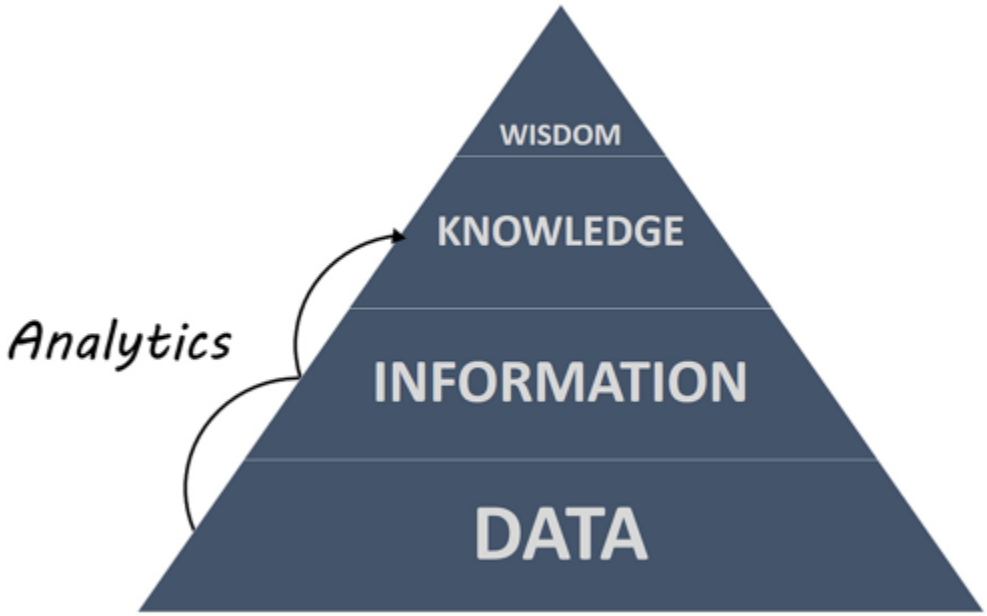
Workshop Tekst- & data mining met RapidMiner



"We are drowning in information but starved for knowledge" (Naisbitt, 1982)

Hugo Benne
VOGIN-IP Lezing 28 maart 2018
Laatst bewerkt op 27 maart 2018

Inhoudsopgave

Workshop tekst- en datamining

(Big) data is hot. Opleidingen voor 'Applied Data Science' en 'Data mining' schieten als paddenstoelen uit de grond en er is veel vraag naar 'data scientists'. Maar wat betekent deze ontwikkeling nu voor informatiespecialisten en wat kunnen wij met dit onderwerp? Dat wordt duidelijk gemaakt in een praktische workshop 'tekstmining'[1] verzorgd voor en door informatiespecialisten.

In de workshop 'Tekst- en datamining' tijdens VOGIN-IP Lezing 2018 gaan we concreet aan de slag met tekstmining. We gebruiken hiervoor de tool RapidMiner. De werkvorm van de workshop is volgens het 'flipped classroom' principe. Dat betekent dat de deelnemers voorafgaand aan de workshop huiswerk krijgen. Deelnemers aan de workshop krijgen tijdig alle hiervoor benodigde informatie toegestuurd. Het huiswerk bestaat uit het doornemen van de theorie (deze syllabus) over data- en tekstmining, het installeren van de software (RapidMiner) op de eigen laptop en het bekijken van enkele video's over RapidMiner. Tijdens de workshop wordt geen theorie meer gegeven maar gaan we direct op de eigen laptop aan de slag met een opdracht, bestaande uit een casus en een tekstcorpus.

NB. Flipped classroom werkt alleen als de deelnemers aan de workshop de ingangsopdrachten vooraf hebben uitgevoerd. Bij deze dus het vriendelijke doch dringende verzoek om die opdrachten van tevoren uit te voeren.

Over de trainer

Hugo Benne (MEd) is een ervaren allround informatie- en mediaprofessional die al ruim 25 jaar actief is in uiteenlopende functies (trainer, customer support engineer, adviseur documentaire informatievoorziening, projectleider, applicatiebeheerder etc.) binnen de sectoren onderwijs, (semi-) overheid en het bedrijfsleven. De laatste jaren werkt hij als docent voor de opleiding HBO-ICT (Information & Media Studies) aan de Faculteit IT & Design van de Haagse Hogeschool. Daarnaast is hij zelfstandig ondernemer en verzorgt hij o.a. trainingen en adviesdiensten binnen het informatie- en mediavakgebied. LinkedIn

[1] https://vogin-ip-lezing.net/workshops-2018

1. Inleiding data- en tekstmining

Datamining is het (geautomatiseerde) proces van het zoeken in grote hoeveelheden data naar valide, nieuwe en mogelijke bruikbare patronen in en verbanden tussen data om daar uiteindelijk nieuwe informatie en dus meerwaarde voor organisaties uit te destilleren.
Die data kunnen bestaan uit binnen een organisatie gegenereerde data of om externe, openbaar toegankelijke data. Met datamining gaan we op zoek naar verbanden in data en kunnen we zo ook inzichten verwerven in iets waar we niet specifiek naar op zoek waren.

Een paar voorbeelden van mining:

- Een supermarkt onderzoekt in de transactionele gegevens uit de kassaregistraties welke producten vaak samen worden verkocht zodat deze artikelen fysiek bij elkaar in de buurt kunnen worden uitgestald in de winkel;

- Een bedrijf zet miningtechnieken op sociale mediaberichten in om uit te zoeken welke personen het meest geïnteresseerd zijn in haar producten en diensten zodat de marketingactiviteiten van het bedrijf op deze doelgroep kunnen worden gericht;

- Een organisatie onderzoekt haar content door middel van tekstmining (miljoenen e-mails, honderdduizenden documenten in het document management systeem) om daar (nieuwe) informatie uit te halen.

De data kunnen verschillende vormen aannemen (zie hoofdstuk over data) en bijvoorbeeld bestaan uit sensordata, transactionele data opgeslagen in relationele databases, sociale mediaberichten, open onderzoeksdata of websites.

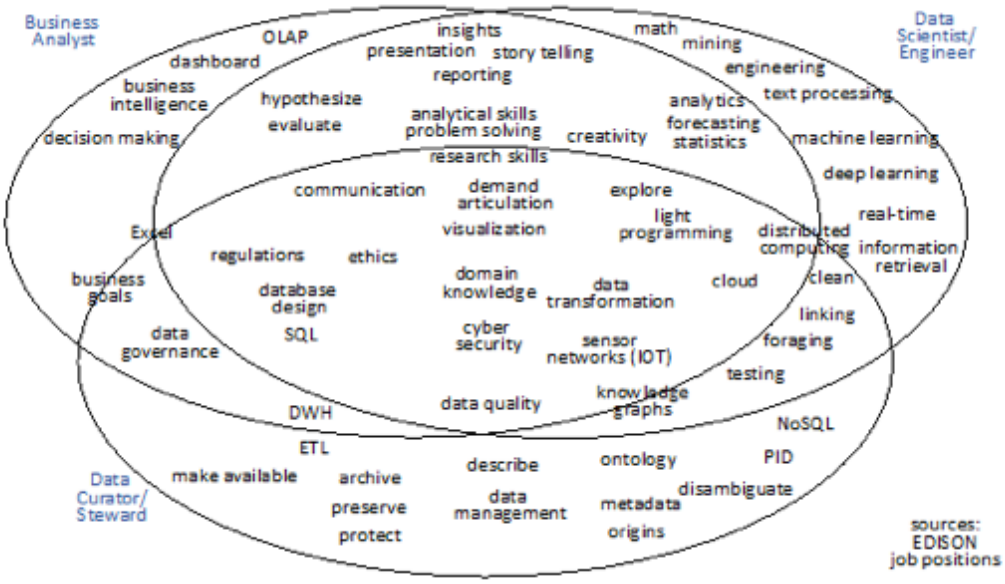
Tekstmining is een vorm van datamining en beslaat het proces van het transformeren van ongestructureerde tekst naar (semi-) gestructureerde tekst, het analyseren van deze tekst en het extraheren van relevante informatie uit deze tekst (Hurwitz, Nugent, Halper, & Kaufman, 2017). Bij tekstmining wordt gezocht naar patronen in (of nieuwe informatie uit) ongestructureerde of semigestructureerde tekst in de vorm van natuurlijke taal. Natural Language Processing (NLP) en statistische methoden maken deel uit van die tekstanalyse.

Merk daarbij op dat tekstmining iets anders is dan het zoeken naar informatie met bijvoorbeeld een internetzoekmachine als Google. Bij het zoeken met een internetzoekmachine gaat het om het zoeken naar reeds bestaande informatie met behulp van zoekwoorden. Dit levert documenten (webpagina's) op waarbij er niets anders op zit dan deze documenten door te nemen om conclusies te kunnen trekken. Bij tekstmining gaat het om het zoeken naar nieuwe en in de dataset impliciet aanwezige informatie. Tekstmining levert geen hele documenten op, maar uit de documenten geëxtraheerde informatie.

1.1. Rol van de informatiespecialist

Door de komst van big data komen er nieuwe kansen voor informatiespecialisten om zich, naast de meer traditionele taken zoals het ontsluiten, metadateren, zoeken naar en organiseren van (met name) ongestructureerde informatie, professioneel te profileren op het gebied van big data en tekstmining.

Het EDISON Data Science Framework (EDISON consortium, 2017), een consortium van zeven organisaties, beschrijft de beroepstaken en competenties van de 'data scientist' professional. Dit framework bestaat uit 5 competentiegebieden: 1) Data Science Analytics, 2) Data Science Engineering, 3) Data Management, 4) Research Methods and Project Management en 5) Business Analytics. Met name het competentiegebied 'Data Management' bevat competenties en vaardigheden die de informatiespecialist bekend zullen voorkomen. Hieronder vallen bijvoorbeeld zaken als 'Make available', 'Data quality', 'Data governance', 'Archive', 'Preserve', 'Ontology' en 'Metadata'.

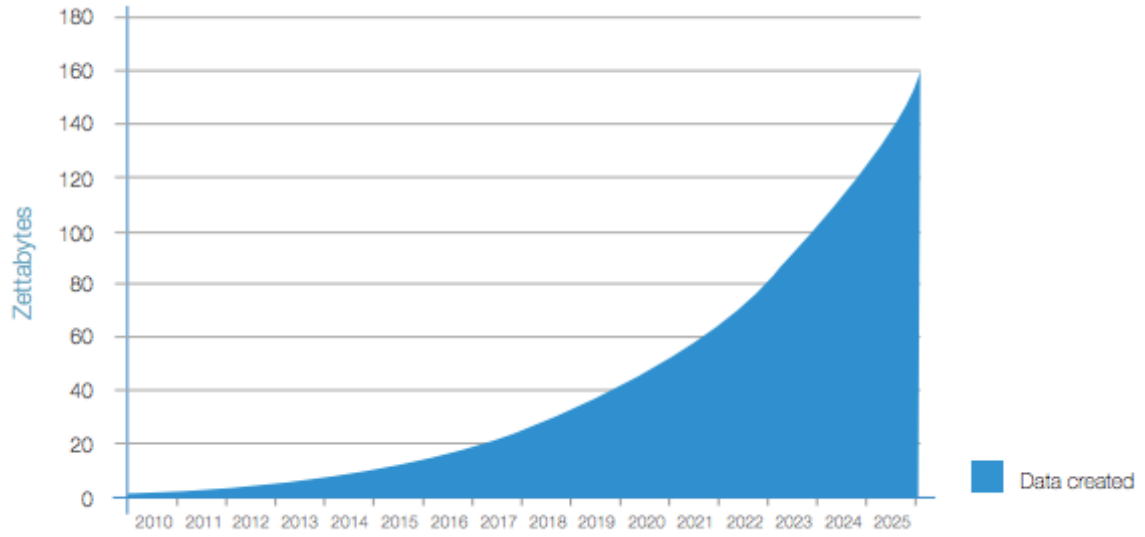


Figuur 1: Beroepstaken, vaardigheden en competenties in data science (Vuurens, 2017)

Het EDISON consortium beschrijft beroepsprofielen die raken aan het werk van de informatiespecialist. In het venndiagram (figuur 1), dat is ontstaan door begrippen en werkzaamheden uit het EDISON framework en uit vacatureteksten op het gebied van data science te groeperen, zijn deze beroepstaken vooral terug te vinden in het onderste deel. Het gaat daarbij om beroepen zoals: 'data stewards', 'digital data curator', 'digital librarians' en 'data archivists'. Binnen het data science vakgebied liggen dus kansen voor informatiespecialisten. Daarnaast is het ook goed als de informatiespecialist inzicht krijgt in dataminingtechnieken, begrijpt wat dit vakgebied precies inhoudt en met data-analisten en data scientists kan communiceren en samenwerken. Informatiespecialisten zijn immers van oudsher bezig met het metadateren, ordenen, vinden, selecteren en classificeren van (met name ongestructureerde) informatie. Dit zijn ook zaken die terug te vinden zijn in het takenpakket van de data scientist.

2. Data

"*We are drowning in information but starved for knowledge*" schreef John Naisbitt in 1982 in zijn boek 'Megatrends' (Naisbitt, 1982) zo'n zeven jaar voordat Tim Berners-Lee het World Wide Web uitvond ("World Wide Web", 2016). De hoeveelheid informatie is daarna exponentieel toegenomen (figuur 2). Volgens IBM genereren we met zijn allen dagelijks 2,5 triljoen (miljard x miljard = 10^{18}) bytes aan data ("IBM - What is big data?", z.d.). Negentig procent van alle beschikbare data is volgens IBM in de laatste twee jaar ontstaan. Deze data kunnen uit van alles bestaan: sensordata, sociale mediaberichten, aankooptransacties, websites, geografische en GPS-gegevens, smartphonedata, etc. Hoe gaan we al deze data opslaan, toegankelijk maken en hier iets zinvols mee doen? Hoe genereren we nieuwe kennis uit deze data? De data waarover we het hier hebben worden de laatste jaren aangeduid als 'big data'. Voordat we verder op het begrip big data ingaan is het noodzakelijk om eerst wat basisbegrippen toe te lichten.



Figuur 2: Global datasphere in zetabytes (Reinsel, Gantz, & Rydning, 2017).

2.1. Data, informatie, kennis

Kennis, informatie en gegevens (data) worden vaak met elkaar verward maar zijn niet hetzelfde. Gegevens vormen de kleinste eenheden in de vorm van losse karakters, tekens of cijfers. Die gegevens zijn op zichzelf vrij 'kaal' en 'betekenisloos'. Wanneer we aan die gegevens betekenis kunnen ontlenen dan spreken we van informatie. Denk bijvoorbeeld aan letters die samen woorden of zinnen kunnen vormen en daardoor dus informatiewaarde krijgen (mits we die taal beheersen). Of wanneer we weten dat een reeks cijfers een telefoonnummer betreft en we daardoor in staat zijn om te handelen, namelijk om op te bellen. Als we aan informatie **E**rvaring, **V**aardigheid en **A**ttitude (houding) toevoegen, wordt informatie kennis (Weggeman, 2001): Kennis = Informatie x E.V.A (zie figuur 3).

Figuur 3: Data, informatie, kennis (Aalten, Becker, Linden, & Sieverts, 2017)

Kennis is het vermogen om informatie om te zetten in nieuwe informatie. Door kennis zijn we in staat om de informatie te interpreteren of in staat om te handelen. Kennis is de belangrijkste grondstof in onze informatiemaatschappij en het belangrijkste productiemiddel in veel organisaties. Er zijn grofweg twee soorten kennis te onderscheiden: expliciete en impliciete kennis (tacit knowledge) (Nonaka, 1995). Expliciete kennis ligt bijvoorbeeld opgeslagen in informatiesystemen en is waarneembaar. Bij impliciete kennis gaat het om kennis die vaak persoonlijk is, in 'hoofden' van mensen zit en in elk geval minder goed waarneembaar is. Bij het data- of tekstminen (ook wel 'knowledge discovery' genoemd) gaat het om het waarneembaar maken (expliciteren) van nog niet eerder waargenomen maar wel in systemen of databases aanwezige kennis.

2.2. Gestructureerde versus ongestructureerde data

Data (en informatie) kan grofweg in drie verschijningsvormen voorkomen:

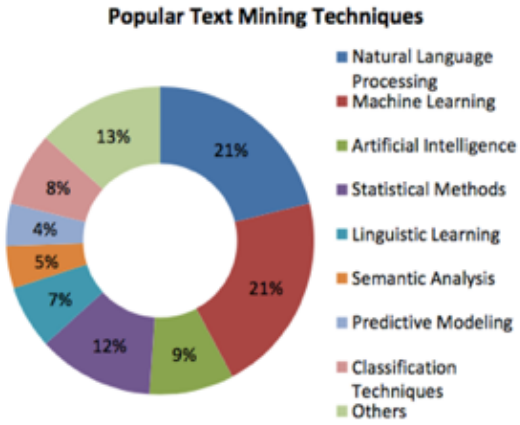
- **Gestructureerde data**: Bij dit type data gaat het om data die liggen opgeslagen in een vast bestandsformaat waarbij de gegevens zijn opgedeeld in records en velden (bv. relationele databases) of kolommen en rijen (spreadsheets). Bij gestructureerde data kan het bijvoorbeeld gaan om sensordata en kassa- of OV-chipkniptransacties. Dit type data is het gemakkelijkst te analyseren. Dit komt omdat bij gestructureerde data direct wordt verwezen naar bepaalde velden en metadata. Met behulp van een querytaal als SQL[2] kan in de database worden gezocht naar bepaalde kenmerken. Zo kunnen bijvoorbeeld met één query alle inwoners binnen een bepaalde stad of postcodegebied worden geselecteerd;

- **Semigestructureerde data**: Dit type data is niet gestructureerd volgens formele datastructuren of –modellen (zoals relationele databases) maar bevat wel metadata, tags of andere coderingen (bijvoorbeeld XML of HTML);

- **Ongestructureerde data**: Dit type data is niet opgeslagen in een van tevoren vastgelegde structuur. Het gaat daarbij bijvoorbeeld om webteksten, audio- en videobestanden, afbeeldingen, tekstbestanden of sociale mediaberichten.

[2] https://nl.wikipedia.org/wiki/SQL

De schattingen over de hoeveelheid ongestructureerde data ten opzichte van de totale hoeveelheid data variëren van ca. 80% tot meer dan 90% (Cai & Zhu, 2015; Dhar, 2013; Khan e.a., 2014). Ook groeit de hoeveelheid ongestructureerde data veel harder dan de hoeveelheid gestructureerde data (Khan e.a., 2014). Aangezien het grootste deel van de beschikbare data ongestructureerd is, is tekstmining dus een belangrijk vorm van datamining die breed wordt toegepast. Volgens het artikel van Kaur en Chopra (2016) gaat het daarbij met name om 'natural language processing,' 'machine learning', 'kunstmatige intelligentie', 'statistische methoden', 'linguïstiek', 'semantische analyse', 'predictive modelling' en om 'classificatie' van informatie (figuur 4).



Figuur 4: Toepassingsgebieden van tekstmining (Kaur & Chopra, 2016).

2.3. Big data

Big data is een verzamelnaam voor allerlei soorten (extern én intern opgeslagen) gestructureerde, semigestructureerde en ongestructureerde data. Big data kan zich beperken tot één informatiesysteem, maar ligt in de meeste gevallen verspreid over meerdere systemen en netwerken. Een mailserver van een organisatie kan dus ook als big data worden gezien. Kenmerk van big data is dat het te groot is om handmatig te verwerken en lange tijd ook te groot om door machines te laten bewerken. Pas de laatste jaren is de techniek snel en geavanceerd genoeg om big data aan te kunnen. De uitdaging van big data ligt niet bij het verzamelen van gegevens, maar juist om er in een gedistribueerde omgeving iets zinvols mee te doen (Jagadish e.a., 2014). Big data wordt vaak getypeerd door de 4, 5, 6 (of meer) V's. Fouad et al. (2015) omschrijven de volgende 6 V's (figuur 5):

Figuur 5: De 6 V's van Big data (Fouad e.a., 2015).

- **Volume**: Verwijst naar de grote hoeveelheden data en de enorme opslagcapaciteit die daarvoor benodigd is;

- **Velocity**: Verwijst naar de ongekende snelheid van de gegevensstroom en de complexiteit om die (niet constante) datastroom te kunnen verwerken;

- **Variety**: Verwijst naar de grote variëteit aan data die kan zijn opgeslagen in de meest uiteenlopende bestandsformaten;

- **Veracity**: Verwijst naar de correctheid, nauwkeurigheid, betrouwbaarheid en zuiverheid van (en de ruis in) de data;

- **Viability**: Verwijst naar het kiezen en filteren van juist die data die levensvatbaar genoeg is om toekomstvoorspellingen mee te kunnen doen;

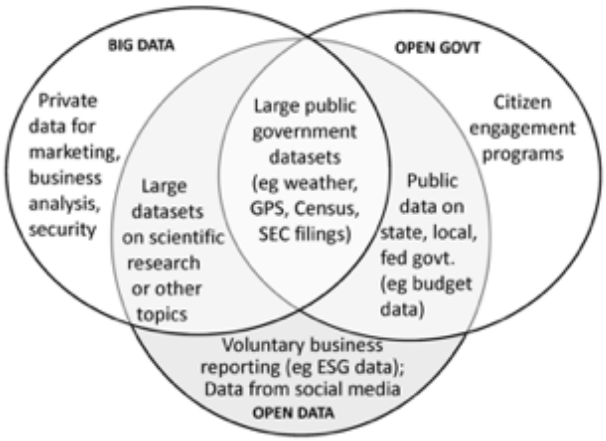
- **Value**: Verwijst naar het doel van big data: het zoeken naar verbanden tussen (of patronen in) informatie die kunnen worden omgezet in nieuwe informatie en waarmee een (economische) (meer-) waarde kan worden gecreëerd.

2.4. Open data

Net als big data is open data ook een verzamelnaam voor allerlei soorten data (gestructureerd, semigestructureerd of ongestructureerd). Open data zijn data die door overheden, erfgoed- en onderzoeksinstellingen openbaar beschikbaar worden gesteld voor hergebruik. Het gaat hierbij om data die zijn bekostigd uit publieke middelen, bijvoorbeeld onderzoekdatasets of statistische data van het CBS. Ook een dataset met daarin een overzicht van de lantaarnpalen in een bepaalde gemeente is een voorbeeld van open data. Verder heeft ook de Nederlandse overheid een depot[3] geopend voor open overheidsdata. Kenmerk van open data is dat deze data vrij zijn van auteursrecht en zijn opgeslagen in een open bestandsformaat waardoor deze data kunnen worden hergebruikt voor onderzoek, het maken van applicaties of andere specifieke doeleinden.

Een mooi voorbeeld van open data is het 'Open Cultuur Data' initiatief[4]. Verschillende erfgoedinstellingen zoals musea en archieven delen hun collecties via dit platform. Het Rijksmuseum heeft op dit platform een dataset van ruim 111.000 museumobjecten (metadata) en afbeeldingen uit haar collectie beschikbaarheid gesteld.

NB. Open data kan ook big data zijn en andersom geldt hetzelfde: big data kan ook (volledig) bestaan uit open data. Figuur 6 geeft een overzicht van de verschillen en overeenkomsten tussen big data, open data en open overheidsdata (Gurin, 2014).



Figuur 6: Big data, Open data, Open overheidsdata (Gurin, 2014)

2.5. Juridische en ethische aspecten van data

Bij het minen van data en de keuze van de datasets rijzen een aantal sociale, ethische en juridische vragen. Mag je de data eigenlijk wel gebruiken om te minen? Mag je de data wel bewaren voor miningactiviteiten of zijn er wettelijke termijnen die bepalen hoe lang de informatie mag worden bewaard? Wie is juridisch gezien eigenaar van de informatie? Is de informatie in open access uitgeven of gepubliceerd onder een Creative Commons-licentie? En hoe zit het met de privacygevoeligheid van de informatie?

[3] https://data.overheid.nl

[4] http://www.opencultuurdata.nl

2.6. Datavisualisatie

Beelden zeggen meer dan duizend woorden. Dat gaat verder dan alleen het tonen van de resultaten van een tekstanalyse in een staafdiagram of een grafiek. Denk bijvoorbeeld aan het visualiseren van informatie door middel van een infographic. Of aan buienradar waar op de kaart van Nederland met één oogopslag is te zien waar neerslag te verwachten is. Ook de Casper website waarmee vluchtbewegingen van vliegtuigen van en naar Schiphol live (met een kleine vertraging) kunnen worden gevolgd[5] is een aansprekend voorbeeld.

Figuur 7: Woordwolk op basis van een TF-IDF frequentietabel (opdracht 2).

Het grafisch weergeven van (patronen in) data is een krachtige manier om te informeren of om teksten te analyseren[6]. Bij tekstmining kan bijvoorbeeld met een woordwolk grafisch worden weergegeven welke (betekenisvolle) woorden het vaakst voorkomen in een tekst. Voor het visualiseren van data bestaan diverse tools, zoals Tableau[7]. Voor het maken van woordwolken kan Wordle[8] of een andere tool worden gebruikt (figuur 7). Ook RapidMiner, de tool die bij deze workshop wordt gebruikt, beschikt over mogelijkheden om (patronen in) data te visualiseren.

[5] http://casperflights.com/unified/?location=eham

[6] Zie voor voorbeelden datavisualisatie: https://informationisbeautiful.net

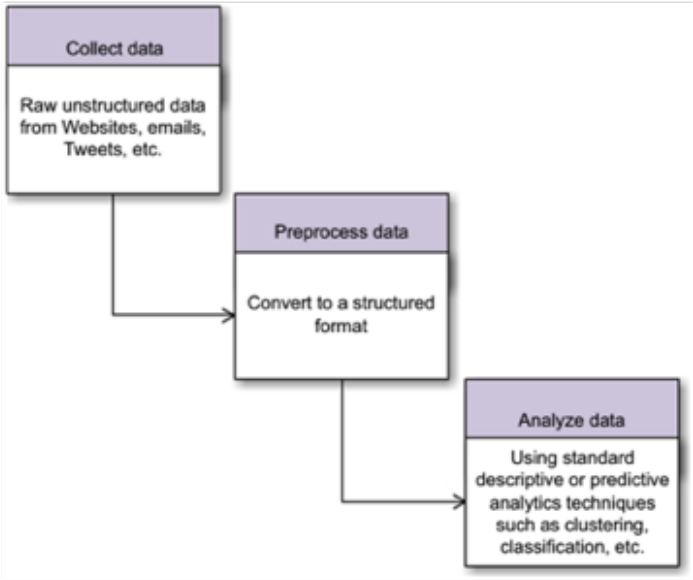
[7] https://www.tableau.com/

[8] http://www.wordle.net/

3. Het tekstminingproces

Het tekstminingproces (figuur 8) is grofweg op te delen in drie subprocessen (Kotu & Deshpande, 2015):

Stap 1: Het bepalen welke data we gaan minen en het inlezen daarvan (**collect**);
Stap 2: Het voorbereiden (opschonen, structureren) van deze data (**pre-process**);
Stap 3: Het analyseren van de data (**analyze**).



Figuur 8: Het miningproces op hoofdlijnen (Kotu & Deshpande, 2015).

3.1. Stap 1: Verzamelen van de data

Voordat we kunnen gaan minen zullen we eerst moeten bepalen wát we willen gaan minen. Hoewel niet precies duidelijk is welke informatie we zullen vinden tijdens het minen (zie inleiding), zijn er wellicht wel vragen die we hiermee hopen te beantwoorden of willen onderzoeken. Zoals bij iedere vorm van research is vraagarticulatie en probleembeschrijving de eerste stap. Wat hoop ik te vinden in de data? Deze vragen bepalen uiteindelijk ook welke datasets we gaan minen. Vragen die daarbij gesteld kunnen worden zijn bijvoorbeeld: Zijn de data native born of gedigitaliseerd? Gestructureerd of niet? Open toegankelijk via het Internet of onderdeel van het deepweb? Big data, sociale media of open data?
Het kan daarbij gaan om allerlei soorten data die zich binnen of buiten de muren van de eigen organisatie bevinden. We kunnen bijvoorbeeld gaan tekstminen binnen:

- Interne content binnen organisaties:
    - Document Management Systemen
    - Customer Relationship Management Systemen
    - Content Management Systemen
    - Netwerkschijven
    - Enterprise Content Management Systemen
    - E-mail

- Externe openbaar beschikbare data:
    - Open Access data
    - Big data
    - Open Data
    - Sociale mediaberichten
    - Webteksten of complete websites

Tekstminen kan plaatsvinden op allerlei verschillende bestandstypes, variërend van open standaarden tot algemene marktstandaarden, zoals bijvoorbeeld:

- Tekstverwerkingsdocumenten of spreadsheets (DOC, ODF, ODT, XLS, PDF, RTF)
- E-mails (PST, MSG, EMAIL, MBOX)
- Sociale media tweets (Twitter, Facebook)
- Webpagina's (HTML, XML)
- Etc.

3.2. Stap 2: Voorbereiden van de data

De volgende stap in het tekstminingproces bestaat uit het voorbereiden (preprocessing) en opschonen (cleansing) van de data. Volgens Forbes (Press, 2016) bestaat zo'n 60% van het werk van een data scientist uit het opschonen en voorbereiden (organiseren) van de data. In dit hoofdstuk worden een aantal typische acties besproken die in deze voorbereidende fase van het miningproces kunnen worden uitgevoerd (tabel 1).

Tabel 1: Voorbereiden van de data (Kotu & Deshpande, 2015).

Table 9.3 A Typical Sequence of Preprocessing Steps to Use in Text Mining

Step	Action	Result
1	Tokenize	Convert each word or term in a document into a distinct attribute
2	Stopword removal	Remove highly common grammatical tokens/words
3	Filtering	Remove other very common tokens
4	Stemming	Trim each token to its most essential minimum
5	n-grams	Combine commonly occurring token pairs or tuples (more than 2)

3.2.1. Transform cases

Voorafgaand aan tokenization kan de (ongestructureerde) tekst worden geconverteerd naar kleine letters (down casing) of juist naar hoofdletters (uppercasing) om hoofdletter-gevoeligheid tijdens het minen uit te sluiten. Zo wordt dan bijvoorbeeld het woord 'Huis' hetzelfde behandeld als 'huis' en niet als twee aparte tokens. Down- of uppercasing is in veel gevallen een goede keuze bij het voorbereiden van de data, maar dit kan wel tot informatieverlies leiden. Zo betekent 'RAM' (Random Access Memory) iets anders dan 'ram' (een mannelijk schaap of een werkwoordsvorm). Ook worden hoofdletters soms gebruikt om een boodschap te benadrukken (zoals in spamberichten). Door down casing verwijder je dus ook kenmerken die misschien gebruikt hadden kunnen worden tijdens de tekstanalyse. By default maakt RapidMiner onderscheid tussen kleine letters en hoofdletters.

3.2.2. Tokenize

Een belangrijke (vaak eerste) stap bij tekstanalyse is om structuur aan te brengen in de ingelezen tekst, dus om de data te transformeren van een ongestructureerde naar een (semi-) gestructureerde tekst. Dit wordt gedaan door alle woorden in een tabel op te slaan als aparte 'tokens' voor verdere verwerking. Tokenization kan worden gedefinieerd als het proces om teksten op te delen in kleinere betekenisvolle entiteiten die tokens worden genoemd. Tokens zijn onafhankelijke en de minimale 'tekstcomponenten' die beschikken over een zekere syntax en semantiek (Sarkar, 2016). Zo wordt dan bijvoorbeeld de zin: "The quick brown fox jumps over the lazy dog"[9] opgesplitst in de volgende tokens (in XML):

```
<sentence>
  <word>The</word>
  <word>quick</word>
  <word>brown</word>
  <word>fox</word>
  <word>jumps</word>
  <word>over</word>
  <word>the</word>
  <word>lazy</word>
  <word>dog</word>
</sentence>
```

3.2.3. Stopwoorden verwijderen

Vervolgens wordt de tekst opgeschoond door betekenisloze stopwoorden (zoals lidwoorden) weg te filteren. In RapidMiner bestaan aparte operatoren voor het verwijderen van stopwoorden uit het Engels, Frans, Arabisch of Duits (helaas niet voor het Nederlands).

3.2.4. Filtering

Onder filtering valt het vastleggen van aan elkaar verwante termen en synoniemen. Dit proces heet '*lexical substitution*' en legt (net als in een thesaurus) semantische relaties tussen termen (bv. voorkeurstermen en niet voorkeurstermen) zodat woorden met dezelfde betekenis op dezelfde manier worden geanalyseerd. Woning wordt dan bijvoorbeeld hetzelfde behandeld als huis en fiets hetzelfde als rijwiel.

Ook het uit de data verwijderen van betekenisloze strings, karakters of opmaakcodes (HTML-codes bijvoorbeeld) valt onder filtering. Verder kan het nodig zijn om bestanden op te schonen waarin bepaalde speciale karakters staan of welke zijn vervuild met strings (stukken tekst) die niet tot de natuurlijke taal behoren die we willen gaan analyseren. In RapidMiner bestaan er diverse operatoren voor filtering. Zo bestaan er operatoren om strings (woorden) kleiner of groter dan een bepaald aantal letters weg te filteren. Bijvoorbeeld het wegfilteren van strings kleiner dan 4 en groter dan 25 karakters.

[9] Deze zin wordt bij NLP en tekstanalyse vaak als voorbeeld gebruikt omdat alle letters uit het (Engelse) alfabet hierin aanwezig zijn.

3.2.5. Stemming

Bij stemming gaat het om het reduceren van een woordenlijst tot een lijst met woorden met hetzelfde 'morfeem'[10] of woordstam. Een morfeem is de kleinste vorm of vervoeging van een woord die kan bestaan zonder dat daarbij betekenisverlies optreedt. Denk aan woorden in de meervoudsvorm, in de vergrotende of overtreffende trap of woorden die grammaticaal zijn vervoegd.

Voorbeeld:
vreemd|***s*** | vreemd|***e*** | vreemd|***er*** | vreemd|***st*** worden behandeld als vreemd (morfeem)

Voor het Engels is het 'Porter algoritme'[11] een veel gebruikt algoritme voor stemming.

3.2.6. n-Grams bepalen

n-Grams zijn woordcombinaties, samengestelde woorden of eigen namen die samen een (semantische) eenheid vormen. n-Grams kunnen bestaan uit 1, 2, 3, 4 of 'n' woorden[12].

Voorbeeld:
"hoger beroepsonderwijs" (2-gram)
"sociaaleconomische status" (2-gram)
"natural language processing" (3-gram)

In RapidMiner kan de 'Generate n-grams (Terms)' operator worden gebruikt om n-grams te generen. In tabel 2 staan voorbeelden van door RapidMiner gegenereerde n-Grams. Het woord 'customer' komt 10 keer voor in 10 verschillende documenten, de 2-gram 'customer_claim' (merk het liggende streepje op bij deze n-grams) komt in totaal twee keer voor in twee verschillende documenten. Het genereren van n-grams is meestal nodig tijdens mining, maar kan het miningproces wel aanzienlijk vertragen.

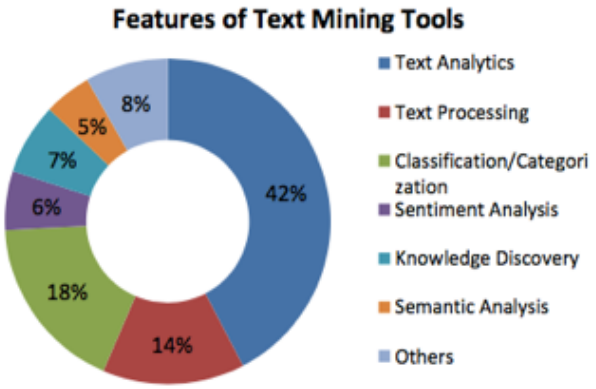
Tabel 2: n-Grams in RapidMiner.

Word	Attribute Name	Total Occurences	Document Occurences
customer	customer	10	10
customer_claim	customer_claim	2	2
customer_place	customer_place	1	1
customer_pleased	customer_pleased	1	1
customer_selected	customer_selected	1	1
customer_service	customer_service	3	3
customer_services	customer_services	2	2

[10] https://nl.wikipedia.org/wiki/Morfeem
[11] http://snowball.tartarus.org/algorithms/porter/stemmer.html
[12] https://en.wikipedia.org/wiki/N-gram

3.3. Stap 3: Analyseren van de data

De laatste stap bestaat uit het analyseren van de data. In het eerder genoemde onderzoek van Kaur en Chopra (2016) is ook gekeken naar de specifieke functies van de door hen onderzochte tekstmining tools. Op basis van deze analyse maakten zij onderscheid in de volgende functies: tekstanalyse, tekstverwerking en tekstbewerking, classificatie, sentimentanalyse, knowledge discovery en semantische analyse (figuur 9).



Figuur 9: Features of text mining tools (Kaur & Chopra, 2016).

- **Tekstanalyse**: Het vinden van bruikbare informatie en patronen in ongestructureerde tekst;
- **Tekstverwerking/Tekstbewerking:** Het (voor-) bewerken en manipuleren van ongestructureerde tekst zodat het kan worden geanalyseerd;
- **Classificatie/Categorisering**: Het herkennen en toekennen van ‘klassen’, ‘typen’ of ‘categorieën’ aan een hoeveelheid data of tekst met als uiteindelijk doel om hierover voorspellingen te kunnen doen;
- **Sentimentanalyse**: Het identificeren van subjectieve elementen (zoals meningen) en sentimenten (positief of negatief qua toon) in een tekst;
- **Knowledge Discovery**: Het vinden van bruikbare nieuwe kennis in grote hoeveelheden tekst**;**
- **Semantische analyse**: Het proces om syntactische structuren (frases, zinnen, paragrafen) in een tekst in verband te brengen met de betekenis van de tekst als geheel (Vrij vertaald naar: Kaur & Chopra, 2016).

3.3.1. Frequentieanalyse

De simpelste vorm van tekstanalyse is het maken van een frequentietabel: een lijst van tokens die aflopend is gesorteerd op het aantal keren dat deze tokens (of n-grams) voorkomen in het document (Term Occurrences) of in het totale tekstcorpus (Term Frequency). Maar zijn de meest voorkomende tokens in een document of tekstcorpus ook de meest betekenisvolle en representatieve woorden voor wat betreft de inhoud?
Om te bepalen hoe belangrijk een woord is binnen de te minen documentenverzameling wordt bij tekstmining de TF-IDF-waarde berekend. TF-IDF is het product van twee numerieke waarden: Term Frequency (TF) en Inverse Document Frequency (IDF)[13] en levert een waarde op tussen de 0 (lage TF-IDF-waarde) en de 1 (hoge TF-IDF-waarde). Woorden met een hoge TF-IDF-waarde zijn dus het meest betekenisvol voor wat betreft de inhoud en dus ook het meest zinvol voor onze tekstanalyse.

De TF-waarde is een numerieke waarde die aangeeft hoe vaak een woord voorkomt in een bepaald corpus[14]. Echter, als een woord vaak voorkomt dan betekent dit nog niet dat het woord 'informatierijk' of 'betekenisvol' en dus belangrijk voor de analyse is. Denk bijvoorbeeld aan veel voorkomende woorden als lidwoorden, voorzetsels, aanwijzende en bijvoeglijke naamwoorden. Deze woorden komen in bijna alle Nederlandstalige documenten voor en geven weinig informatie over de inhoud van het document. De IDF-waarde corrigeert dit door voor de woorden een score te berekenen met betrekking tot de vraag hoe 'zeldzaam' en 'bijzonder' het woord is binnen het tekstcorpus dat we minen. Woorden die in bijna alle Nederlandstalige documenten voorkomen zoals 'daar' of 'wat' krijgen dus een lage IDF-waarde (dichterbij de waarde 0). Woorden die weinig voorkomen in het corpus krijgen een hoge IDF-waarde (dichterbij de waarde 1).

Het bepalen van de TF-IDF-waarde is een typisch onderdeel tijdens de voorbereidende fase. Met deze waarde kan een frequentietabel worden gemaakt met tokens die representatief zijn voor de inhoud van het document. Een frequentietabel kan worden gevisualiseerd door middel van een woordwolk (tagcloud). De tokens/woorden met de hoogste TF-IDF-waarde worden dan het grootst weergegeven in de woordwolk.

3.3.2. Classificatie

Bij het classificeren van data wordt met behulp van statistische (probabilistische) en linguïstische technieken automatisch categorieën toegekend aan een datacluster met als doel om hierover voorspellingen te kunnen doen. Bijvoorbeeld door sociale mediaberichten te analyseren, te classificeren en te 'voorspellen' of de tekst door een man of door een vrouw is geschreven. Of bij het inlezen van e-mail te voorspellen of een mail spam is of niet.

[13] https://en.wikipedia.org/wiki/Tf%E2%80%93idf

[14] Een corpus kan bestaan uit één of meerdere documenten. Zie hoofdstuk 7.

3.3.2.1. Naïve Bayes algoritme

Voor het classificeren van data bestaan verschillende algoritmes. Een binnen de information retrieval en tekstmining veel gebruikt algoritme is het Naïve Bayes[15] algoritme dat is gebaseerd op het theorema van Bayes[16]. Het theorema van Bayes drukt de kans dat een bepaalde mogelijkheid ten grondslag ligt aan een gebeurtenis uit in de voorwaardelijke kansen op de gebeurtenis bij elk van de afzonderlijke mogelijkheden. Bij toepassing van het theorema wordt uitgegaan van een reeds bekende kans P(A) op een gebeurtenis A, de zogenaamde a-priori-kans, op basis van eerder onderzoek. Na waarneming van een gerelateerde gebeurtenis B is kennis verkregen over de kans van optreden van A. Deze nieuwe kans wordt de a-posteriori-kans genoemd en is juist de voorwaardelijke kans P(A|B) ("Theorema van Bayes", 2018). In een formule ziet dit er als volgt uit:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

In opdracht 1 (spamdetectie) wordt het Naïve Bayes algoritme gebruikt om te bepalen of een Sms-bericht 'Ham' is of 'Spam'. Dit algoritme wordt veel toegepast bij het classificeren van (met name) grote hoeveelheden data en werkt in veel gevallen sneller en minstens zo goed als meer geavanceerde algoritmes en voorspellingsmodellen. Victor Lavrenko heeft een playlist[17] op YouTube gepubliceerd over dit algoritme (Lavrenko, 2015). Wie meer wil weten over de achtergrond en de werking van dit algoritme kan zijn video's bekijken.

Classificatietechnieken kunnen ook worden gebruikt om de data automatisch te 'verrijken' door hieraan gemeenschappelijke kenmerken (metadata) toe te kennen. De hierbij toegekende categorieën zullen bijna altijd worden toegekend op basis van een van te voren opgestelde informatietaal, zoals een classificatie, een taxonomie of een thesaurus (Aalten e.a., 2017). Bij deze taak kan de informatieprofessional een belangrijke rol spelen.

3.3.2. Sentimentanalyse

Bij sentimentanalyse (ook wel 'opinion mining' genaamd) gaat het om het vinden van subjectieve elementen in natuurlijk taal, zoals meningen. Met behulp van algoritmes kunnen teksten (bijvoorbeeld berichten in de sociale media) worden gecategoriseerd als positief, negatief of neutraal. Ook dit vakgebied is nog volop in ontwikkeling en wordt bij bedrijven bijvoorbeeld gebruikt om sociale mediaberichten te monitoren (webcare) met als doel om snel in te grijpen als eventuele reputatieschade dreigt.

[15] https://en.wikipedia.org/wiki/Naive_Bayes_classifier
[16] https://nl.wikipedia.org/wiki/Theorema_van_Bayes
[17] https://www.youtube.com/playlist?list=PLBv09BD7ez_6CxkuiFTbL3jsn2Qd1IU7B

4. Beschikbare tools voor tekstmining

Voor het minen van tekst zijn diverse tools en softwarepakketten beschikbaar. In het artikel van Kaur en Chopra (2016) is een overzicht te vinden van de beschikbare miningsoftware waarbij deze tools met elkaar worden vergeleken op bijvoorbeeld het type tool of de aangeboden functionaliteit. Daarnaast bestaan er ook diverse online tools voor tekstmining, tekst- en sentimentanalyse, waaronder:

Tabel 3: Overzicht van online tekstmining tools.

Online tool	**URL**
Find Keyword	http://www.find-keyword.com/
Google N-gram viewer	https://books.google.com/ngrams/
Opendover	http://www.opendover.nl/
Lingpipe	http://alias-i.com/lingpipe/index.html
Ranks	https://www.ranks.nl/
SentiWordNet	http://sentiwordnet.isti.cnr.it/
Tagcrowd	https://tagcrowd.com/
Text analyser	https://www.usingenglish.com/resources/text-statistics.php
Text Analysis Online	http://textanalysisonline.com/
Textalyser	http://textalyser.net/
Textfixer	https://www.textfixer.com/
Voyant	http://voyant-tools.org/
Woordwolk	https://www.woordwolk.nl/
Wordle	http://www.wordle.net/

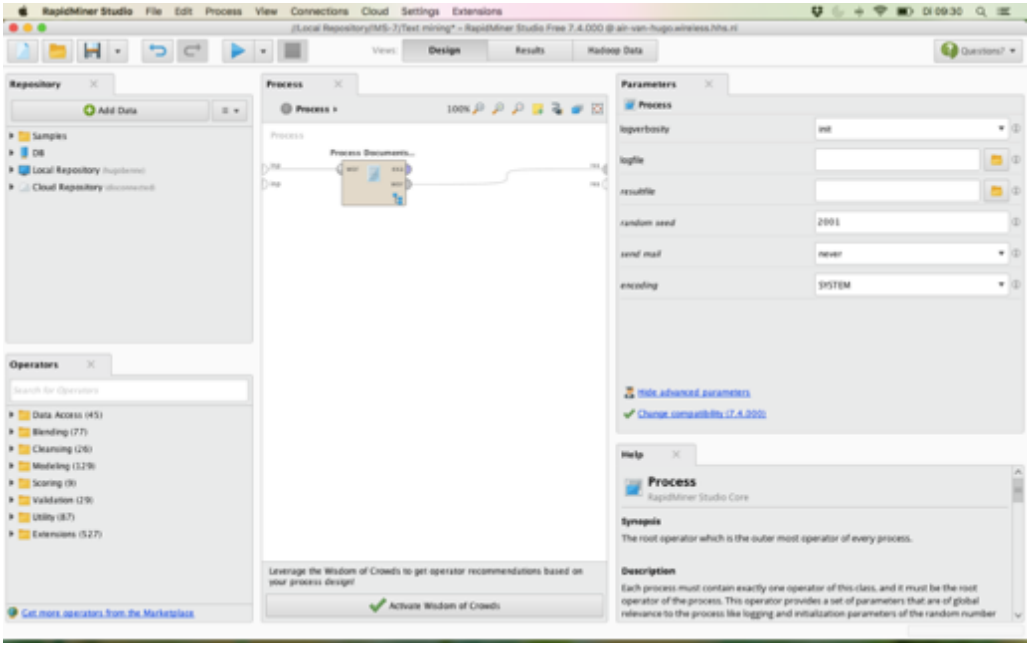
Zie verder ook deze overzichtspagina's (secundaire bronnen) met informatie over en links naar (online) tekstmining tools:

- Digital tools for textual analysis:
  https://folgerpedia.folger.edu/Digital_tools_for_textual_analysis
- DIRT (Digital Research Tools)
  http://dirtdirectory.org/categories/text-mining
- Duke University: Introduction to text analysis
  https://guides.library.duke.edu/text_analysis
- Free online text tools:
  https://www.online-utility.org/text/index.jsp
- Tapor
  http://tapor.ca/home

Iedere tool heeft zijn specifieke kenmerken en functionaliteit. In deze workshop gebruiken we RapidMiner als miningtool omdat deze tool van alle onderzochte tools in de breedte het meest compleet is en bovendien gratis is te gebruiken voor datasets tot 10.000 records.

5. RapidMiner

RapidMiner is een tool voor datamining en beschikt over diverse add-ons voor tekstmining. Deze tool is gratis beschikbaar voor datasets tot 10.000 records en is geschikt voor zowel Windows als Apple computers. Er bestaat een speciale onderwijslicentie voor RapidMiner die de beperking niet heeft[18]. In afbeelding 1 zie je het overzichtsscherm van RapidMiner. Linksboven is een vak waar de datasets worden weergegeven. Dit noemen we de repository. Linksonder de operatoren (functies) die binnen het miningproces kunnen worden toegepast. In het midden (het ontwerpgrid) kan het miningproces worden ontworpen en de operatoren worden geplaatst. Onder de verschillende tabbladen worden de resultaten weergegeven. Rechtsboven kunnen de parameters en variabelen voor de toegepaste operatoren worden ingevoerd. Rechtsonder bevindt zich een helpfunctie. Met de blauwe 'play' button in de toolbar kan een proces worden gestart en uitgetest.



Afbeelding 1: RapidMiner Studio.

5.1. Installatie RapidMiner

- Installeer voorafgaand aan de workshop het programma RapidMiner (versie 8.1 of hoger) op je laptop of Mac, zie https://rapidminer.com/
- Geef bij het downloaden van RapidMiner aan dat je gebruik wilt maken van de volgende extensions:
    - Tekst Analysis by AYLIEN (voor Windows/MacOS). Voor deze extension moet je je registreren bij AYLIEN. Vervolgens krijg je een token om de extension te activeren.
    - Tekst Processing
    - WordNet Extension (alleen op MacOs)
    - Web Mining

[18] https://rapidminer.com/educational-program/

- Installeren van de extensions kan binnen RapidMiner via de menuoptie ‘Extensions’ en dan kiezen voor ‘Marketplace’;
- Mocht je werkzaam zijn in het (hoger) onderwijs, dan is het aan te raden om een (gratis) onderwijslicentie aan te vragen en te installeren;
- **Alleen voor MacOS**: Download de WordNet 3.0 Englisch dictionary bestanden van https://wordnet.princeton.edu/download/current-version
  Kies daarbij de versie voor ‘Unix-like systems’. Pak het ZIP-bestand uit en installeer het op je laptop;
- Voor een aantal opdrachten is het noodzakelijk om over een Twitter-account te beschikken, RapidMiner 8.1 of hoger geïnstalleerd te hebben of over een onderwijslicentie op RapidMiner te beschikken. Dit staat bij de opdrachten aangegeven.

5.2. RapidMiner documentatie

Voor het uitvoeren van de opdrachten kan gebruik worden gemaakt van de volgende documentatie over RapidMiner:

RapidMiner Documentation : Getting started with RapidMiner Studio. (2017). Geraadpleegd 18 december 2017, van https://docs.rapidminer.com/latest/studio/getting-started/

RapidMiner Documentation : Operators manual. (2017). Geraadpleegd 18 december 2017, van http://docs.rapidminer.com/studio/operators/

RapidMiner Documentation : What's new in RapidMiner Studio 8.0. (2017). Geraadpleegd 18 december 2017, van http://docs.rapidminer.com/studio/releases/

RapidMiner Studio Manual. (2017). Geraadpleegd 18 december 2017, van http://docs.rapidminer.com/studio/operators/rapidminer-studio-operator-reference.pdf

5.3. RapidMiner Video tutorials

Bekijk deze video: https://www.youtube.com/watch?time_continue=28&v=kq61oFXD4YI

En eventueel deze video tutorials over RapidMiner op YouTube:

RapidMiner Playlist YouTube. (z.d.). Geraadpleegd 18 december 2017, van https://www.youtube.com/playlist?list=PLssWC2d9JhOZLbQNZ80uOxLypglgWqbJA

5.4. WordNet

WordNet is een lexicologische database voor het Engels, zie https://wordnet.princeton.edu/

In casus 2 en 3 van de Mining workshop wordt WordNet gebruikt voor analyseren van sentimenten en ‘subjectieve elementen’ in de Engelse natuurlijke taal.

Wat we hiervoor nodig hebben is:
- De WordNet Extension voor RapidMiner;
- De WordNet 3.0 database bestanden.

De extension is te installeren binnen RapidMiner (zie Marketplace).

De databasebestanden zelf zijn te downloaden van
https://wordnet.princeton.edu/download/current-version

Windows:
WordNet 3.0 werkt alleen onder MacOS en Unix. Gebruik op Windows computers de Tekst analysis by AYLIEN extension in plaats van de WordNet extension.

MacOs/Unix:
Als je op een Mac werkt dan heb je de keuze voor de WordNet 3.0 of de AYLIEN extension.

6. Opdrachten workshop

Tijdens de workshop gaan we aan de slag met de vier casussen. Casus 1 gaat over het classificeren van data en spamdetectie (§6.1). Casus 2 gaat over filmreviews en sentimentanalyse (§6.2). Casus 3 gaat over sentimentanalyse binnen Twitterberichten (§6.3). Casus 4 tot slot gaat over frequentie-analyse, visualisatie en woordwolken (§6.4).

6.1. Casus 1: Ham of Spam? Spamdetectie

Bij deze casus gebruiken we een dataset[19] van 5574 Sms-berichten die zijn verzameld door de Universiteit van Californië. Er zijn drie versies van deze dataset beschikbaar: klein, middel en groot (150, 500 of 5574 berichten). De kleinste datasets zijn handig om tijd te besparen tijdens het ontwerpen, het uittesten en debuggen van het proces. De bestanden zijn gecodeerd volgens de UTF8 karakterset en bestaan uit twee kolommen (velden). Het eerste veld (HamORSpam) geeft een vooraf door mensen toegekende 'classificatie' van het sms-bericht. Een bericht kan gemarkeerd zijn als 'ham' (gewenst bericht) of als 'spam' (ongewenst spambericht). Het HamOrSpam-veld wordt niet gebruikt voor de classificatie zelf maar dient om achteraf te kunnen controleren (valideren) hoe betrouwbaar het model is geweest in haar voorspelling. Het tweede veld (MailText) bevat de (ongestructureerde) tekst van het Sms-bericht waarop we het classificatiemodel gaan toepassen.

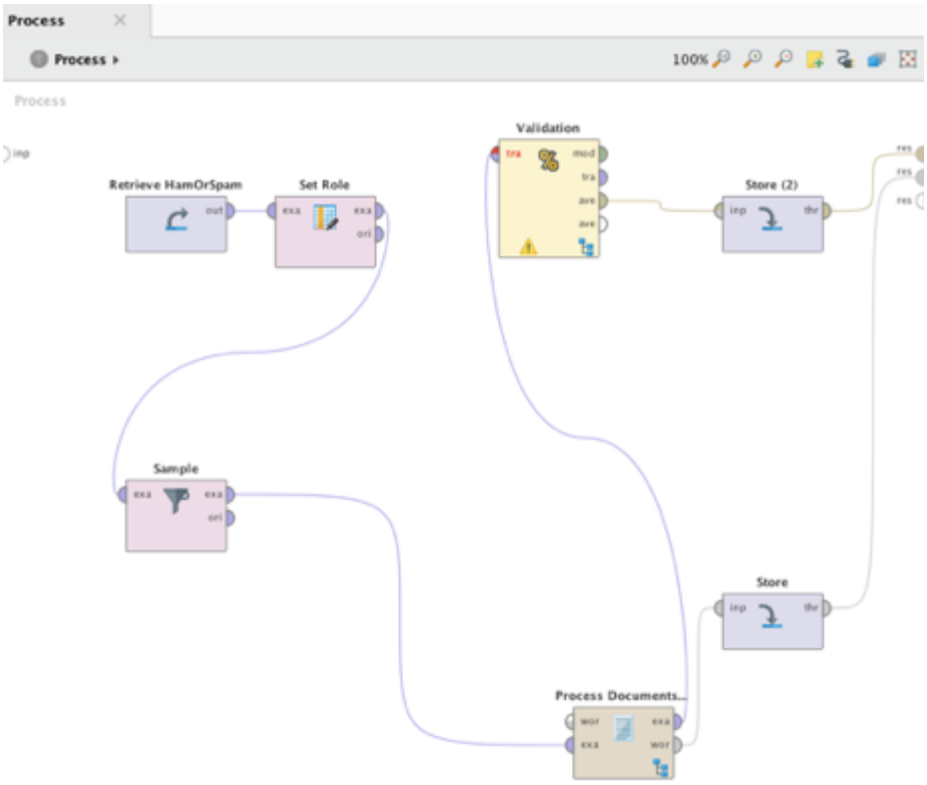
NB. Gebruik bij deze opdracht de casusbeschrijving van Hofmann en Klinkenberg (hoofdstuk 13 over spamdetectie) als achtergrondmateriaal (2014, pp. 199–211).

6.1.1. Opdracht 1: Spamdetectie

1. Plaats de kleinste dataset (150) in het lokale repository (linksboven) en noem deze dataset 'HamORSpam (150)';
2. Sleep de 'HamORSpam' dataset op het ontwerpgrid of gebruik hiervoor de 'Retrieve' operator;
3. Plaats de 'Set role' operator op het grid. Hiermee kan worden aangegeven van welk type het veld is. Dit doe je door op 'Set role' te klikken en bij 'Edit list' (bij Parameters) de velden te definiëren. Definieer het veld 'HamORSpam' (de eerste kolom) van het type 'label' en 'MailText (de tweede kolom) van het type 'regular';
4. Plaats de 'Sample' operator op het grid. Hiermee wordt de grootte bepaald van de set waarop we de classificatie gaan baseren. Klik op 'Balance data'. Voeg bij 'Edit list' in de Sample operator een 'spam'-klasse toe met als grootte '747' en een 'ham'-klasse van dezelfde grootte;
5. Plaats de 'Process Documents from data' operator op het ontwerpgrid. Dubbelklik op deze operator zodat je dit subproces (afbeelding 3) kunt vullen. Plaats hierin de 'Tokenize' operator. Vink bij parameters de opties 'Create word vector', 'Add meta information', 'Keep text' en 'Select attributes and weights' aan. Klik op 'Edit list' en geef attribute 'MailText' 1.0 mee als weight en attribute HamORSpam 0.0 als weight. Keer vervolgens terug naar het hoofdproces via het kruimelpad;

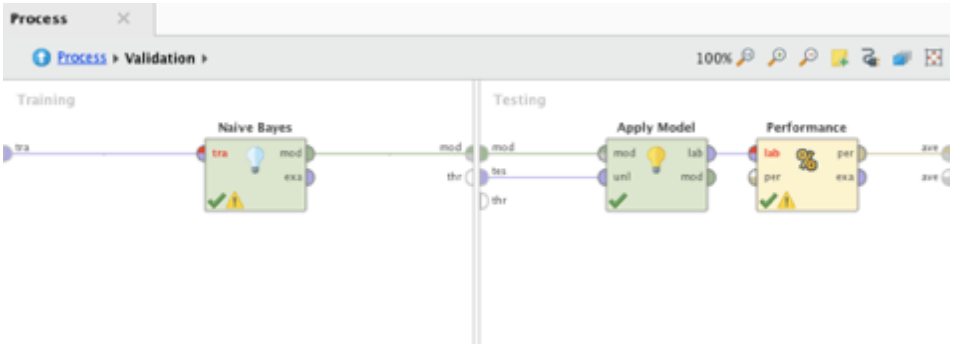
[19] http://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection

6. Plaats achter de ‘Process documents from data’ operator een ‘Store’-operator. Vul bij parameters de naam van het bestand in. Hierin wordt de woordenlijst opgeslagen zodat deze kan worden gebruikt voor verdere analyse;
7. Plaats de ‘Validation (split validation)’ operator op het grid. Deze operator dient om de validiteit van het model te testen op basis van het ‘HamORSpam’ veld;
8. Dubbelklik op de ‘Validation’ operator. Plaats in de linkerkolom de ‘Naïve Bayes’ operator. Met deze operator wordt volgens het Naïve Bayes algoritme[20] het bericht geclassificeerd als ‘ham’ of ‘spam’;
9. Plaats in de rechterkolom de ‘Apply Model’ (afbeelding 3). Plaats hiernaast de ‘Performance’ operator. Met deze operator wordt gemeten hoe betrouwbaar het model is. Dit wordt gedaan door de waarde uit het veld ‘HamORSpam’ te vergelijken met de voorspelde waarde volgens het Naïve Bayes model;
10. Plaats achter de ‘Validation’ operator een ‘Store’ operator. Vul bij parameters de naam van het bestand in. Hierin wordt het model opgeslagen zodat het gebruikt kan worden voor toekomstige spamdetectie;
11. Verbind de operatoren met elkaar zoals weergegeven in afbeelding 2 (hoofdproces) en afbeelding 3 (subproces);
12. Druk vervolgens op de ‘play’ button om het minen te starten;
13. Als het miningproces foutloos verloopt: verander de dataset in HamORSpam (full);
14. Bestudeer op het resultatentablad de Wordlist, de Exampleset en de PerformanceVector;
15. Bekijk in het linker menu de statistieken en grafieken voor deze dataset.



Afbeelding 2: Hoofdproces spamdetectie.

[20] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

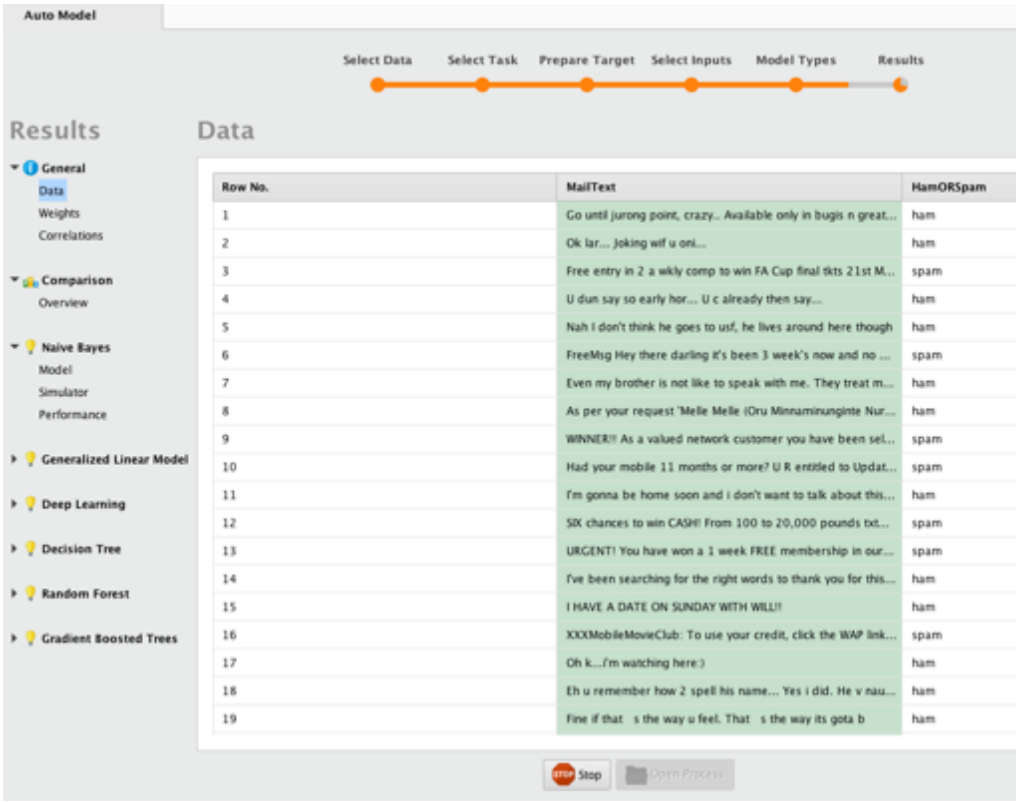


Afbeelding 3: Subproces 'Classificatie'.

6.1.2. Opdracht 2: Auto Model

Deze opdracht kan alleen worden uitgevoerd als er gebruik wordt gemaakt van RapidMiner versie 8.1 of hoger en je beschikt over een onderwijslicentie. Vanaf deze versie bestaat er de mogelijkheid om via een wizard een dataset te analyseren volgens verschillende (standaard) miningtechnieken en -modellen zoals het Naïve Bayes model, Decision Tree, Deep Learning, Random Forest, Comparison, Generalized Linear Model en Gradient Boosted Trees (afbeelding 4).

1. Klik op het 'Auto Model' tabblad;
2. Lees het HamOrSpam (Full) bestand in;
3. Volg de stappen binnen de 'Auto Model'-wizard (kies voor Predict);
4. Start het proces. Dit kan even duren;
5. Bestudeer de resultaten.



Afbeelding 4: Auto Model-functie in RapidMiner (versie 8.1. of hoger).

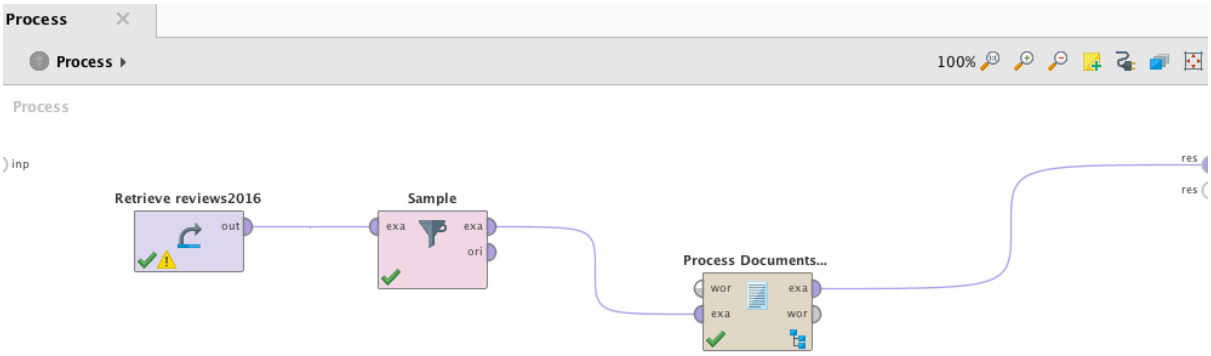
6.2. Casus 2: Sentimentanalyse filmreviews

In het Excelbestand 'Reviews (full).xls' bevinden zich 1226 filmreviews voor 17 verschillende speelfilms. Het doel van deze opdracht is om aan de hand van een analyse te bepalen wat het sentiment is van de review en dit sentiment te vergelijken met het cijfer dat de reviewer aan de film heeft toegekend. Het bestand (karakterset UTF8) is als volgt opgebouwd:

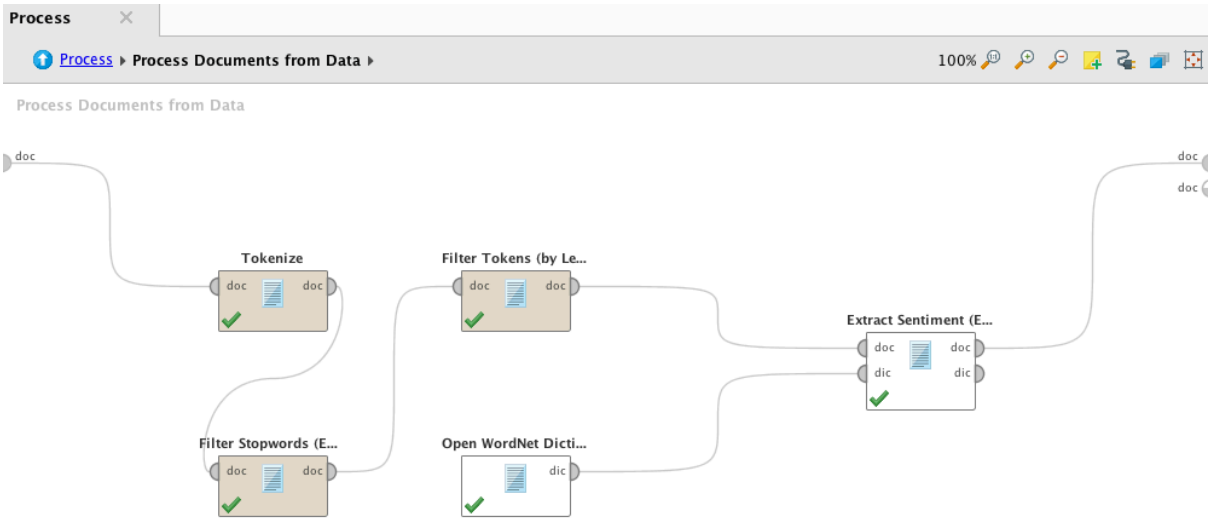
Kolom	Veldnaam (attribuut)	Omschrijving
A	Film	De naam van de speelfilm
B	ID	Het unieke ID (recordnummer) van de film.
C	Cijfer 1	De waardering die de reviewer aan deze film heeft gegeven op een schaal van 1 tot 10. Als er -1 staat ontbreekt dit cijfer.
D	Cijfer 2	Hoeveel lezers dit een waardevolle review vonden.
E	Cijfer 3	Het totaal aantal mensen dat een oordeel gaf over deze review.
F	Titel	Titel van de review.
G	Review	De reviewtekst.

6.2.1. Sentimentanalyse met WordNet

1. Plaats het bestand met filmreviews in het lokale RapidMiner repository (linksboven) en noem deze dataset 'Reviews';
2. Sleep de 'Reviews (full)' dataset op het ontwerpgrid of gebruik hiervoor de 'Retrieve' operator;
3. Sleep de 'Sample' operator op het grid en stel de 'Sample size' in op 300 (absolute);
4. Plaats de 'Proces Documents from data' operator op het grid. Selecteer 'TF-IDF' bij 'Vector creation'. Vink de opties 'Create word vector', 'Add meta information', 'Keep text' en 'Select attributes and weights' aan;
5. Klik op 'Edit list' en geef 'Review' 0.95 mee als weight en 'Titel' 0.5 als weight;
6. Klik op Proces Documents from data' operator om het subproces (afbeelding 6) aan te maken waarin de 'preprocessing' plaatsvindt. Gebruik hiervoor de operatoren: Tokenize, Filter Tokens (by length), Filter Stopwords (English;
7. Plaats hier ook de 'Open WordNet dictionary' en geef bij de parameters het pad aan naar de WordNet 3.0 dictionary bestanden op je laptop;
8. Plaats hierachter de 'Extract Sentiment (English)' operator en verbind de operatoren in dit subproces met elkaar volgens afbeelding 6;
9. Keer via het kruimelpad weer terug naar het hoofdproces;
10. Verbind de operatoren met elkaar zoals staat aangegeven in afbeelding 5;
11. Druk op de play-button om het minen te starten;
12. Bestudeer op het resultatentabblad en de exampleset;
13. Bekijk ik het linker menu de statistieken en grafieken voor deze dataset.
14. Vergelijk de uitkomst van de sentimentanalyse met de cijfers (Cijfer 1) die de reviewers voor deze film hebben gegeven. Komt de uitkomst van de sentimentanalyse overeen met het door de reviewer gegeven cijfer? Of zijn er juist grote verschillen?



Afbeelding 5: Hoofdproces sentimentanalyse met WordNet.

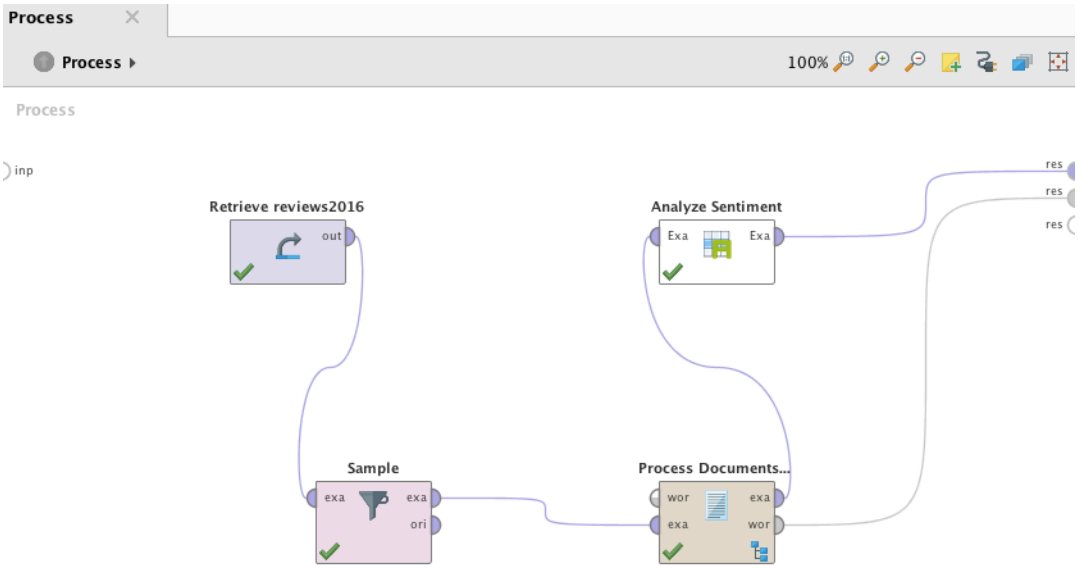


Afbeelding 6: Subproces sentimentanalyse met WordNet.

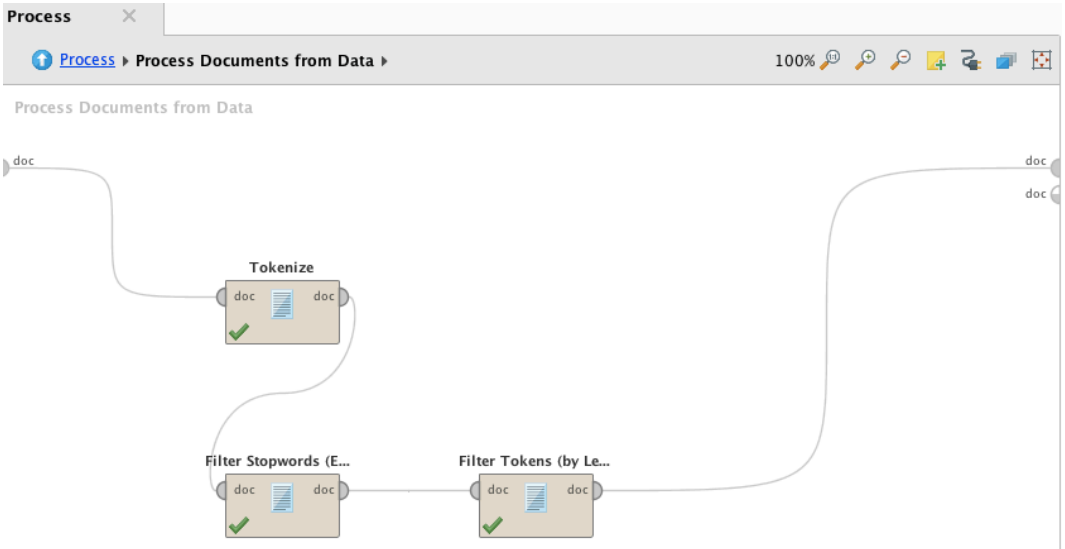
6.2.2. Sentimentanalyse met AYLIEN

1. Plaats het bestand met filmreviews in het lokale RapidMiner repository (linksboven) en noem deze dataset ‘Reviews’;
2. Sleep de ‘Reviews (full)’ dataset op het ontwerpgrid of gebruik hiervoor de ‘Retrieve’ operator;
3. Sleep de ‘Sample’ operator op het grid en stel de ‘Sample size’ in op 300 (absolute);
4. Plaats de ‘Proces Documents from data’ operator op het grid. Selecteer ‘TF-IDF’ bij ‘Vector creation’. Vink de opties ‘Create word vector’, ‘Add meta information’, ‘Keep text’ en ‘Select attributes and weights’ aan;
5. Klik op ‘Edit list’ en geef ‘Review’ 0.95 mee als weight en ‘Titel’ 0.5 als weight;
6. Klik op Proces Documents from data’ operator om het subproces (afbeelding 8) aan te maken waarin de ‘preprocessing’ plaatsvindt. Gebruik hiervoor de operatoren: Tokenize, Filter Tokens (by length), Filter Stopwords (English;
7. Keer via het kruimelpad weer terug naar het hoofdproces;
8. Voeg de AYLIEN operator ‘Analyze Sentiment’ toe op het grid en verbind de operatoren met elkaar zoals aangegeven in afbeelding 7;
9. Druk op de play-button om het minen te starten;
10. Bestudeer op het resultatentabblad en de exampleset;

11. Bekijk ik het linker menu de statistieken en grafieken voor deze dataset.
12. Vergelijk de uitkomst van de sentimentanalyse met de cijfers (Cijfer 1) die de reviewers voor deze film hebben gegeven. Komt de uitkomst van de sentimentanalyse overeen met het door de reviewer gegeven cijfer? Of zijn er juist grote verschillen?



Afbeelding 7: Hoofdproces sentimentanalyse met AYLIEN.



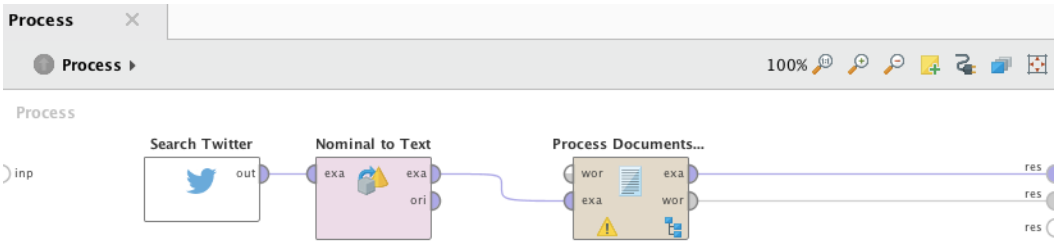
Afbeelding 8: Subproces sentimentanalyse met AYLIEN.

6.3. Casus 3: Sentimentanalyse Twitter

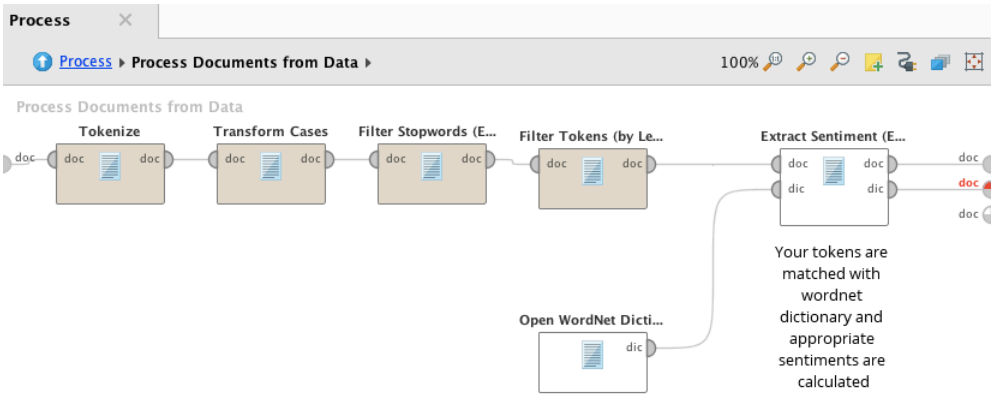
In deze casus gaan we de sentimenten in Twitterberichten analyseren. Daarvoor is het wel nodig om te beschikken over een Twitteraccount.

6.3.1. Sentimentanalyse Twitter met WordNet

1. Plaats de 'Search Twitter' operator op de grid;
2. Verbind (via parameters) de Twitterverbinding met je eigen Twitteraccount;
3. Definieer de zoekquery naar eigen voorkeur en limiteer daarbij het aantal zoekresultaten tot 400;
4. Plaats hierachter de 'Nomimal to Text' operator (attribute Text);
5. Plaats hierachter de 'Proces Documents from Data' operator;
6. Selecteer 'TF-IDF' bij 'Vector creation'. Vink de opties 'Create word vector', 'Add meta information' en 'Keep text' aan;
7. Klik op Proces Documents from data' operator om het subproces (afbeelding 10) aan te maken waarin je de 'preprocessing' gaat uitvoeren. Gebruik hiervoor de operatoren: Tokenize, Transform Cases, Filter Tokens (by length) en Filter Stopwords (English);
8. Plaats hier ook de 'Open WordNet dictionary' en geef bij de parameters het pad aan naar de WordNet 3.0 dictionary bestanden op je laptop;
9. Plaats hierachter de 'Extract Sentiment (English)' operator en verbind de operatoren in dit subproces met elkaar volgens afbeelding 9;
10. Keer terug naar het hoofdproces en druk op de play-button;
11. Bestudeer de sentimenten voor de berichten die zijn geanalyseerd.



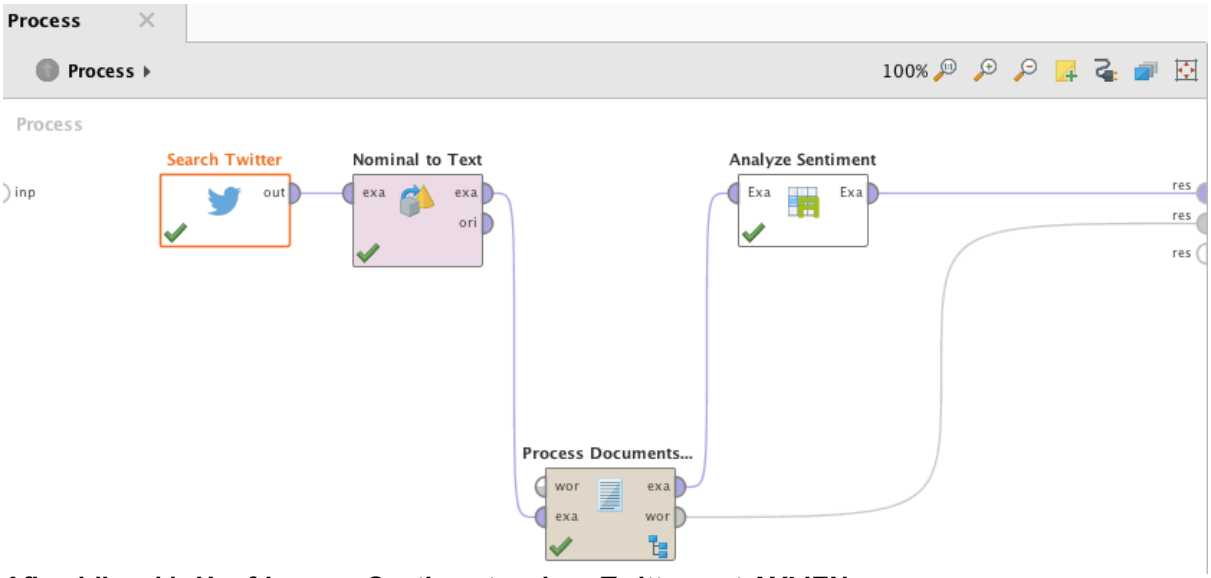
Afbeelding 9: Hoofdproces Sentimentanalyse Twitter met WordNet.



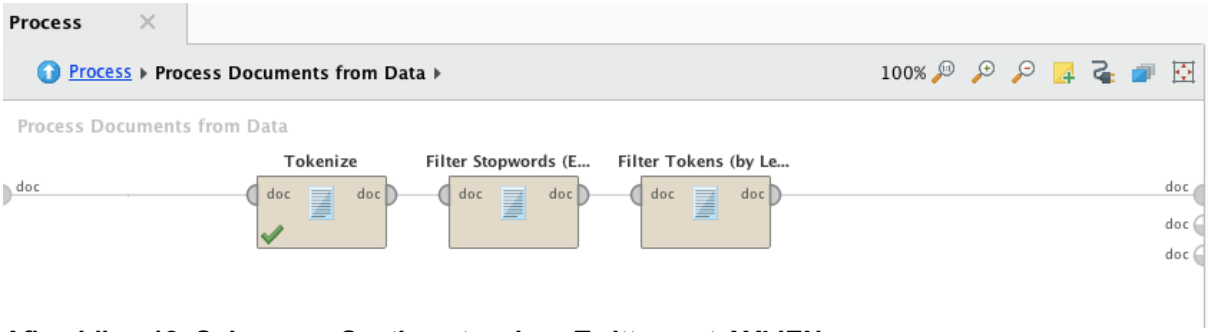
Afbeelding 10: Subproces Sentimentanalyse Twitter met WordNet.

6.3.2. Sentimentanalyse Twitter met AYLIEN

1. Plaats de ‘Search Twitter’ operator op de grid;
2. Verbind (via parameters) de Twitterverbinding met je eigen Twitteraccount;
3. Definieer de zoekquery naar eigen voorkeur en limiteer daarbij het aantal zoekresultaten tot 400;
4. Plaats hierachter de ‘Nomimal to Text’ operator (attribute Text);
5. Plaats hierachter de ‘Proces Documents from Data’ operator;
6. Plaats hierachter de ‘Analyze Sentiment’ operator van AYLIEN;
7. Selecteer ‘TF-IDF’ bij ‘Vector creation’. Vink de opties ‘Create word vector’, ‘Add meta information’ en ‘Keep text’ aan;
8. Klik op Proces Documents from data’ operator om het subproces (afbeelding 12) aan te maken waarin je de ‘preprocessing’ gaat uitvoeren. Gebruik hiervoor de operatoren: Tokenize, Filter Tokens (by length) en Filter Stopwords (English);
9. Keer terug naar het hoofdproces en druk op de play-button;
10. Bestudeer de sentimenten voor de berichten die zijn geanalyseerd.



Afbeelding 11: Hoofdproces Sentimentanalyse Twitter met AYLIEN.



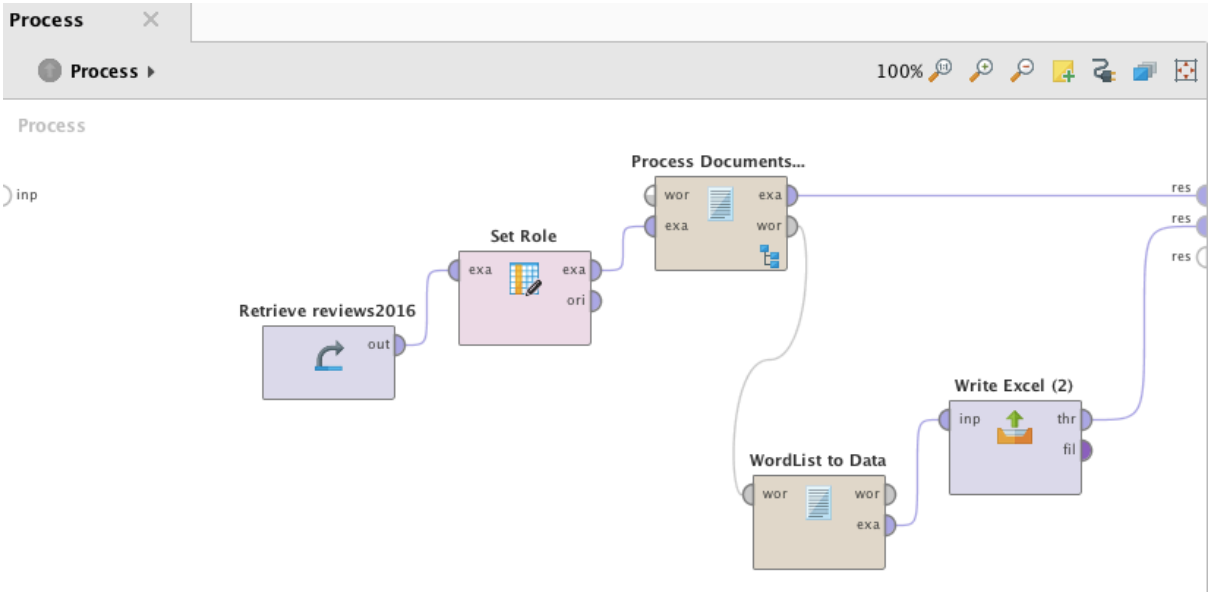
Afbeelding 12: Subproces Sentimentanalyse Twitter met AYLIEN.

6.4. Casus 4: Woordwolk filmreviews

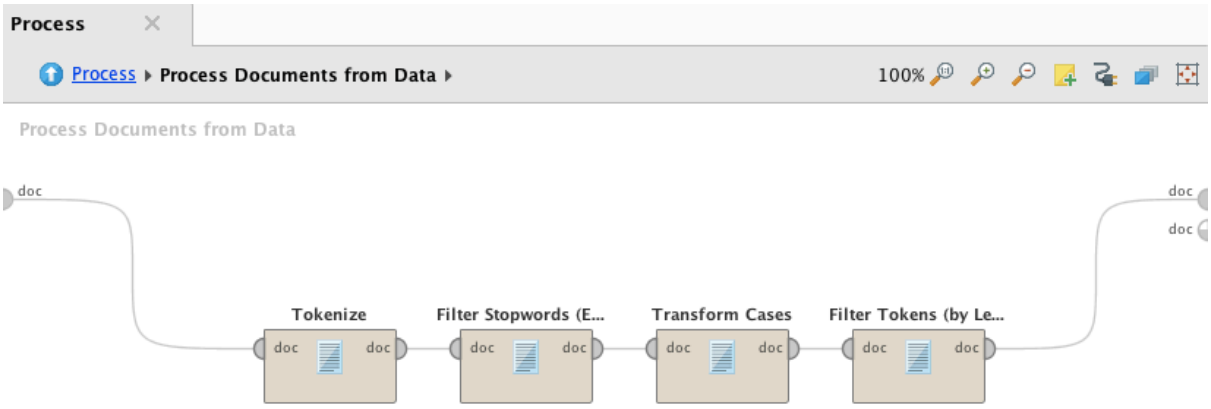
In deze casus maken we (bijvoorbeeld met Wordle[21]) een woordwolk met behulp van de TF-IDF-waardes uit de filmreviews opdracht.

1. Plaats het bestand met filmreviews in het lokale RapidMiner repository (linksboven) en noem deze dataset 'Reviews';
2. Sleep de 'Reviews' dataset op het ontwerpgrid of gebruik hiervoor de 'Retrieve' operator;
3. Plaats de 'Set role' operator op het grid. Definieer 'Titel' en 'Review' van het type 'label', 'Cijfer 1', 'Cijfer 2' en 'Cijfer 3' van het type 'weight', 'Film' van het type 'regular' en 'ID' van het type 'id';
4. Plaats de 'Proces Documents from data' operator op het grid. Selecteer 'TF-IDF' bij 'Vector creation'. Vink de opties 'Create word vector', 'Add meta information', 'Keep text' en 'Select attributes and weights' aan;
5. Klik op 'Edit list' en geef 'Review' 0.95 mee als weight en 'Titel' 0.5 als weight;
6. Klik op Proces Documents from data' operator om het subproces (afbeelding 14) aan te maken waarin je de 'preprocessing' gaat uitvoeren. Gebruik hiervoor de operatoren: Tokenize, Transform Cases, Filter Tokens (by length) en Filter Stopwords (English);
7. Keer via het kruimelpad weer terug naar het hoofdproces;
8. Plaats vervolgens de 'Wordlist to Data' op het grid;
9. Plaats hierachter de 'Write Excel' operator en geef een pad aan (bij parameters) waar de uitvoer moet worden weggeschreven;
10. Verbind de operatoren met elkaar zoals staat aangegeven in afbeelding 13;
11. Druk op de play-button om het minen te starten;
12. Open de uitvoer in het Excelbestand en bestudeer deze uitvoer;
13. Gebruik deze uitvoer om met een (online) tool (bv. Wordle) een woordwolk te maken. Zie figuur 7 als voorbeeld. Zorg daarbij dat de woorden in de woordwolk in dezelfde verhouding in grootte verschillen als de TF-IDF-waardes aangeven.

[21] http://www.wordle.net/advanced



Afbeelding 13: Hoofdproces Termfrequenties



Afbeelding 14: Subproces Termfrequenties

7. Gebruikte terminologie en verklarende woordenlijst

De totale verzameling documenten die we gaan tekstminen noemen we een corpus of tekstcorpus (meervoud 'corpora')[22]. Een corpus kan bestaan uit één of meerdere (tekst-) documenten. Ieder document kan weer bestaan uit hoofdstukken, alinea's, paragrafen, zinnen en woorden. Al deze tekstonderdelen (entiteiten) kunnen apart of in samenhang worden gemined en geanalyseerd. Een zin in een tekst of een regel/record in een dataset kan in de context van tekstmining als apart 'documenten' worden gezien. De kleinste entiteiten zijn de losse woorden (tokens) die in de tekst worden gescheiden door middel van spaties. Hieronder worden een aantal gebruikte termen en begrippen toegelicht:

Tabel 4: Gebruikte terminologie en verklarende woordenlijst.

Begrip	Verklaring
Attribute (attribuut)	Een kolom- of veldnaam.
Classificatie	Het herkennen en toekennen van 'klassen', 'typen' of 'categorieën' aan een hoeveelheid data of tekst met als uiteindelijk doel om hierover voorspellingen te kunnen doen.
Cluster	Hoeveelheid data of objecten die over dezelfde eigenschap beschikken.
Clustering	Clustering is het groeperen van een hoeveelheid data of objecten met vergelijkbare eigenschappen.
Clusteranalyse	Zie clustering.
Corpus	De totale verzameling documenten die we gaan minen noemen we de tekstverzameling, corpus of tekstcorpus (meervoud 'corpora'). Een corpus kan bestaan uit één of meerdere tekstdocumenten.
CSV	Comma Separated Values. Een gestructureerd bestandsformaat waarbij de waarden (velden) zijn gescheiden door een komma.
Datamining	Datamining is het proces waarbij grote hoeveelheden data in een databank (automatisch) worden geanalyseerd en waarbij men op zoek gaat naar statische verbanden (patronen) binnen datasets. Datamining wordt ook wel knowledge discovery genoemd.
Entiteit	Opzichzelfstaande en bij elkaar horende eenheid.
Dataset	Een (bij elkaar horende) verzameling (meestal gestructureerde) gegevens, bijvoorbeeld een Excelbestand of relationele database.
Filtering	Het wegfilteren van betekenisloze woorden, letters, cijfers, karakters uit een corpus of woordenlijst.
Flipped classroom	Bij flipped classroom vindt de klassikale instructie (overdracht van theorie) voorafgaand aan de les plaats. Tijdens het contactmoment of les zelf wordt dan geen theorie meer gegeven, maar gaan de trainees aan de slag met een (verdiepende) opdracht of casus.

[22] https://en.wikipedia.org/wiki/Text_corpus

Gestructureerde informatie	Data die ligt opgeslagen in een vast bestandsformaat waarbij de gegevens zijn opgedeeld in records en velden (bv. relationele databases) of kolommen en rijen (bv. spreadsheets).
HTML	HyperText Markup Language
IDF	Inverted Document Frequency. Een numerieke (logaritmische) waarde tussen de 0 en 1 die aangeeft hoe 'zeldzaam' en 'bijzonder' het woord is binnen het totale tekstcorpus dat we minen. Als het getal dicht bij de 0 ligt is het een algemeen woord met weinig betekenis. Als het getal dicht bij de 1 ligt is het woord specifiek en betekenisvol.
Knowledge discovery	Zie datamining.
Label attribute	Het 'doel' attribuut. De veld- of kolomnaam die je wilt gaan vergelijken met andere attributen of wilt gaan analyseren.
Morfeem	Een morfeem is de kleinste vorm of vervoeging van een woord die in een taal kan bestaan zonder dat daarbij betekenisverlies optreedt.
n-Gram	n-Grams zijn woordcombinaties, samengestelde woorden of eigen namen die samen een (semantische) entiteit of eenheid vormen. n-Grams kunnen bestaan uit 1, 2, 3, 4 of 'n' woorden.
Ongestructureerde informatie	Dit type data is niet opgeslagen in een van tevoren vastgelegde structuur. Het gaat daarbij bijvoorbeeld om websites, audio- en videobestanden, afbeeldingen, tekstbestanden of sociale mediaberichten.
Opinion mining	Zie sentimentanalyse.
Role	Het veldtype behorende bij een attribuut.
Sentimentanalyse	Het identificeren van subjectieve elementen (zoals meningen) en sentimenten (positief of negatief qua toon) in een tekst.
Semigestructureerde informatie	Data die niet is gestructureerd maar wel metadata, tags of andere coderingen bevat (denk bijvoorbeeld aan HTML opmaakcodes).
Stemming	Bij stemming gaat het om het reduceren van een woordenlijst tot een lijst met woorden met dezelfde 'morfeem' of woordstam.
SMS	Short Message Service
SQL	Structured Query Language. Querytaal voor relationele database management systemen (RDBMS).
Tekstcorpus	Zie corpus.
Term frequency (TF)	Het totaal aantal keer dat een woord voorkomt in het tekstcorpus of in de dataset.
Term occurrences	Het totaal aantal keer dat een woord voorkomt in een document of record.
Text mining	Datamining op grote hoeveelheden (ongestructureerde) data zoals tekstbestanden.
TF-IDF	Het product van TF*IDF. Met de berekende waardes kan een vectordiagram worden gemaakt met woorden die betekenisvol zijn en dus zinvol om te analyseren.
Tokenization	Het opdelen van een hoeveelheid ongestructureerde tekst in tokens.

Tokens	De kleinste semantische entiteit in een hoeveelheid natuurlijk taal. Tokens zijn meestal losse woorden of n-Grams.
UTF	Universal Character Set Transformation Format. Een universele en internationale karakterset die de meeste talen ondersteunt.
XML	Extensible Markup Language

Gebruikte literatuur

Aalten, J. van, Becker, P., Linden, M. van der, & Sieverts, E. (2017). *Maak het vindbaar op schijven, sites en SharePoint (opentextbook)*. Geraadpleegd van https://udocstore.nl/docs/9789492388001/maak-het-vindbaar-op-schijven--sites-en-sharepoint

Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, *14*.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, *56*(12), 64–73.

EDISON consortium. (2017). *EDISON Data Science FrameworK (EDSF)*. Geraadpleegd van http://edison-project.eu/

Fouad, M., Oweis, N., Gaber, T., Ahmed, M., & Snasel, V. (2015). Data Mining and Fusion Techniques for WSNs as a Source of the Big Data. In *Procedia Computer Science* (Vol. 65). https://doi.org/10.1016/j.procs.2015.09.023

Gurin, J. (2014). Big Data and Open Data: How Open Will the Future Be. *I/S: A JOURNAL OF LAW AND POLICY FOR THE INFORMATION SOCIETY*, *10*(3), 691–704.

Hofmann, M., & Klinkenberg, R. (2014). RapidMiner data mining use cases and business analytics applications. Geraadpleegd van http://0proquest.safaribooksonline.com/9781482205503

Hurwitz, J., Nugent, A., Halper, F., & Kaufman, M. (2017). *Big data voor dummies*. (W. Zefat, Vert.). BBNC Uitgevers: Amersfoort.

IBM - What is big data? Bringing big data to the enterprise. (z.d.). Geraadpleegd 2 januari 2018, van https://www.ibm.com/software/in/data/bigdata/

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. . (2014). Big data and its technical challenges. *Communications of the ACM*, *57*(7), 86–94.

Kaur, A., & Chopra, D. (2016). Comparison of text mining tools (pp. 186–192). Gepresenteerd bij Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO), 2016 5th International Conference on, IEEE.

Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Mahmoud Ali, W. K., Alam, M., … Gani, A. (2014). Big Data: Survey, Technologies, Opportunities, and Challenges. *The Scientific World Journal*, *2014*, 18.

Kotu, V., & Deshpande, B. (2015). *Predictive analytics and data mining : concepts and practice with RapidMiner*. Waltham (MA) [etc.]: Elsevier/Morgan Kaufmann. Geraadpleegd van https://doc.lagout.org/Others/Data%20Mining/Predictive%20Analytics%20and%20Data%20Mining_%20Concepts%20and%20Practice%20with%20RapidMiner%20%5BKotu%20%26%20Deshpande%202014-12-03%5D.pdf

Lavrenko, V. (2015). Naïve Bayes Classifier - YouTube. Geraadpleegd 21 januari 2018, van https://www.youtube.com/playlist?list=PLBv09BD7ez_6CxkuiFTbL3jsn2Qd1IU7B

Naisbitt, J. (1982). *Megatrends : ten new directions transforming our lives*. New York: Warner Books.

Nonaka, I. (1995). *The knowledge-creating company : how Japanese companies create the dynamics of innovation*. New York: Oxford University Press.

Press, G. (2016). Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. Geraadpleegd 17 januari 2018, van https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

Reinsel, D., Gantz, J., & Rydning, J. (2017). *Data Age 2025: The Evolution of Data to Life-Critical - Don't Focus on Big Data; Focus on the Data That's Big* (White paper) (p. 25 p.). Framingham (MA): International Data Consortium (IDC). Geraadpleegd van https://www.seagate.com/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf

Sarkar, D. (2016). *Text Analytics with Python A Practical Real-World Approach to Gaining Actionable Insights from your Data*. Berkeley, CA: Apress.

Theorema van Bayes. (2018, januari 11). In *Wikipedia*. Geraadpleegd van https://nl.wikipedia.org/w/index.php?title=Theorema_van_Bayes&oldid=50755433

Vuurens, J. (2017). *[Beroepstaken, vaardigheden en competenties in Data Science]*. Gepresenteerd bij Opleidingsdag HBO-ICT, Haagse Hogeschool, Den Haag.

Weggeman, M. (2001). *Kennismanagement inrichting en besturing van kennisintensieve organisaties.* Schiedam: Scriptum.

World Wide Web. (2016, mei 26). In *Wikipedia, the free encyclopedia*. Geraadpleegd van https://en.wikipedia.org/w/index.php?title=World_Wide_Web&oldid=722203188