CR1M3 F1GH7ING M3T 5TAT1STI3K

Naam student: Mark de Ree

Studentnummer: 07007965

Opleiding: Bachelor Bedrijfswiskunde (voltijd)

Onderwijsinstelling: De Haagse Hogeschool, Delft

Adres: Rotterdamseweg 137

Postcode en plaats: 2628 AL Delft

Afstudeercoach: mvr. ir. H.L. Liem

Studiejaar: 2011/2012

Afstudeerbedrijf: NCIM Groep

Adres: Overgoo 11

Postcode en plaats: 2266 JZ Leidschendam

Bedrijfsbegeleider: dhr. Paul Ligthart MSc

Bedrijfsbegeleider: dhr. dr. Micha Streppel

Voorwoord

De afgelopen maanden heb ik tijdens mijn afstudeerstage bij de NCIM Groep mogen meewerken aan het Forensic Dashboard project. Ik heb de vrijheid gekregen om te zoeken naar een interessant afstudeeronderwerp dat moest voldoen aan de eisen van de NCIM Groep en van mijn opleiding. Deze gekregen vrijheid heb ik enorm op prijs gesteld. Nu heb ik kunnen werken aan een applicatie met een flinke dosis wiskunde. Een leuke en uitdagende opdracht. Het succes van mijn afstudeerstage heb ik te danken aan verschillende personen binnen en buiten de NCIM Groep. Ik wil van deze gelegenheid gebruik maken om die mensen hier in het bijzonder te noemen en te bedanken. Ik dank bij dezen:

* Paul Ligthart en Micha Streppel, mijn begeleiders vanuit de NCIM Groep. Mede door de wekelijkse feedback, sturing en inzichten is het project goed verlopen.
* Harold Kasperink, de opdrachtgever en Executive Vice President van de NCIM Groep. Hij heeft een grote bijdrage geleverd bij het valideren van de applicatie door de Enron testcase bij het project te betrekken.
* Annette Bleeker, projectleider van alle interne projecten en dus ook van het Forensic Dashboard. Zij heeft mij ondersteund zodat ik een product kon maken die voor zowel de NCIM Groep als de Haagse Hogeschool van waarde kunnen zijn.
* Jason Lun en Gijs Pannebakker voor hun feedback over het maken van een goede en overzichtelijke flowchart. Deze moet duidelijk zijn, zodat mijn standalone applicatie in een later traject goed kan worden geïmplementeerd in het Forensic Dashboard.
* Stephan Ligtvoet voor de soepele samenwerking tijdens het project. Ook al hadden we ieder ons eigen deelproject, ik kon hem vertellen en laten zien waar ik mee bezig was.
* Milko Zuijdendorp voor informatie, een demonstratie en inzicht in Xiraf.
* Cathy Liem voor de begeleiding vanuit de Haagse Hogeschool. Besprekingen verliepen goed waarin ik goede feedback kreeg. Als ik een vraag had kon ik haar mailen, waarop ik vrijwel direct antwoord kreeg.
* Mijn onlangs overleden opa Cees Verdegaal. Aan jou draag ik deze thesis op. Jammer dat je het niet meer mee hebt kunnen maken.

Mijn afstudeerstage bij de NCIM Groep heeft ongeveer vier maanden geduurd. Ik ben trots op het eindresultaat dat u voor u ziet en hoop dat ik hiermee een nuttige maatschappelijke en wetenschappelijke bijdrage heb kunnen leveren.

# Management samenvatting

Xiraf is een database van digitale sporen en heeft als doel om de enorme hoeveelheid data doorzoekbaar te maken door middel van filters en data analyses om uiteindelijk misdaadzaken sneller op te lossen. Wat betreft data analyses is er genoeg ruimte voor verbetering. Deze data analyses kunnen betrekking hebben op foto’s, video’s, Microsoft Office bestanden, e-mail berichten enzovoort.

In deze scriptie wordt de nadruk gelegd op data analyse die betrekking heeft op de e-mail berichten in Xiraf. Deze data analyse is een e-mail content analyse. Dit houdt in dat er wordt gekeken naar de inhoud van de e-mail. Iedere e-mail is anders en voor rechercheurs is het belangrijk om e-mails te vinden die betrekking hebben op de zaak waaraan zij werken. Om deze paar belangrijke e-mails te vinden is deze e-mail content analyse gemaakt.

De e-mail content analyse is gebaseerd op het online e-mail spamfilter en maakt gebruik van Bayesiaanse statistiek. Het online e-mail spamfilter filtert spam e-mails van legitieme e-mails door onder andere te kijken naar de verschillende woorden die in de e-mail staan. Enkele woorden die vaak in spam berichten voorkomen zijn ‘viagra’ en ‘sex’. In legitieme e-mails komen deze woorden nauwelijks voor. Door gebruik te maken van Bayesiaanse statistiek kan het online e-mail spamfilter van iedere inkomende e-mail berekenen wat de kans is dat het een spam e-mail is. De e-mail content analyse werkt vrijwel op dezelfde manier als het online e-mail spamfilter. Het verschil is echter, het onderwerp van de e-mail. In de e-mail content analyse moet niet worden gezocht naar spam berichten, maar moet worden gezocht naar e-mails die van belang zijn voor de zaak. Woorden als ‘geweer’, ‘vermoord’, ‘kinderporno’ en ‘fraude’ zijn in dit geval belangrijk, afhankelijk van de soort zaak waaraan wordt gewerkt.

Voordat de content e-mail analyse zijn werking goed kan doen, moet eerst informatie worden doorgegeven aan het filter. Deze informatie wordt doorgegeven door e-mails te markeren als zaak gerelateerd of niet zaak gerelateerd. Het filter kijkt naar de woorden die in deze e-mails voorkomen en kan daaruit opmaken welke woorden belangrijk zijn voor de zaak en welke woorden dit niet zijn. Des te meer e-mail berichten zijn gemarkeerd, des te nauwkeuriger het filter zal werken. Een woord dat relatief vaak voorkomt in zaak gerelateerde e-mails en relatief weinig in niet zaak gerelateerde e-mails, kan worden gezien als een belangrijk woord.

Om de werking van de e-mail content analyse te testen is gebruik gemaakt van een testcase. Deze testcase heeft betrekking op het fraude schandaal van het Amerikaanse bedrijf ‘Enron Corporation’ in 2001. Dit was destijds het grootste en meest complexe fraude schandaal in de Amerikaanse historie. Vanwege de omvang van de fraude en de impact op de zakenwereld, is een set van ongeveer een half miljoen e-mail berichten van het hoger management publiekelijk gemaakt ter onderzoek. Als testcase moet de e-mail content analyse de belangrijke e-mails, waarin aspecten van het fraude schandaal in voorkomen, kunnen scheiden van de niet belangrijke e-mails.

Uit de testcase is gebleken dat de e-mail content analyse werkt zoals deze zou moeten. Tijdens het draaien van de analyse treden er geen fouten op en de heuristiek en de formules worden correct in de applicatie toegepast. Dit zorgt ervoor dat voor alle e-mail berichten kan worden berekend of de e-mail is gerelateerd aan de zaak of niet.

Voor de testcase is een beperkte hoeveelheid e-mails gemarkeerd. Er zijn slechts 414 e-mails gemarkeerd van de 238.281 e-mails. Ondanks deze beperkte hoeveelheid gemarkeerde e-mails heeft dit ervoor gezorgd dat ongeveer drie kwart van alle e-mails uitgefilterd werden als niet zaak gerelateerd. Bij een grotere hoeveelheid gemarkeerde e-mails zou de e-mail content analyse nog nauwkeuriger zijn werk kunnen doen, waardoor er nog meer e-mails uitgefilterd kunnen worden. In verband met het gebrek aan nauwkeurigheid en de enorme onbalans van e-mail verkeer tussen de verschillende medewerkers van Enron Corporation, heeft de e-mail content analyse niet kunnen achterhalen wie de hoofdpersonen waren achter het fraude schandaal.

Er wordt aanbevolen om te onderzoeken hoeveel e-mails er gemarkeerd moeten worden, wil de e-mail content analyse nauwkeurige resultaten geven en hoe deze resultaten fluctueren indien er meer of minder e-mails zijn gemarkeerd. Daarnaast dient ook gekeken te worden naar de meegegeven waarde voor de sterkte van zeldzame woorden en een variabele hoeveelheid woorden dat maximaal gebruikt mag worden om de zaak gerelateerdheid van een e-mail te bepalen. Hierin is het ook interessant hoe de analyse zijn resultaten geeft naarmate hier waarden veranderen.

Er zijn nog vele manieren om de e-mail content analyse doelgerichter en nauwkeuriger te laten werken. Voor de doelgerichtheid kan gebruik gemaakt worden van een zogenaamde woordenboost. Dit houdt in dat bepaalde woorden zwaarder worden meegerekend in de e-mail analyse. Deze woordenlijst kan door de gebruiker zelf worden samengesteld. Om de e-mail analyse nauwkeurigere resultaten te laten geven kan er gebruik gemaakt worden van de analyse van extra eigenschappen van de e-mail. Zoals tekst dat voorkomt in het onderwerp van de e-mail. De meeste online e-mail spamfilters maken ook gebruik van meerdere kenmerken in een e-mail om te bepalen of een e-mail spam is of niet. Daarnaast kunnen ook nog zaakprofielen worden toegepast ter nauwkeurigheid van de resultaten. Dit houdt in dat er per misdaadprofiel (moord, fraude, zedenmisdrijf) een eigen filter bestaat. In ieder verschillend profiel moet immers worden gezocht op een andere set woorden in de e-mails.

Inhoudsopgave

[Management samenvatting 5](#_Toc319921662)

[Hoofdstuk 1 Inleiding 9](#_Toc319921663)

[Hoofdstuk 2 Probleemstelling 11](#_Toc319921664)

[Hoofdstuk 3 Literatuuronderzoek 13](#_Toc319921665)

[3.1 Xiraf 13](#_Toc319921669)

[3.1.1 Algemene beschrijving 13](#_Toc319921670)

[3.1.2 De verwerkingsfase 13](#_Toc319921671)

[3.1.3 De analysefase 14](#_Toc319921672)

[3.2 E-mail filtering 16](#_Toc319921673)

[3.2.1 Spamfilters 16](#_Toc319921674)

[3.2.2 Het Bayesiaanse Spamfilter 16](#_Toc319921675)

[3.2.3 Een variant op het Bayesiaanse Spamfilter 17](#_Toc319921676)

[3.2.4 De stelling van Bayes 18](#_Toc319921677)

[3.2.5 De kansberekening voor een enkele term 19](#_Toc319921678)

[3.2.6 De kansberekening voor meerdere termen 24](#_Toc319921679)

[3.2.7 Zeldzame termen 29](#_Toc319921680)

[3.2.8 Toevoeging van een heuristiek 32](#_Toc319921681)

[Hoofdstuk 4 De e-mail content analyse applicatie 35](#_Toc319921682)

[4.1 De werking van de applicatie 35](#_Toc319921684)

[4.2 De werking van de e-mail content analyse 39](#_Toc319921685)

[Hoofdstuk 5 De Enron testcase 53](#_Toc319921686)

[5.1 Aanpak 53](#_Toc319921688)

[5.2 Resultaten 54](#_Toc319921689)

[5.2.1 Totale analysegegevens: 54](#_Toc319921690)

[5.2.2 Gemarkeerde analysegegevens 55](#_Toc319921691)

[5.2.3 Resultaten na de gelopen analyse 55](#_Toc319921692)

[5.2.4 Grafieken 56](#_Toc319921693)

[Hoofdstuk 6 Conclusies en aanbevelingen 63](#_Toc319921694)

[6.1 Conclusies 63](#_Toc319921696)

[6.2 Aanbevelingen 64](#_Toc319921697)

[6.2.1 Variabele hoeveelheid doorgenomen e-mails 64](#_Toc319921698)

[6.2.2 Een variabele sterkte bij de kansberekening voor zeldzame woorden 64](#_Toc319921699)

[6.2.3 Variabel aantal e-mails bij de heuristiek 64](#_Toc319921700)

[6.3 Uitbreidingen 65](#_Toc319921701)

[6.3.1 Uitbreiding classificatie: Woordenboost 65](#_Toc319921702)

[6.3.2 Uitbreiding classificatie: Headers 65](#_Toc319921703)

[6.3.3 Zaakprofielen 66](#_Toc319921704)

[6.3.4 Overeenkomende e-mail berichten 66](#_Toc319921705)

[6.3.5 Woordsynoniemen 67](#_Toc319921706)

[Literatuurlijst 69](#_Toc319921707)

[Bijlagen 71](#_Toc319921708)

[Bijlage 1 Formules 71](#_Toc319921709)

[Bijlage 2 De gehele e-mail analyse flowchart 77](#_Toc319921710)

[Bijlage 3 Enron Corporation 81](#_Toc319921711)

# Hoofdstuk 1 Inleiding

Sinds 3 augustus 2011 hebben de NCIM Groep, een Nederlands IT-bedrijf, en het Nederlands Forensisch Instituut (NFI) een partnershipovereenkomst.(1) Dit Nederlandse IT-bedrijf gaat zorgdragen voor de verdere nationale en internationale uitrol en ondersteuning van de softwaretoepassing Xiraf. Een softwaretoepassing ontwikkeld door het NFI. Met het sluiten van het partnerschap zet het NFI de trend voort waarbij het, behalve leverancier van forensische diensten ook leverancier wordt van forensische technologie.

Xiraf wordt op dit moment gebruikt door verschillende politie- en opsporingsdiensten.(1) Xiraf is een geavanceerd zoek- en verwerkingssysteem voor digitale sporen. Xiraf maakt digitale sporen eenvoudig doorzoekbaar voor forensische onderzoekers. Voorbeelden van sporen die worden gebruikt en gecatalogiseerd zijn tekstdocumenten, fotomateriaal, videomateriaal, web historie, e-mails en chat-berichten. Deze sporen zijn afkomstig van bijvoorbeeld computers, harde schijven, mobieltjes en digitale foto- en videocamera’s. Na het catalogiseren van bewijsstukken kunnen forensische onderzoekers het bewijs doorzoeken middels een web interface. Eén voordeel hiervan is dat dit ervoor zorgt dat Xiraf op afstand te benaderen is. Pim Takkenberg, teamleider Team High Tech Crime bij het Korps Landelijke Politiediensten (KLPD) geeft nog een voordeel: “Dat is handig als je onderzoek verricht met meerdere partijen. Onlangs gaven we in een internationaal onderzoek nog partijen uit drie andere landen toegang tot de data.”(1)

De praktijk kan een dergelijke tool goed gebruiken, want de hoeveelheid digitale data neemt dagelijks toe. Elke zaak die wordt onderzocht bevat al snel meer dan zeven miljoen sporen, die allemaal in de database moeten worden opgeslagen.(2) Deze datahoeveelheiden zijn soms zo groot dat de server een dag moet ‘stampen’ om alle informatie te vinden en doorzoekbaar te maken. Deze data wordt in één database gestopt waarin de gebruiker bijvoorbeeld op zoektermen of op datum kan zoeken.(3)

Xiraf kan in het kort worden omschreven als:

Een database van digitale sporen met als doel om de enorme hoeveelheid data doorzoekbaar te maken door middel van filters en data analyses om uiteindelijk misdaadzaken sneller op te lossen.

De nationale en internationale uitrol van Xiraf wordt bemoeilijkt door concurrentie van andere software pakketten. Enkele van deze concurrenten zijn Forensic Toolkit (FTK), EnCase en Nuix. Dit zijn allemaal geavanceerde computer forensische software. Met deze software kunnen bijzondere opsporingsdiensten en IT-security professionals grondige en effectieve computeronderzoeken uitvoeren. Deze software pakketten omvatten daarnaast nog de mogelijkheid tot het filteren van bestanden en het uitvoeren van geavanceerde zoekfunctionaliteiten. Forensic Toolkit en EnCase biedt de mogelijkheid om alle verschillende soorten bestanden te onderzoeken, net als XIRAF. Nuix beperkt zich tot en specialiseert zich in e-mail berichten.

Een voordeel van Xiraf is de geautomatiseerde extractie van digitale sporen. Bij concurrerende software pakketten moeten onderzoekers eerst aangeven wat onderzocht moet worden. Een ander voordeel is de open structure van Xiraf. Dit betekent dat er tegenaan geprogrammeerd kan worden.

In hoofdstuk 2 komt de probleemstelling aan de orde. In hoofdstuk 3 komt het literatuuronderzoek ter sprake. Voordat de standalone applicatie en prototype gerealiseerd kan worden, moet eerst onderzoek worden gedaan naar de materie, Xiraf, waarmee gewerkt wordt. Daarnaast wordt onderzoek gedaan naar de werking van e-mail filtering en de mogelijkheden tot het toepassen van een e-mail filter in Xiraf. Het vierde hoofdstuk staat in het teken van de geprogrammeerde applicatie. Hierin is te lezen hoe het programma werkt en hoe de geïmplementeerde analyse werkt voor het filteren van de e-mails. Hierbij wordt gebruik gemaakt van een flowchart. In het vijfde hoofdstuk staat beschreven hoe de applicatie is getest en wat de resultaten van deze tests zijn. In het zesde hoofdstuk staan conclusies, aanbevelingen en mogelijke uitbreidingen van het project. Tot slot volgt de literatuurlijst en de bijlagen. In bijlage I staan alle formules waarvan gebruik gemaakt is gedurende het project. In bijlage II staat de flowchart in zijn geheel. Deze flowchart geeft weer hoe de analyse van het e-mail filteren werkt. In bijlage III staat informatie over Enron Corporation. Enron Corporation staat bekend van één van de grootste en meest complexe fraudeschandalen uit de Amerikaanse historie en ging in 2001 ten onder. In deze bijlage staat beschreven hoe Enron Corporation ten onder ging en aan welke praktijken Enron deelnam. Op basis van deze informatie is grote set e-mails gemarkeerd als zaak of niet zaak gerelateerd.

# Hoofdstuk 2 Probleemstelling

In dit hoofdstuk komt de probleemstelling aan de orde die geleidt heeft tot het formuleren van de afstudeeropdracht.

Zoals beschreven is Xiraf een database die grote hoeveelheden digitale sporen kan inladen en ordenen. Klanten van Xiraf zijn echter alleen geïnteresseerd in de digitale sporen die voor het onderzoek relevant zijn. Hierin kan Xiraf geen ondersteuning bieden.

Voor het vinden van de relevante digitale sporen dienen extra analyses geïmplementeerd te worden door slim gebruik te maken van impliciet aanwezige gegevens. In Xiraf zijn er op dit moment geen analyses beschikbaar bij e-mail berichten. Er zijn verschillende analyses mogelijk, zoals een analyse voor data mining of tekst analyse. Met behulp van deze analyses is het mogelijk om de e-mails een rating te geven die de mate van relevantie weergeeft.

In deze scriptie wordt een wiskundige toepassing ontwikkeld waarna deze als standalone applicatie (prototype) wordt gerealiseerd. Deze applicatie is gemaakt om de werking van de analyse te controleren. De applicatie is standalone omdat het niet gekoppeld is aan Xiraf of andere externe software waarvan de applicatie afhankelijk is. Het was namelijk niet mogelijk om de applicatie met Xiraf te linken op het moment dat het ontwikkelen van de applicatie begon. Bij het maken van deze applicatie is echter wel rekening gehouden met een integratie met Xiraf die in de toekomst zal volgen. De applicatie bevat een flinke dosis wiskunde en daarnaast sluit hij aan op de doelstellingen van Xiraf zelf. Dit is tevens de afbakening waarvan gebruik is gemaakt tijdens het afstudeertraject.

# Hoofdstuk 3 Literatuuronderzoek

In dit hoofdstuk wordt het literatuuronderzoek beschreven. Voordat een wiskundige toepassing kan worden ontwikkeld moet bekend zijn met welke materie gewerkt wordt en waarmee rekening gehouden moet worden. Als dit eenmaal bekend is kan een wiskundige toepassing worden geformuleerd en deze worden uitgewerkt voor implementatie. In de eerste paragraaf van dit hoofdstuk wordt uitgelegd hoe Xiraf werkt en hoe het programma er op het ogenblik uit ziet. In de tweede paragraaf wordt de wiskundige toepassing beschreven en uitgewerkt.



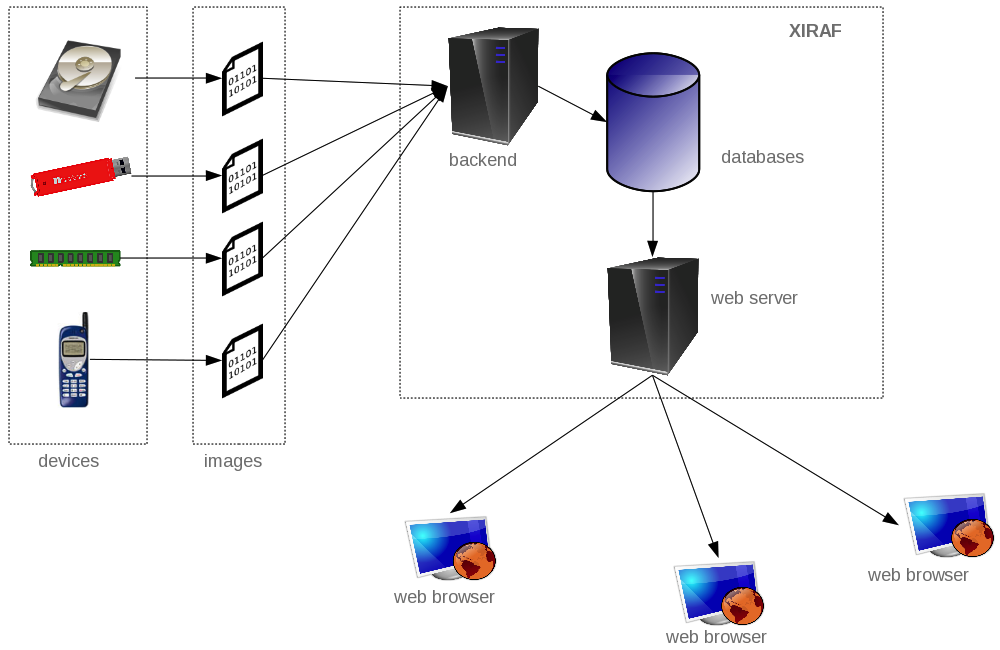
## Xiraf

### Algemene beschrijving

Zoals in de inleiding al is beschreven is Xiraf een geavanceerd zoek- en verwerkingssysteem voor digitale sporen. Xiraf maakt digitale sporen eenvoudig doorzoekbaar voor forensische onderzoekers. Voorbeelden van sporen die worden gebruikt en gecatalogiseerd zijn tekstdocumenten, fotomateriaal, videomateriaal, web historie, e-mails en chat-berichten. Deze sporen zijn afkomstig van bijvoorbeeld computers, harde schijven, mobieltjes en digitale foto- en videocamera’s. Na het catalogiseren van bewijsstukken kunnen forensische onderzoekers het bewijs doorzoeken middels een web interface. Een Xiraf project kan dus worden onderverdeeld in twee fasen. De eerste fase is de verwerkingsfase. In deze fase wordt alle data ingeladen in Xiraf. De tweede fase is de analysefase. In deze fase kan de Xiraf gebruiker (een forensisch onderzoeker) alle ingeladen data filteren om informatie te ontdekken die belangrijk is voor het onderzoek.

### De verwerkingsfase

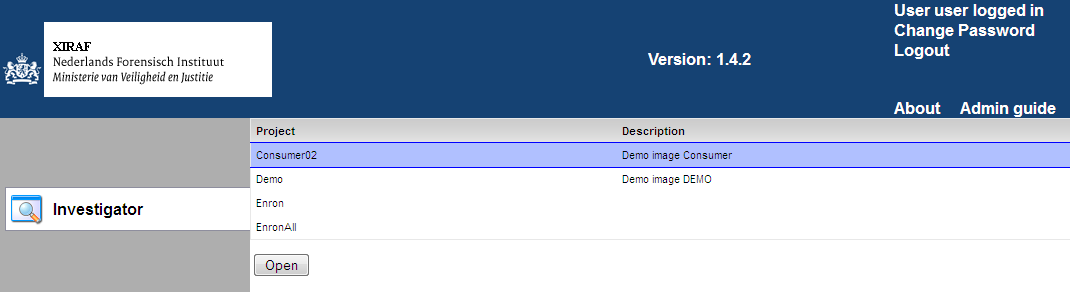
In ieder forensisch onderzoek wordt tussen alle verzamelde sporen gezocht naar bewijsmateriaal, om te gebruiken in de rechtszaal. Deze sporen kunnen digitaal en analoog (fysiek) zijn. Xiraf heeft betrekking op de digitale sporen. Voordat alle data van deze apparaten in de databases van Xiraf wordt geladen, worden images gemaakt van deze apparaten. Een image is een computerbestand waarin een exacte kopie van een opslagmedium is opgeslagen. Deze images of kopieën worden vervolgens in Xiraf geladen waarin ze worden gecatalogiseerd en georganiseerd. Na het inladen, catalogiseren en organiseren van alle data is deze op te vragen door gebruikers via de webbrowser. Een afbeelding van de werking van de verwerkingsfase is te zien in figuur 3.1.



Figuur 3.1: De verwerking van data in Xiraf.(4)

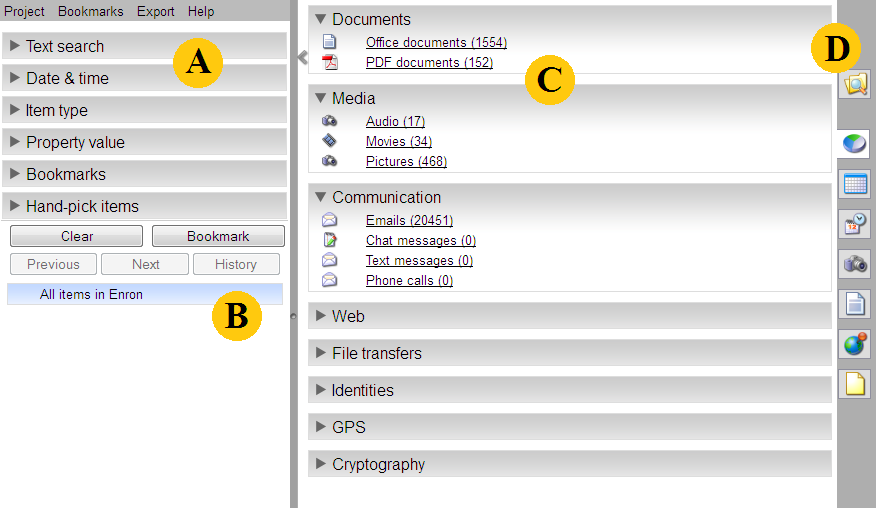
### De analysefase

Vanwege het feit dat Xiraf een web gebaseerde applicatie is, kan deze via een internet browser worden bereikt. Op de startpagina dient te worden ingelogd voordat er gebruik gemaakt kan worden van Xiraf. Nadat er is ingelogd kan er in Xiraf gewerkt worden. Er wordt een scherm getoond zoals te zien is in figuur 3.2.



**Figuur 3.2:** Na het inloggen kan een project worden geselecteerd om te onderzoeken.

In de figuur 3.2 is te zien dat er vier projecten beschikbaar zijn voor de ingelogde gebruiker. De zichtbare projecten kunnen per gebruiker verschillen en worden beheerd door de administrator. Na het selecteren en openen van het gewenste project kan het onderzoek beginnen. In figuur 3.3 is te zien hoe dit scherm eruit ziet, waarin onderzocht kan worden.



**Figuur 3.3:** Na het openen van een project kan alle data in dat project worden onderzocht.

Het onderzoek scherm is opgedeeld in drie kolommen. In de linker kolom staan de filters die gebruikt kunnen worden om de hoeveelheid digitale sporen significant de reduceren. Deze filters zijn te zien bij (**A**). Bij (**B**) vormt zich een lijst van toegevoegde filters. Zoals in figuur 3.3 te zien is, zijn er geen filters actief. Het enige aanwezige filter is ‘All items in Enron’. Hierbij is ‘Enron’ de naam van het geselecteerde project.

In de middelste kolom staan alle digitale sporen die voldoen aan het filter. Deze sporen zijn in verschillende categorieën ingedeeld. Bij (**C**) zijn drie categorieën geopend: ‘Documents’, ‘Media’ en ‘Communication’. In het bovenstaande voorbeeld heeft de categorie ‘Documents’ 1554 Office documenten en 152 PDF documenten die voldoen aan het filter.

Deze middelste kolom kan de digitale sporen op verschillende manieren presenteren. In de rechterkolom, bij (**D**), kan een keuze gemaakt worden betreffende de visualisatie van de digitale sporen. Standaard is de Quick Start visualisatie geselecteerd. Dit geeft de digitale sporen aan op de categorie bestandstype, zoals te zien in figuur 3.3. Als gekozen wordt voor een andere visualisatie, dan past de middelste kolom zich aan deze keuze aan. Enkele voorbeelden van andere visualisaties zijn de geografische voorstelling en de voorstelling met een tijdlijn. De geografische voorstelling wordt voornamelijk gebruikt bij het zoeken naar foto en video materiaal die zijn opgenomen met een camera met ingebouwde gps-functionaliteiten. Met een tijdlijn kan e-mail verkeer gemakkelijker worden uitgelezen.

## E-mail filtering

### Spamfilters

Waarschijnlijk de bekendste toepassing van e-mail filtering betreft het spamfilter. Een spamfilter is een stuk software dat in e-mails zoekt naar tekens van spam in content, eigenschappen en kenmerken van de e-mail. Normaal gezien leest het spamfilter de e-mail in, besluit of de e-mail spam is en zal op basis daarvan een actie uitvoeren. In het geval dat het spamfilter een e-mail als spam markeert kunnen verschillende acties optreden, afhankelijk van het soort spamfilter. Het ene spamfilter zal de e-mail in een ‘Junk’-map of ‘Ongewenste mail’-map zetten, terwijl een ander spamfilter de e-mail ongemerkt verwijderd. In het geval dat een e-mail wordt gemarkeerd als legitieme mail (ook wel ham genoemd, het tegenovergestelde van spam) wordt deze in de inbox geplaatst.

Spamfilters zoeken naar patronen in e-mails om spam te herkennen. Enkele voorbeelden zijn:

* Content gerelateerde patronen – In spam e-mails worden andere woorden gebruikt dan in ham e-mails. Woorden als ‘viagra’ of ‘sex’ komen bijvoorbeeld veel vaker voor in spam e-mails. Als één van deze termen zelfs wordt gebruikt in het onderwerp van de e-mail, is het bericht al verdacht.
* Bijlage gerelateerde patronen – In de hoop de ontvanger van een e-mail een valse bijlage te openen, maken afzenders in de bijlage vaak gebruik van een dubbele extensie met vele spaties. Met een bijlage als ‘Leuke mop.txt .exe’ hoopt de afzender dat de ontvanger denkt dat het een onschuldig txt-bestand is, terwijl het in werkelijkheid een exe-bestand is.

### Het Bayesiaanse Spamfilter

Er zijn wereldwijd vele verschillende spamfilters ontwikkeld en er zijn vele soorten technieken om spam van legitieme e-mail te filteren. Veel van deze ontwikkelde spamfilters maken gebruik van de Bayesiaanse statistiek, sommige spamfilters in combinatie met andere technieken. Heinz Tschabitscher(5), een e-mail enthousiast die zich sinds 1997 bezig houdt met alle aspecten van e-mail en andere vormen van online communicatie, heeft een top 12 gemaakt van zijn persoonlijk beste Windows Spamfilters.(6) Van deze 12 spamfilters maken er 9 gebruik van de Bayesiaanse statistiek. Door het bedrijf GFI worden spamfilters die gebruik maken van de Bayesiaanse statistiek beoordeeld als het meest effectief.(7) Het voordeel van spamfilters die gebruik maken van Bayesiaanse statistiek is de nauwkeurige beoordeling van e-mails, daarentegen moet het filter wel eerst getraind worden. Dit betekent dat de gebruiker eerst van een grote hoeveelheid e-mails moet aangeven of deze als spam of als ham beschouwd moeten worden. Het filter probeert hier dan uit te leren welke eigenschappen van de e-mails vooral bij spam voorkomen, en welke niet. Op basis hiervan worden de e-mails gefilterd.

### Een variant op het Bayesiaanse Spamfilter

In Xiraf, de database voor digitale sporen, worden onder andere e-mails ingeladen. Tijdens forensisch onderzoek wordt deze database van e-mails doorgenomen op zoek naar informatie gerelateerd aan de zaak. Deze e-mails worden door de forensische onderzoekers beoordeeld op basis van de content. Is de e-mail belangrijk voor de zaak of niet? Het Bayesiaanse spamfilter doet bijna exact hetzelfde. Er wordt gekeken naar de content van de e-mail en afhankelijk daarvan wordt besloten om de e-mail te bewaren of niet. Door een dergelijk filter in Xiraf te implementeren kan het beoordelingsproces voor een groot deel worden geautomatiseerd. Het filter zal dan de content van de e-mails doornemen en middels een kansberekening bepalen of deze e-mail belangrijk is voor de zaak of niet.

De beoordeling van het filter is afhankelijk van informatie die in het verleden door de gebruiker is ingegeven. Deze informatie kan gezien worden als gemarkeerde e-mails. In het filter in het online e-mailaccount worden de e-mails gemarkeerd als spam of als ham. Er wordt een database opgesteld van woorden die voorkwamen in die e-mails. Sommige woorden komen vaker voor in spam e-mails dan in ham e-mails, woorden zoals ‘viagra’ en ‘sex’. Als een nieuwe e-mail binnenkomt kan een statistische kans worden berekend dat de e-mail spam is of niet, afhankelijk van de woorden die erin voorkomen. Hiervoor geldt echter wel dat er een grote hoeveelheid e-mails vooraf moeten zijn doorgenomen, zodat het filter beschikking heeft over een grote hoeveelheid informatie. Dit is nodig om het filter nauwkeurig te laten werken. De grootte van deze hoeveelheid e-mails wordt verderop in de scriptie besproken. In het filter in Xiraf moet ook eerst een grote hoeveelheid e-mails worden doorgenomen. Deze e-mails worden echter niet gemarkeerd als spam of ham. De e-mails in Xiraf kunnen worden gemarkeerd als zaak gerelateerd of niet zaak gerelateerd. Net als in het online e-mailaccount wordt een database gemaakt van woorden die voorkomen in mails die zaak gerelateerd zijn of niet. Als deze database van woorden groot genoeg is om nauwkeurig te bepalen of een e-mail zaak gerelateerd is of niet, dan kan dit voor alle aanwezige e-mails in Xiraf worden berekend.

De hoeveelheid e-mails die gemarkeerd moeten worden om de database van woorden op te stellen kan variëren. Bij het vormen van een database van woorden voor e-mails van een online e-mailaccount meldt de ene bron dat er een tijdperiode aan vooraf gaat van twee weken om de database op te stellen,(7) twee andere bronnen melden dat een hoeveelheid van honderden mailtjes moeten worden doorgewerkt.(8,9) Daarnaast wordt er ook aangeraden dat de hoeveelheid gemarkeerde e-mails per categorie gelijk aan elkaar moeten zijn.(10,11)

In het geval van een forensisch onderzoek in Xiraf kan het vinden van een zaak gerelateerde e-mail worden gezien als het zoeken naar een speld in een hooiberg. Stel dat de allereerste zaak waarin het filter gebruikt kan worden een e-mail database heeft van 10.000 e-mails en hiervan zijn 10 e-mails belangrijk voor de zaak. Vanwege het feit dat dit de allereerste zaak is waarin het filter gebruikt kan worden, is de database van woorden nog leeg. Er zijn namelijk nog geen e-mails gemarkeerd als zaak of niet zaak gerelateerd. In deze zaak zullen eerst e-mails gemarkeerd moeten worden om deze database van informatie te voorzien. Los van het feit dat de hoeveelheid e-mails per categorie (zaak gerelateerd en niet zaak gerelateerd) enorm van elkaar verschillen, betreft deze hele set van e-mails niet genoeg informatie om e-mails in e-mail databases in opvolgende zaken betrouwbaar te categoriseren als zaak of niet zaak gerelateerd. Met slechts 10 e-mails die zijn gemarkeerd als zaak gerelateerd kan er nog niet betrouwbaar automatisch e-mails worden geanalyseerd. Voor het goed laten functioneren van de analyse moeten van beide categorieën enkele honderden mails zijn gemarkeerd. Deze database van woorden kan in toekomstige zaken gebruikt blijven worden.

Aangezien er verschillende soorten zaken zijn (zedenzaken, moordzaken, fraudezaken) kan er gekozen worden om voor iedere soort zaak, ook wel profiel, een eigen database van woorden te maken. In fraudezaken, moordzaken en zedenzaken moeten er immers op andere kernwoorden gezocht worden. Eén algemene database zal minder nauwkeurige resultaten geven in vergelijking met profielgerichte databases. Het is ook mogelijk om voor iedere zaak een nieuwe database te maken. Er zijn altijd gegevens die specifiek voor een enkele zaak van toepassing zijn, zoals namen of locaties. Dit betekent echter wel dat er voor iedere zaak een grote hoeveelheid e-mails moet worden doorgenomen voordat het filter weet op welke informatie het moet zoeken. Het is een afweging die gemaakt moet worden tussen de nauwkeurigheid van de e-mail content analyse en de tijdsbesteding voor rechercheurs. Aangezien de werking voor iedere database hetzelfde is, wordt in deze scriptie de focus gelegd op het genereren van een enkele database die nodig is voor de beoordeling van e-mails. Daarnaast zal het filter zich beperken tot de Bayesiaanse statistiek op de content van de e-mails. In latere stadia kunnen andere technieken worden toegevoegd en gecombineerd, zoals in verschillende online spamfilters het geval is.

### De stelling van Bayes

Voor het bepalen of een e-mail gemarkeerd kan worden als zaak gerelateerd of als niet zaak gerelateerd wordt gebruik gemaakt van de stelling van Bayes. Deze stelling geeft de relatie weer tussen de waarschijnlijkheden van A en B en de conditionele waarschijnlijkheden van A gegeven B en B gegeven A. Hierbij staat de term ‘Pr’ voor ‘Probability’, de Engelse term voor kans of waarschijnlijkheid.

De algemene formule is als volgt:

(3.1)

Met de variabelen:

De kans op A onder voorwaarde van B;

De kans op B onder voorwaarde van A;

De kans op A; en

De kans op B.

Deze algemene formule zal de basis vormen voor de variant op het Bayesiaanse spamfilter. Er moet worden berekend wat de kans is dat een bericht zaak gerelateerd is, onder voorwaarde dat er een bepaald woord in staat. Of zelfs, wat is de kans dat een bericht zaak gerelateerd is, onder voorwaarde dat er een set van woorden in staat.

Met de algemene formule kan de kans worden berekend dat een bericht zaak gerelateerd is, onder de voorwaarde dat er een bepaald woord in staat. Neem hiervoor de variabele A als ‘zaak gerelateerde e-mail’ en de variabele B als ‘een willekeurig gekozen woord’. Deze algemene formule gaat echter niet in deze vorm werken op een set van woorden. Om de validiteit van de formules te controleren wordt er gebruik gemaakt van drie verschillende literatuurbronnen(12,13,14). Deze zullen met elkaar worden vergeleken of ze ieder hetzelfde berekenen en indien dit niet het geval is, welke formule het beste past bij onze toepassing.

* De formules van de eerste bron(12) worden in het vervolg de BSF-formules genoemd;
* De formules van de tweede bron(13) worden in het vervolg de BI-formules genoemd;
* De formules van de derde bron(14) worden in het vervolg de VU-formules genoemd;

### De kansberekening voor een enkele term

Ter validatie van de formules worden deze van de verschillende bronnen met elkaar vergeleken. Allereerst de formules om de kans op een zaak gerelateerde e-mail te berekenen, onder voorwaarde dat er één bepaald woord in staat. Hierbij zijn de formules overgenomen in de vorm zoals ze in hun bron vermeldt staan.

In de verschillende formules wordt echter gebruik gemaakt van verschillende variabelen. In de algemene formule, zoals te zien is in paragraaf 3.2.4, wordt gebruik gemaakt van de variabelen A en B. Om de leesbaarheid te verhogen wordt in de formules gebruik gemaakt van vaste variabelen. Deze variabelen zijn als volgt gedeclareerd:

**+** Het plus-teken geeft aan dat we te maken hebben met een zaak gerelateerde e-mail

**–** Het min-teken geeft aan dat we te maken hebben met een niet zaak gerelateerde e-mail

**W** De letter W staat voor een willekeurig gekozen woord.

* De BSF-formule voor de kansberekening dat een e-mail zaak gerelateerd is onder voorwaarde dat er een willekeurig gekozen woord in staat:

(3.2)

* De BI-formule voor de kansberekening dat een e-mail zaak gerelateerd is onder voorwaarde dat er een willekeurig gekozen woord in staat:

(3.3)

* De VU-formule voor de kansberekening dat een e-mail zaak gerelateerd is onder voorwaarde dat er een willekeurig gekozen woord in staat:

(3.4)

Uit de bovenstaande formules is te zien dat de BI-formule en de VU-formule met elkaar overeenkomen, er wordt in de bron echter gebruik gemaakt van een andere schrijfwijze. Tevens komen deze formules exact overeen met de algemene formule, formule 3.1, te zien in

paragraaf 3.2.4. De schrijfwijze van de BSF-formule verschilt in die mate dat niet gelijk te zien is of deze ook overeenkomt. Om te controleren of dezelfde berekening wordt gedaan moet worden bekeken welke variabelen er nodig zijn om de verschillende waarschijnlijkheden uit te rekenen. Eerst bekijken we met een voorbeeld of de uitkomsten overeenkomen.

**Voorbeeld:**

Het filter is getraind met 1000 berichten, waarvan er 100 zaak gerelateerd zijn en 900 niet zaak gerelateerd. Tijdens het trainen is een database van woorden gemaakt. Naast de lijst van woorden heeft het ook per woord bijgehouden hoe vaak deze is voorgekomen in zaak gerelateerde e-mail en in niet zaak gerelateerde e-mail. Een klein deel van deze database is in tabel 3.1 te zien.

**Tabel 3.1:** Een gedeelte uit de database.

|  |  |  |
| --- | --- | --- |
| **Woord** | **Aantal keer in zaak gerelateerde e-mail** | **Aantal keer in niet zaak gerelateerde e-mail** |
| geweer | 60 | 9 |
| boeken | 2 | 80 |
| bende | 35 | 25 |
| twee | 8 | 140 |

Het filter komt in Xiraf een mail tegen met daarin de volgende tekst:

*“Gisteren ben ik met nog twee bende leden een geweer wezen kopen. Heb jij er al een?”*

Moet het filter deze e-mail nu categoriseren onder zaak gerelateerde e-mails of niet zaak gerelateerde e-mails? Stel dat we alleen kijken naar het woord ‘geweer*’* dat voorkomt in de e-mail en we willen weten of de e-mail zaak gerelateerd is of niet. In tabel 3.2 zijn de formules opgesplitst in de verschillende kansen.

**Tabel 3.2:** Berekening van de deelkansen.

|  |  |  |
| --- | --- | --- |
| **Kansberekening** | **Beschrijving** | **Uitkomst** |
|  | De kans dat een e-mail zaak gerelateerd is als het woord ‘geweer’ erin voorkomt. Dit is wat we gaan berekenen. | ? |
|  | De kans dat het woord ‘geweer’ voorkomt in een zaak gerelateerde e-mail. Dit is het aantal keer dat het woord ‘geweer’ is voorgekomen in een zaak gerelateerde e-mail gedeeld door het totaal aantal zaak gerelateerde e-mails. |  |
|  | De kans dat een e-mail zaak gerelateerd is. Dit is het aantal zaak gerelateerde e-mails gedeeld door het totaal aantal e-mails. |  |
|  | De kans dat het woord ‘geweer’ voorkomt in een niet zaak gerelateerde e-mail. Dit is het aantal keer dat het woord ‘geweer’ is voorgekomen in een niet zaak gerelateerde e-mail gedeeld door het totaal aantal niet zaak gerelateerde e-mails. |  |
|  | De kans dat een bericht niet zaak gerelateerd is. Dit is het aantal niet zaak gerelateerde e-mails gedeeld door het totaal aantal e-mails. |  |
|  | De kans dat het woord ‘geweer’ voorkomt. Dit is het totaal aantal keer dat het woord ‘geweer’ is voorgekomen gedeeld door het totaal aantal e-mails. |  |

Met behulp van tabel 3.2 kan de kansberekening worden uitgevoerd voor de BSF-formule, BI-formule en VU-formule. Afhankelijk van de uitkomst zal het filter in Xiraf de e-mail gaat categoriseren als zaak gerelateerd of niet. De uitkomst van de BSF-formule is te zien in formule 3.5 en de uitkomst van de BI-formule en VU-formule is te zien in formule 3.6.

(3.5)

(3.6)

Bij een kans groter dan 0,50 of 50% kan de e-mail worden gecategoriseerd als zaak gerelateerd. Als de kans kleiner of gelijk is aan 0,50 of 50% kan de e-mail worden gecategoriseerd als niet zaak gerelateerd. In het geval dat het filter een e-mail tegenkomt met daarin alleen het woord ‘geweer’, dan zal het filter deze e-mail markeren als zaak gerelateerd. Voor de voorbeeldzin kan nog niet worden gezegd of deze zaak gerelateerd is of niet, er komen namelijk meer woorden in deze zin voor dan alleen het woord ‘geweer’. Hoe dit berekend moet worden komt in paragraaf 3.2.6 aan de orde.

Aan de hand van het voorbeeld is te zien dat de uitkomst hetzelfde is voor alle drie de formules. Dit kan toeval zijn. Om te controleren of de formules exact hetzelfde zijn, worden de deelkansen omschreven in de volgende variabelen:

*a* Het totaal aantal e-mails gemarkeerd als zaak gerelateerd;

(In het voorbeeld zou de waarde voor *a* gelijk zijn aan 100)

*b* Het totaal aantal e-mails gemarkeerd als niet zaak gerelateerd;

(In het voorbeeld zou de waarde voor *b* gelijk zijn aan 900)

*c* Het aantal zaak gerelateerde e-mails waarin het woord voorkomt; en

(In het voorbeeld zou de waarde voor *c* gelijk zijn aan 60)

*d* Het aantal niet zaak gerelateerde e-mails waarin het woord voorkomt.

(In het voorbeeld zou de waarde voor *d* gelijk zijn aan 9)

Met behulp van deze vier variabelen kunnen alle kansberekeningen worden uitgevoerd. In tabel 3.2 zijn de verschillende deelkansen al berekend. In tabel 3.3 wordt gebruik gemaakt van dezelfde deelkansen, maar wordt deze omgeschreven in de vier variabelen die zojuist zijn gedeclareerd. Daarnaast wordt gecontroleerd of de formules overeenkomen met de uitkomsten uit tabel 3.2.

**Tabel 3.3:** Berekening van de deelkansen, omschreven in variabelen.

|  |  |  |
| --- | --- | --- |
| **Kansberekening** | **Formule** | **Controle voorbeeld 1** |
|  | ? | ? |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Nu de deelkansen zijn teruggebracht naar variabelen kan onderzocht worden of er sprake was van toeval of dat de BSF-formule in zijn geheel overeenkomt met de BI-formule en VU-formule. Hiervoor worden de BSF-formule en de overeenkomende BI-formule en VU-formule herhaald. De BSF-formule zag er als volgt uit:

(3.7)

De BI-formule en VU-formule zagen eruit als volgt:

(3.8)

Het verschil tussen beide formules zit in de noemer. Bij de controle of beide formules aan elkaar gelijk zijn moet het volgende gelden:

(3.9)

Door middel van de variabelen in tabel 3.3 kan formule 3.9 omgeschreven worden in variabelen en vervolgens worden herleidt. Deze herleiding is te zien in de formules 3.10 en 3.11. Dit komt er als volgt uit te zien:

(3.10)

(3.11)

Hieruit kan geconcludeerd worden dat de BSF-formule, BI-formule en VU-formule aan elkaar gelijk zijn. Het is nu mogelijk om deze enkele formule te herleiden in variabelen. Deze herleiding is te zien in de formules B1.12 tot en met B1.15 in Bijlage 1. De uitkomst van deze herleiding is te zien in formule 3.12.

(3.12)

Hieruit is af te lezen dat de kans op een zaak gerelateerde e-mail afhankelijk is van slechts twee variabelen, in plaats van de vier die voorheen zijn gedeclareerd. Door middel van het voorbeeld kan worden gecontroleerd of deze uitkomst correct is. Volgens het voorbeeld had de variabele *c* – het aantal zaak gerelateerde e-mails waarin het woord voorkomt – een waarde van 60. De variabele *d* – het aantal niet zaak gerelateerde e-mails waarin het woord voorkomt – had een waarde van 9. Als deze gegevens worden ingevoerd in formule 3.12 komt daar het volgende uit:

(3.13)

### De kansberekening voor meerdere termen

Aangezien e-mails nagenoeg nooit één enkel woord bevatten, heeft de herleidde formule 3.12 op dit punt geen waarde. Het filter heeft een formule nodig dat kan bepalen of een e-mail zaak gerelateerd is afhankelijk van meerdere woorden. In dezelfde bronnen(12,13,14) als in paragraaf 3.2.5 zijn ook formules te vinden die rekening houden met meerdere woorden. Ook hier zijn de formules overgenomen in de vorm zoals ze in hun bron vermeld staan.

Om de leesbaarheid te verhogen wordt in de formules weer gebruik gemaakt van vaste variabelen. Deze variabelen zijn als volgt gedeclareerd:

**+** Het plus-teken geeft aan dat we te maken hebben met een zaak gerelateerde e-mail

**–** Het min-teken geeft aan dat we te maken hebben met een niet zaak gerelateerde e-mail

De letter staat voor het i-de woord uit de verzameling van woorden in de e-mail.

* De BSF-formule voor de kansberekening dat een e-mail zaak gerelateerd is onder voorwaarde dat er meerdere woorden in voorkomen:

(3.14)

Waarvoor geldt:

(3.15)

* De BI-formule voor de kansberekening dat een e-mail zaak gerelateerd is onder voorwaarde dat er meerdere woorden in voorkomen:

(3.16)

Waarvoor geldt:

(3.17)

(3.18)

(3.19)

* De VU-formule voor de kansberekening dat een e-mail zaak gerelateerd is onder voorwaarde dat er meerdere woorden in voorkomen:

(3.20)

Waarvoor geldt:

(3.21)

(3.22)

Hier is al veel minder goed te zien of deze formules ook hetzelfde betekenen. Merk op dat alle drie de formules een andere schrijfwijze gebruiken voor de kansberekening dat de e-mail zaak gerelateerd is of niet, afhankelijk van meerdere termen. De BSF-formule gebruikt zelfs alleen het symbool *p*. Dit heeft te maken met de (on)afhankelijkheid van de verschillende woorden die voorkomen in een e-mail. In paragraaf 3.2.5 was een voorbeeldzin gegeven:

*“Gisteren ben ik met nog twee bende leden een geweer wezen kopen. Heb jij er al een?”*

Als wordt gekeken naar de schrijfwijze in de VU-formule: , wordt daarmee bedoeld dat een e-mail de woorden ‘Gisteren’ **en** ‘ben’ **en** ‘ik’ **en** ‘met’ enz. bevatten? Stel dat er een miljoen e-mails zijn verwerkt, maar dat geen van deze e-mails ál deze woorden bevat. In dat geval kan er geen uitspraak worden gedaan over de zaak gerelateerdheid van de e-mail. Indien er gebruik gemaakt wordt van **of** in plaats van **en**, dan geeft dit ook problemen. Als er wordt gezocht naar e-mails waarin de woorden ‘Gisteren’ **of** ‘ben’ **of** ‘ik’ **of** ‘met’ enz. voorkomen, dan vallen bijna alle e-mails hierbinnen. Voor de kansberekening moeten de woorden onafhankelijk van elkaar worden bepaald. De schrijfwijze in de formules verschillen weliswaar van elkaar, maar de bedoeling is hetzelfde. Vanwege het feit dat bij iedere formule hetzelfde bedoeld wordt, wordt in het vervolg voor iedere formule gebruik gemaakt van de schrijfwijze .

Om te controleren of de drie formules hetzelfde berekenen, worden de formules uitgeschreven in deelkansen waarna deze worden vereenvoudigd in variabelen. Dit maakt het makkelijker om de formules met elkaar te vergelijken.

* De BSF-formule kan worden herschreven in:

(3.23)

* De BI-formule kan worden herschreven in:

(3.24)

* De VU-formule kan worden herschreven in:

(3.25)

Wederom komen de BI-formule en de VU-formule exact met elkaar overeen. De BSF-formule is wederom anders geformuleerd. Om te controleren of alle formules met elkaar overeen komen worden deze weer omgeschreven in variabelen. Deze variabelen zijn in paragraaf 3.2.5 gedefinieerd. In tabel 3.4 wordt herhaald van welke variabelen gebruik gemaakt worden en de betekenis van deze variabelen. In tabel 3.5 worden de deelkansen herhaald met hoe deze omschreven kunnen worden in variabelen.

**Tabel 3.4:** Benodigde variabelen.

|  |  |
| --- | --- |
| **Variabele** | **Betekenis** |
| *a* | Het totaal aantal e-mails gemarkeerd als zaak gerelateerd. |
| *b* | Het totaal aantal e-mails gemarkeerd als niet zaak gerelateerd. |
|  | Het aantal zaak gerelateerde e-mails waarin het i-de woord voorkomt. |
|  | Het aantal niet zaak gerelateerde e-mails waarin het i-de woord voorkomt. |

**Tabel 3.5:** Deelkansen, beschrijving en omschreven in variabelen.

|  |  |  |
| --- | --- | --- |
| **Kansberekening** | **Beschrijving** | **Formule** |
|  | De kans dat een e-mail zaak gerelateerd is als het woord erin voorkomt. Deze formule is in paragraaf 3.2.5 herleid. |  |
|  | De kans dat het woord voorkomt in een zaak gerelateerde e-mail. Dit is het aantal keer dat het woord is voorgekomen in een zaak gerelateerde e-mail gedeeld door het totaal aantal zaak gerelateerde e-mails. |  |
|  | De kans dat een e-mail zaak gerelateerd is. Dit is het aantal zaak gerelateerde e-mails gedeeld door het totaal aantal e-mails. |  |
|  | De kans dat het woord voorkomt in een niet zaak gerelateerde e-mail. Dit is het aantal keer dat het woord is voorgekomen in een niet zaak gerelateerde e-mail gedeeld door het totaal aantal zaak gerelateerde e-mails. |  |
|  | De kans dat een e-mail niet zaak gerelateerd is. Dit is het aantal niet zaak gerelateerde e-mails gedeeld door het totaal aantal e-mails. |  |

Met behulp van deze tabellen kan de BSF-formule worden omgeschreven in:

(3.26)

Formule 3.26 is verder te vereenvoudigen. Deze uitwerkingen zijn te zien in de formules B1.28 tot en met B1.34 in Bijlage 1. De uitkomst van deze vereenvoudiging is te zien in formule 3.27.

(3.27)

Zoals te zien is in formule 3.27 wordt er wederom gebruik gemaakt van slechts twee variabelen. Nu de BSF-formule is vereenvoudigd wordt hetzelfde gedaan voor de BI-formule/VU-formule. Hierbij wordt weer gebruik gemaakt van tabel 3.4 en 3.5. De uitwerkingen zijn te zien in de formules B1.36 tot en met B1.47 in Bijlage 1. De uitkomst van deze vereenvoudiging is te zien in formule 3.28.

(3.28)

In tegenstelling tot de vereenvoudigde BSF-formule worden hier alle vier de variabelen in gebruikt. Deze extra variabelen zijn wegingsfactoren. In het geval dat er meer e-mails zaak gerelateerd zijn of andersom, zorgen deze wegingsfactoren ervoor dat e-mails in de juiste categorie worden ingedeeld, zaak gerelateerd of niet zaak gerelateerd. Aangezien de BSF-formule geen gebruik maakt van deze wegingsfactoren wordt er aangeraden om gebruik te maken van even grote hoeveelheden e-mails per categorie.(10,11) De BSF-formule werkt dus alleen in het geval dat er evenveel zaak als niet zaak gerelateerde e-mails zijn. De BI- en VU-formule werken altijd. Tijdens een forensisch onderzoek mag er nooit worden aangenomen dat er evenveel zaak als niet zaak gerelateerde e-mails zijn, dus wordt doorgewerkt met de BI- en VU-formule.

Ondanks de juiste herleidingen klopt formule 3.28 niet. Dit kan verduidelijkt worden in een voorbeeld. In dit voorbeeld wordt alleen gebruik gemaakt van neutrale woorden. Dit zal ervoor moeten zorgen dat de kans dat een e-mail zaak gerelateerd is, gelijk is aan 0,50 of 50%. In dit voorbeeld wordt gebruik gemaakt van de verhouding 100 zaak gerelateerde e-mails en 900 niet zaak gerelateerde e-mails. In tabel 3.6 staat de data:

**Tabel 3.6:** De benodigde data voor het voorbeeld

|  |  |  |
| --- | --- | --- |
| **Woord** | **Aantal keer in zaak gerelateerde e-mail** | **Aantal keer in niet zaak gerelateerde e-mail** |
| de | 80 | 720 |
| een | 98 | 882 |
| is | 90 | 810 |
| op | 75 | 675 |

Zoals in tabel 3.6 te zien is zijn alle woorden neutraal. Het woord ‘de’ is in 80 van de 100 (80%) zaak gerelateerde e-mails voorgekomen en in 720 van de 900 (80%) niet zaak gerelateerde e-mails. Het woord ‘een’ is in 98 van de 100 (98%) zaak gerelateerde e-mails voorgekomen en in 882 van de 900 (98%) niet zaak gerelateerde e-mails. De verhouding zaak gerelateerd en niet zaak gerelateerd is bij ieder woord gelijk aan elkaar. Als deze gegevens worden ingevuld in formule 3.28, komt daar het volgende uit:

(3.29)

Zoals in formule 3.29 te zien is, komt de uitkomst niet overeen met de verwachtte 0,50. Dit komt door een fout in de wegingsfactor. In formule 3.28 wordt gebruik gemaakt van de wegingsfactoren en . Als men en gebruikt, wordt de wegingsfactor wel juist toegepast. In formule 3.30 wordt van deze wegingsfactoren gebruik gemaakt. Dit leidt tot de uitkomst die werd verwacht bij een e-mail met alleen neutrale woorden:

(3.30)

Dat deze situatie niet alleen optreedt in dit voorbeeld maar in alle gevallen, is te zien in de formules 3.31, 3.32 en 3.33. Hierbij vormen formules 3.31 en 3.32 de startsituatie en formule 3.33 de uitkomst. De uitwerkingen zijn te vinden onder formules B1.50 tot B1.55 in Bijlage I.

Aangezien er sprake is van neutrale woorden met een uitkomst van 0,50 kan worden gesteld:

(3.31)

Waarvoor geldt:

(3.32)

Na enkele vereenvoudigingen komt hier de volgende uitkomst uit:

(3.33)

Hiermee is bewezen dat de wegingsfactor gelijk moet zijn aan en en niet van en . De te gebruiken formule is dus als volgt:

(3.34)

In het geval dat er evenveel e-mails zaak als niet zaak gerelateerd zijn, zal formule 3.34 dus eenzelfde uitkomst moeten geven als de BSF-formule. Dit is gemakkelijk aan te tonen. Neem aan dat er in een onderzoek in totaal y-aantal e-mails zaak gerelateerd zijn. Het totaal aantal niet zaak gerelateerde e-mails is dan ook y-aantal groot. Dit betekend dat voor de variabele a en b allebei y kan worden ingevuld. De verbeterde BI- en VU-formule zullen er hierdoor als volgt uit komen te zien:

(3.35)

(3.36)

Formule 3.36 komt nu exact overeen met de BSF-formule, zoals te zien is in formule 3.27.

### Zeldzame termen

Bij het maken van een statistische berekening kan het probleem ontstaan dat er een woord in de betreffende e-mail voorkomt dat nog niet in voorgaande zaak en/of niet zaak gerelateerde e-mails is voorgekomen. Laten we bijvoorbeeld gebruik maken van het woord *vermoorden*. Stel dat er 20 e-mails in het verleden zijn gevonden met het woord *vermoorden* erin. Al deze 20 e-mails werden gemarkeerd als zaak gerelateerd. Als de passende waarden moeten worden ingevuld in de onderstaande formule, vormt er een probleem.

(3.37)

Met de variabelen:

Het totaal aantal e-mails gemarkeerd als zaak gerelateerd;

Het totaal aantal e-mails gemarkeerd als niet zaak gerelateerd;

Het aantal zaak gerelateerde e-mails waarin het i-de woord voorkomt;

Het aantal niet zaak gerelateerde e-mails waarin het i-de woord voorkomt; en

Het aantal woorden in de e-mail.

Ondanks dat er geen gegevens bekend zijn over de waarden van a, b, n of andere woorden in de e-mail, is al duidelijk dat deze verdeling van 20 tegen 0 een probleem zal veroorzaken. De kans dat de e-mail zaak gerelateerd is komt door deze verdeling uit op exact 100%. Gezien het feit dat er sprake is van een statistische analyse mag en kan de uitkomst nooit exact 100% zijn, hooguit dicht tegen de 100% aan. Formule 3.37 kan worden geschreven in de vorm . Hiervoor zal gelden:

(3.38)

(3.39)

Aangezien er in het verleden geen enkele niet zaak gerelateerde e-mail is voorgekomen met daarin het woord *vermoorden*, zal y gelijk zijn aan 0. De verhouding die dan nog overblijft is , en dat is gelijk aan 1, of 100%. Het kan ook andersom voorkomen, dat een woord nog geen enkele keer in zaak gerelateerde e-mails is voorgekomen. In dit geval verandert de formule in . Dat is gelijk aan 0 of 0%. Er is ook nog de mogelijkheid dat een woord helemaal nieuw is in de database. Dit zal nog grotere problemen veroorzaken. In de andere gevallen wordt een e-mail als zaak gerelateerd (100%) of als niet zaak gerelateerd (0%) gemarkeerd door de analyse. In het geval van een nieuw woord zal de analyse de berekening doen. Delen door 0 bestaat niet, hierdoor zal het programma crashen en de analyse niet kunnen uitvoeren.

Voor het omgaan met zeldzame termen is er een formule gemaakt.(12,15) Deze is als volgt:

(3.40)

Met de variabelen:

De gecorrigeerde kans dat een e-mail zaak gerelateerd is onder voorwaarde dat woord W erin voorkomt;

De kans dat een e-mail zaak gerelateerd is onder voorwaarde dat woord W erin voorkomt;

De kans dat een e-mail zaak gerelateerd is;

Het aantal keer dat het woord W minstens moet zijn voorgekomen voordat er met vertrouwen kan worden gezegd dat de e-mail zaak of niet zaak gerelateerd is, in plaats van dat er sprake is van toeval;

Het aantal e-mails waarin het woord is voorgekomen.

De variabele wordt als constante gebruikt. In beide bronnen(12,15) over zeldzame termen wordt een suggestie gedaan over de in te vullen waarde. In de ene bron(12) wordt de suggestie gesteld om deze op 3 te zetten, in de andere bron(15) is deze 1. Voor dit project is gekozen om de middenweg te kiezen, .

In formule 3.40 wordt omschreven als het aantal e-mails waarin het woord is voorgekomen. Deze kan worden omgeschreven in . Het aantal zaak en niet zaak gerelateerde e-mails waarin het woord voorkomt bij elkaar opgeteld.

Als de formule weer wordt omgeschreven in de bekende variabelen , , , en , dan geeft dit de onderstaande formule. Er is sprake van en en niet en , omdat de gecorrigeerde kans betrekking heeft op een enkel woord en niet op meerdere woorden.

(3.41)

(3.42)

(3.43)

Zeldzame termen in e-mails zorgen ervoor dat er bij implementatie geen gebruik gemaakt kan worden van de verbeterde BI- en VU-formule of de vereenvoudigde vorm ervan. Het probleem van zeldzame termen zorgt ervoor dat er per woord moet worden gekeken wat de kans is dat deze in een zaak gerelateerde e-mail voorkomt. Oftewel . De BSF-formule is wel hierop ingespeeld.

Om toch gebruik te maken van een formule die zowel rekening houdt met zeldzame woorden als ongelijkheden in zaak en niet zaak gerelateerde e-mails, moet de verbeterde BI- en VU-formule worden geschreven in termen van , zoals in de BSF-formule. Bij de analyse van de BSF-formule, de BI-formule en de VU-formule voor meerdere termen is gebleken dat er bij de BSF-formule geen gebruik wordt gemaakt van een wegingsfactor. Door echter terug te beredeneren kan deze wegingsfactor in de BSF-formule worden toegevoegd. Dit is nodig om te kunnen werken met zeldzame termen en de wegingsfactor samen. De verwerkingen hiervan zijn te zien in de formules B1.65 tot en met B1.72. De uitkomst van deze verwerkingen is te zien in formule 3.44.

(3.44)

De formule die te zien is in formule 3.44 is de verbeterde BI- en VU-formule, geschreven in de vorm van de BSF-formule. Dit zorgt ervoor dat er per woord geanalyseerd kan worden en rekening gehouden kan worden met zeldzame termen. Daarnaast is de wegingsfactor in de formule meegenomen.

De formule waarvan gebruik gemaakt kan worden is:

(3.45)

Met de variabelen:

Het totaal aantal e-mails gemarkeerd als zaak gerelateerd;

Het totaal aantal e-mails gemarkeerd als niet zaak gerelateerd;

Het aantal zaak gerelateerde e-mails waarin het i-de woord voorkomt;

Het aantal niet zaak gerelateerde e-mails waarin het i-de woord voorkomt;

Het totaal aantal woorden in de e-mail;

De kans dat de e-mail zaak gerelateerd is onder voorwaarde dat het i-de woord erin voorkomt.

Waarvoor geldt:

(3.46)

(3.47)

### Toevoeging van een heuristiek

Een probleem dat nu kan optreden is de zogenaamde ‘overflow’ en ‘underflow’. In alle programmeertalen, zo ook C# en VBA, komen data types voor. Voor het maken van berekeningen in het algoritme zijn data types nodig die zowel hele hoge als hele kleine waarden kunnen bevatten. Deze zijn er uit formule 3.45 als volgt uit te halen:

(3.48)

(3.49)

(3.50)

(3.51)

De eerste twee ( en ) kunnen extreem hoge waarden krijgen. Na verloop van tijd zal de woordendatabase gevuld zijn door de vele gecategoriseerde e-mails. Neem aan dat er op een gegeven moment 10.000 e-mails zijn gecategoriseerd, hiervan bijvoorbeeld 500 () als zaak gerelateerd en 9.500 () als niet zaak gerelateerd. Tijdens de analyse worden de e-mails één voor één geanalyseerd en bepaald of deze e-mail zaak of niet zaak gerelateerd is. Deze e-mails kunnen kort en lang zijn. Er kan een e-mail tussen zitten met bijvoorbeeld 400 () woorden. Als deze gegevens in de formules worden ingevuld, komt hier het volgende uit:

(3.52)

(3.53)

Het derde en vierde item ( en ) kunnen extreem lage waarden krijgen. Bij het analyseren van de database kan er een e-mail tussen zitten met bijvoorbeeld 400 () woorden. Stel dat deze e-mail niet zaak gerelateerd is, de kans op een zaak gerelateerde e-mail per woord ligt tussen de 0 en 0,5. Bij een extreem niet zaak gerelateerde e-mail ligt de gemiddelde kans bijvoorbeeld op een waarde van 0,2. In dit geval wordt het volgende berekend:

(3.54)

(3.55)

In verschillende programmeertalen is *Double* het data type dat de grootste dataverzameling voor getallen kan hebben. Deze getallenverzameling loopt van tot voor positieve getallen en van tot voor negatieve getallen.(16) Twee van de vier berekende waarden vallen buiten deze getallenverzameling. Dit zal er voor zorgen dat er een overflow of underflow in het programma optreedt. Het grootste probleem zit bij de formules 3.52 en 3.53, de overflow. Bij een overflow is een getal zo hoog dat het programma dit niet meer begrijpt. In het geval van een underflow is de waarde zo klein dat deze wordt gezien als 0.

Om dit probleem te omzeilen wordt er gebruik gemaakt van een heuristiek. In deze heuristiek is de achterliggende gedachte dat neutrale woorden als ‘de’, ‘een’, ‘is’, ‘in’ of equivalenten daarvan in andere talen genegeerd kunnen worden, omdat deze nauwelijks tot geen waarde toevoegen aan de beoordeling van de e-mails. Sommige bayesiaanse spamfilters maken gebruik van deze heuristiek.(12,17) Door gebruik te maken van deze heuristiek kan de hoeveelheid woorden die worden gebruikt in de formule drastisch worden verminderd. Dit leidt weer tot het voorkomen van overflows en underflows. Om deze neutrale woorden te detecteren kan gebruik gemaakt worden van de volgende formules:

(3.56)

(3.57)

In het geval dat er een woord is voorgekomen in zowel zaak als niet zaak gerelateerde e-mails, kan hier geen gebruik gemaakt worden van . Dit kan niet, omdat hierin geen wegingsfactoren zijn meegenomen. Een willekeurig gekozen woord kan bijvoorbeeld in 20 zaak gerelateerde e-mails zijn voorgekomen en ook in 20 niet zaak gerelateerde e-mails. De formule geeft dan een uitkomst van 0,5. Een neutraal woord dus? Nee, dat hoeft niet zo te zijn. Er zijn bijvoorbeeld 40 zaak gerelateerde e-mails doorgenomen en 1000 niet zaak gerelateerde e-mails. Het willekeurig gekozen woord komt dus in 50% van alle zaak gerelateerde e-mails voor en in 2% van alle niet zaak gerelateerde e-mails. Hierin is duidelijk te zien dat het woord geen neutraal woord is. In het geval van een zeldzaam woord kan er wederom gebruik gemaakt worden van , de formule voor de kansberekening dat een woord zaak gerelateerd is.

In het geval van een neutraal woord, zal de individuele kans dat het woord voorkomt in een zaak gerelateerde e-mail rond de 0,50 liggen. In de formule zal een waarde uitkomen dicht bij of exact zijn aan 0. Indien er sprake is van een woord die duidelijk in zaak of niet zaak gerelateerde e-mails voorkomen, dan geeft de formule een waarde dicht tegen de 0,5 als uitkomst. Door middel hiervan kan bijvoorbeeld een top 10 gemaakt worden voor de belangrijkste woorden. Deze woorden hebben de meeste invloed op de uitkomst van de zaak of niet zaak gerelateerde kans van de e-mail. Het algoritme maakt gebruik van de top 30 belangrijkste woorden die voorkomen in de e-mail, om op basis daarvan te bepalen of de e-mail gecategoriseerd moet worden als zaak of als niet zaak gerelateerd. Er is gekozen voor 30 woorden op basis van de volgende berekeningen:

(3.58)

(3.59)

(3.60)

Deze berekeningen laten zien hoeveel e-mails er door de gebruikers gecategoriseerd mogen worden voordat er een overflow optreed. In het geval dat gebruik gemaakt wordt van de top 50 belangrijkste woorden (formule 3.60) mogen er maximaal 1,46 miljoen e-mails door de gebruikers worden gecategoriseerd naar zaak en niet zaak gerelateerd. In het geval van 30 woorden zijn dit ruim 18 en een half miljard e-mails.

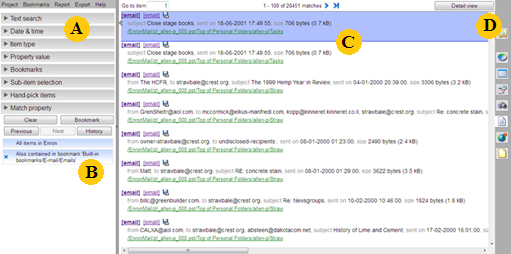
# Hoofdstuk 4 De e-mail content analyse applicatie

In dit hoofdstuk komt de gemaakte standalone applicatie ter sprake. Deze applicatie is gemaakt om de formules te valideren of ze werken in een applicatie. In de eerste paragraaf wordt de werking en interface van de applicatie behandeld. In de tweede paragraaf wordt ingegaan op de werking van de e-mail content analyse. Hierin staat beschreven welke stappen gevolgd worden om de analyse correct uit te voeren.



## De werking van de applicatie

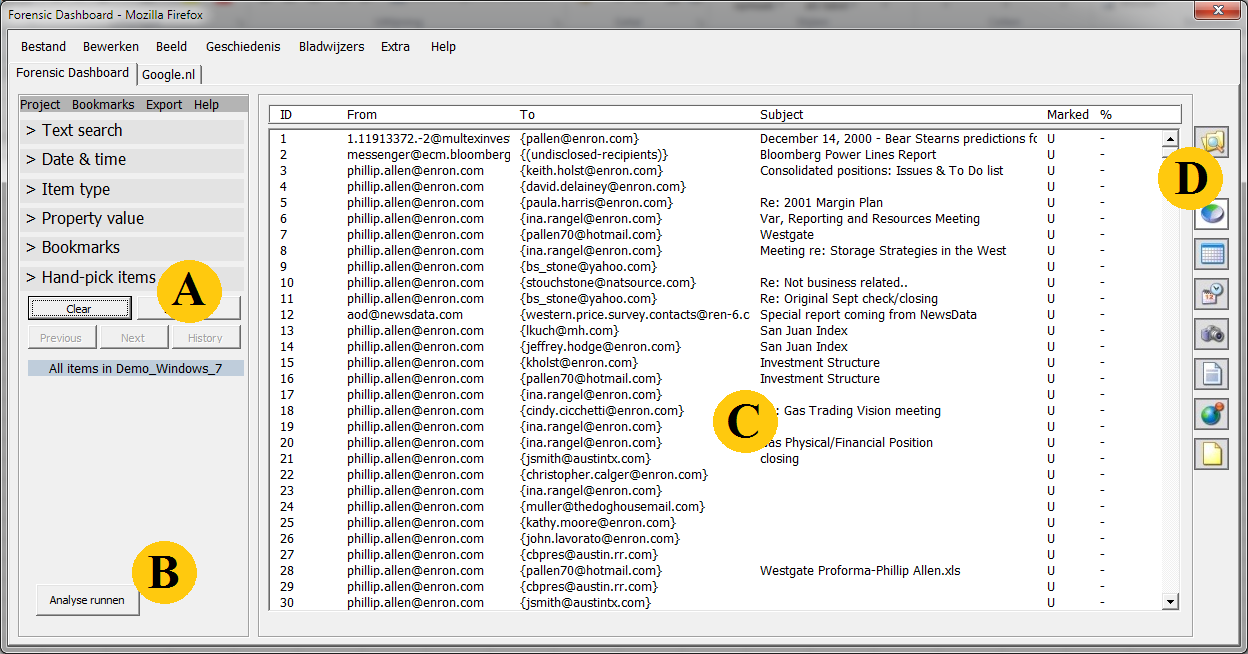
Voor het gebruik en de herkenning van de applicatie is een interface gebouwd die overeenkomt met de interface van Xiraf versie 1.4.2. In hoofdstuk 3 was in figuur 3.2 en 3.3 te zien hoe deze algemene interface eruit ziet. Aangezien de applicatie betrekking heeft op het onderdeel e-mail, wordt in figuur 4.1 weergegeven hoe deze interface eruit ziet in Xiraf.



**Figuur 4.1**: De interface in Xiraf bij onderzoek naar e-mail in Summary-view

Figuur 4.1 komt bijna exact overeen met figuur 3.3 in hoofdstuk 3. In de linker kolom staan de filters die gebruikt kunnen worden om de hoeveelheid e-mails significant te reduceren. Deze filters zijn te zien bij (**A**). Bij (**B**) vormt zich een lijst van toegevoegde filters. De twee actieve filters zijn ‘All items in Enron’ en ‘Also contained in bookmark Built-in bookmarks/E-mail/Emails’. Deze filters zorgen ervoor dat alle e-mail berichten in de middelste kolom komen te staan van de zaak ‘Enron’. Te zien bij (**C**). In de rechterkolom, bij (**D**), kan een keuze gemaakt worden betreffende de visualisatie van de e-mail berichten. Deze visualisatie is de opzet van de middelste kolom. Een voorbeeld van een andere visualisatie is de voorstelling met een tijdlijn. Dit zorgt ervoor dat alle e-mail berichten op volgorde van datum worden geplaatst en onder hun betreffende datum.

In figuur 4.2 is de interface te zien die is gemaakt met behulp van Visual Basic for Application (VBA) in Microsoft Excel 2010. Hierin is het middelste gebied het belangrijkst.



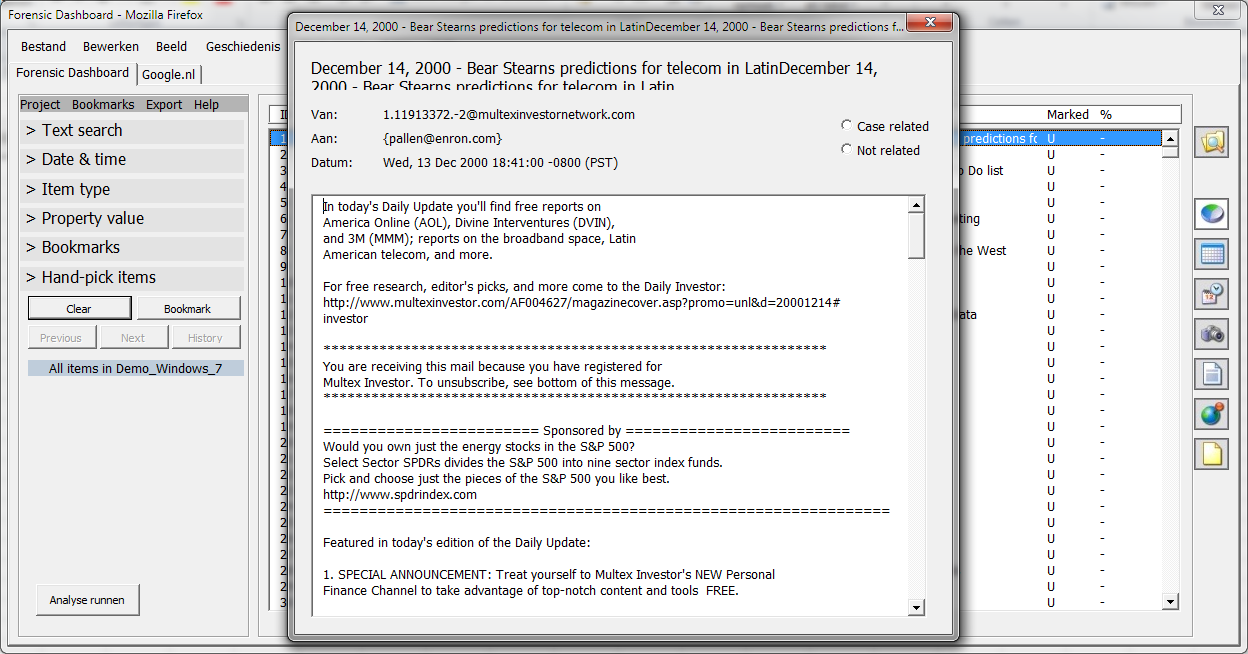
**Figuur 4.2**: De nagemaakte interface van Xiraf in Microsoft Excel 2010.

Net als figuur 4.1 staan in de linker kolom de filters bij (**A**). Deze filters zijn niet belangrijk voor de applicatie, dus staat dit er slechts voor het uiterlijk. Er kunnen geen filters worden toegevoegd door op balken op knopjes in deze regio in te drukken. Bij (**B**) is een wijziging te zien ten opzichte van figuur 4.1. Er staat hier een knop met de tekst ‘Analyse runnen’. Als hierop wordt gedrukt worden alle ingeladen e-mails, te zien bij (**C**), op content geanalyseerd. Bij (**D**) staan weer de verschillende mogelijkheden van visualiseren. Net als de filters is dit toegevoegd voor het uiterlijk en gebeurd er niets als er op één van deze plaatjes gedrukt wordt.

Bij het opstarten van de interface wordt informatie ingeladen over een grote set e-mails. Deze informatie is afkomstig uit Microsoft SQL Server Management Studio. Microsoft SQL Server Management Studio een programma dat te vergelijken is met Microsoft Access. Het kan grote hoeveelheden informatie opslaan, onder andere in databases. Daarnaast is het in Microsoft SQL Server Management Studio mogelijk om gemakkelijk en snel informatie uit te lezen. In Microsoft SQL Server Management Studio is een database gemaakt met informatie over 238.281 e-mails. Deze worden bij het opstarten van de applicatie allemaal ingeladen. Deze set e-mails is afkomstig uit een totale set van ongeveer een half miljoen e-mails. Deze grote groep e-mails heeft vroeger deel uitgemaakt van het onderzoek naar het fraudeschandaal van Enron Corporation in 2001. Het is waarschijnlijk de enige set ‘echte’ e-mails die publiekelijk gemaakt is. In deze set van een half miljoen e-mails kwamen veel dubbele e-mails voor. Na het filteren van deze dubbele e-mails zijn er nog 238.281 e-mails overgebleven. Meer hierover is te lezen in hoofdstuk 5.

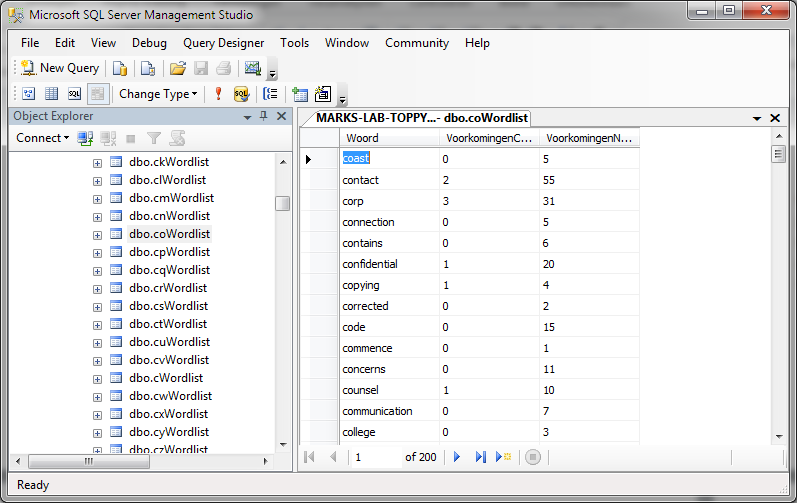
De ingeladen informatie in de interface is verdeeld over 6 kolommen. In de eerste kolom (vanaf links) zijn de ID nummers van de e-mails weergegeven. In de tweede kolom staat de afzender van de e-mail en in de derde kolom naar wie de e-mail is verstuurd. In de vierde kolom staat het onderwerp van de e-mail en in de vijfde kolom staat de markering van de e-mail. Deze markering kan de letter ‘U’ (Unknown), ‘C’ (Case Related) en ‘N’ (Not Related) hebben. Deze markering is gegeven door de gebruiker. In de zesde en laatste kolom komen percentages te staan. Deze percentages representeren de kans dat een e-mail zaak gerelateerd is. In figuur 4.2 zijn nog geen percentages gegeven, omdat de analyse nog niet is gestart.

Op iedere regel in de interface staat informatie over een e-mail. Als de gebruiker dubbelklikt op een e-mail wordt deze ingeladen en geopend in een nieuw scherm. Een voorbeeld hiervan is te zien in figuur 4.3. De gebruiker kan hier doornemen wat er in de e-mail staat en afhankelijk daarvan de e-mail markeren als zaak gerelateerd of als niet zaak gerelateerd.



**Figuur 4.3**: Een willekeurig geopende e-mail uit de database.

Op het moment dat een e-mail wordt gemarkeerd als zaak gerelateerd of als niet zaak gerelateerd worden de woordendatabases bijgewerkt en de database die bijhoudt hoeveel e-mails er zijn gemarkeerd als zaak gerelateerd of niet zaak gerelateerd. Deze databases worden ook in Microsoft SQL Server Management Studio bijgehouden. In figuur 4.4 is een voorbeeld gegeven hoe zo een woordendatabase eruit ziet.

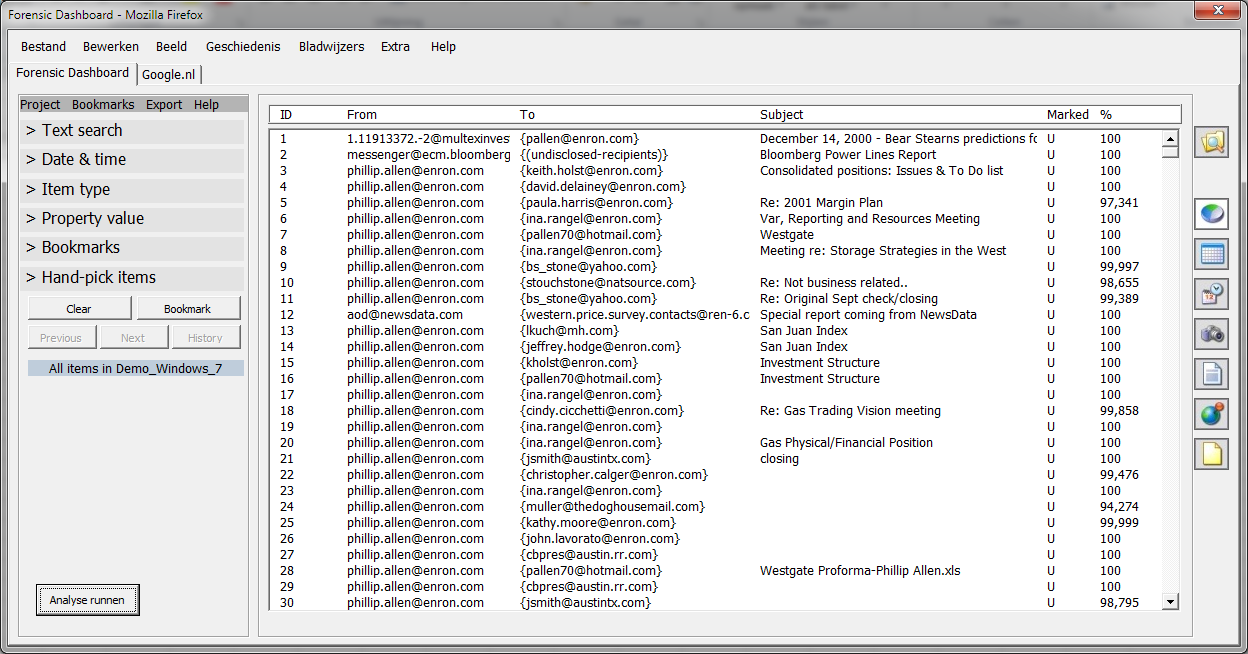


**Figuur 4.4**: Een woordendatabase in Microsoft SQL Server Management Studio.

Zoals in figuur 4.4 te zien is, is in 2 zaak gerelateerde e-mails het woord ‘contact’ voorgekomen. In 55 niet zaak gerelateerde e-mails is dit woord voorgekomen. Deze gegevens zijn gebaseerd op 414 gemarkeerde e-mails, waarvan 19 als zaak gerelateerd en 395 als niet zaak gerelateerd. Als er een nieuwe e-mail wordt gemarkeerd als zaak gerelateerd en het woord ‘contact’ komt in deze e-mail voor. Dan zal de waarde 2 worden opgehoogd naar 3. Indien er in die e-mail een woord voorkomt dat nog niet in zijn bijbehorende woordendatabase voorkomt, dan wordt deze hieraan toegevoegd en daarbij ook meteen een waarde van 1 voor het aantal zaak gerelateerde e-mails waarin het woord is voorgekomen en 0 voor het aantal niet zaak gerelateerde e-mails waarin het woord is voorgekomen.

Deze woordendatabases zijn gebaseerd op de eerste twee karakters van een woord. Dit is gedaan om te zorgen voor een betere performance tijdens de analyse. Het woord ‘geweer’ komt bijvoorbeeld in de database genaamd ‘geWordlist’. Net als woorden als ‘gevecht’, ‘gevangen’, ‘geest’ enzovoort. Woorden die ook beginnen met de letters ‘ge’. In totaal kunnen er 702 (26x27) databases worden aangemaakt. Er zijn 676 (26x26) combinaties voor de eerste twee karakters, waarbij allebei de karakters en letter is van a tot z. De andere 26 databases is voor woorden als ‘e-commerce’. Dit woord begint met een letter waarna een symbool volgt. Het woord ‘e-commerce’ komt in de ‘eWordlist’.

Nu er e-mails door de gebruiker zijn gemarkeerd als zaak gerelateerd of niet zaak gerelateerd, kan de analyse worden uitgevoerd. Deze wordt uitgevoerd over de gehele database van e-mails die vooraf zijn ingeladen. Het enige dat verandert voor de gebruiker is de laatste kolom in de interface. Na de analyse staan hierin de percentages, de kans dat de e-mail zaak gerelateerd is. Dit is te zien in figuur 4.5. Veel van deze e-mails krijgen een percentage van 0 of van 100 procent. Bedenk echter dat deze waarden zijn afgerond. Bij het berekenen van kansen is nooit iets exact 0% of exact 100%, slechts een waarde hier heel dicht tegen aan.

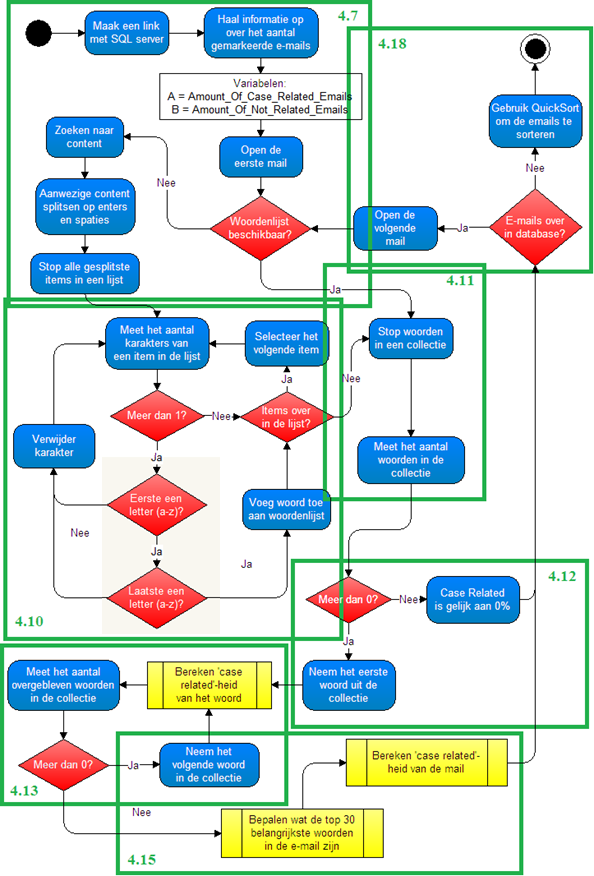


**Figuur 4.5**: De interface na het inladen en analyseren van de e-mails.

Na de analyse worden alle e-mails automatisch gesorteerd. Deze sortering is gebaseerd op de berekende percentages dat een e-mail zaak gerelateerd is en sorteert deze percentages van hoog naar laag. De e-mails die het meest waarschijnlijk gerelateerd zijn aan de zaak komen dus bovenin de database te staan en de e-mails die het meest waarschijnlijk niet gerelateerd zijn aan de zaak komen onderin. Voor het sorteren wordt gebruik gemaakt van een algoritme genaamd Quick Sort. Dit is één van de snelste algoritmes om een grote hoeveelheid waarden te sorteren.(25)

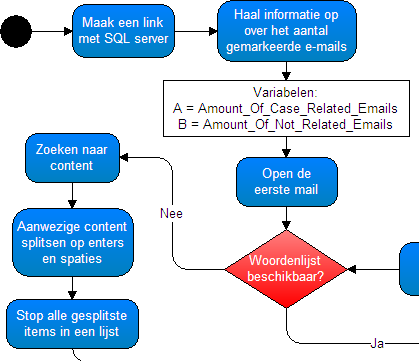
## De werking van de e-mail content analyse

In deze paragraaf wordt besproken hoe het algoritme werkt voor het analyseren van de e-mails. Er wordt hierbij gebruik gemaakt van activity diagrams, ook wel flowcharts. Allereerst wordt het hoofd activity diagram weergegeven. Deze is te zien in figuur 4.6. Dit activity diagram is in verschillende gebieden ingedeeld. Deze zijn afgebakend door een groen vierkant. Hoe de e-mail analyse werkt wordt per gebied behandeld. Het beginpunt van het activity diagram is bij het geheel zwarte rondje. Het zwarte rondje met de witte ring is het eindpunt.



**Figuur 4.6**: Het activity diagram met aangegeven gebieden.

Het beginpunt van het activity diagram in figuur 4.6 is links bovenin te zien. Dit punt bevindt zich in het gebied 4.7, dit betekend dat dit gebied apart te zien is in figuur 4.7.



**Figuur 4.7**: Het startgedeelte uit het activity diagram. In dit deel wordt in SQL Server opgezocht of de e-mail voorheen al is geanalyseerd.

De analyse begint met het creëren van een link met Microsoft SQL Server Management Studio. Alle informatie die nodig is om de e-mail content analyse uit te voeren is in Microsoft SQL Server Management Studio te vinden. In Microsoft SQL Server Management Studio zijn 3 soorten databases gemaakt:

* E-mail informatie database

In deze database staan de 238.281 e-mails met informatie over deze e-mails. Deze informatie is: ID-nummer, bron (de locatie van de e-mail op de harde schijf), afzender, aantal ontvangers, ontvangers, onderwerp, datum van versturen, relatie van de e-mail aan de zaak (U = Unknown, C = Case Related, N = Not Related), het aantal woorden en de lijst van woorden. Een voorbeeld hiervan is verderop te zien in figuur 4.9.

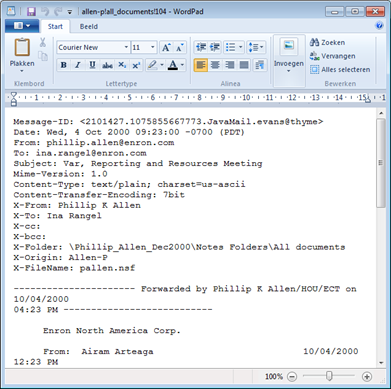
* Woorden databases

Een voorbeeld hiervan was al gegeven in figuur 4.4. Deze databases zijn gesplitst op de eerste 2 karakters van een woord. Naast de woorden, wordt ook bijgehouden in hoeveel zaak gerelateerde e-mails het woord is voorgekomen en in hoeveel niet zaak gerelateerde e-mails het woord is voorgekomen.

* ‘Aantal markeringen’ database

Dit is de kleinste soort database van de drie. In deze database wordt bijgehouden hoeveel e-mails zijn gemarkeerd als zaak gerelateerd en hoeveel e-mails zijn gemarkeerd als niet zaak gerelateerd.

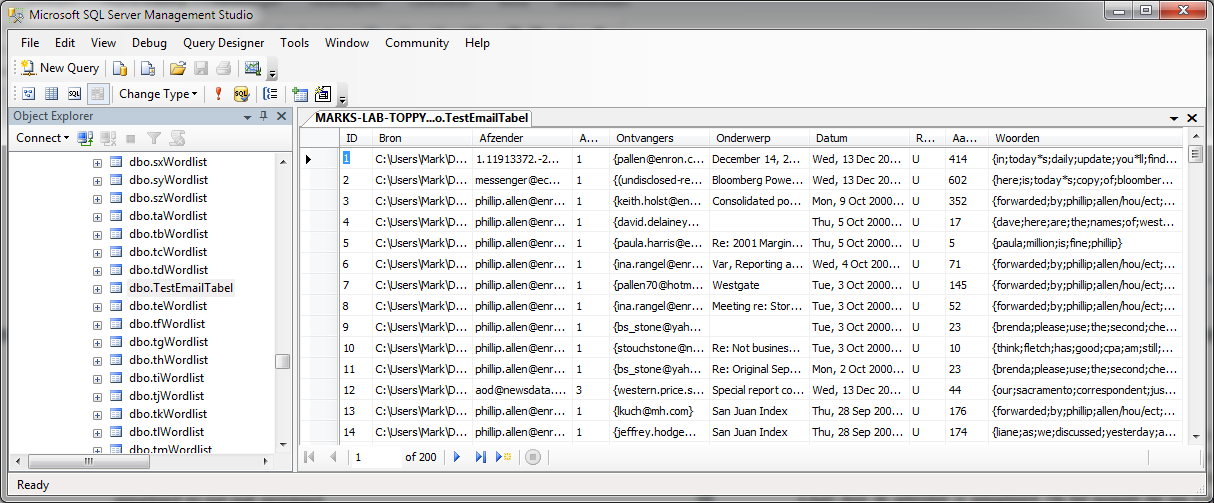
Na het linken van Microsoft Excel 2010 met de Microsoft SQL Server Management Studio wordt uitgelezen hoeveel e-mails zijn gemarkeerd als zaak gerelateerd (in de variabele: Amount\_Of\_Case\_Related\_Emails) en hoeveel e-mails zijn gemarkeerd als niet zaak gerelateerde (in de variabele: Amount\_Of\_Not\_Related\_Emails). Vervolgens wordt de eerst e-mail uit de database geopend. Deze e-mail berichten staan op de harde schijf en zijn te openen in verschillende tekstverwerkers. In figuur 4.8 is een voorbeeld gegeven van een e-mail bericht, geopend in WordPad.



**Figuur 4.8**: Een voorbeeld van een origineel e-mail bericht.

Al deze e-mails beginnen met eigenschappen over de e-mail. Enkele eigenschappen zijn de datum waarom de e-mail is verstuurd, wie de afzender is van de e-mail en naar wie deze e-mail gestuurd is. Deze eigenschappen zijn ook te vinden in Microsoft SQL Server Management Studio. Er staan echter ook eigenschappen in als de bestandsnaam van een mogelijke bijlage. Deze informatie is niet in Microsoft SQL Server Management Studio opgenomen, omdat deze informatie niet relevant is voor het onderzoek. De op internet te downloaden e-mail dataset bevat namelijk geen bijlagen.

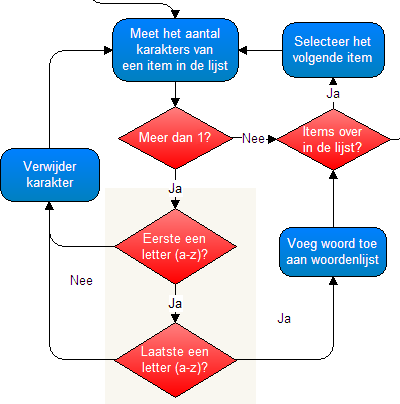
In Microsoft SQL Server Management Studio wordt bij iedere e-mail een lijst van woorden opgeslagen die in de e-mail aanwezig zijn. Een voorbeeld hiervan is te zien in figuur 4.9. Deze lijst van woorden kan op twee verschillende manieren in Microsoft SQL Server Management Studio worden geladen. De eerste manier is op het moment dat een e-mail als zaak of niet zaak gerelateerd wordt gemarkeerd. Tijdens het markeren van een e-mail moet een lijst van woorden worden opgesteld die aanwezig zijn in die e-mail. Deze lijst is nodig om de woorden databases bij te werken op het moment dat een e-mail wordt gemarkeerd. Deze woordenlijst wordt ook aangemaakt op het moment dat de e-mail content analyse wordt afgespeeld. Dit gebeurd als er nog geen woordenlijst beschikbaar is.



**Figuur 4.9**: Een gedeelte van de e-mail informatie database in SQL Server.

Na het openen van de e-mail wordt gecontroleerd of er in Microsoft SQL Server Management Studio een woordenlijst beschikbaar is voor de betreffende e-mail. De eerste keer dat deze analyse gestart wordt, hebben alleen de gemarkeerde e-mails beschikking over een woordenlijst. De meeste e-mail berichten beschikken waarschijnlijk niet over een dergelijke woordenlijst. Dit is afhankelijk van de hoeveelheid e-mails die zijn gemarkeerd. Wat er gebeurd als een e-mail wel een woordenlijst ter beschikking heeft, komen we later op.

Als de e-mail nog geen woordenlijst ter beschikking heeft, wordt de e-mail regel voor regel doorgenomen. Zoals in figuur 4.8 te zien is, bevatten de bovenste regels in de e-mail informatie over deze e-mail. Vervolgens volgt er een witregel waarna de content van de e-mail te zien is. De analyse kan deze witregel herkennen en hierdoor weten wanneer deze de content van de e-mail heeft bereikt. Deze content wordt gesplitst op enters en spaties. Al deze gesplitste items worden in een lijst gestopt waarna deze items worden gecontroleerd. Hoe dit in zijn werking gaat is te zien in figuur 4.10.



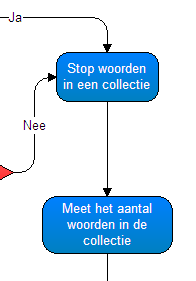
**Figuur 4.10**: Een gedeelte uit het activity diagram, hierin worden alle woorden die in de e-mail voorkomen gevalideerd en in een woordenlijst gezet.

De reden waarom deze items eerst moeten worden gecontroleerd voordat de items in de woordenlijst komen, is dat gelijke woorden ook daadwerkelijk bij elkaar horen in de database van woorden. Een zin sluit normaal gewoonlijk af met een punt en in deze zin is het mogelijk dat er komma’s voorkomen. Bij het splitsen van woorden op enters en komma’s blijven deze nog aan de woorden vastgeplakt.

Bij het doornemen van de lijst met items wordt allereerst gecontroleerd of deze meer dan één karakter heeft, minimaal twee dus. Een woord moet namelijk minimaal twee karakters bevatten wil deze in een woorden database worden opgenomen. Deze woorden databases zijn namelijk gesplitst op de eerste twee karakters van het woord. Indien een item begint of eindigt met een karakter dan geen letter van a tot z is, wordt dit karakter verwijderd. Pas als het voldoet aan de volgende eisen, mag het item worden gezien als een woord en wordt deze in de woordenlijst toegevoegd:

* Minimaal 2 karakters lang
* Het eerste karakter is een letter, a tot z
* Het laatste karakters is een letter, a tot z

Als de hele lijst van items is doorgenomen gaan we door in het algoritme. Het vervolg is te zien in figuur 4.11.

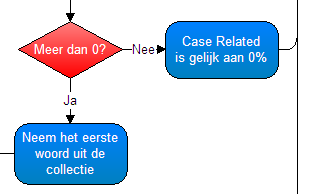


**Figuur 4.11**: Een gedeelte uit het activity diagram, hierin wordt de woordenlijst gefilterd op unieke woorden en de hoeveelheid unieke woorden geteld.

Het blok ‘Stop woorden in een collectie’ kan op twee manieren worden bereikt:

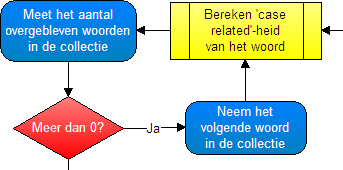
* De e-mail heeft in Microsoft SQL Server Management Studio al beschikking over een lijst van woorden die in de e-mail staan. Via deze manieren hoeft er geen lijst van woorden gemaakt te worden, zoals hiervoor behandeld.
* De e-mail heeft in Microsoft SQL Server Management Studio geen beschikking over een lijst van woorden. Deze lijst van woorden is gemaakt na het controleren van gesplitste items in de e-mail.

In Visual Basic for Applications (VBA) bestaat een zogenaamde ‘collectie’. Dit kan vergeleken worden met een array, een lijst van items. Bij een collectie is het mogelijk om alleen unieke items aan de collectie toe te voegen. In sommige e-mails staan enkel cijfercombinaties, een telefoonnummer bijvoorbeeld. Dit zorgt ervoor dat geen enkel item is geaccepteerd als woord, er is in dit geval dus geen content qua tekst in de e-mail aanwezig. Voor een content analyse is de zaak gerelateerdheid dan 0%. Dit is te zien in figuur 4.12.



**Figuur 4.12**: Een gedeelte uit het activity diagram, hierin wordt gecontroleerd hoeveel unieke woorden in de e-mail voorkomen.

Nadat deze e-mail een beoordeling heeft gekregen van 0%, wat betreft zaak gerelateerdheid, is de content analyse voor deze e-mail afgerond en kan de volgende e-mail worden geanalyseerd. Indien er wel content aanwezig is in de e-mail wordt het algoritme verder doorlopen. Dit vervolg is te zien in figuur 4.13.



**Figuur 4.13**: Een gedeelte uit het activity diagram, hierin wordt voor ieder woord berekend in welke mate het woord zaak gerelateerd is.

In dit vervolg wordt voor ieder woord in de collectie berekend wat de kans is dat dit woord voorkomt in een zaak gerelateerde e-mail en wat de belangrijkheid is van het woord. Met de belangrijkheid wordt bedoeld in wat voor mate het woord zaak gerelateerd of niet zaak gerelateerd is. Neutrale woorden horen een lage belangrijkheid te krijgen en woorden die zeer zaak gerelateerde of zeer niet zaak gerelateerd zijn een hoge belangrijkheid. Deze waarden zijn nodig om de heuristiek te laten werken. Het gele blok in figuur 4.13 houdt een sub proces in. Dit proces is te zien in figuur 4.14.

Voor het berekenen van de kans dat het woord zaak gerelateerd is en de belangrijkheid van het woord, moet er informatie gehaald worden uit Microsoft SQL Server Management Studio. Er wordt allereerst gezocht of het woord al bestaat in de bijbehorende woorden database. Mocht het woord niet gevonden worden, dan mag het woord geen effect hebben op de kans voor de gehele e-mail. De kans dat het woord zaak gerelateerd is (in de variabele: Chance\_For\_Word\_To\_Be\_Case\_Related) wordt gelijk gesteld aan 0,5 en de belangrijkheid van het woord (in de variabele: Word\_Importance) wordt gelijk gesteld aan 0. Indien het woord wel in de database wordt gevonden, dan zoeken we het aantal voorkomingen voor het woord op. Het aantal zaak gerelateerde e-mails waarin het woord is voorgekomen wordt in de variabele Amount\_Of\_Case\_Related\_Occurenences gestopt en voor niet zaak gerelateerde e-mails in de variabele Amount\_Of\_Not\_Related\_Occurences. In figuur 4.14 worden deze variabelen ook gelijk gesteld aan de variabelen C en D om de formules overzichtelijk te houden. Indien het woord in zowel zaak gerelateerde als niet zaak gerelateerde e-mails is voorgekomen wordt gebruik gemaakt van de volgende twee formules:

(4.01)

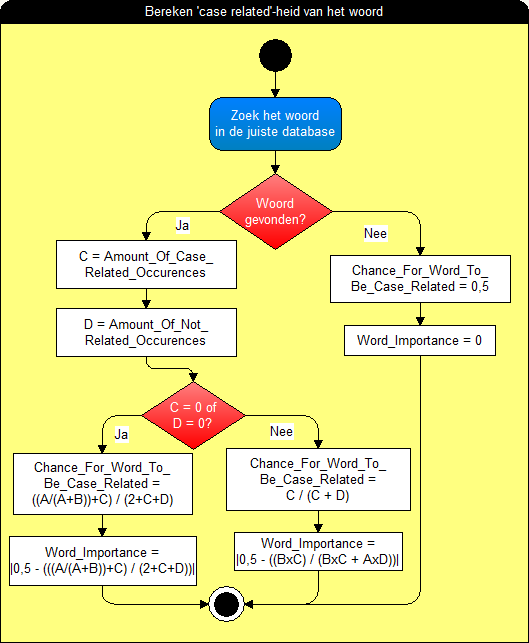
(4.02)

De variabelen A en B in formule 4.02 waren de waarden voor het aantal gemarkeerde e-mails als zaak gerelateerd en het aantal gemarkeerde e-mails als niet zaak gerelateerd. Deze zijn in het begin van het algoritme al gedefinieerd.

Indien het woord alleen nog maar is voorgekomen in zaak gerelateerde e-mail of alleen nog maar in niet zaak gerelateerde e-mail, moet er gebruik gemaakt worden van twee andere formules. Deze zijn als volgt:

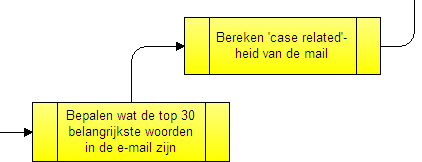
(4.03)

(4.04)



**Figuur 4.14**: In dit sub proces wordt berekend in welke mate het woord zaak gerelateerd is en hoe belangrijk het woord is bij de kansberekening voor de gehele e-mail.

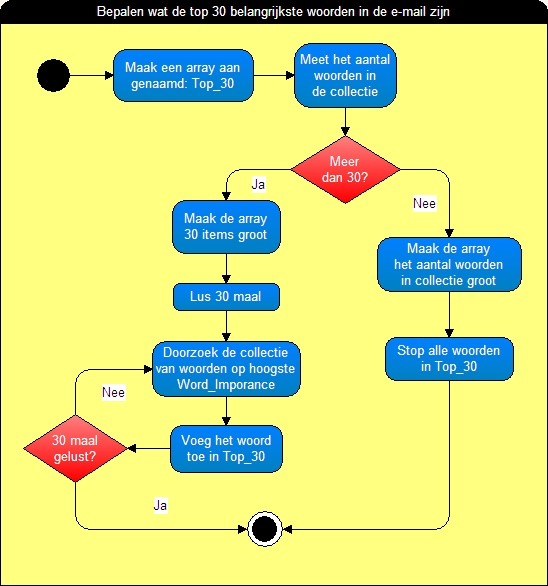
Als alle woorden in de collectie een waarde hebben meegekregen voor belangrijkheid en zaak gerelateerdheid gaat het algoritme door naar de volgende instructie. Dit is te zien in figuur 4.15.



**Figuur 4.15**: Een gedeelte uit het activity diagram, hierin wordt gebruik gemaakt van de heuristiek, waarna wordt berekend in welke mate de e-mail zaak gerelateerd is.

In het eerste sub proces wordt bepaald wat de 30 belangrijkste woorden zijn die in de e-mail voorkomen. Dit heeft dus betrekking op de toegevoegde heuristiek om te voorkomen dat er een overflow of underflow optreedt in het programma. Als bekend is welke 30 woorden het belangrijkste zijn om te bepalen of een e-mail zaak of niet zaak gerelateerd is, kan deze kans worden berekend. In figuur 4.16 is weergegeven hoe wordt bepaald wat de belangrijkste 30 woorden zijn uit de collectie van woorden.

Allereerst wordt een array (een lijst items) aangemaakt, die we voor het gemak ‘Top\_30’ noemen. Bij het maken van zo een array moet worden aangegeven hoeveel items deze kan opslaan. Vanwege het feit dat het aantal woorden in de collectie variabel is, moet eerst worden gekeken hoeveel woorden in deze collectie zitten. Indien er meer dan 30 woorden in de collectie zitten, geven we de array een grootte van 30 items. Indien dit er minder zijn, bijvoorbeeld 20, dan geven we de array een grootte van 20 items. Al deze 20 woorden kunnen meteen in de array worden toegevoegd, want ze behoren allemaal tot de top 30 van hoogste belangrijkheid. In het geval dat er meer dan 30 woorden in de collectie zitten, dan komt het algoritme in een lus terecht. In deze lus wordt 30 maal naar het woord gezocht met de hoogste waarde van Word\_Importance. Om ervoor te zorgen dat niet iedere keer hetzelfde woord wordt gevonden, wordt de Word\_Importance veranderd naar 0 nadat het woord is toegevoegd aan de Top\_30.

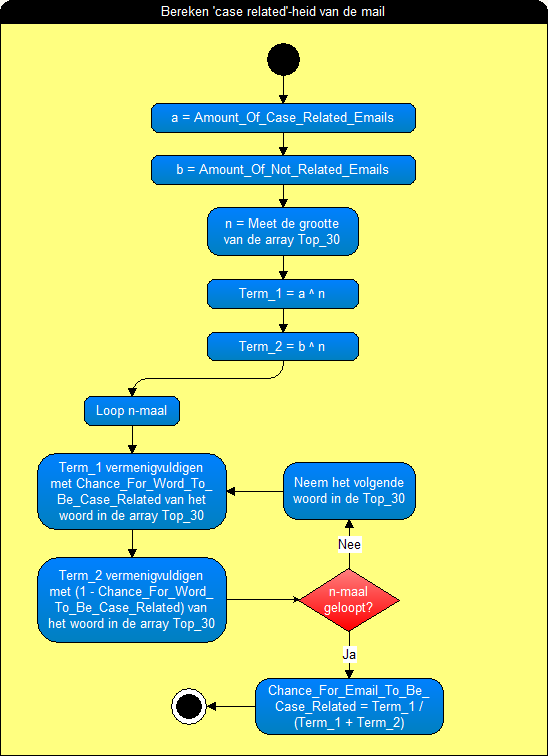


**Figuur 4.16**: In dit sub proces wordt een lijst opgesteld van woorden die het belangrijkst zijn voor de beoordeling van de e-mail.

Deze Top\_30 wordt ook in het volgende sub proces gebruik om te berekenen wat de kans is dat de e-mail zaak gerelateerd is. Dit is te zien in figuur 4.17 en beschrijft in stappen hoe de formule wordt uitgevoerd, die in hoofdstuk 3 is gemaakt. Deze formule ging als volgt:

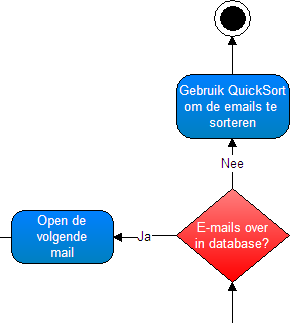
(4.05)

De waarden voor de variabele a en b zijn in het begin van de analyse al ingelezen, deze staan in figuur 4.17 ter definitie. De variabele Term\_1 staat voor de teller in formule 4.05. Deze is gelijk aan het eerste gedeelte van de sommatie in de noemer van de breuk. De variabele Term\_2 is gelijk aan het tweede gedeelte van de sommatie in de noemer van de breuk.



**Figuur 4.17**: In dit sub proces wordt berekend wat de kans is dat de e-mail gerelateerd is aan de zaak.

Nadat dit sub proces is doorlopen is berekend wat de kans is, dat de e-mail zaak gerelateerd is. Dan komen we aan bij het laatste gedeelte van het algoritme. Dit is te zien in figuur 4.18. Er wordt gecontroleerd of er nog e-mails in de database staan die nog geanalyseerd moeten worden. Indien dit het geval is, wordt de volgende e-mail geopend en geanalyseerd. Indien dit niet het geval is, wordt het sorteeralgoritme Quick Sort aangeroepen om de database te sorteren op zaak gerelateerdheid. E-mails met een hoge zaak gerelateerdheid komen bovenaan en e-mails met een lage zaak gerelateerd komen onderaan. Als de e-mails zijn gesorteerd is de analyse klaar. Dit is te zien aan de zwarte cirkel met de witte ring. Dit staat voor het einde van het proces.



**Figuur 4.18**: Een gedeelte uit het activity diagram, hierin wordt gecontroleerd of er nog e-mails overgebleven zijn ter analyse.

# Hoofdstuk 5 De Enron testcase

Enron Corporation – bekend van één van de grootste en meest complexe fraudeschandalen uit de Amerikaanse historie – ging in 2001 ten onder met een schuld van ongeveer 20 miljard Amerikaanse dollars. Na het faillissement werd een grootschalig onderzoek ingesteld. In dit onderzoek is onder andere gebruik gemaakt van het e-mail verkeer binnen Enron om erachter te komen wie schuldig zijn aan fraude en samenzwering. In dit hoofdstuk wordt gebruik gemaakt van de openbaar gemaakte e-mail dataset om te controleren of de grote spelers van het fraudeschandaal gevonden kunnen worden door gebruik te maken van de ontwikkelde analyse. In de eerste paragraaf wordt beschreven van welke aanpak gebruik gemaakt is om erachter te komen welke personen betrokken waren bij het fraudeschandaal van Enron. In de tweede paragraaf worden de resultaten gegeven.



## Aanpak

In het verleden zijn verschillende analyses uitgevoerd op de Enron e-mail dataset. Deze dataset omvat e-mailberichten van ongeveer 150(23) gebruikers, voornamelijk van het hoger management van Enron. Dit zijn ongeveer een half miljoen(23) e-mails zonder bijlagen. Voor zover bekend de enige ‘echte’ openbare e-mail set. Veel van de analyses op deze set e-mails betreffen link- of social network analyse. Jeffrey Heer heeft een social network analyse gedaan op deze Enron e-mail dataset.(24) Dit heeft hij gedaan om erachter te komen wie de hoofdpersonen waren bij het fraudeschandaal van Enron. Hierin is het voornamelijk belangrijk wie e-mails met elkaar uitwisselen en het onderwerp van deze e-mails, waarbij de nadruk wordt gelegd op de personen die informatie met elkaar uitwisselen. Nu is het de vraag of de beschreven variant op het bayesiaanse spamfilter in deze scriptie, ook kan aanwijzen wie de meest betrokken personen waren bij het fraudeschandaal van Enron.

In de e-mail dataset van ongeveer een half miljoen e-mails komen dubbele berichten voor die vooraf nog verwijderd moeten worden. Als de ene persoon een e-mail stuurt naar een andere persoon, dan staat exact dezelfde e-mail bij twee personen in hun e-mail account. Voor de analyse zijn alleen unieke e-mails van belang. Na het uitfilteren van alle dubbele e-mails zijn er 238.281 e-mails overgebleven. Het filter moet deze e-mails analyseren op zaak gerelateerdheid. Voordat dit kan gebeuren moet het filter leren welke woorden in e-mails belangrijk zijn en welke niet. Om het filter dit te leren moet een significante hoeveelheid e-mails worden doorgenomen en worden gemarkeerd als zaak gerelateerd of als niet zaak gerelateerd. Des te meer e-mails worden gemarkeerd, des te betrouwbaarder het filter resultaten zal geven. Wel moet rekening worden gehouden met welke e-mails worden gekozen om te markeren. In de gehele dataset zitten e-mails van ongeveer 150 verschillende gebruikers. Het zou bijvoorbeeld niet handig zijn om de e-mails van 10 willekeurige personen te markeren. Met slechts een paar personen die zijn berecht voor fraude en samenzwering is er een vrij grote kans dat deze 10 personen geen van allen hierbij betrokken zijn. Hierdoor is ervoor gekozen om van alle unieke e-mails een groep willekeurig gekozen e-mails te onderzoeken om te markeren. Als dit is gebeurd kan het filter in werking worden gesteld om zo de waarschijnlijk zaak gerelateerde e-mails te scheiden van de waarschijnlijk niet zaak gerelateerde e-mails.

Op dit punt zijn alle e-mails door het filter gecategoriseerd als zaak gerelateerd of als niet zaak gerelateerd. Alleen de zaak gerelateerde e-mails zijn vanaf dit moment belangrijk. Iedere e-mail die is gecategoriseerd als zaak gerelateerd heeft een afzender en mogelijk meerdere ontvangers. Al deze personen zijn voor deze testcase ‘verdacht’. Nu is het de vraag welke persoon de meeste zaak gerelateerde e-mails heeft verstuurd of ontvangen. Deze persoon zal hoogst waarschijnlijk betrokken zijn bij het fraudeschandaal van Enron. Er zijn echter meerdere personen betrokken bij het fraudeschandaal van Enron, dus wordt een grafiek gemaakt van de top 10 meest geadresseerde personen in zaak gerelateerde e-mails. Met behulp hiervan kan worden gecontroleerd of het filter zijn werk heeft gedaan.

## Resultaten

Om te valideren of de heuristiek en formules correct werken in een applicatie is gebruik gemaakt van de hierboven genoemde testcase. Het doel is om te kijken of de e-mail content analyse e-mails kan opsporen die te maken hebben met het fraude schandaal.

### Totale analysegegevens:

* Totaal aantal e-mails: 517.424
* Aantal unieke e-mail berichten: 238.281
* Aantal e-mail berichten uitgefilterd: 279.143

De online te downloaden e-mail dataset bevat ongeveer een half miljoen e-mails. Om precies te zijn: 517.424. In deze dataset komen vele e-mails meerdere keren voor. Als een gebruiker een e-mail stuurt naar 10 collega’s, dan komt dezelfde e-mail 11 keer voor in de dataset. Alleen unieke e-mail berichten zijn nodig voor de e-mail content analyse. Bij het opsporen van dubbele e-mails is gecontroleerd of het onderwerp, de datum en tijd van versturen en de afzenders en ontvangers met elkaar overeenkomen. Na het uitfilteren van dubbele e-mails zijn 238.281 unieke e-mail berichten overgebleven. Er zijn dus 279.143 e-mails uitgefilterd.

### Gemarkeerde analysegegevens

* Aantal e-mails gemarkeerd: 414
* Aantal e-mails gemarkeerd als zaak gerelateerd: 19
* Percentage e-mails gemarkeerd als zaak gerelateerd: 4,58%
* Aantal e-mails gemarkeerd als niet zaak gerelateerd: 395
* Percentage e-mails gemarkeerd als niet zaak gerelateerd: 95,42%

Voordat de content e-mail analyse zijn werking goed kan doen, moet eerst informatie worden doorgegeven aan het filter. Deze informatie wordt doorgegeven door e-mails te markeren als zaak gerelateerd of niet zaak gerelateerd. Het filter kijkt naar de woorden die in deze e-mails voorkomen en kan daaruit opmaken welke woorden belangrijk zijn voor de zaak en welke woorden dit niet zijn. Des te meer e-mail berichten zijn gemarkeerd, des te nauwkeuriger het filter zal werken. Een woord dat relatief vaak voorkomt in zaak gerelateerde e-mails en relatief weinig in niet zaak gerelateerde e-mails, kan worden gezien als een belangrijk woord. In totaal zijn 414 e-mail berichten doorgenomen, hiervan zijn 19 e-mails gemarkeerd als zaak gerelateerd. Dit is 4,58% van alle gemarkeerde e-mail berichten. De overige 395 e-mails zijn gemarkeerd als niet zaak gerelateerd. Dit is 95,42% van alle gemarkeerde e-mails berichten. Met deze informatie zal de e-mail content analyse in werking worden gesteld. Er is gebruik gemaakt van de informatie uit Bijlage 3 voor het markeren van e-mails.

### Resultaten na de gelopen analyse

* Aantal e-mails geanalyseerd: 238.281
* E-mails geanalyseerd als zaak gerelateerd: 60.389
* Percentage zaak gerelateerde e-mails na de e-mail content analyse: 25,34%
* E-mails geanalyseerd als neutraal: 1.663
* Percentage neutrale e-mails na de e-mail content analyse: 0,70%
* E-mails geanalyseerd als niet zaak gerelateerd: 176.229
* Percentage niet zaak gerelateerde e-mails na de e-mail content analyse: 73,96%

Voor alle e-mail berichten is een kans uitgerekend, namelijk de kans dat de e-mail zaak gerelateerd is. Bij een uitkomst van boven de 50% wordt de e-mail gezien als zaak gerelateerd. Bij een uitkomst van exact 50% wordt de e-mail gezien als neutraal en met een uitkomst onder de 50% wordt de e-mail gezien als niet zaak gerelateerd. Na het analyseren van de 238.281 unieke e-mail berichten zijn er 60.389 e-mails beoordeeld als zaak gerelateerd. Dit zijn 25,34% van alle geanalyseerde e-mail berichten. Van de 238.281 geanalyseerde e-mails zijn er 1.663 beoordeeld als neutraal, 0,70% van het totaal aantal geanalyseerde e-mails. Er zijn 176.229 e-mails beoordeeld als niet zaak gerelateerd, een percentage van 73,96% van het totaal.

* Aantal e-mails gemarkeerd: 414
* Aantal e-mails gemarkeerd als zaak gerelateerd: 19
  + Aantal e-mails hiervan geanalyseerd als zaak gerelateerd: 19 (100%)
  + Aantal e-mails hiervan geanalyseerd als neutraal: 0 (0%)
  + Aantal e-mails hiervan geanalyseerd als niet zaak gerelateerd: 0 (0%)
* Aantal e-mails gemarkeerd als niet zaak gerelateerd: 395
  + Aantal e-mails hiervan geanalyseerd als zaak gerelateerd: 64 (16%)
  + Aantal e-mails hiervan geanalyseerd als neutraal: 0 (0%)
  + Aantal e-mails hiervan geanalyseerd als niet zaak gerelateerd: 331 (84%)

Er zijn in totaal 414 e-mails gemarkeerd, waarvan er 19 als zaak gerelateerd en 395 als niet zaak gerelateerd. In een perfecte verwerking van de e-mail content analyse, zouden alle gemarkeerd e-mails als zaak gerelateerd ook in deze categorie moeten zijn geanalyseerd. Hetzelfde geldt voor de verwerking van de niet zaak gerelateerd e-mails. De 19 e-mails die zijn gemarkeerd als zaak gerelateerd, zijn in de analyse ook allemaal beoordeeld als zaak gerelateerd. Van de 395 als niet zaak gerelateerde e-mails, zijn er 64 (16%) geanalyseerd als zaak gerelateerd. De overige 331 (84%) e-mails zijn geanalyseerd als niet zaak gerelateerd.

### Grafieken

Na de gelopen analyse is het mogelijk om voor de verschillende e-mail adressen te meten hoe vaak deze zijn voorgekomen in e-mails die zijn beoordeeld als zaak gerelateerd, neutraal of niet zaak gerelateerd. Het aantal keer dat een e-mail adres voorkomt kan worden gesplitst naar afzenders en ontvangers, maar ook samen worden genomen. Bij een nauwkeurige e-mail content analyse moeten de e-mail adressen van personen die te maken hebben gehad met het fraude schandaal bij Enron naar voren komen.

In figuur 5.1 is een grafiek te zien die weergeeft welke e-mail adressen de meeste zaak gerelateerde e-mails heeft verstuurd. Jeff Dasovich heeft het meeste aantal e-mails verstuurd die na de e-mail content analyse als zaak gerelateerd zijn beoordeeld. Jeff Dasovich heeft in totaal 5.195 e-mails verstuurd, waarvan er 1.903 zijn beoordeeld als zaak gerelateerd, 12 als neutraal en 3.280 als niet zaak gerelateerd. Daarnaast is Jeff Dasovich ook de persoon die het meeste aantal e-mails heeft verstuurd van iedereen. De aantallen van de andere personen uit figuur 5.1 zijn te zien in tabel 5.1.

**Figuur 5.1**: De top 10 meest voorkomende e-mail adressen voor uitgaand e-mailverkeer.

In tabel 5.1 zijn de resultaten van e-mail adressen toegevoegd van verschillende veroordeelde medewerkers van Enron Corporation.(20,26,27,28,29) De e-mail adressen van Jeff Skilling, Kenneth Lay, Andrew Fastow, Richard Causey, Dan Boyle, Mark Koenig, Ben Glisan, Michael Kopper, Tim Belden, Larry Lawyer en Ken Rice. Hetzelfde gebeurt ook voor tabel 5.2 en tabel 5.3.

**Tabel 5.1**: De gegevenstabel die hoort bij de grafiek in figuur 5.1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Nr** | **E-mail adres** | **Zaak gerelateerd** | **Neutraal** | **Niet zaak gerelateerd** | **Totaal** |
| 1 | jeff.dasovich@enron.com | 1.903 | 12 | 3.280 | 5.195 |
| 2 | kay.mann@enron.com | 1.564 | 16 | 3.055 | 4.635 |
| 3 | chris.germany@enron.com | 1.363 | 40 | 1.936 | 3.339 |
| 4 | matthew.lenhart@enron.com | 1.218 | 12 | 890 | 2.120 |
| 5 | tana.jones@enron.com | 1.129 | 1 | 2.946 | 4.076 |
| 6 | gerald.nemec@enron.com | 663 | 3 | 1.430 | 2.096 |
| 7 | sara.shackleton@enron.com | 653 | 1 | 3.651 | 4.305 |
| 8 | louise.kitchen@enron.com | 508 | 5 | 785 | 1.298 |
| 9 | john.arnold@enron.com | 491 | 11 | 844 | 1.346 |
| 10 | john.lavorato@enron.com | 467 | 5 | 472 | 944 |
| … | … | … | … | … | … |
| 144 | tim.belden@enron.com | 59 | 1 | 164 | 224 |
| 496 | jeff.skilling@enron.com | 17 | 0 | 42 | 59 |
| 1.072 | mark.koenig@enron.com | 7 | 0 | 22 | 29 |
| 2.379 | kenneth.lay@enron.com | 3 | 0 | 21 | 24 |
| 3.159 | andrew.fastow@enron.com | 2 | 0 | 5 | 7 |
| 3.497 | richard.causey@enron.com | 2 | 0 | 24 | 26 |
| 8.174 | larry.lawyer@enron.com | 1 | 0 | 8 | 9 |
| 19.634 | ben.glisan@enron.com | 0 | 0 | 7 | 7 |
| 19.648 | dan.boyle@enron.com | 0 | 0 | 7 | 7 |
| 19.671 | ken.rice@enron.com | 0 | 0 | 7 | 7 |
| -- | michael.kopper@enron.com | -- | -- | -- | -- |

In figuur 5.2 is een grafiek te zien die weergeeft welke e-mail adressen de meeste zaak gerelateerde e-mails heeft ontvangen. Net als in figuur 5.1 voert Jeff Dasovich de lijst aan. Hij heeft het meeste aantal e-mails ontvangen die na de e-mail content analyse als zaak gerelateerd zijn beoordeeld. Jeff Dasovich heeft in totaal 8.399 e-mails ontvangen, waarvan er 1.950 zijn beoordeeld als zaak gerelateerd, 32 als neutraal en 6417 als niet zaak gerelateerd. Daarnaast is Jeff Dasovich ook de persoon die het meeste aantal e-mails heeft ontvangen van iedereen. De aantallen van de andere personen zijn te zien in tabel 5.2. Wederom inclusief de veroordeelde personen van het fraude schandaal bij Enron. Merk op dat het e-mail adres klay@enron.com vele malen is gebruikt, in plaats van kenneth.lay@enron.com.

**Figuur 5.2**: De top 10 meest voorkomende e-mail adressen voor inkomend e-mailverkeer.

**Tabel 5.2**: De gegevenstabel die hoort bij de grafiek in figuur 5.2.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Nr** | **E-mail adres** | **Zaak gerelateerd** | **Neutraal** | **Niet zaak gerelateerd** | **Totaal** |
| 1 | jeff.dasovich@enron.com | 1.950 | 32 | 6.417 | 8.399 |
| 2 | richard.shapiro@enron.com | 1.677 | 15 | 5.886 | 7.578 |
| 3 | louise.kitchen@enron.com | 1.647 | 14 | 4.110 | 5.771 |
| 4 | tana.jones@enron.com | 1.503 | 8 | 5.903 | 7.414 |
| 5 | sara.shackleton@enron.com | 1.229 | 4 | 6.094 | 7.327 |
| 6 | klay@enron.com | 1.210 | 9 | 333 | 1.552 |
| 7 | mark.taylor@enron.com | 1.112 | 7 | 4.814 | 5.933 |
| 8 | james.steffes@enron.com | 1.084 | 4 | 3.620 | 4.708 |
| 9 | gerald.nemec@enron.com | 1.073 | 6 | 3.380 | 4.459 |
| 10 | john.lavorato@enron.com | 1.036 | 4 | 2.716 | 3.756 |
| … | … | … | … | … | … |
| 16 | tim.belden@enron.com | 851 | 3 | 2.642 | 3.496 |
| 36 | kenneth.lay@enron.com | 565 | 8 | 1.465 | 2.038 |
| 80 | jeff.skilling@enron.com | 372 | 4 | 1.337 | 1.713 |
| 202 | mark.koenig@enron.com | 237 | 0 | 591 | 828 |
| 233 | richard.causey@enron.com | 220 | 0 | 689 | 909 |
| 738 | andrew.fastow@enron.com | 97 | 0 | 296 | 393 |
| 1.069 | ben.glisan@enron.com | 69 | 0 | 278 | 347 |
| 1.178 | ken.rice@enron.com | 64 | 3 | 242 | 309 |
| 2.968 | larry.lawyer@enron.com | 20 | 0 | 218 | 238 |
| 3.485 | michael.kopper@enron.com | 16 | 0 | 77 | 93 |
| 5.886 | dan.boyle@enron.com | 8 | 0 | 40 | 48 |

In figuur 5.3 is een grafiek te zien die weergeeft welke e-mail adressen het meest geadresseerd zijn in zaak gerelateerde e-mails. Net als in figuur 5.1 en 5.2 voert Jeff Dasovich de lijst aan. Zijn e-mail adres kwam het vaakst voor al geadresseerde in e-mails die na de e-mail content analyse als zaak gerelateerd zijn beoordeeld. De exacte aantallen zijn te zien in tabel 5.3.

**Figuur 5.3**: De top 10 meest voorkomende e-mail adressen gesorteerd op zaak gerelateerd geanalyseerde e-mails.

**Tabel 5.3**: De gegevenstabel die hoort bij de grafiek in figuur 5.3.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Nr** | **E-mail adres** | **Zaak gerelateerd** | **Neutraal** | **Niet zaak gerelateerd** | **Totaal** |
| 1 | jeff.dasovich@enron.com | 3.853 | 44 | 9.697 | 13.594 |
| 2 | tana.jones@enron.com | 2.632 | 9 | 8.849 | 11.490 |
| 3 | kay.mann@enron.com | 2.294 | 18 | 6.025 | 8.337 |
| 4 | louise.kitchen@enron.com | 2.155 | 19 | 4.895 | 7.069 |
| 5 | chris.germany@enron.com | 2.092 | 49 | 3.785 | 5.926 |
| 6 | sara.shackleton@enron.com | 1.882 | 5 | 9.745 | 11.632 |
| 7 | richard.shapiro@enron.com | 1.788 | 16 | 6.270 | 8.074 |
| 8 | gerald.nemec@enron.com | 1.736 | 9 | 4.810 | 6.555 |
| 9 | matthew.lenhart@enron.com | 1.694 | 33 | 1.739 | 3.466 |
| 10 | mark.taylor@enron.com | 1.542 | 20 | 6.266 | 7.828 |
| … | … | … | … | … | … |
| 14 | klay@enron.com | 1.210 | 9 | 333 | 1.552 |
| 25 | tim.belden@enron.com | 910 | 4 | 2.806 | 3.720 |
| 55 | kenneth.lay@enron.com | 568 | 8 | 1.486 | 2.062 |
| 120 | jeff.skilling@enron.com | 389 | 4 | 1.379 | 1.772 |
| 241 | mark.koenig@enron.com | 244 | 0 | 613 | 857 |
| 274 | richard.causey@enron.com | 222 | 0 | 713 | 935 |
| 807 | andrew.fastow@enron.com | 99 | 0 | 301 | 400 |
| 1.204 | ben.glisan@enron.com | 69 | 0 | 285 | 354 |
| 1.299 | ken.rice@enron.com | 64 | 3 | 249 | 316 |
| 3.163 | larry.lawyer@enron.com | 21 | 0 | 226 | 247 |
| 3.824 | michael.kopper@enron.com | 16 | 0 | 77 | 93 |
| 5.882 | dan.boyle@enron.com | 8 | 0 | 47 | 55 |

Aan de hand van deze resultaten zal in hoofdstuk 6 een conclusie worden getrokken.

# Hoofdstuk 6 Conclusies en aanbevelingen



## Conclusies

Het onderzoek in deze scriptie heeft zich toegesplitst op de werking van online spamfilters voor e-mail berichten en de mogelijkheden om een dergelijk filter toe te passen in een andere situatie. De kern, de vele formules en de heuristiek, zijn uitgebreid gecontroleerd en toepasbaar gemaakt naar de situatie dat Xiraf er gebruik van kan maken. Dit leidt ertoe dat de e-mail content analyse werkt zoals deze zou moeten. Tijdens het draaien van de analyse treden er geen fouten op en de heuristiek en de formules worden correct in de applicatie toegepast. Dit zorgt ervoor dat voor alle e-mail berichten kan worden berekend of de e-mail is gerelateerd aan de zaak of niet.

Uit de testcase in hoofdstuk 5 is gebleken dat een heleboel e-mails al worden uitgefilterd met een beperkte hoeveelheid gemarkeerde e-mails. Er zijn slechts 414 e-mails gemarkeerd van de 238.281 e-mails. Ondanks deze beperkte hoeveelheid gemarkeerde e-mails heeft dit ervoor gezorgd dat ongeveer drie kwart van alle e-mails uitgefilterd werden als niet zaak gerelateerd. Bij een grotere hoeveelheid gemarkeerde e-mails zou de e-mail content analyse nog nauwkeuriger zijn werk kunnen doen, waardoor er nog meer e-mails uitgefilterd kunnen worden.

In de testcase zijn de hoofdpersonen van het fraudeschandaal van Enron niet naar voren gekomen door de e-mail content analyse. Dit komt voornamelijk door de beperkte hoeveelheid e-mails van de hoofdpersonen van het fraudeschandaal, die in de e-mail dataset aanwezig zijn. Daarnaast waren niet genoeg e-mails gemarkeerd om nauwkeurig de niet zaak gerelateerde e-mails uit te filteren. Een persoon als Jeff Dasovich, de persoon die de grootste hoeveelheid e-mails heeft verstuurd en ontvangen, wordt hierdoor gezien als de persoon die de meeste zaak gerelateerde e-mails in zijn mailbox heeft. Dit komt slechts vanwege zijn veelvuldig gebruik van e-mailverkeer, waar de hoofdpersonen van het fraudeschandaal veel minder gebruik maakte hiervan.

## Aanbevelingen

### Variabele hoeveelheid doorgenomen e-mails

Tijdens de afstudeerperiode heeft de nadruk gelegen op het correct toepassen en formuleren van de heuristiek en formules. Dit kostte relatief veel tijd en hierdoor is een beperkte hoeveelheid aan e-mails doorgenomen om te markeren als zaak gerelateerd of niet zaak gerelateerd. Deze hoeveelheid is te beperkt gebleken om in de testcase de hoofdpersonen naar de voorgrond te brengen, mede door de onbalans van e-mail gebruik door de medewerkers. Om te weten te komen of de e-mail content analyse de hoofdpersonen van het fraudeschandaal kan ontmaskeren moeten meer e-mails worden gemarkeerd om de analyse nauwkeuriger te laten verlopen.

Het is hierin dan ook een aanbeveling om te onderzoeken hoeveel e-mails er minimaal doorgenomen moeten worden voordat het filter een acceptabele nauwkeurige uitkomst geeft. Daarnaast kan ook worden onderzocht hoe de resultaten veranderen naarmate er nog meer e-mail berichten worden doorgenomen.

### Een variabele sterkte bij de kansberekening voor zeldzame woorden

In de situatie dat er een grote onbalans is tussen de hoeveelheid zaak gerelateerde e-mails en niet zaak gerelateerde e-mails is het aan te raden om onderzoek te doen naar de sterkte die meegegeven moet worden voor zeldzame woorden. Deze sterkte bepaald hoe vaak een woord minstens moet voorkomen, voordat gezegd kan worden dat het woord zaak of niet zaak gerelateerd is. Vanwege de onbalans tussen de hoeveelheid zaak gerelateerde en niet zaak gerelateerde e-mails, is het goed mogelijk dat een woord nog niet is voorgekomen in een gemarkeerde zaak gerelateerde e-mail. Dit woord kan echter al vele keren zijn voorgekomen in niet zaak gerelateerde e-mails. Door de onbalans moet hier eventueel met twee maten worden gemeten. Is een woord nog niet voorgekomen in niet zaak gerelateerde e-mails en al enkele keren in zaak gerelateerde e-mails (van de toch al beperkte hoeveelheid e-mails dat in deze categorie wordt gemarkeerd), dan behoord hier deze drempel lager te liggen.

### Variabel aantal e-mails bij de heuristiek

Evenals het aantal doorgenomen e-mails en een andere kracht die wordt meegegeven bij zeldzame woorden kan de heuristiek ook verschillende resultaten geven, afhankelijk van de hoeveelheid mee te nemen woorden. De heuristiek in de huidige staat zorgt ervoor dat in de e-mail content analyse de 30 meest belangrijke woorden worden meegenomen bij de beoordeling van de e-mail. Met belangrijk wordt bedoeld de mate van zaak of niet zaak gerelateerd. Neutrale woorden als: de, een, is, van, of, zullen weinig effect hebben op de beoordeling van het filter. Het kan interessant zijn hoe de beoordeling van de e-mail varieert indien gebruik gemaakt wordt van de top 10, 20, 40 of 80 woorden in de heuristiek.

## Uitbreidingen

### Uitbreiding classificatie: Woordenboost

Tijdens het project is het idee geopperd om een zogenaamde ‘woordenboost’ toe te voegen in het filter. Het idee was als volgt: de eindgebruikers, technische rechercheurs van wetshandhaving instanties, krijgen de mogelijkheid om een database op te stellen van voor gedefinieerde verdachte woorden. Denk hierbij aan woorden als pistool, geweer, vermoorden, fraude, minderjarig enzovoort. De woorden hebben waarschijnlijk al een grotere kans om voor te komen in een zaak gerelateerde e-mail dat een niet zaak gerelateerde e-mail. De kans dat het woord voorkomt in een zaak gerelateerde e-mail is dus groter dan 50%. Stel dat het woord ‘vermoord’ een kans heeft van 80% en het woord ‘pistool’ een kans van 72% dat het woord voorkomt in een zaak gerelateerde e-mail, berekend door filter. De woordenboost maakt het mogelijk om deze percentages te verhogen. Komt het woord ‘vermoord’ voor in de database van voor gedefinieerde verdachte woorden, dan wordt de kans bijvoorbeeld geboost naar een kans van 90%. Het gemiddelde van de huidige kans en 100%. Het woord ‘pistool’ kan op deze wijze worden geboost naar een kans van 86%. Door deze woorden een extra boost te geven zal de zaak gerelateerdheid van de e-mail ook stijgen.

Tijdens implementatie van deze feature moet er met de heuristiek rekening gehouden worden. In een e-mail met meer dan 30 woorden, wordt de heuristiek toegepast. Het is mogelijk dat een woord vooraf buiten deze top 30 valt, maar na de correctie van de woordenboost wel thuis hoort in de top 30 belangrijke woorden.

Deze werkwijze van de woordenboost hoeft niet per se volgens de hierboven genoemde wijze te werken. Er is over deze functionaliteit nagedacht en dit leek een goede manier om de zaak gerelateerdheid van een e-mail te verhogen afhankelijk van de gedefinieerde woordenlijst.

### Uitbreiding classificatie: Headers

Vele hedendaagse spamfilters maken gebruiken van de bayesiaanse statistiek om een e-mail te categoriseren als spam of ham. Deze is echter niet gelimiteerd tot de content, ook wel tekst, van de e-mail. Deze spamfilters letten ook informatie beschikbaar in de headers. Enkele voorbeelden van headers, ook wel koppen, van een e-mail zijn onderwerp, afzender, bijlages etc. Stuurt bijvoorbeeld een vriend – die in uw contactenlijst staat – een e-mail met daarin een tiental keer het woord ‘viagra’ als grap, dan zou het filter op content beslissen deze e-mail te verwijderen of te verplaatsen naar de map voor ongewenste e-mail. Aangezien u de afzender in uw contactenlijst heeft staan beslist het filter om de e-mail toch in uw inbox te plaatsen.

De online e-mail filters letten ook op toegevoegde bijlagen. In het geval van een spam e-mail met een valse toegevoegde bijlage, maken afzenders in de bijlage soms gebruik van een dubbele extensie met vele spaties. Met een bijlage als ‘Leuke mop.txt .exe’ hoopt de afzender dat de ontvanger denkt dat het een onschuldig txt-bestand is, terwijl het in werkelijkheid een exe-bestand is.

De online e-mail filters maken dus gebruik van meerdere technieken dan alleen de content analyse om spam e-mail te onderscheiden van de legitieme e-mails. Zo zijn er ook mogelijkheden om het filter in de geprogrammeerde applicatie uit te breiden en te verbeteren.

### Zaakprofielen

Er zijn verschillende soorten misdrijven waarmee de recherche mee te maken heeft. Er zijn zedenzaken, moordzaken, fraudezaken en ga zo maar verder. In ieder van deze zaken is er een andere set woorden belangrijk. Bij zedenzaken zijn woorden als ‘bloot’, ‘naakt’, ‘fotos’ en ‘minderjarig’ belangrijk. In moordzaken en fraudezaken zijn er andere woorden belangrijk voor het oplossen van de zaak. Om de effectiviteit van het filter de waarborgen zou het beter zijn om aan ieder zaakprofiel een eigen woordendatabase te koppelen. In deze woordendatabase staat de informatie dat het filter nodig heeft om van een e-mail te berekenen wat de kans is dat deze zaak gerelateerd is.

Om dit mogelijk te maken moet aan iedere zaak in Xiraf en Forensic Dashboard worden meegegeven welk profiel hoort bij de betreffende zaak. Hieraan kan het filter dan herkennen met welke woordendatabase gewerkt moet worden. Deze functionaliteit is echter (nog) niet aangebracht in Xiraf en Forensic Dashboard, dus is het op dit moment niet mogelijk om deze functionaliteit te in voegen.

### Overeenkomende e-mail berichten

Tijdens een recherche onderzoek zal de applicatie zoeken naar e-mail berichten die mogelijk belangrijk zijn voor het onderzoek. Door gebruik te maken van de bayesiaanse statistiek zou het mogelijk zijn om te zoeken naar overeenkomende e-mail berichten.

Een rechercheur heeft tijdens het onderzoek een belangrijk e-mail bericht gevonden en zou graag willen weten of er andere e-mail berichten zijn die hier op lijken. Bijvoorbeeld een reply-mail of doorstuur-mail met dus overeenkomende content. De woordendatabase kan worden gezien als de content van die ene e-mail waarvan de rechercheur wil weten of er overeenkomstige e-mail berichten zijn. Het filter zal weer alle e-mail berichten in de database moeten doornemen en zoeken of er berichten zijn waarin veel woorden of een hoog percentage van de woorden overeenkomen. Eventueel kan gebruik gemaakt woorden van een set van woorden. In de e-mail van de rechercheur komt bijvoorbeeld voor: ‘in deze e-mail staat’. Komt deze set van woorden ook in een ander e-mail bericht voor? Dit zijn allemaal mogelijkheden om te zoeken naar overeenkomende e-mail berichten. Hoe exact wordt bepaald welke e-mails overeenkomen en in wat voor mate deze met elkaar overeenkomen moet nog nader worden onderzocht.

### Woordsynoniemen

Verschillende woorden worden door verschillende personen en in verschillende situaties anders geschreven. In het geval dat een woord aan het begin van een zin staat, dan begint de eerste letter van dit woord waarschijnlijk geschreven met een hoofdletter. Een woord als cocaïne kan ook op verschillende manieren worden geschreven. De ene persoon schrijft dit woord zonder een trema op de i, terwijl een andere persoon dit wel doet. Daarnaast kan er altijd sprake zijn van typfouten waardoor woorden net iets anders worden geschreven dan zou moeten.

In Xiraf is reeds de mogelijkheid om te zoeken naar woorden in e-mails. Ik ga er in dit geval van uit dat dit filter zoekt naar exacte matches. Zou het interessant zijn om te zoeken naar een woord dat mogelijk verkeerd of anders is geschreven? Enkele mogelijkheden waarop het woord cocaïne geschreven kan worden:

* Cocaïne;
* cocaine;
* coocaïne;
* c0caine;
* Coca1ne;

Enzovoort.

Deze analyse zal ook enige raakvlakken hebben met dat van de applicatie om e-mails te categoriseren als zaak of niet zaak gerelateerd. De applicatie zoekt naar woorden in een e-mail om hieruit te beslissen of deze overeenkomt met een e-mail die vooraf is gemarkeerd als zaak of niet zaak gerelateerd. In dit geval gaan we op zoek naar woorden die overeenkomen met het opgegeven woord. Is er een letter teveel geschreven maar komt de rest overeen? Is er een letter anders geschreven terwijl de rest van het woord exact overeenkomt?

# Literatuurlijst

1. Nederlands Forensisch Instituut (2011). *NFI en NCIM Groep sluiten partnership*. Verkregen op 9 februari, 2012 van http://www.forensischinstituut.nl/nieuws/2011/ nfi-en-ncim-groep-sluiten-partnership.aspx
2. De Vries, W. (2011). *NFI neemt infrastructuur kinderporno-opsporingstool op de schop*. Verkregen op 9 februari, 2012 van http://tweakers.net/nieuws/76020/nfi-neemt-infrastructuur-kinderporno-opsporingstool-op-de-schop.html
3. Nederlands Forensisch Instituut (2011). *Nieuwe tool maakt aanpak cybercrime makkelijker*. Verkregen op 9 februari, 2012 van http://www.nfi-info.nl/artikelen/ artikelen\_item/t/nieuwe\_tool\_maakt\_aanpak\_cybercrime\_makkelijker
4. Nederlands Forensisch Instituut. *XIRAF*. Verkregen op 9 februari, 2012 van http://www.holmes.nl/NFIlabs/XIRAF/index.html
5. About.com Email. *Heinz Tschabitscher*. Verkregen op 9 februari, 2012 van http://email.about.com/bio/Heinz-Tschabitscher-1215.htm
6. About.com Email. *Top 12 Windows Spam Filters*. Verkregen op 9 februari, 2012 van http://email.about.com/od/windowsspamfightingtools/tp/anti-spam.htm
7. GFI. *Why Bayesian filtering is the most effective anti-spam technology*. Verkregen op 9 februari, 2012 van http://www.gfi.com/whitepapers/why-bayesian-filtering.pdf
8. Wikipedia, the free encyclopedia. *Spam (post)*. Verkregen op 9 februari, 2012 van http://nl.wikipedia.org/wiki/Spam\_(post)
9. Aasli Consulting. *Spam filtering using statistical filters*. Verkregen op 9 februari, 2012 van http://www.aasli.com/spam/index.php
10. Process software. *Introduction to Bayesian Filtering. Using Bayes’ Formula to keep spam out of your Inbox*. Verkregen op 9 februari, 2012 van http://www.process.com/ precisemail/bayesian\_filtering.htm
11. The Apache SpamAssasin Project. *Effective Training*. Verkregen op 9 februari, 2012 van http://spamassassin.apache.org/full/3.2.x/doc/sa-learn.html#effective\_training
12. Wikipedia, the free encyclopedia. *Bayesian spam filtering*. Verkregen op 9 februari, 2012 van http://en.wikipedia.org/wiki/Bayesian\_spam\_filtering
13. Wikipedia, the free encyclopedia. *Bayesian inference*. Verkregen op 9 februari, 2012 van http://en.wikipedia.org/wiki/Bayesian\_inference
14. Universiteit van Amsterdam & Vrije Universiteit Amsterdam (2007). *Spamfilter*. Verkregen op 9 februari, 2012 van http://www.w2cinformatica.nl/compcrim/ filter.html
15. Robinson G. (2003). *A Statistical Approach to the Spam Problem*. Verkregen op 9 februari, 2012 van http://www.linuxjournal.com/article/6467
16. Microsoft. *Data Type Summary*. Verkregen op 9 februari, 2012 van http://msdn.microsoft.com/en-us/library/47zceaw7.aspx
17. Techopedia. *Bayesian Filter*. Verkregen op 9 februari, 2012 van http://www.techopedia.com/definition/1635/bayesian-filter
18. VPRO Tegenlicht. *Enron tijdlijn*. Verkregen op 14 februari, 2012 van http://tegenlicht.vpro.nl/nieuws/2007/november/enron-tijdlijn.html
19. Docunet. *Enron, The Smartest Guys In The Room.* Verkregen op 14 februari, 2012 van http://docunet.nl/enron-the-smartest-guys-in-the-room
20. Wikipedia, the free encyclopedia. *Enron Scandal*. Verkregen op 15 februari, 2012 van http://en.wikipedia.org/wiki/Enron\_scandal
21. Time.com U.S. *Enron’s Collapse*. Verkregen op 14 februari, 2012 van http://www.time.com/time/interactive/0,31813,2013797,00.html
22. Enron Creditors Recorvery Corp. *About ECRC*. Verkregen op 15 februari, 2012 van http://www.enron.com/index\_option\_com\_content\_tast\_section\_id\_1\_Itemid\_2.htm
23. William. W. Cohen. *Enron Email Dataset*. Verkregen op 15 februari, 2012 van http://www.cs.cmu.edu/~enron
24. Heer J. (2004). *Exporing Enron, Visualizing ANLP Results*. Verkregen op 15 februari, 2012 van http://hci.stanford.edu/jheer/projects/enron/v1
25. Squirm. *Sorting*. Verkregen op 17 februari, 2012 van http://www.xtremevbtalk.com/ showthread.php?t=78889
26. Associated Press. *Former Enron exac gets 27-months sentence*. Verkregen op 17 februari, 2012 van www.msnbc.msn.com/id/19293341/ns/business-corporate\_ scandals/t/former-enron-exec-gets--month-sentence/
27. BBC News. *Top Enron official pleads guilty*. Verkregen op 17 februari, 2012 van http://news.bbc.co.uk/2/hi/business/3601628.stm
28. Celebrity Wonder. *Enron: The Smartest Guys in the Room: timeline of events*. Verkregen op 17 februari, 2012 van http://celebritywonder.ugo.com/movie/ 2005\_Enron:\_The\_Smartest\_Guys\_in\_the\_Room\_timeline\_of\_events.html
29. The New York Times. *Former Enron Executive Pleads Guilty to Tax Violations*. Verkregen op 17 februari, 2012 van http://www.nytimes.com/2002/11/27/business/ former-enron-executive-pleads-guilty-to-tax-violations.html

# Bijlagen

## Bijlage 1 Formules

(B1.1)

(B1.2)

(B1.3)

(B1.4)

(B1.5)

(B1.6)

(B1.7)

(B1.8)

(B1.9)

(B1.10)

(B1.11)

(B1.12)

(B1.13)

(B1.14)

(B1.15)

(B1.16)

(B1.17)

(B1.18)

(B1.19)

(B1.20)

(B1.21)

(B1.22)

(B1.23)

(B1.24)

(B1.25)

(B1.26)

(B1.27)

(B1.28)

(B1.29)

(B1.30)

(B1.31)

(B1.32)

(B1.33)

(B1.34)

(B1.35)

(B1.36)

(B1.37)

(B1.38)

(B1.39)

(B1.40)

(B1.41)

(B1.42)

(B1.43)

(B1.44)

(B1.45)

(B1.46)

(B1.47)

(B1.48)

(B1.49)

(B1.50)

(B1.51)

(B1.52)

(B1.53)

(B1.54)

(B1.55)

(B1.56)

(B1.57)

(B1.58)

(B1.59)

(B1.60)

(B1.61)

(B1.62)

(B1.63)

(B1.64)

(B1.65)

(B1.66)

(B1.67)

(B1.68)

(B1.69)

(B1.70)

(B1.71)

(B1.72)

(B1.73)

(B1.74)

(B1.75)

(B1.76)

(B1.77)

(B1.78)

(B1.79)

(B1.80)

(B1.81)

(B1.82)

(B1.83)

(B1.84)

(B1.85)

(B1.86)

(B1.87)

(B1.88)

(B1.89)

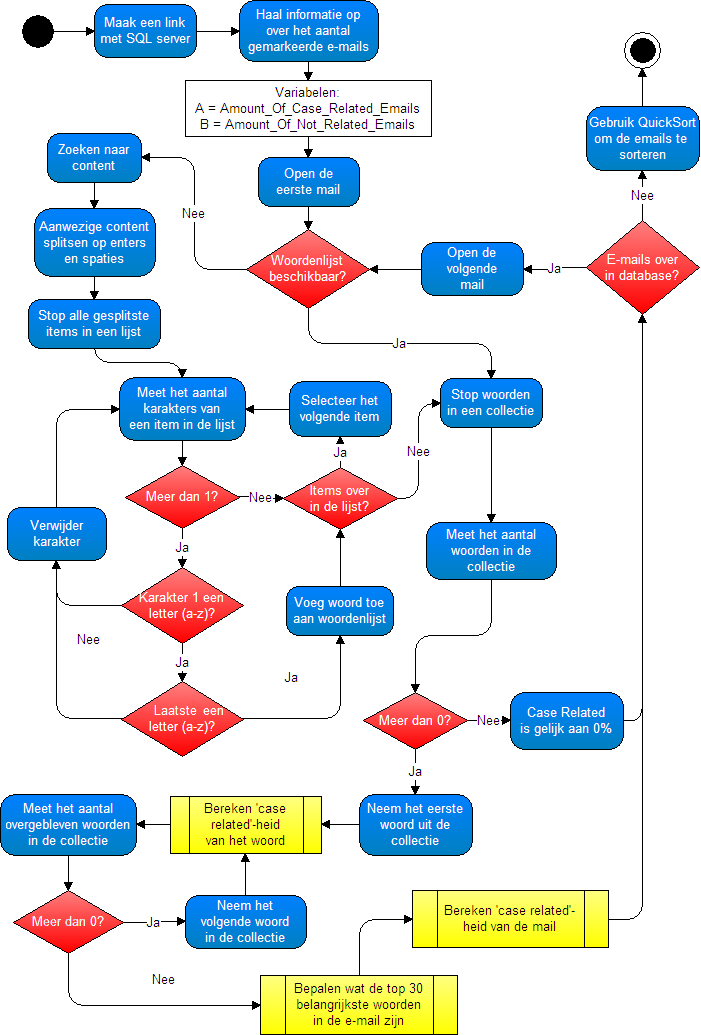
(B1.90)

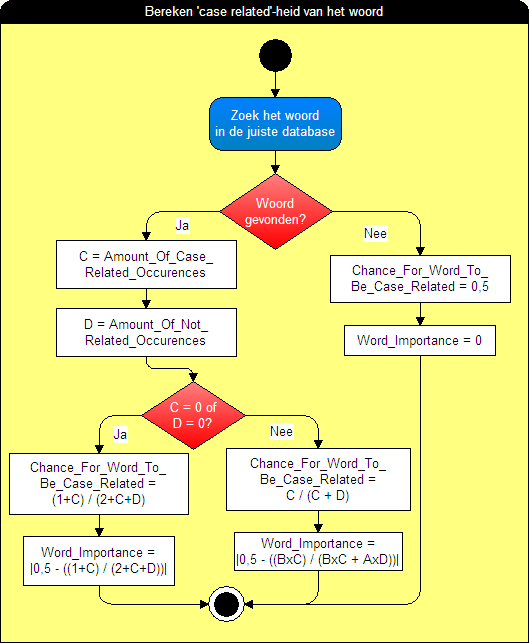
(B1.91)

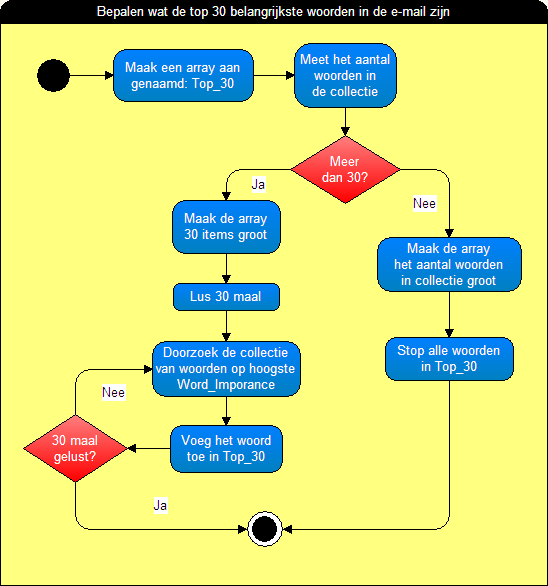
(B1.92)

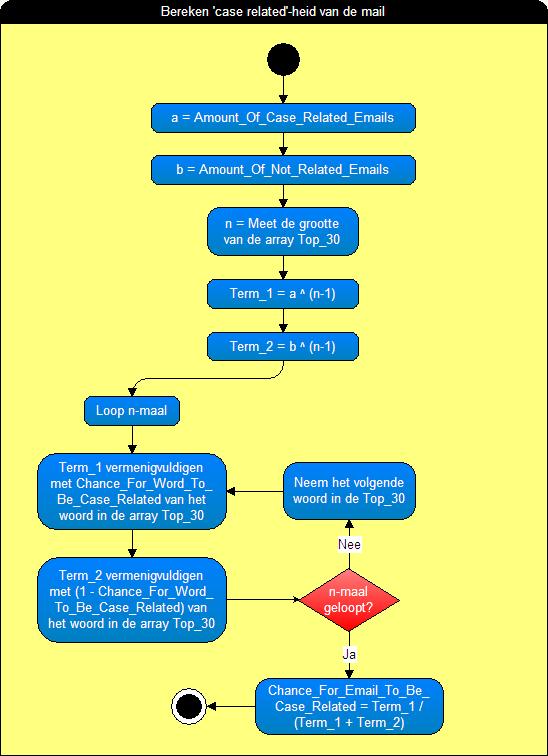
(B1.93)

## Bijlage 2 De gehele e-mail analyse flowchart



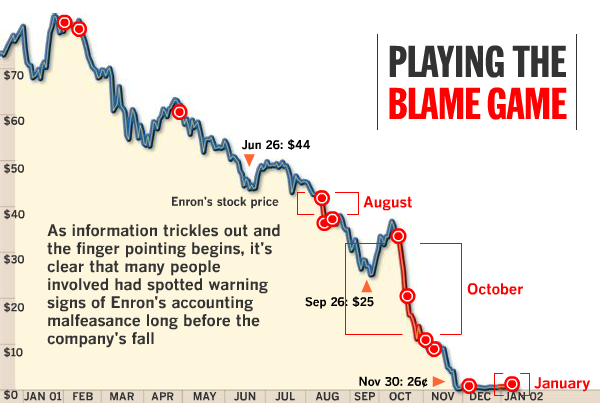






## Bijlage 3 Enron Corporation

Enron Corporation is opgericht in 1985 door Kenneth Lay als Amerikaans energiebedrijf, dat vooral gas levert. In 1989 begint het met het verhandelen van aardgasproducten. Hiermee wordt Enron de grootste aardgashandelaar van Noord-Amerika en Groot-Brittannië.(18) Twee jaar later vraag Enron goedkeuring om de boekhouding te doen volgens het mark-to-market principe, waarop kort daarna toestemming wordt gegeven.(19) Dit maakte het mogelijk voor Enron om potentiele toekomstige winsten te boeken op producten of diensten op de dag dat een deal is gemaakt. Enron kon hierdoor enorme winsten in de administratie boeken ook al kwam er nauwelijks een cent binnen. Ondanks interne financiële problemen breidt Enron uit naar verschillende andere markten, waaronder de handel in elektriciteit. Deze financiële problemen binnen Enron zorgen ervoor dat verkopers op creatieve wijze geld moeten zien te verdienen. Eén van deze creatieve oplossingen heeft in 2000 en 2001 gezorgd voor de energie crisis in Californië. Door verschillende blackouts te genereren vlogen de energieprijzen in Californië omhoog en wist Enron miljarden dollars extra te verdienen.(19) Toch bleef Enron diep in de rode cijfers.



**Figuur B.1**: De beurskoers van Enron Corp. van januari 2001 tot januari 2002.(21)

Ondanks de financiële problemen deed Enron het uitstekend op Wall Street. Door het mark-to-market boekhouden heeft Enron namelijk nooit aan de buitenwereld verliezen hoeven rapporteren. Op 5 maart 2001 kwam het artikel *Is Enron Overpriced?* in Fortune Magazine.(20) Hierin werd erop gewezen dat analisten en investeerders niet exact wisten hoe Enron zijn geld verdiend en dat er binnen Enron sprake was van ‘vage transacties’. Vanaf dit moment begon het schip, genaamd Enron, te zinken. In de opvolgende maanden vertrokken verschillende leden van het hogere management en verkochten hun aandelen voor vele miljoenen dollars. De waarden van de aandelen van het zevende grootste bedrijf van Amerika kelderde dit jaar van ruim 80 dollar naar een paar cent.

Op 2 december 2011 werd Enron bankroet verklaard. Op dat moment was dit het grootste faillissement in de Amerikaanse historie. Duizenden mensen raakten werkloos en velen van hun raakten ook hun spaarplan kwijt dat afhankelijk was van de beurswaarde van Enron. Er werd een grootschalig onderzoek opgezet om te onderzoeken hoe het mogelijk was dat Enron failliet ging en wie de schuldigen hiervan waren. In de opvolgende jaren hebben 16 mensen bekend schuldig te zijn aan misdaden binnen het bedrijf en vijf anderen werden schuldig gevonden. Tegenwoordig bestaat Enron Corporation als Enron Creditors Recorvery Corp. (ECRC). Hun missie is het om de gedupeerde crediteuren van het voormalige Enron Corporation terug te betalen wat ze zijn verloren door het faillissement. Als dit is gebeurd wordt ECRC opgeheven.(22)

Enkele veroordeelde ex-Enron medewerkers zijn: Jeff Skilling, Kenneth Lay, Andrew Fastow, Richard Causey, Dan Boyle, Mark Koenig, Ben Glisan, Michael Kopper, Tim Belden, Larry Lawyer en Ken Rice.(20,26,27,28,29)

Na het faillissement werd een grootschalig onderzoek ingesteld. In dit onderzoek is onder andere gebruik gemaakt van het e-mail verkeer binnen Enron om erachter te komen wie schuldig zijn aan fraude en samenzwering.