maincontent"		32
		31

Bijlage 5 Verdeling aantal links

Aantal links:	Aantal:
0	951
1	544
2	412
3	245
4	183
5	140
6	189
7	142
8	95
9	73
10	63
10 < x <= 15	146
15 < x <= 25	118
25 < x <= 50	101
50 < x <= 200	36
200 < x <= 400	20
400 < x <= 1000	20
> 1000	12
Totaal:	3.490

Bijlage 6 Lijst meest voorkomende links

Link:	Aantal:
http://www.cashbazar.com/	964
http://www.organicgreens.us/	273
http://www.home-based-business-team.com/	268
http://www.end-your-debt.com/	235
http://www.theinquirer.net/?article=32550	205
http://www.homeequityhelp.net/	190
http://trimyourdebt.com/	165
http://www.clickpress.com/	160
http://www.bazuji.com/havingitall.html	160
http://www.digitalroom.com/Poster-Printing.html	160
http://www.rankingyourwaytothebank.com/	153
http://www.copywritingcourse.com/	139
http://www.unlockthegame.com/	128
http://www.newyorkmetrotechnologyjobs.com/	120
http://www.senn-sational.com/freeresources.htm	120
http://www.theinquirer.net/default.aspx?article=32550	118
http://www.copywritingcourse.com/keyword	106
https://www.dellbatteryprogram.com/	106
http://www.trimyourdebt.com/	99
http://www.elevatingyourbusiness.com/	99
http://www.everypleasures.com/	99
http://www.streetsmartsmarketing.com/free-ebook.htm	93
http://www.communication-newsletter.com/	90

Bijlage 7 Lijst meest voorkomende links na filtering

Link:	Aantal:
http://www.theinquirer.net/?article=32550	205
http://www.theinquirer.net/default.aspx?article=32550	117
https://www.dellbatteryprogram.com/	106
http://miluping.com/dellnews/category/uncategorized/	68
https://www.dellbatteryprogram.com/Default.aspx	60
http://technorati.com/tag/dell	58
http://www.dellbatteryprogram.com/	44
#comment	43
http://www.engadget.com/2006/07/28/another-dell-laptop-ignites/	41
#	37
	35
http://www.engadget.com/2006/06/22/dude-your-dell-is-on-fire/	35
http://bl-technology.desertscapeit.com/?cat=1	32
http://www.engadget.com/2006/07/31/dell-laptop-number-3-explodes/	30
#Top	28
http://www.engadget.com/2006/08/14/dell-recalls-4-1-million-batteries/	27
http://technorati.com/tag/laptop	27
#respond	26
http://elliottback.com	23
http://technorati.com/tag/Sony	22
http://www.dell.com	22
http://www.dellbatteryprogram.com	22
#comments	21
http://www.gizmodo.com/gadgets/laptops/dell-laptop-explodes-in-	21

flames-182257.php	
http://forumz.tomshardware.com/hardware/Dude-Dell-freaking-blew-ftopict192887.html	21
http://www.engadget.com/2006/09/20/dell-battery-explodes-at-yahoo-hq-hundreds-evacuat/	20
https://support.apple.com/ibook_powerbook/batteryexchange/	20

Bijlage 8 Code: Bekomen resultaten General Inquirer

```
include('connectie.php');

set_time_limit(0);

for ($i = 0; $i < 14310; $i++) {
	$filename =  "$i.txt";
	$url = "txt/" . $filename;

	$sFile = @file_get_contents($url);
	if (empty($sFile))
		{
			// Web page empty/access failure
			echo "Bestand niet gevonden.";
	} else {
		$sFile = preg_replace('/[^a-zA-Z0-9\.\?\'!;:,\s]/','',$sFile);
		$sFile = preg_replace('/\s/','+',$sFile);

		$gedaan = false;
		while ($gedaan == false) {
			if (strlen($sFile) > 2000) {
				$hulp = substr($sFile, 1700, 300);
				$hulppos = stripos($hulp,".");
				if ($hulppos == false) {
					$begin = 2000;
				} else {
					$begin = 1700 + $hulppos + 1;
				}
				$array[] =
				"http://www.webuse.umd.edu:9090/GI?sentence=" .
				substr($sFile, 0, $begin);
				$sFile = substr ($sFile, $begin);
			} else {
				$array[] =
				"http://www.webuse.umd.edu:9090/GI?sentence=" .
				$sFile;
				$gedaan = true;
			}
		}

		for ($a = 0; $a < count ($array); $a++) {
			$sFile = @file_get_contents($array[$a]);
			if (empty($sFile))
				{
					// Web page empty/access failure
					echo "<br/>Url niet gevonden.";
			} else {
```

```
$pos = stripos($sFile,"<TR> <TH> tag </TH> <TH> N </TH> <TH> % </TH>
<TH> words </TH> </TR>");
                        if ($pos == false) {
                                echo "<br />Geen resultaat.<br />";
                        } else {
                                $array2 = explode("<TR>",$sFile);

                                for ($x = 2; $x < count ($array2); $x++) {
                                        $arrayklein = explode("<TD>",
                                        $array2[$x]);
                                        $categorie =
                                        trim(strip_tags($arrayklein[1]));
                                        $n = trim(strip_tags($arrayklein[2]));
                                        $perc = trim(strip_tags($arrayklein[3]));
                                        $woorden =
                                        trim(strip_tags($arrayklein[4]));

                                        $query = "INSERT INTO HARVARD ( idurl, i,
                                        categorie, n, perc, woorden) VALUES ( $i,
                                        $a, '$categorie', $n, $perc, '$woorden')";
                                        mysql_query($query);
                                }
                        }
                }
        }
        unset($array);
        unset($array2);
    }
}


//waarden optellen naar database
include('connectie.php');

set_time_limit(0);


$sql = "SELECT distinct(categorie) as cat FROM harvard ORDER BY categorie ASC";
$resultquery = mysql_query($sql);
$aantal = mysql_num_rows ($resultquery);

if($aantal == NULL){
    echo "Geen rijen aangetroffen.";
} else {
    while ($row = mysql_fetch_assoc($resultquery)){
        $categorie = $row["cat"];
        $categorie = preg_replace('/\*/','',$categorie);
        $query = "ALTER TABLE `harvardoverzicht` ADD $categorie INT( 10 )";
        mysql_query($query);
    }
```

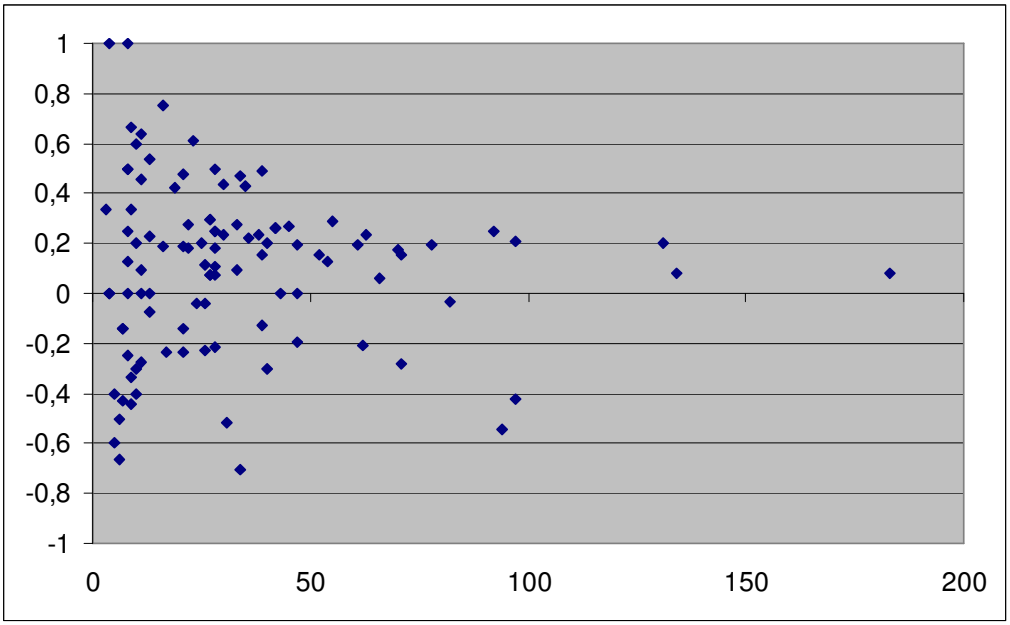
```
}

for ($i = 1; $i < 14291; $i++) {
$sql = "SELECT idurl, categorie, sum(n) as aantal FROM harvard where idurl = $i
GROUP BY idurl, categorie ORDER BY idurl, categorie";
	$resultquery = mysql_query($sql);
	$aantal = mysql_num_rows ($resultquery);

	if($aantal == NULL){
		echo "Geen rijen aangetroffen.";
	} else {
		$query = "INSERT INTO harvardoverzicht (id) VALUES ($i)";
		mysql_query($query);

		while ($row = mysql_fetch_assoc($resultquery)){
			$categorie = $row["categorie"];
			$aantal = $row["aantal"];
			$categorie = preg_replace('/\*/','',$categorie);

			$query = "UPDATE harvardoverzicht SET $categorie = $aantal
			WHERE idurl = $i";
			mysql_query($query);
		}
	}
}
```

Bijlage 9 Lijst met meest voorkomende pos en neg woorden

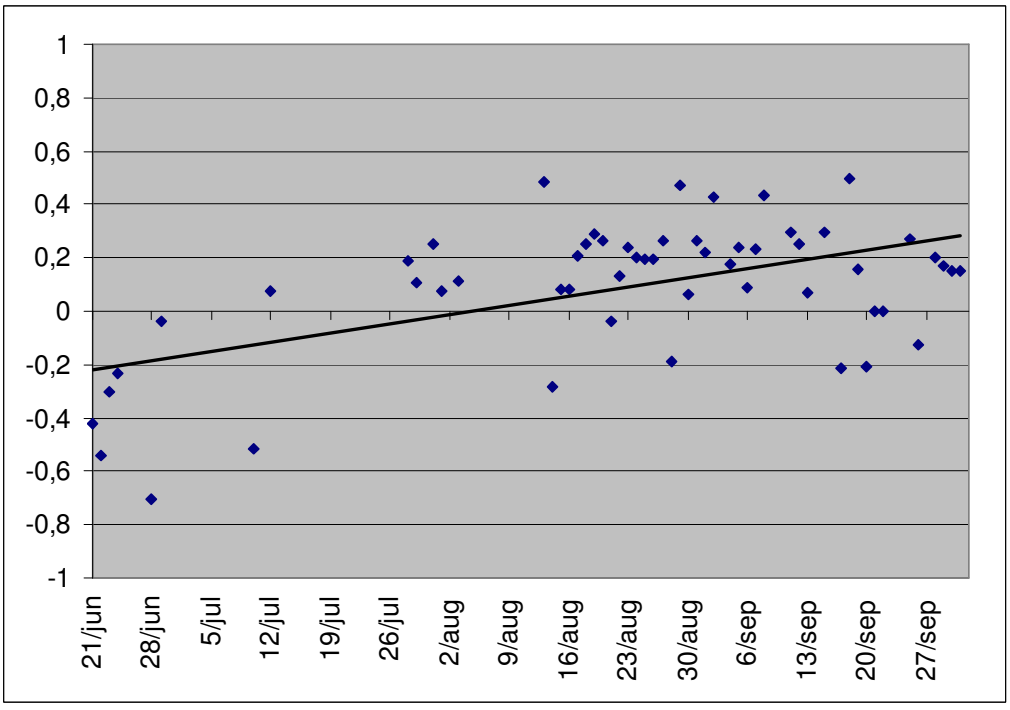
Positieve woorden	Aantal	Negatieve woorden	Aantal
FREE#1	1858	EXPLODE	1920
HOME	1689	PROBLEM	1606
GOOD#1	1406	NEED#1	821
WELL#2	1247	BAD	732
REAL#1	1166	HARD#1	575
GIVE#1	961	CHEAP	571
BEST	917	NEED#2	537
SAFETY#2	833	EXPLOSION	472
GREAT#1	662	COST#1	384
PROVIDE#1	645	COOL	383
LIKE#2	644	AGAINST	366
OFFER#1	641	HIT#1	363
BETTER#1	640	BURN#1	356
SECURITY#1	606	WAR	353
ACTUAL#2	605	BURN#2	351
FRIEND	581	HAZARD#1	295
HELP#2	566	LOW#1	278
LYRIC	510	ACCIDENT	254
LOVE#1	497	DAMAGE#1	243
SHARE#1	491	INJURY	224
COMMISSION	485	HATE#1	219
CREATE	478	CONCERN#2	217
ALLOW#1	406	BREAK#1	211
HEALTH	406	HELL	210
SUPPORT#2	401	TROUBLE#1	210

ABLE	400	KILL#1	201
LAW	366	LOST#1	200
HOPE#1	362	DAMN	199
OPEN#1	351	LOSE#1	198
LOVE#2	345	MISS#1	192
LIGHT#1	342	DEATH	191
PRO	327	HURT#1	191
MAJOR#1	325	CHARGE#6	184
FAVORITE	319	FIGHT#2	181
NICE#1	319	AVOID	180
PORTABLE	314	EXPLOSIVE	179
CONSIDER#1	311	POOR#1	178
SERIOUS	310	COST#2	172
FUN#1	309	CUT#1	172
SAVE#1	299	DEAD	171
EASY#1	295	DROP#4	171
SOLUTION	290	THREAT	171
ORDER#2	281	DEFECTIVE	169
EXPERIENCE#1	274	SECRET	168
IMPORTANT	269	DANGEROUS	165
CARE#1	268	FEAR#1	165
TRUE#1	256	RIVAL#1	165
CONTACT#1	253	SORRY#1	162
REPAIR#1	251	ATTACK#1	161
SPECIAL	246	CHARGE#7	154

Bijlage 10 Invloed dagen met weinig posts (score)

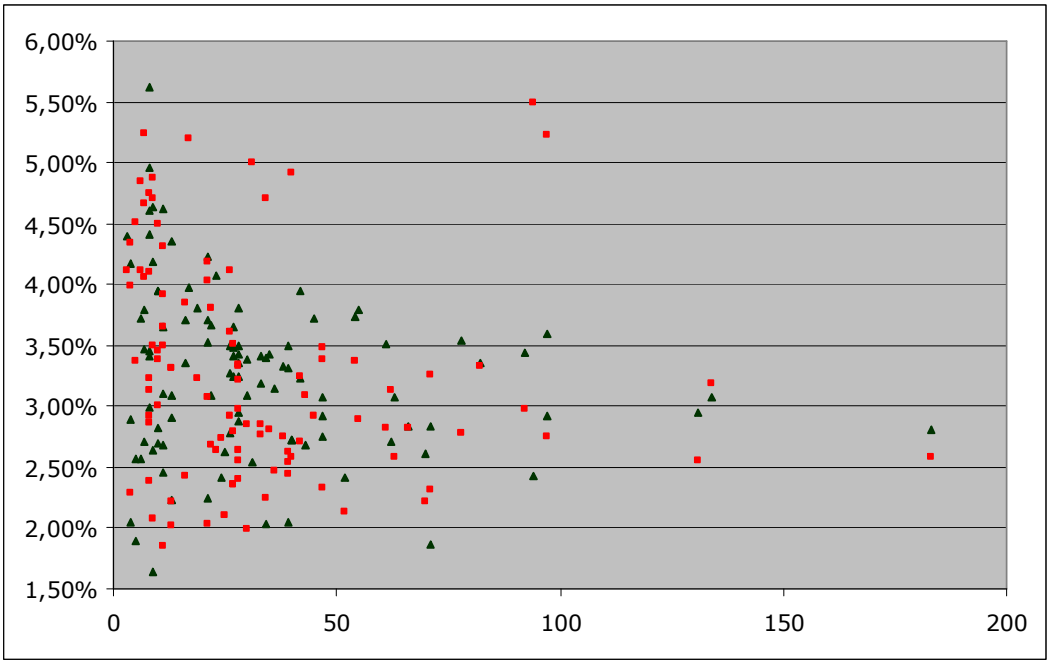


Figuur 24: X-as: aantal blogposts per dag/ Y-as: score

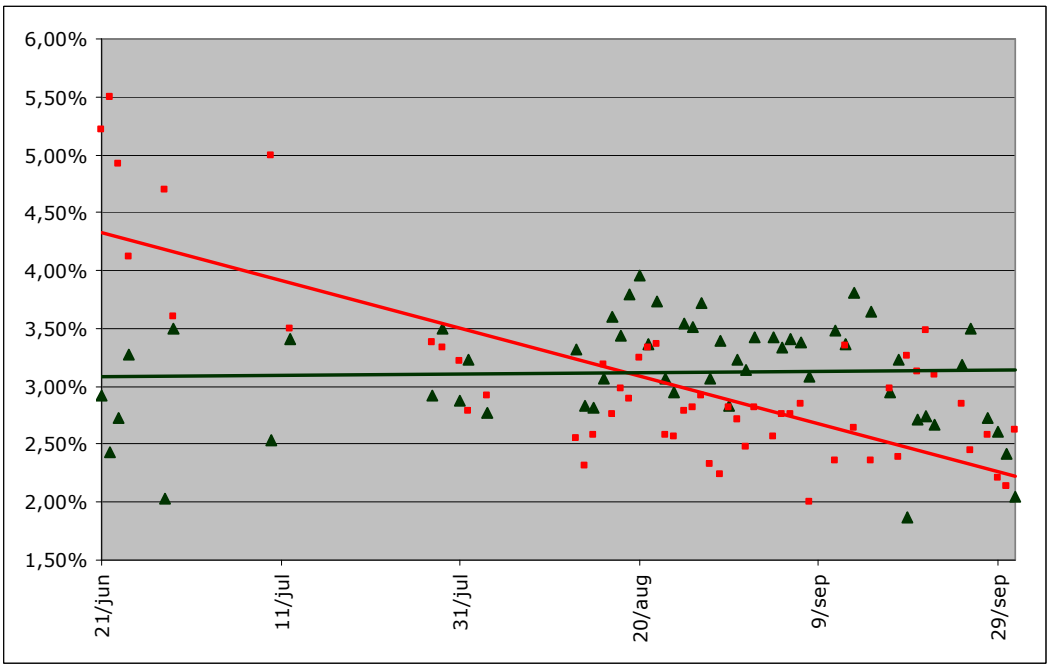


Figuur 25: gemiddelde waarde per dag voor dagen met meer dan 25 posts

Bijlage 11 Invloed dagen met weinig posts (%neg/pos woorden)



Figuur 26: X-as: aantal blogposts per dag / Y-as: score



Figuur 27: gemiddelde score per dag voor blogposts met meer dan 25 posts

Bijlage 12 Code: Indelen woordgroepen door General Inquirer

```
include('connectie2.php');

set_time_limit(0);

$query = "SELECT id, woord FROM pos";
$resultqry = mysql_query($query);
$aantal = mysql_num_rows ($resultqry);

if($aantal == NULL){
	echo "Geen rijen aangetroffen.";
} else {
	while ($row = mysql_fetch_assoc($resultqry)){
		$id= $row["id"];
		$woord= $row["woord"];

			$woord = preg_replace('/[^a-zA-Z0-9\.\?\'!;:,\s]/','',$woord);
			$woord = preg_replace('/\s/','+',$woord);

			$url = "http://www.webuse.umd.edu:9090/GI?sentence=" .
			$woord;

			$sFile = @file_get_contents($url);
			if (empty($sFile)) {
					// Web page empty/access failure
					echo "<br/>Url niet gevonden.";
			} else {
				$pos = stripos($sFile,"<TR> <TH> tag </TH> <TH> N
				</TH> <TH> % </TH> <TH> words </TH> </TR>");
				if ($pos == false) {
					echo "<br />Geen resultaat.<br />";
				} else {
					$array2 = explode("<TR>",$sFile);

					for ($x = 2; $x < count ($array2); $x++) {
						$arrayklein = explode("<TD>",
						$array2[$x]);
						$categorie =
						trim(strip_tags($arrayklein[1]));
						$n = trim(strip_tags($arrayklein[2]));
						$perc = trim(strip_tags($arrayklein[3]));
						$woorden =
						trim(strip_tags($arrayklein[4]));

						if ($categorie == 'Pos'){
							$query = "UPDATE pos SET pos = $n
							WHERE id = $id";
```

```
                                        mysql_query($query);
                                    }
                                    if ($categorie == 'Neg'){
                                        $query = "UPDATE pos SET neg = $n
                                        WHERE id = $id";
                                        mysql_query($query);
                                    }
                                }
                            }
                        }
                    }
                }
```

Bijlage 13 Lijst van meest voorkomende clusters

Clusters	Indeling	Aantal
fire	NEUTRAAL	1632
company	NEUTRAAL	863
well	POSITIEF	731
problem	NEGATIEF	536
right	NEUTRAAL	435
home	POSITIEF	430
kind	POSITIEF	270
buy	NEUTRAAL	259
matter	NEUTRAAL	250
courtesy	POSITIEF	241
point	NEUTRAAL	207
even	NEUTRAAL	203
good	POSITIEF	200
cost	NEGATIEF	193
share	POSITIEF	191
too	NEUTRAAL	187
consumer product safety commission	NEUTRAAL	185
mind	NEUTRAAL	171
turn	NEUTRAAL	169
fun	POSITIEF	166
game	NEUTRAAL	153
deal	NEUTRAAL	152
free	POSITIEF	152
ill	NEGATIEF	150
better	POSITIEF	146
live	NEUTRAAL	144
bit	NEUTRAAL	143
hell	NEGATIEF	143

safety	POSITIEF	142
hard	NEGATIEF	136
explosion	NEGATIEF	136
light	POSITIEF	132
hand	NEUTRAAL	131
service	NEUTRAAL	131
bad	NEGATIEF	129
quality	NEUTRAAL	127
easy	POSITIEF	126
safe	POSITIEF	123
fire hazard	NEGATIEF	123

Bijlage 14 Vergelijking woordgroepen

Tabel 6: enkele statistieken over concepten met woord uit Harvard-IV-4 dictionary categorie 'Negativ'

Zelf ingevuld:			
NEGATIEF:	690	48,80%	
POSITIEF:	145	10,25%	
NEUTRAAL:	579	40,95%	
	1.414	100,00%	
General Inquirer:			
NEGATIEF:	12.346	62,12%	
POSITIEF:	1.011	5,09%	
NEUTRAAL:	6.516	32,79%	
	19.873	100,00%	
Zelf <-> General Inquirer			
General Inquirer	Zelf		
NEGATIEF:	NEGATIEF:	490	66,22%
	POSITIEF:	51	6,89%
	NEUTRAAL:	199	26,89%
		740	100,00%
POSITIEF:	NEGATIEF:	12	13,79%
	POSITIEF:	47	54,02%
	NEUTRAAL:	28	32,18%
		87	100,00%
NEUTRAAL:	NEGATIEF:	188	32,03%

	POSITIEF:	47	8,01%
	NEUTRAAL:	352	59,97%
		587	100,00%
Komen			
Overeen:	889	62,87%	
Niet:	525	37,13%	

Tabel 7: enkele statistieken over concepten met woord uit Harvard-IV-4 dictionary categorie 'Positiv'

Zelf ingevuld:			
NEGATIEF:	142	5,60%	
POSITIEF:	1.188	46,83%	
NEUTRAAL:	1.207	47,58%	
	2.537	100,00%	
General Inquirer:			
NEGATIEF:	461	1,68%	
POSITIEF:	20.594	74,99%	
NEUTRAAL:	6.408	23,33%	
	27.463	100,00%	
Zelf <-> General Inquirer			
General			
Inquirer	Zelf		
NEGATIEF:	NEGATIEF:	22	44,90%
	POSITIEF:	13	26,53%
	NEUTRAAL:	14	28,57%
		49	100,00%
POSITIEF:	NEGATIEF:	46	2,45%
	POSITIEF:	1.053	56,01%
	NEUTRAAL:	781	41,54%

		1.880	100,00%
NEUTRAAL:	NEGATIEF:	74	12,17%
	POSITIEF:	122	20,07%
	NEUTRAAL:	412	67,76%
		608	100,00%
Komen			
Overeen:	1.487	58,61%	
Niet:	1.050	41,39%	

Auteursrechterlijke overeenkomst

Opdat de Universiteit Hasselt uw eindverhandeling wereldwijd kan reproduceren, vertalen en distribueren is uw akkoord voor deze overeenkomst noodzakelijk. Gelieve de tijd te nemen om deze overeenkomst door te nemen, de gevraagde informatie in te vullen (en de overeenkomst te ondertekenen en af te geven).

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Kennisontsluiting ten behoeve van management uit de blogosfeer
Richting: **Handelsingenieur in de beleidsinformatica** Jaar: **2007**
in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Ik ga akkoord,

Antoon KEUNEN

Datum: **07.06.2007**