Onderzoek Exploreren van antwoorden

Djimaro Talahatu - 563631
Hogeschool van Arnhem en Nijmegen, faculteit ICA, HBO-ICT Software Development
Docentbegeleider: Dennis Breuker
Bedrijfsbegeleider: Alex Bijsterveld
Opdrachtgever: Eva de Schipper (Cito)
3 mei 2020, Apeldoorn
Versie 0.1

Inhoudsopgave

1 Definities

Woord	Definitie
Token	Een token is het kleinste element van een computer programma dat een betekenis heeft voor de compiler.
Corpora/corpus	In linguistics is een corpus een collectie van linguistische data (meestal zit het in een database). Deze data wordt gebruikt voor onderzoek en in lessen. Corpus wordt ook wel text corpus genoemd.
Vector	In de context van Natural Language Processing zijn vectoren een lijst van woorden die iets beschrijven.
Teksteigenschappen	Dit zijn bijvoorbeeld het aantal woorden van een tekst, het aantal unieke, type woorden of meest gebruikte woorden.

2 Inleiding

In dit hoofdstuk wordt er besproken wie de opdrachtgever is, waarom het onderzoek wordt uitgevoerd en worden de hoofd- en deelvragen geformuleerd.

2.1 Waarom dit onderzoek

Deze opdracht wordt namens Luminis uitgevoerd voor Cito. Bij dit project is Cito de opdrachtgever ook wel bekend van de Cito-toets. Cito laat zien wat leerlingen in hun mars hebben door het ontwikkelen van toetsen en examens en de resultaten vervolgens te meten met hulpmiddelen zoals meetinstrumenten. Om te laten zien wat leerlingen kunnen, moet de kwaliteit van toetsen en examens goed zijn. De kwaliteit van toetsen en examens onder ander wordt bepaald door onder ander de antwoorden die leerlingen geven op een onderdeel van een toets of examen.

Het doel van Cito is het ervoor zorgen dat iedere leerling gelijke kansen krijgt. Dit is, sinds de oprichting, de drijfveer voor medewerkers van Cito. De primaire focus van 'goed en eerlijk toetsen' is verbreed naar het objectief meten van mogelijkheden en talenten.

Toetsen en examens bestaan uit open en gesloten vragen. Bij een open vraag kan de leerling op zijn manier antwoord geven. Bij een gesloten vraag kan de leerling kiezen uit een set van antwoorden.

Er zijn meetinstrumenten die de kwaliteit van gesloten vragen kunnen meten. Bij een gesloten vraag is er een beperkte set van antwoorden beschikbaar. Deze set van antwoorden ondersteunen de interpretatie van de vraag. Hierdoor is het gemakkelijk om de antwoorden van leerlingen te analyseren met behulp van meetinstrumenten.

Het is moeilijker om de kwaliteit van open vragen te beoordelen. Dit geldt nog meer voor Nederlandse open vragen. Hiervoor zijn er nog minder hulpmiddelen beschikbaar.

De kwaliteit van Nederlandse open vragen worden op dit moment minder goed gewaarborgd, omdat het niet duidelijk is op welke eigenschappen er gelet moet worden. Voor dit project heeft Cito een open vraag(Zie Bijlage A) met daarbij antwoorden(Zie Bijlage B) gegeven. Deze open vraag gaat over het maken van een samenvatting in minder dan 150 woorden. Om te bepalen welke eigenschappen er nodig zijn om de kwaliteit van de open vraag te waarborgen is er onderzoek nodig.

Om de kwaliteit van de eerder genoemde vraag te waarborgen moet er een expert worden ingeschakeld, omdat het niet duidelijk was wat het begin punt is. Bij dit project is de focus gezet op het achterhalen van de kwaliteit van een open vraag aan de hand van de antwoorden. Omdat het ging over het maken van een samenvatting ging ik ervan uit dat er een expert nodig was op het gebied van schrijfvaardigheid. Om deze reden is de taalcoach van de HAN ingeschakeld. Uit deze interview bleek dat er geen theorie beschikbaar is om te bepalen waar op gelet moet worden. Voor het interview (Zie Bijlage C) Dit komt omdat een goede tekst kan bestaan uit veel woorden en weinig woorden met korte en lange zinnen. Er is dus geen theorie die aangeeft of een samenvatting goed is. Er werd wel geadviseerd om te kijken naar een aantal eigenschappen. Deze eigenschappen zijn te vinden in het interview.

De eigenschappen zijn bekend er kan dus worden gezocht naar kant en klare oplossingen die de eigenschappen van Nederlandse antwoorden inzichtelijk kunnen maken. Er zijn applicaties gevonden die tekst kunnen verwerken en eigenschappen inzichtelijk kunnen maken, maar niet voor Nederlandse teksten. Hierom is er gekeken naar wat deze applicaties met elkaar gemeen hebben. Wat deze applicaties met elkaar gemeen hebben is dat ze gebruik maken van Natural Language Processing. Nu is de vraag "Hoe kan Natural Language Processing de teksteigenschappen op het gebied van schrijfvaardigheid op basis van digitale antwoorden van studenten op een open vraag inzichtelijk maken?".

2.2 Hoofdvraag

Hoe kan Natural Language Processing de teksteigenschappen op het gebied van schrijfvaardigheid op basis van digitale antwoorden van studenten op een open vraag inzichtelijk maken?

2.3 Deelvragen

Wat zijn teksteigenschappen op het gebied van schrijfvaardigheid?

Wat is Natural Language Processing?Bij deze deelvraag horen de volgende subdeelvragen:

- Wat zijn de globale stappen voor het verwerken van teksten die geschreven zijn in natuurlijke taal in Natural Language Processing?
- Welke technieken horen er bij data preprocessing in Natural Language Processing?

Bij de deelvraag "Wat is data representation in Natural Language processing?"hoort de subdeelvraag "Welke language modellen horen er bij data representation?".

Waaruit bestaat Natural Language Processing?

Welke onderdelen van Natural Language Processing kunnen de teksteigenschappen op het gebied van schrijfvaardigheid op basis van digitale antwoorden van studenten op een open vraag inzichtelijk maken?

Nadat de bovenstaande vragen zijn beantwoord wordt er antwoord gegeven op de hoofdvraag.

3 Wat zijn teksteigenschappen op het gebied van schrijfvaardigheid?

In dit hoofdstuk worden de teksteigenschappen beschreven die te maken hebben met schrijfvaardigheid. Eerst worden de termen besproken die te maken hebben met het interview die in de inleiding is besproken. Daarna wordt het interview besproken.

3.1 Termen

Hier worden de termen besproken die te maken hebben met het interview.

3.1.1 Wat zijn woordsoorten?

Bij taalkundig ontleden wordt er van elk woord in de zin bepaald welke woordsoort het woord heeft. De woordsoort geeft aan tot welke categorie een bepaald woord hoort. Voorbeelden van woordsoorten zijn: werkwoorden, zelfstandige naamwoorden en bijvoeglijke naamwoorden. (Onzetaal, z.d.)

3.1.2 Wat is TF-IDF?

TF-IDFstaat voor Term Frequentie Inverted Document Frequentie. Term Frequentie geeft aan hoe vaak een woord voorkomt in een tekst. Inverted Document Frequentie geeft aan hoe vaak het woord voorkomt in andere documenten. TF-IDF geeft aan hoe vaak een woord van een tekst gebruikt wordt in andere documenten.

Hierbij zijn woorden die het minst vaak voorkomen in andere teksten het belangrijkst.

3.1.3 Wat is Lexical Density?

Zelfstandig naamwoord, werkwoord, bijvoeglijk naamwoord en bijwoorden zijn content woorden.

Het meten van structuur en complexiteit wordt met Lexical Density gedaan worden door content woorden te delen door het totaal aantal woorden. (Analyzemywriting, z.d.-a)

3.2 Interview

Hier wordt het interview besproken. Eerst wordt er besproken dat er geen theorie is over teksteigenschappen , die te maken hebben met schrijfvaardigheden. Daarna wordt er besproken waar er, volgens de taalcoach, op gelet moet worden bij schrijfvaardigheid.

3.2.1 Geen theorie

Tijdens het interview (Zie Bijlage A) met de taalcoach is gebleken dat er geen theorie is over de teksteigenschappen in het domein van schrijfvaardigheid.

De taalcoach geeft aan dat een tekst kan bestaan uit lange, korte zinnen, veel woorden, weinig woorden, unieke woorden en voegwoorden. Dit geldt voor slecht en goed geschreven teksten. Op

dit moment is er geen theorie is over de teksteigenschappen die aantonen dat een leerling schrijfvaardig is.

3.2.2 Waar moet er wel op gelet worden?

De taalcoach gaf aan dat het een goed begin is om naar de volgende teksteigenschappen te kijken:

- Toont het totaal aantal woorden dat er gebruikt zijn.
- Toont welke woorden het meest zijn gebruikt. (Zonder lidwoorden en eventueel andere veel gebruikte woorden)
- Toont per meest gebruikte woord hoe vaak het woord is gebruikt.
- Per woordsoort moet er getoond worden hoeveel er van elke soort gebruikt is.
- Toont welke volgens TF-IDF de meest belangrijke woorden zijn.
- Toont wat de Lexical Density is.

4 Wat is Natural Language Processing?

Hier wordt er besproken wat Natural Language Processing is. Eerst wordt er besproken wat Natural Language is en daarna wat Natural Language Processing is.

4.1 Wat is Natural Language?

Natuurlijke taal (Natural Language) verwijst naar de manier hoe mensen met elkaar communiceren; tekst en praten. Dagelijks zijn mensen bezig met teksten en praten. Het communiceren kan gebeuren door borden te lezen, e-mails, webpagina's, praten, gebarentaal en nog veel meer. Dit bevat allemaal belangrijke data. Hiervoor zijn er methodes om natuurlijke taal te begrijpen en erover te redeneren.(Brownlee, 2017)

4.2 Wat is Natural Language Processing?

Natural Language Processing is het manipuleren van natuurlijke taal door de computer. Het kan zo simple zijn als het tellen van woorden en zo complex als het analyseren van schrijfstijlen. Natural Language Processing is betrokken bij het begrijpen van menselijke statement(s), zodat er door de computer een nuttig antwoord kan worden gegeven.(Brownlee, 2017)

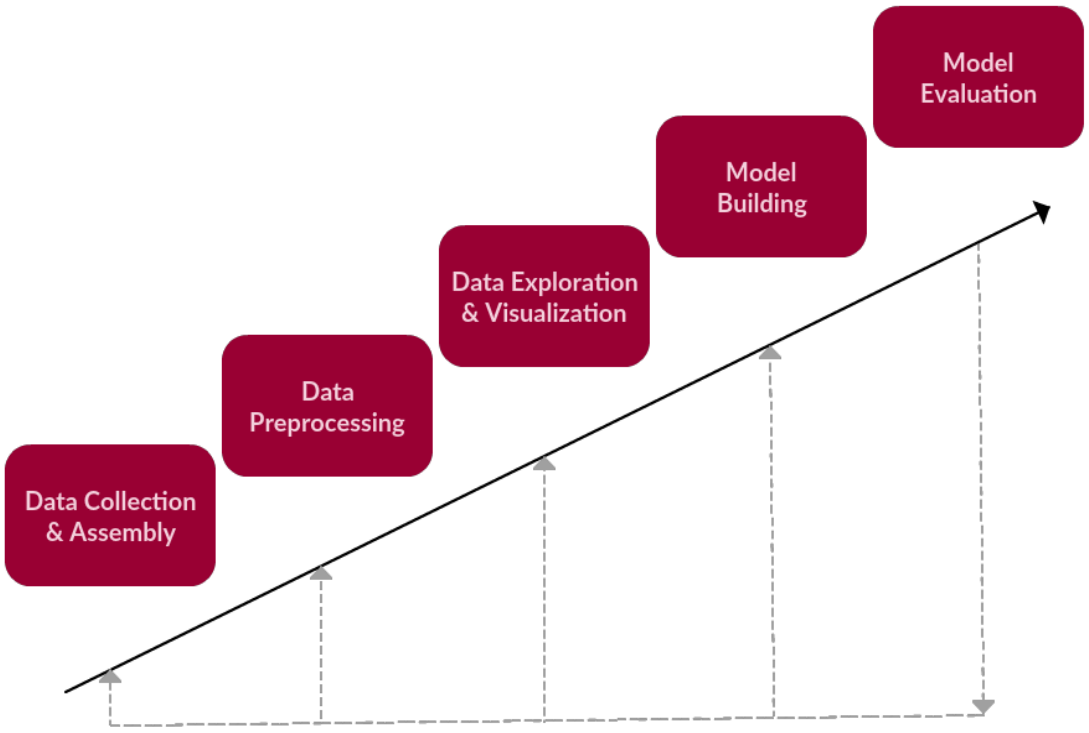
Omdat het gaat om digitale tekstuele antwoorden wordt er alleen gekeken naar Natural Language Processing onderdelen die aansluiten op digitale teksten.

5 Wat zijn de globale stappen voor het verwerken van teksten die geschreven zijn in natuurlijke taal in Natural Language Processing?

Hier wordt er besproken wat de globale stappen zijn van Natural Language Processing voor het verwerken van tekstuele data. Eerst wordt er een overzicht van de globale stappen getoond. Daarna wordt er per globale stap een korte uitleg gegeven.

5.1 Overview

Hieronder staat een overzicht van de globale stappen voor het uitvoeren van Natural Language taken.



Figuur 5.1: Natural Language globale stappen overview

De stappen van model building en evaluation worden hier uitgelegd. (Mayo, 2018a)

5.1.1 Data collection en Assembly

Data collection en Assembly gaat over het maken of verkrijgen van de corpus. De corpus wordt ook wel tekstuele data genoemd. De tekstuele data kan bijvoorbeeld e-mails, wikipedia of financiële

rapportages zijn.

5.1.2 Data preprocessing

Hier worden de voorbereidings taken op de corpus uitgevoerd. data preprocessing is het volgende onderdeel van het verwerken van tekstuele data. Het onderdeel gaat over het omzetten van tekst naar een andere vorm dat analyseerbaar en voorspelbaar is voor een taak. (Ganesan, 2019) Hierna kan de data gebruikt worden voor een Natural Language Processing taak.

5.1.3 Data Exploration en Visualization

Het doel van deze stap is inzicht krijgen in de data. Het doel wordt behaald door de data te visualiseren. Het visualiseren kan gedaan worden door het visualiseren van het aantal woorden, hoe de woorden verdeeld zijn en nog meer technieken.

5.1.4 Model building

Hier wordt er een model gebouwd. Bij deze stap worden er modellen getraind en getest.

5.1.5 Model Evaluation

Hier wordt er gecontroleerd of het model voldoet aan de eisen die de maker stelt. (Mayo, 2017)

6 Welke technieken horen er bij data preprocessing in Natural Language Processing?

Hier wordt er besproken welke technieken er bij de globale stap Data Preprocessing hoort. Eerst wordt Tokenization besproken, daarna Normalization, gevolgd door Noise Removal. Als laatst wordt er een overzicht getoond van Tokenization, Normalization en Noise Removal.

6.1 Tokenization en Segmentation

Tokenization is een stap waarbij lange stukken tekst wordt opgesplitst. Grote stukken tekst worden afgebroken in zinnen. Zinnen worden afgebroken in woorden. Bij Tokenization wordt ook tekst Segmentation genoemd. Segmentation verwijst naar het afbreken van een grote stuk tekst die groter zijn dan enkele woorden. Bijvoorbeeld zinnen.

6.2 Normalization

Normalization verwijst naar een series van gerelateerde taken die de data op dezelfde manier omzet. De tekst kan bijvoorbeeld de volgende behandelingen krijgen:

- De tekst wordt omgezet naar alleen hoofdletters of kleine letters.
- Weghalen van interpunctie.
- Nummers worden omgezet naar de woord variant.
- etc...

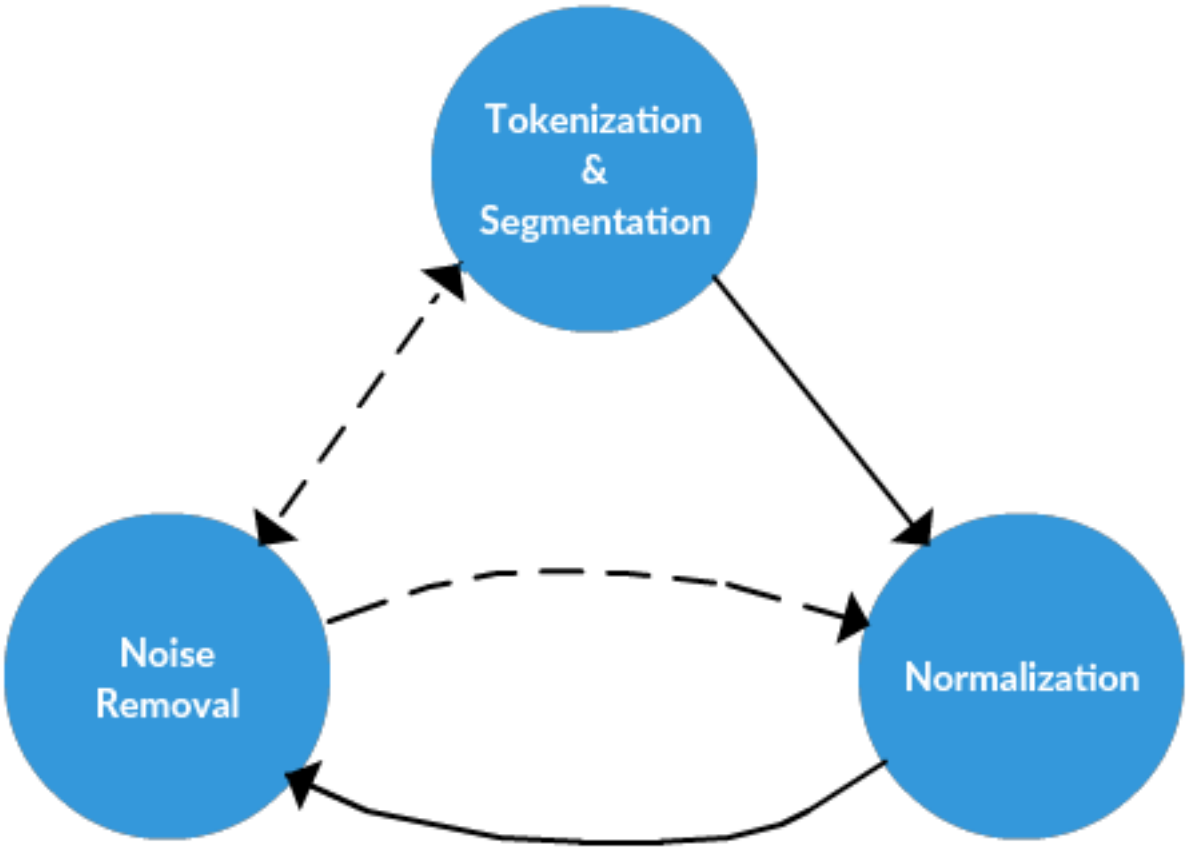
6.3 Noise Removal

Noise Removal is het weghalen van ongewenste inhoud. Wat en wanneer de ongewenste inhoud wordt weggehaald is afhankelijk van de Natural Language Processing taak.

Als de tekst onderdeel is van een html pagina dan wordt de html code vaak weggehaald. (Ganesan, 2019)

6.4 Overzicht technieken

Eerder is aangegeven dat de volgorde waarin de stappen worden uitgevoerd afhankelijk is van de taak. Dit betekent dus dat de stappen die horen bij data preprocessing niet lineair zijn. Hieronder staan de stappen gevisualiseerd hoe ze uitgevoerd kunnen worden.



Figuur 6.2: Text preprocessing framework

(Mayo, 2018b)

7 Wat is data representation in Natural Language Processing?

Hier wordt er verteld wat data representation is.

Data representation is een Natural Language Processing proces om tekst om te zetten naar getallen.

Language modellen zijn modellen die waarschijnlijkheid, frequentie of andere cijfers toewijzen aan woorden.

Deze modellen worden dan ook vaak gebruikt bij het omzetten van woorden naar getallen.

8 Welke language modellen horen er bij data representation in Natural Language Processing?

Hier wordt er verteld welke language modellen er horen bij data representation. Eerst wordt er een overzicht getoond van language modellen die gebruikt worden bij data representation. Daarna wordt er per techniek uitgelegd wat het is, waar het vaak voor gebruikt wordt en wordt er kort uitgelegd hoe het werkt.

8.1 Overview language modellen

Er zijn een aantal bestaande language modellen beschikbaar:

- Hot encoding model
- N-grams language model
- TF-IDF
- Bag of words
- Vector semantics
- Word2Vec
- GloVe

(Mayo, 2018b) De output van Hot encoding, N-grams language model en TF-IDF wordt vaak gebruikt bij lineaire machine learning algoritmes.

8.2 Hot encoding model

Hot encoding model is het representeren van woorden met een array van vectors.

Hierbij representeert elke vector een uniek woord. Om een uniek woord te representeren wordt er een component van de vector op 1 gezet. De rest van het component wordt op 0 gezet.

Omdat een vector veel componenten heeft kost dit veel geheugen. Als alternatief hierop kan er gebruik worden gemaakt van N-grams language models en TF-IDF.

8.3 N-grams language model

N-grams model wordt gebruikt voor het voorspellen van woorden.

Het voorspellen doet N-grams model door het maken van setjes van 1, 2, of meer woorden. Door deze woorden te vergelijken maakt het model een voorspelling.

8.4 TF-IDF

TF-IDF staat voor term frequentie inverted document frequency.

Term frequency wordt berekend op basis van hoevaak het woord in een document voorkomt.

Inverted document frequentie is de hoeveelheid hoe weinig een woord voorkomt in andere documenten. Woorden die minder vaak voorkomen krijgen een hoger gewicht. Deze woorden zijn dan ook belangrijker dan andere woorden.

8.5 Bag of words

Bag of words model wordt gebruikt voor het categoriseren, classificeren of bevestigen van tekst.

Bag of words is een ongeorganiseerde set van woorden. Hier wordt er bijgehouden hoe vaak elk woord voorkomt.

8.6 Vector semantics

Vector semanticswordt gebruikt voor het omgaan met woorden die hetzelfde betekenen.

De definitie van een woord wordt bepaald door te tellen welke andere woorden erbij horen. Het woord wordt gerepresenteerd door een vector, een lijst van nummers en coördinaten. Deze representatie wordt ook (word) embedding genoemd.

8.7 Word2Vec

Dense vectors zijn getrainde classifiers die kunnen voorspellen of een woord(en) in de buurt komt van een andere woord(en). Als de woorden (context woorden) in de buurt komen van het woord waar het om gaat (doelwoord) en in de juiste context worden gebruikt dan is dit positief overeenkomstig. Dit wordt ook wel dot product genoemd.

Bij Word2Vec wordt er gebruik gemaakt van twee modellen: CBOW model en Skip-gram model.

8.7.1 CBOW model

CBOW model staat voor continous bag of words model.

Met dit model wordt het doel woord bepaald aan de hand van context woorden.

8.7.2 Skip-gram model

Met dit model worden de context woorden voorspeld aan de hand van het doel woord.

8.8 Global Vectors (GloVe)

GloVe optimaliseert de word embedding direct zodat het dot product van twee woorden vectors gelijk zijn aan de log van hoe vaak deze twee woorden bij elkaar voorkomen. (Kana, 2019)

9 Waaruit bestaat Natural Language Processing?

Hier worden de subcategorieën van Natural Language Processing (NLP) uitgelegd. Eerst worden de grootste categorieën NLP en Natural Language Understanding uitgelegd. Daarna wordt er een overzicht getoond van NLP en NLU. Vervolgens worden de subcategorieën van NLP uitgelegd.

9.1 NLP

Het proces om ongestructureerde data om te zetten naar gestructureerde data.

9.2 NLU

De computer begrijpt wat een persoon zegt waardoor een mens een interactie in natuurlijke taal kan aangaan met een computer. (Expert System Team, 2019)

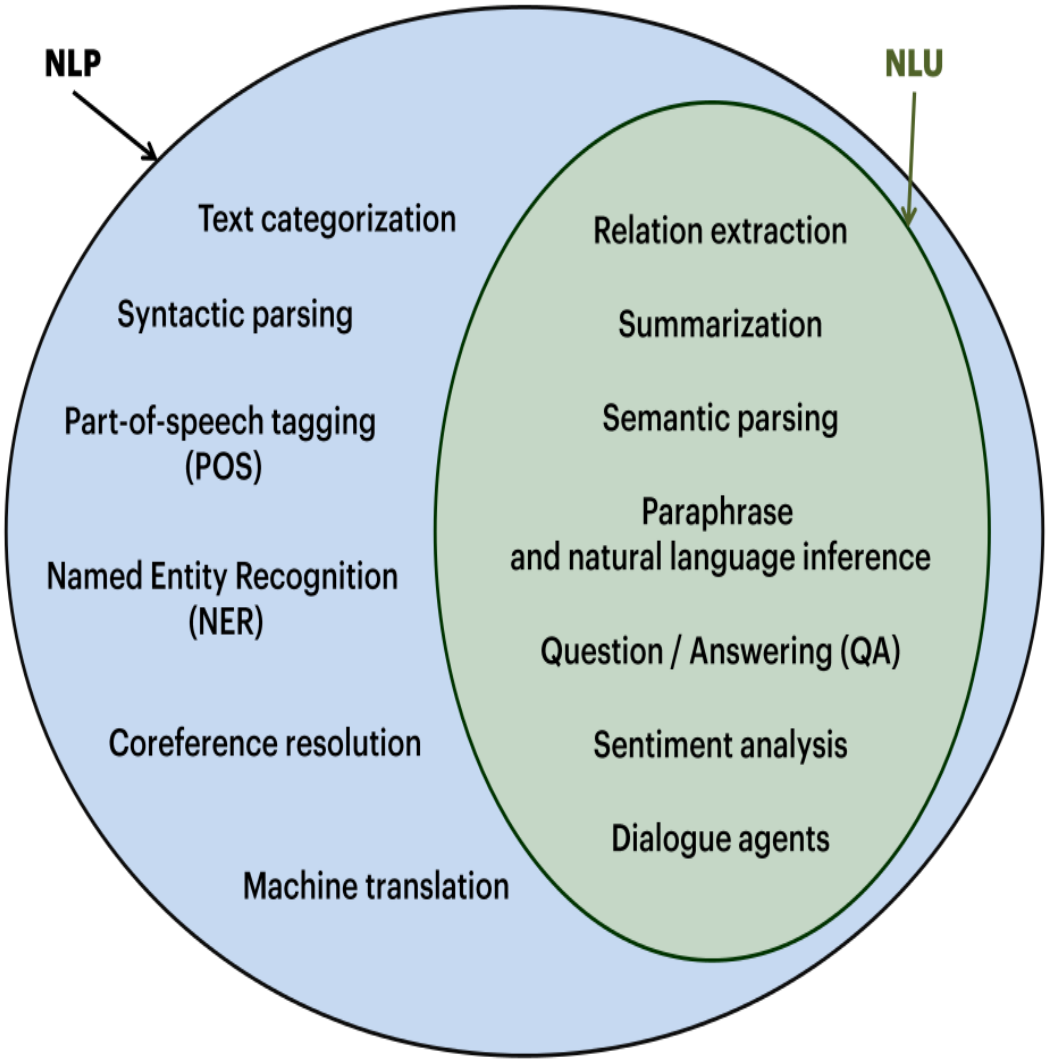
9.3 NLP of NLU

Bij deze opdracht gaat het over het omzetten van ongestructureerde data naar gestructureerde data zodat de onderzoekers inzicht krijgen. Bij deze opdracht is er geen interactie met de computer in natuurlijke taal.

NLP sluit goed aan op de opdracht en NLU niet. Hierom wordt Natural Language Understanding niet verder uitgelegd.

9.4 Overzicht subcategorieën NLP en NLU

Hieronder staat een overzicht over Natural Language Processing en de bijbehorende categorieën. Onder het overzicht staat er per subcategorie van NLP een uitleg.



Figuur 9.3: Natural Language Processing subset

(Olivier, z.d.-a)

9.4.1 Text categorization

Is het proces om tekst te categoriseren in georganiseerde groepen. (Monkeylearn, z.d.)

9.4.2 Syntactic parsing

Syntactic parsing is een techniek die de relatie tussen tokens toont. (Bengfort, z.d.-a)

9.4.3 Part of Speech tagging (POS)

Part of Speech tagging is het proces om zinnen om te zetten naar een overzicht. Elke woord heeft een bijbehorende tag. Bijvoorbeeld "huis"heeft als tag zelfstandignaamwoord. (Geeksforgeeks, z.d.)

9.4.4 Named Entity Recognition (NER)

Named Entity Recognition vind en classificeert atomische elementen in tekst. Het classificeren gebeurt op basis van voorgedefinieerde categorieën. Bijvoorbeeld namen van personen, organisaties en plaatsen. (Wordlift, z.d.)

9.4.5 Coreference resolution

Coreference resolution vind alle referencies die verwijzen naar dezelfde entiteit in een tekst. (Stanford, z.d.-a)

9.4.6 Machine translation

Vertaald een natuurlijke taal naar een andere natuurlijke taal. (Stanford, z.d.-b)

10 Welke onderdelen van Natural Language Processing kunnen de teksteigenschappen op het gebied van schrijfvaardigheid op basis van digitale antwoorden van studenten op een open vraag inzichtelijk maken?

Hier wordt er per teksteigenschap gekeken naar welke onderdeel van Natural Language processing de eigenschap inzichtelijk kan maken. Voor een overzicht van de teksteigenschappen (Zie teksteigenschappen).

Eerst wordt er besproken wat er geldt voor alle teksteigenschappen. Daarna wordt er per teksteigenschap besproken wat dit precies inhoud en welke onderdelen deze teksteigenschap inzichtelijk kan maken.

10.1 Voordat de teksteigenschappen worden besproken

Hier wordt er besproken wat er voor elke teksteigenschap geld.

Bij elke teksteigenschap wordt er gebruik gemaakt van Tokenization en Noise Removal.

Bij bijna elke teksteigenschap wordt er gebruik gemaakt van Normalization behalve bij de volgende teksteigenschap(pen):

- Toont het totaal aantal woorden dat er gebruikt zijn.

10.2 Welke onderdelen van Natural Language Processing sluiten aan op de teksteigenschappen?

Hier wordt er per teksteigenschap besproken wat de teksteigenschap inhoud en welke onderdelen van NLP hierop aansluiten.

10.2.1 Toont het totaal aantal woorden dat er gebruikt zijn.

10.2.1.1 Wat houd deze teksteigenschap in?

Bij deze teksteigenschap moet er gekeken worden naar de volgende punt(en):

- Dit onderdeel van NLP moet alle woorden kunnen van een tekst kunnen tellen.

10.2.1.2 Welke onderdelen van Natural Language Processing lossen dit probleem op?

Er is gezocht naar onderdelen die ongewenst inhoud kan weghalen. Bij Tokenization wordt interpunctie zoals ",“ ook als een token gezien. Noise Removal kan ongewenste inhoud zoals komma's weghalen. (Ganesan, 2019) Hierdoor is het mogelijk om alleen de woorden te tellen.

10.2.2 Toont welke woorden het meest zijn gebruikt. (Zonder lidwoorden en eventueel andere veel gebruikte woorden)

10.2.2.1 Wat houd deze teksteigenschap in?

Bij deze teksteigenschap moet er gekeken worden naar de volgende punten:

- Hoe vaak een woord per betekenis voorkomt in de tekst.
- Woorden zoals lidwoorden en eventueel andere veel gebruikte woorden moeten eruit gefilterd kunnen worden.

Stel de volgende tekst:"Ik ga naar huis en dan naar de dokter, omdat ik mij naar voel". In deze context is het woord "naar" drie keer gebruikt met twee betekenissen. Het gaat hier om hoe vaak het woord per betekenis is gebruikt. In dit geval is het woord "naar" één keer gebruikt met als betekenis onaangenaam". Ook is "naam"twee keer gebruikt met als betekenis in de richting van.

10.2.2.2 Welke onderdelen van Natural Language Processing lossen dit probleem op?

Er is gezocht naar oplossingen die de meeste woorden per betekenis kan tellen. Hieruit kwamen de volgende resultaten: GloVe, Word2Vec, Vector semantics of Bag of words.

10.2.3 Toont per meest gebruikte woord hoe vaak het woord is gebruikt

10.2.3.1 Wat houd deze teksteigenschap in?

Bij deze teksteigenschap moet er gekeken worden naar de volgende punten:

- Hoe vaak een woord per betekenis voorkomt in de tekst.
- Woorden zoals lidwoorden en eventueel andere veel gebruikte woorden moeten eruit gefilterd kunnen worden.

Stel de volgende tekst:"Ik ga naar huis en dan naar de dokter, omdat ik mij naar voel". In deze context is het woord "naar"drie keer gebruikt met twee betekenissen. Het gaat hier om hoevaak het woord per betekenis is gebruikt. In dit geval is het woord "naar"één keer gebruikt met als betekenis önaangenaam". Ook is "naam"twee keer gebruikt met als betekenis in de richting van.

10.2.3.2 Welke onderdelen van Natural Language Processing lossen dit probleem op?

Er is gezocht naar oplossingen die de meeste woorden per betekenis kan tellen. Hieruit kwamen de volgende resultaten: GloVe, Word2Vec en Vector semantics.

10.2.4 Per woordsoort moet er getoond worden hoeveel er van elke soort gebruikt is.

10.2.4.1 Wat houd deze teksteigenschap in?

Tellen welke Woordsoorten er zijn gebruikt.

10.2.4.2 Welke onderdelen van Natural Language Processing lossen dit probleem op?

Het bepalen van woordsoorten gebeurt op basis van de antwoorden van leerlingen. De tekst moet per woord een woordsoort krijgen. Hierom is er gekeken naar welke onderdelen van Natural Language Processing er aansluiten op het bepalen van woordsoorten.

Hierop sluit Named Entity Recognition en Part of Speech tagging op aan.

Named Entity Recognition classificeert niet elk woord van een zin. Bij Part of Speech tagging wordt elk woord van een zin getagd.

Het verschil tussen NER en POS is het taggen en classificeren. Een eigenschap van taggen is dat één woord meerdere tags kan hebben. Bij classificeren kan één woord maximaal bij één categorie horen.

Stel de volgende zin: "Can you google it?"In dit geval is "googleëen werkwoord.

Stel de volgende zin: "Can you search it on google?"In dit geval is het een zelfstandig naamwoord.

Bij NER is dit niet mogelijk. Bij NER kan "google" alleen een werkwoord zijn of alleen een zelfstandig naamwoord. NER sluit niet aan op de eerder genoemde woordsoort teksteigenschap "Hoe vaak is elke woordsoort gezegd". Hierdoor is er voor POS gekozen.
(Malik, 2019)

10.2.5 Toont welke volgens TF-IDF de meest belangrijke woorden zijn.

10.2.5.1 Wat houd deze teksteigenschap in?
Het tellen hoe vaak een woord met dezelfde betekenis voorkomt in andere documenten.

10.2.5.2 Welke onderdelen van Natural Language Processing lossen dit probleem op?

Er is gezocht naar onderdelen van Natural Language Processing die woorden kunnen tellen en kunnen tellen hoevaak deze woorden in andere documenten voorkomen. TF-IDF sluit hierop aan.

10.2.6 Toont aan wat de Lexical Density is.

10.2.6.1 Wat houd deze teksteigenschap in?
Lexical Density is het aantal content words gedeeld door het totaal aantal woorden.

Content woorden zijn zelfstandige naamwoorden, werkwoorden, bijvoeglijke naamwoorden en bijwoorden. Dit valt onder de categorie woordsoorten.

10.2.6.2 Welke onderdelen van Natural Language Processing lossen dit probleem op?

Er is gezocht naar onderdelen van Natural Language Processing die de content woorden kunnen

tellen en het totaal aantal woorden kunnen tellen.

Er is gezocht naar oplossingen die content woorden kunnen tellen. Hierom is er gekeken naar welke onderdelen van Natural Language Processing aansluiten op het bepalen van woordsoorten. Hierop sluit Named Entity Recognition en Part of Speech tagging op aan.

Named Entity Recognition classificeert niet elk woord van een zin. Bij Part of Speech tagging wordt elk woord van een zin getagd.

Het verschil tussen NER en POS is het taggen en classificeren. Een eigenschap van taggen is dat één woord meerdere tags kan hebben. Bij classificeren kan één woord maximaal bij één categorie horen.

Stel de volgende zin: "Can you google it?"In dit geval is "googleëen werkwoord. Stel de volgende zin: "Can you search it on google?"In dit geval is het een zelfstandig naamwoord.

Bij NER is dit niet mogelijk. Bij NER kan "googleälleen een werkwoord zijn of alleen een zelfstandig naamwoord. NER sluit niet aan op de eerder genoemde woordsoort teksteigenschap "Hoevaak is elke woordsoort gezegd". Hierom is er voor POS gekozen. (Malik, 2019)

Voor het tellen van woorden is er gezocht naar onderdelen die ongewenste inhoud kan weghalen. Bij Tokenization wordt interpunctie zoals ",ook als een token gezien. Noise Removal kan ongewenste inhoud zoals komma's weghalen. (Ganesan, 2019) Hierdoor is het mogelijk om alleen de woorden te tellen.

11 Conclusie

In dit onderzoek is er gezocht naar een antwoord op de vraag: 'Hoe kan Natural Language Processing de teksteigenschappen op het gebied van schrijfvaardigheid op basis van digitale antwoorden van studenten op een open vraag inzichtelijk maken?'.

Eerst is er een interview gehouden met de opdrachtgever en een taalcoach om de teksteigenschappen te achterhalen. Voor het analyseren van teksten op het gebied van schrijfvaardigheid is het belangrijk om te kijken naar verschillende teksteigenschappen:

- Toont het totaal aantal woorden dat er gebruikt zijn.
- Toont welke woorden het meest zijn gebruikt. (Zonder lidwoorden en eventueel andere veel gebruikte woorden)
- Toont per meest gebruikte woord hoe vaak het woord is gebruikt
- Per woordsoort moet er getoond worden hoeveel er van elke soort gebruikt is.
- Toont welke volgens TF-IDF de meest belangrijke woorden zijn.
- Toont wat de Lexical Density is.

Daarna is er gezocht naar onderdelen van Natural Language Processing die de teksteigenschappen inzichtelijk kunnen maken.

Bij elke teksteigenschap wordt er gebruik gemaakt van Tokenization en Noise Removal.

Bij bijna elke teksteigenschap wordt er gebruik gemaakt van Normalization behalve bij "Toont het totaal aantal woorden dat er gebruikt zijn."

Bij de teksteigenschap "Toont het totaal aantal woorden dat er gebruikt zijn.ïs er gezocht naar onderdelen van Natural Language Processing die de meeste woorden per betekenis kan tellen. Deze teksteigenschap kan inzichtelijk worden gemaakt door gebruik te maken van Tokenization en Noise Removal.

Bij de teksteigenschap "Toont welke woorden het meest zijn gebruikt. (Zonder lidwoorden en eventueel andere veel gebruikte woorden)ïs er gezocht naar onderdelen van Natural Language Processing die de meeste woorden per betekenis kan tellen. Deze teksteigenschap kan inzichtelijk worden gemaakt door gebruik te maken van de volgende resultaten: GloVe, Word2Vec en Vector semantics.

Bij de teksteigenschap "Toont per meest gebruikte woord hoe vaak het woord is gebruiktïs er gezocht naar onderdelen van Natural Language Processing die de meeste woorden per betekenis kan tellen. Deze teksteigenschap kan inzichtelijk worden gemaakt door gebruik te maken van de volgende resultaten: GloVe, Word2Vec en Vector semantics.

Bij de teksteigenschap "Per woordsoort moet er getoond worden hoeveel er van elke soort gebruikt is.ïs er gezocht naar onderdelen van Natural Language Processing die woordsoorten kunnen

toewijzen aan woorden. Deze teksteigenschap kan inzichtelijk worden gemaakt door gebruik te maken van de volgende resultaat:

Bij de teksteigenschap "Toont welke volgens TF-IDF de meest belangrijke woorden zijn.ïs er gezocht naar onderdelen van Natural Language Processing die TF-IDF woorden kan tonen.Deze teksteigenschap kan inzichtelijk worden gemaakt door gebruik te maken van de volgende resultaat: TF-IDF

Toont wat de Lexical Density is. Deze teksteigenschap kan inzichtelijk worden gemaakt door gebruik te maken van Tokenization, Noise Removal en Part of Speech tagging.

12 Literatuurlijst

Aimultiple. (2020, 19 januari). 100+ AI Use Cases / Applications in 2020. Geraadpleegd op 27 januari 2020, van https://blog.aimultiple.com/ai-use-cases/

Analyzemywriting. (z.d.). Lexical Density. Geraadpleegd op 16 februari 2020, van http://www.analyzemywriting.com/lexical_density.html

Bengfort, B. (z.d.). Syntax Parsing with CoreNLP and NLTK. Geraadpleegd op 16 februari 2020, van https://www.districtdatalabs.com/syntax-parsing-with-corenlp-and-nltk

Brownlee, J. (2017, 22 september). What is Natural Language Processing? Geraadpleegd op 20 januari 2020, van https://machinelearningmastery.com/natural-language-processing/

Builtin. (z.d.). What is Artificial Intelligence?. Geraadpleegd op 27 januari 2020, van https://builtin.com/artificial-intelligence

Cito. (z.d.). Meerjarenbeleidsplan. Geraadpleegd op 16 januari 2020, van https://www.cito.nl/kennis-en-innovatie/cito-kennisorganisatie/cito-meerjarenbeleidsplan

Dar, P. (2019, 18 maart). 8 Excellent Pretrained Models to get you Started with Natural Language Processing (NLP). Geraadpleegd op 22 januari 2020, van https://www.analyticsvidhya.com/blog/2019/03/pretrained-models-get-started-nlp/

Expert System Team. (2019, 22 januari). Natural Language Understanding: What is it and how is it different from NLP. Geraadpleegd op 18 februari 2020, van https://expertsystem.com/natural-language-understanding-different-nlp/

Ganesan, K. (2019, 10 april). All you need to know about text preprocessing for NLP and Machine Learning. Geraadpleegd op 16 februari 2020, van https://www.kdnuggets.com/2019/04/text-preprocessing-nlp-machine-learning.html

Geeksforgeeks. (z.d.). NLP | Part of Speech – Default Tagging. Geraadpleegd op 16 februari 2020, van https://www.geeksforgeeks.org/nlp-part-of-speech-default-tagging/

Javatpoint. (z.d.). Subsets of Artificial Intelligence. Geraadpleegd op 27 januari 2020, van

https://www.javatpoint.com/subsets-of-ai

Kana, M. (2019, 15 juni). Subsets of Artificial Intelligence. Geraadpleegd op 18 februari 2020, van https://towardsdatascience.com/representing-text-in-natural-language-processing-1eead30e57d8

Malik, U. (2019, 27 maart). Python for NLP: Parts of Speech Tagging and Named Entity Recognition. Geraadpleegd op 31 januari 2020, van https://stackabuse.com/python-for-nlp-parts-of-speech-tagging-and-named-entity-recognition/

Mayo, M. (2018a, 20 juni). Natural Language Processing Nuggets: Getting Started with NLP. Geraadpleegd op 16 februari 2020, van https://www.kdnuggets.com/2018/06/getting-started-natural-language-processing.html

Mayo, M. (2018b, 7 november). Data Representation for Natural Language Processing Tasks. Geraadpleegd op 16 februari 2020, van https://www.kdnuggets.com/2018/11/data-representation-natural-language-processing.html

Mayo, M. (2017, 29 november). Data Representation for Natural Language Processing Tasks. Geraadpleegd op 16 februari 2020, van https://www.kdnuggets.com/2018/11/data-representation-natural-language-processing.html

Monkeylearn. (z.d.). What is Text Classification. Geraadpleegd op 16 februari 2020, van https://monkeylearn.com/what-is-text-classification/

Olivier, P. (z.d.). Introduction to NLP (Part I). Geraadpleegd op 31 januari 2020, van https://www.ekino.com/articles/introduction-to-nlp-part-i

Onzetaal. (z.d.). Woordsoorten (taalkundig ontleden). Geraadpleegd op 31 januari 2020, van https://onzetaal.nl/taaladvies/woordsoorten-taalkundig-ontleden/

Stanford. (z.d.-a). Coreference Resolution. Geraadpleegd op 14 februari 2020, van https://nlp.stanford.edu/projects/coref.shtml

Stanford. (z.d.-b). Machine Translation. Geraadpleegd op 14 februari 2020, van https://nlp.stanford.edu/projects/mt.shtml

Wordlift. (z.d.). Named-entity recognition. Geraadpleegd op 14 februari 2020, van https://wordlift.io/blog/en/entity/named-entity-recognition/

13 Bijlagen

Bijlage A : Schrijftaak
Bijlage B : AntwoordenLeerlingen
Bijlage C : Interview