Troonredes door de jaren heen

N.T de Visscher
10667474

Bachelor thesis
Credits: 12 EC

Bachelor Opleiding Informatiekunde

University of Amsterdam
Faculty of Science
Science Park 904
1098 XH Amsterdam

Supervisor
Dr. M. J. Marx

ILPS, IvI
Faculty of Science
University of Amsterdam
Science Park 904
1098 XH Amsterdam

2016-06-26

Contents

Abstract

1 Introduction

Dit onderzoek is een replicatie-onderzoek aan de hand van een eerder uitgevoerd onderzoek dat als paper is gepubliceerd. Het paper in kwestie beschrijft een onderzoek naar de "State of the Union" speeches uit Amerika van Rule et al. [2015]. In dat onderzoek werd gekeken naar wat er veranderde binnen de speeches sinds ze voor het eerst werden gehouden en hoe de speeches er hedendaags uitzien. Er werd gekeken naar de verschuiving in taalgebruik en inhoud in de speeches. Dit werd gedaan om een beeld te krijgen van wat de belangrijke onderwerpen in een speech waren en welke thema's behandeld werden over de jaren. Hieruit werd gekeken naar het moment dat de verschuiving naar het moderne Amerikaanse politieke bewustzijn plaatsvond. In het onderzoek werd gekeken naar de speeches van 1790 t/m 2014 welke jaarlijks uitgesproken worden door de president.

De State of the Union is niet de enige speech die al lange tijd jaarlijks wordt gehouden. Zo ook de Nederlandse troonredes. Deze troonredes worden sinds 1814 jaarlijks door de koning(in) voorgedragen. De troonredes kunnen gezien worden als een Nederlandse versie van de State of the Union. De troonredes bevatten echter vaak reflecties op het afgelopen jaar in plaats van een samenvatting over de staat van het land. De staat van het land wordt ook besproken maar dit is meer gericht op een kijk naar de positie van Nederland in de wereld. Dit is echter niet altijd het geval geweest. Zo was de inhoud vroeger anders dan nu, net als het taalgebruik.

Tijdens het onderzoek zal er gekeken worden naar de veranderingen in de troonredes over de jaren. Het gaat hier vooral om de onderwerpen en thema's die behandeld worden. Er zal allereerst niet specifiek worden gekeken naar de verandering in taalgebruik en grammatica, maar dit kan wel gebruikt worden om een beeld te vormen over de veranderingen van de troonredes.

Om het onderzoek een duidelijke richting te geven zal gepoogd worden de volgende onderzoeksvraag te beantwoorden:

- Hoe kan "word co-occurence" als tekstanalyse techniek worden gebruikt om de verschuiving in onderwerpen/thema's over 200 jaar troonredes weer te geven?

We beantwoorden deze vraag met behulp van de volgende deelvragen:

1. Zijn de methodes gebruikt in Rule et al. [2015] toepasbaar op het Nederlandse troonrede corpus?

2. Zijn de resultaten representatief voor de werkelijkheid?

2 Related Work

Deze sectie bestaat uit een aantal "blokken", waarin je per blok de relevante literatuur beschrijft. Neem alleen literatuur op die van belang is voor jouw onderzoeksvraag en deelvragen.Typisch heb je 1 blok voor je hoofdvraag en per deelvraag **RQi** een blok.

2.1 Het gebruik van co-occurrences in tekst analyse

2.1.1 co-occurrence extractie

Men kan op verschillende manieren co-occurrences uit teksten halen. Om deze manieren te verduidelijken gebruiken we de volgende voorbeeldzin: "Het is erg heet". Vanuit deze zin kunnen op de volgende manieren co-occurences worden gehaald. Door enkel co-occurrences te gebruiken van woorden die exact naast elkaar in een zin staan, "Het,is" & "is,erg" & "erg,heet". Ook kan het door woorden in een zin volgens de zinsvolgorde met elkaar te koppelen zelfs als er andere woorden tussen, hierdoor zouden "Het,heet" & "Het,erg" & "is,heet" ook co-occurrences zijn. Of het kan door alle mogelijke combinaties van woorden in een zin te vormen, dit gebeurt veelal op alfabetische woordvolgorde. Volgens de laatste manier wordt geen rekening meer gehouden met de volgorde van de woorden. De manier die het beste is om de co-occurrences uit de tekst te halen is afhankelijk van wat men uiteindelijk wil kunnen zeggen met de co-occurrences. [Shimohata et al., 1997]

2.2 Induceren van het onderwerp van een tekst

2.2.1 Trefwoord extractie

Er bestaan verschillende manieren om trefwoorden uit teksten te halen. De meeste eenvoudige manier is door simpelweg te kijken naar de waarschijnlijkheid dat een woord voorkomt. Dit wordt gedaan door te kijken naar de frequentie dat een woord voorkomt in een tekst en dit te delen door het totaal aantal woorden in de tekst. Verder kan men kijken naar een geheel corpus, de meest bekende methode hiervoor is TF-IDF (term frequency–inverse document frequency)[Ramos, 2003]. Hierbij wordt gekeken naar het voorkomen van woorden in een tekst tegenover het voorkomen in het gehele corpus. De TF-IDF score van een woord is hoog voor een specifieke tekst uit een corpus als deze vaak

voorkomt in die tekst, maar verder weinig in de rest van het corpus. Door de frequentie van het voorkomen van een woord in een tekst te compenseren met het voorkomen van het woord in de gehele tekst wordt rekening gehouden met woorden die in het algemeen veel worden gebruikt, zoals stopwoorden [Aggarwal and Zhai, 2012]. Daarnaast zijn er methodes gericht op individuele teksten zoals in Matsuo and Ishizuka [2004]. Hier wordt een algoritme gebruikt om trefwoorden uit een enkele tekst te halen zonder gebruik te maken van een corpus.

2.2.2 Invloed van context

Om een onderwerp of meerdere onderwerpen toe te kunnen wijzen aan een tekst moet men rekening houden met meerdere aspecten. De meest belangrijke hiervan is de context van de tekst. De context bepaald namelijk op welke manier woorden geïnterpreteerd worden en wat ze betekenen voor de tekst. Zo zijn er woorden die enkel voor specifieke domeinen betekenis hebben. Er zijn al meerdere methodes ontwikkeld om rekening te houden met de context van een tekst, maar de meeste hiervan hebben hiervoor een corpus nodig om het context netwerk te kunnen bouwen. Om van een individuele tekst de context te kunnen bepalen kan zeer lastig zijn. Er kan rekening gehouden worden met de context van een corpus door gebruik te maken van woordenboeken voor het specifieke domein dat het corpus bestrijkt. Hiervoor moet men echter wel zelf bepalen wat het domein van het corpus is. [Aggarwal and Zhai, 2012]

3 Methodology

3.1 Description of the data

De dataset die gebruikt gaat worden om deze vragen te beantwoorden is een verzameling van troonredes sinds 1814. De gehele dataset bestaat uit 165 verschillende troonredes welke in totaal uit 219236 woorden bestaan, deze set zullen we vanaf nu het corpus noemen. Deze troonredes zijn terug te vinden op www.troonredes.nl [Herko Coomans, 2015] . Om een duidelijker beeld te geven van de troonredes een korte uitleg van wat de troonredes nu precies zijn en waar ze vandaan komen.

De troonrede wordt jaarlijks door de koning(in) uitgesproken op Prinsjesdag, de 3de dinsdag van september. voor 1904 werden de troonredes uitgesproken in de vergaderzaal van de 2de kamer. Sinds 1904 worden de troonredes voor de ridderzaal op het binnenhof in Den Haag uitgesproken. De troonrede wordt live uitgezonden op televisie en is na te lezen op de website van de Nederlandse overheid. De eerste troonrede werd in 1814 als een algehele toespraak voor de Staten-generaal gehouden. De troonredes worden vooral gebruikt om wets- en beleidsveranderingen door te geven, als beschouwing op het afgelopen jaar en de staat van het land. In recentere jaren heeft deze beschouwing zich ook uitgebreid naar gebeurtenissen door de hele wereld die invloed uitoefenen op de Nederlandse staat. Ook wordt sinds 1918 het regeerprogramma voor het komende jaar in de troonredes behandelt. Sinds 1848 worden de troonredes geschreven door ministers vanwege een grondwetsherziening waardoor de ministers verantwoordelijk werden voor al het doen en laten van de koning(in). Hiermee werd het kabinet verantwoordelijk voor de uitspraken die gedaan worden in de troonredes. [Rijksoverheid, 2016] Dit zorgt ervoor dat de troonredes een beeld geven van wat de Nederlandse regering op dat moment belangrijk vindt.

3.1.1 Missing data

Er zijn verscheidene jaren waarvan er geen data beschikbaar is, dit doordat er niet elk jaar een troonrede gegeven is. Dit heeft verschillende redenen, zoals oorlogen, angst voor rellen, onvrede over het kabinet of gezondheidsredenen omtrent de koning(in). Dit zorgt ervoor dat niet elk jaar sinds 1814 wordt gerepresenteerd door een troonrede. Omdat de inhoud van de troonredes als verzamelde dataset wordt gebruikt heeft dit minimaal invloed op de uiteindelijke uitkomsten van

het onderzoek dat er jaren missen. Door de missende jaren is het weergeven van de verschuiving in onderwerpen mogelijk niet volledig representatief, omdat er geen uitspraken gedaan kunnen worden over de missende jaren. In tabel 1 een overzicht van de jaren waar geen troonredes van zijn en de reden dat er in die jaren geen troonrede is gegeven:

Jaren	reden
1888-1890	Verhindering wegens ziekte van de koning en interne beleidsproblemen
1905-1924	Eerste wereldoorlog en politieke instabiliteit
1940-1947	Tweede wereldoorlog en politieke instabiliteit

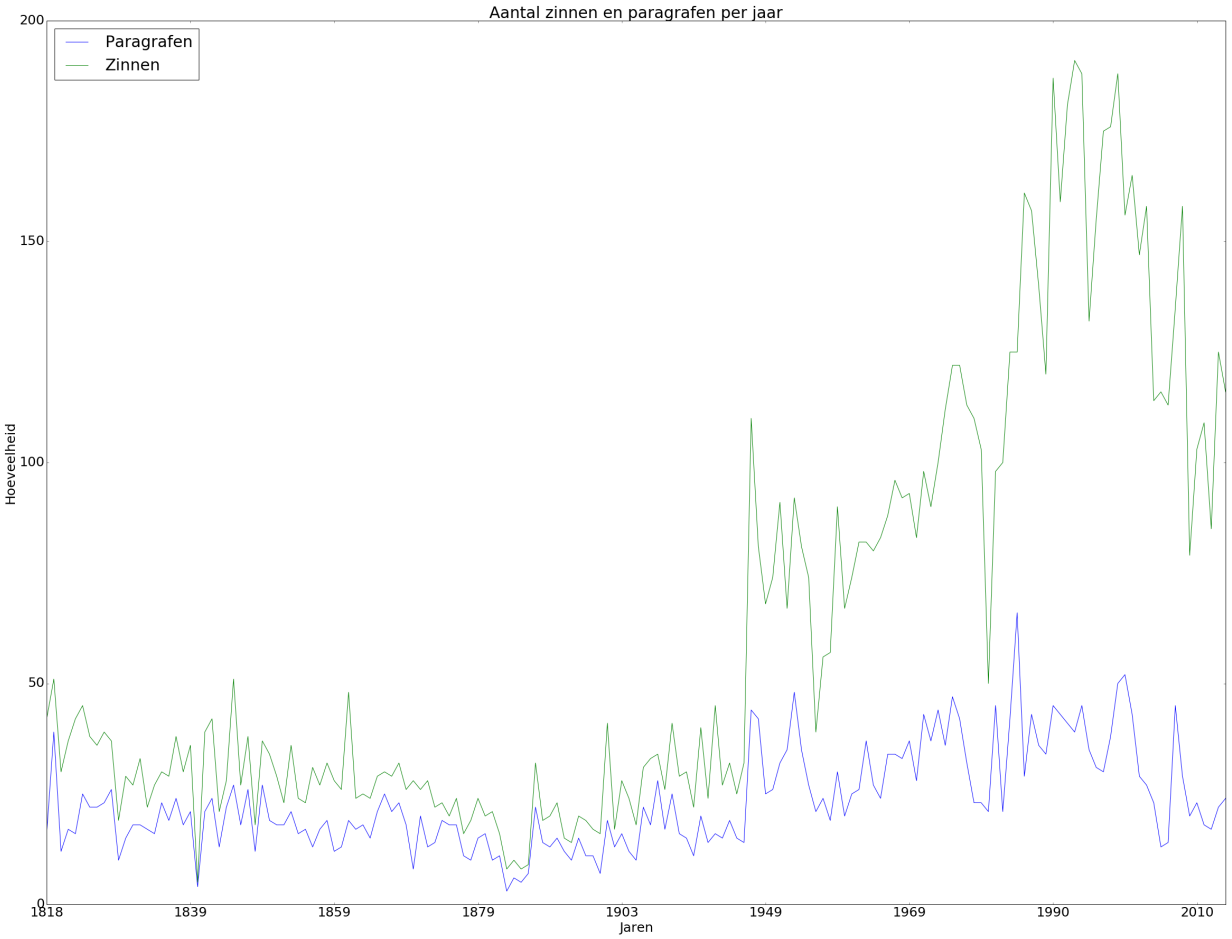
tabel 1: Missende troonredes

3.1.2 Quality of the data

Hier moet nog een deel komen over het feit of het ingescand is of ingetypt.

3.2 Wat plotjes en tabelletjes

De grafiek in figuur 1 geeft een beeld van de grootte van de troonredes over de jaren aan de hand van het aantal zinnen en paragrafen per jaar.



figuur 1: Weergave van het aantal zinnen en paragrafen per jaar.

3.3 Methods

Om uit het corpus te kunnen achterhalen of er overkoepelende onderwerpen zijn en wat deze mogelijk zijn wordt er gebruik gemaakt van tekstanalyse technieken. Specifiek wordt er gebruik gemaakt van POS-methodes en co-occurence aanpakken. Callon et al. [1991]

Allereerst wordt er met behulp van POS-methodes(Pattern of Speech) gekeken naar het volledige corpus. Deze methodes kijken naar het corpus als geheel en proberen hier patronen uit te halen. Er is specifiek gebruik gemaakt van "Pattern", een Python module, die getraind is voor Nederlandse tekst. Deze module is getraind op verschillende soorten Nederlandse teksten en kan woorden uit de tekst onderverdelen in verschillende categorieën, zoals lidwoorden en zelfstandige naamwoorden. Met behulp van deze module is het corpus gefilterd zodat enkel de zelfstandige naamwoorden zijn overgebleven. De reden dat we enkel zelfstandige naamwoorden gebruiken is het feit dat deze representatief zijn voor de tekst, omdat vanuit de zelfstandige naamwoorden afgeleid kan worden wat het doel en inhoud van de tekst is.

Hierna wordt de data van de troonredes gelemmatiseerd. Hierdoor worden de woorden herleid tot hun lemma(stam), waardoor ze kunnen worden geanalyseerd als één term. Enkele voorbeelden hiervan: "steden" wordt herleid tot "stad" en "mensen" tot "mens". Doordat alle woorden hierdoor tot de stam zijn herleidt is er een duidelijk beeld van alle unieke woorden die gebruikt worden. Hierdoor zal de relatie tussen woorden sneller duidelijk worden, omdat er nu geen onderscheid wordt gemaakt tussen de verschillende vervoegingen van unieke woorden. Met behulp van deze gelemmatiseerde termen wordt een co-occurence matrix opgesteld. Deze matrix wordt aangemaakt door bij te houden hoe vaak een combinatie van 2 termen samen in een paragraaf voorkomen. De matrix geeft uiteindelijk voor alle combinaties van 2 termen de frequentie dat ze samen in een paragraaf voorkomen weer.

Voor alle combinaties uit deze matrix wordt een nabijheidsscore berekend via de volgende definitie:

$$S(W_1, W_2) \equiv^{\text{def}} \frac{\sum_{c \epsilon W \setminus W_1, W_2\}, PMI(W_1,c)>0} min(PMI(W_1,c), PMI(W_2,c))}{\sum_{c \epsilon W \{W_1, W_2\}, PMI(W_1,c)>0} PMI(W_1,c)}$$

Hierbij is het doel om uit te vinden welke termen relevant zijn en hoe deze zich over tijd met andere termen associëren. De score geeft het verwantschap tussen twee termen "W1" & "W2" aan. Dit verwantschap

wordt bepaald aan de hand van PMI(Pointwise Mutual Information) tussen twee termen [Bouma, 2009]. Hierbij geeft de PMI tussen twee termen "X" & "Y" aan hoeveel de termen ons over elkaar kunnen vertellen. Hierbij gaat het om het verschil in de kans dat term "X" of "Y" individueel voorkomt in een paragraaf in de tekst en de kans dat termen "X" & "Y" samen voorkomen in een paragraaf in de tekst. Wiskundig gezien wordt PMI als volgt berekend:

$$PMI(X,Y) \equiv^{\log \frac{p(X,Y)}{p(X)}}$$

Hierbij is P(X,Y) de kans dat termen "X" & "Y" samen in een paragraaf voorkomen en P(X) de kans dat term "X" in een paragraaf voorkomt. Dit wordt berekend door het aantal keer dat de term voorkomt in de tekst te delen door het aantal paragrafen. De PMI is een score die positief of negatief kan zijn. Goede collocatie paren van termen hebben een hoge PMI score omdat ze net iets minder samen voorkomen dan dat ze individueel voorkomen. Een score van 0 betekent dat de termen onafhankelijk van elkaar zijn en er geen uitspraak gedaan kan worden over het voorkomen van term "X" als "Y" voorkomt en vice versa.

Deze vergelijking voor de nabijheidsscore maakt gebruikt van de termen uit het gehele corpus om context te bepalen. Hiervoor wordt voor elke combinatie termen (W1,W2) gekeken naar de woorden waarmee ze samen in een paragraaf voorkomen. De som van PMI(W,C) voor alle termen (C) uit het corpus waarvoor geldt dat PMI(W,C) >0 wordt voor beide termen berekend. Dit betekent dus dat er enkel wordt gekeken naar combinaties van termen die onafhankelijk van elkaar zijn, met een score van 0, of positief. Als hieruit volgt dat PMI(W1,C) gelijk is aan PMI(W2,C) kan men stellen dat als W1 in een paragraaf voorkomt W2 ook voorkomt en vice versa. Een nabijheidsscore van 0 geeft aan dat de woorden nooit samen in een paragraaf voorkomen. Aan de hand van deze score wordt een gewogen semantisch netwerk gevormd met de termen als nodes en de gewichten op de edges. Om de termen binnen dit netwerk te kunnen analyseren wordt gebruik gemaakt van een community detectie algoritme. Het specifieke algoritme dat gebruikt wordt is dat van Blondel et al. [2008]. Het doel van dit algoritme is om vanuit het gewogen netwerk clusters te vormen van samenhangende subsets van termen. Om ervoor te zorgen dat enkel de meest relevante termen worden meegegeven aan het community detectie algoritme is een drempelwaarde bepaald. Deze drempelwaarde geeft aan vanaf welke waarde voor de nabijheidsscore de co-occurence combinaties door

het algoritme verwerkt worden. De drempelwaarde ligt tussen de 0 en 1 en wordt bepaald door vanaf 0 de score op te hogen. Dit is een handmatig proces waarbij de drempelwaarde is bereikt als er binnen het netwerk een component van meer dan twee nodes verdwijnt. Hiermee worden de zwak verbonden nodes uit het netwerk verwijderd. De drempelwaarde die gevonden is voor ons netwerk is =0.65. Het resulterende netwerk is door het community detectie algoritme verwerkt om de verschillende subsets te kunnen identificeren. Vanuit deze subsets zijn onderwerpen bepaald aan de hand van de termen die er onder vallen. Afhankelijk van de termen binnen deze clusters kunnen deze clusters gezien worden als indicatief voor onderwerpen/thema's binnen het corpus. Dit induceren van onderwerpen moet een overzicht geven van de verschillende termen die tot een onderwerp behoren en hoe sterk de samenhang is tussen termen binnen een onderwerp. Ook zou het een beeld moeten geven van de samenhang tussen verschillende onderwerpen.

4 Evaluation

Met een subsectie voor elke deelvraag.

In hoeverre is je vraag beantwoord?

Een mooie graphic/visualisatie is hier heel gewenst.

Hou het kort maar krachtig.

4.1 Zijn de methodes gebruikt in Rule et al. [2015] toepasbaar op het Nederlandse troonrede corpus?

De Nederlandse troonrede is net zoals het de Amerikaanse State of the Union een toespraak die jaarlijks vanuit de overheid worden gegeven. In beide gevallen wordt de voordracht door één positie voorgedragen, voor Nederland is dit de koning(in) en voor de VS is dit de president. Voor beide teksten geldt dat ze zijn opgebouwd uit verschillende paragrafen en er veelal voor elke paragraaf één onderwerp centraal staat. Beide hebben een centraal motief gericht op het meedelen van de huidige staat van het land en mogelijk voornemens voor het komende jaar. Omdat de opbouw en het doel van beide teksten zoveel op elkaar lijken zijn de methodes die gebruikt zijn binnen Rule et al. [2015] goed toepasbaar op het corpus van de Nederlandse troonredes. Hierbij hoeven er maar kleine aanpassingen gemaakt te worden, waarbij het grootste verschil ligt in het feit dat de teksten in verschillende talen zijn geschreven. Voor het Nederlands betekent gaf dit enkele problemen omdat woorden over de jaren op verschillende manieren zijn geschreven en de taal over de jaren is verandert. Zo werd het woord "de" vroeger als "den" geschreven. Hierdoor zijn enkele woorden die vroeger anders geschreven werden niet opgevangen door de POS-tagging. Op het moment dat dit duidelijk werd is het handmatig aangepast. Op deze manier is er net als in Rule et al. [2015] het corpus gereduceerd tot de relevante zelfstandige naamwoorden. Hierbij is dus gebruik gemaakt van modules die op de Nederlandse taal getraind waren. Hierna zijn de 1000 meest voorkomende termen gebruikt worden om de co-occurence matrix op te stellen dat gebruikt is om de nabijheidsscore te berekenen waarmee het semantische netwerk is gevormd.

4.2 Zijn de resultaten representatief voor de werkelijkheid?

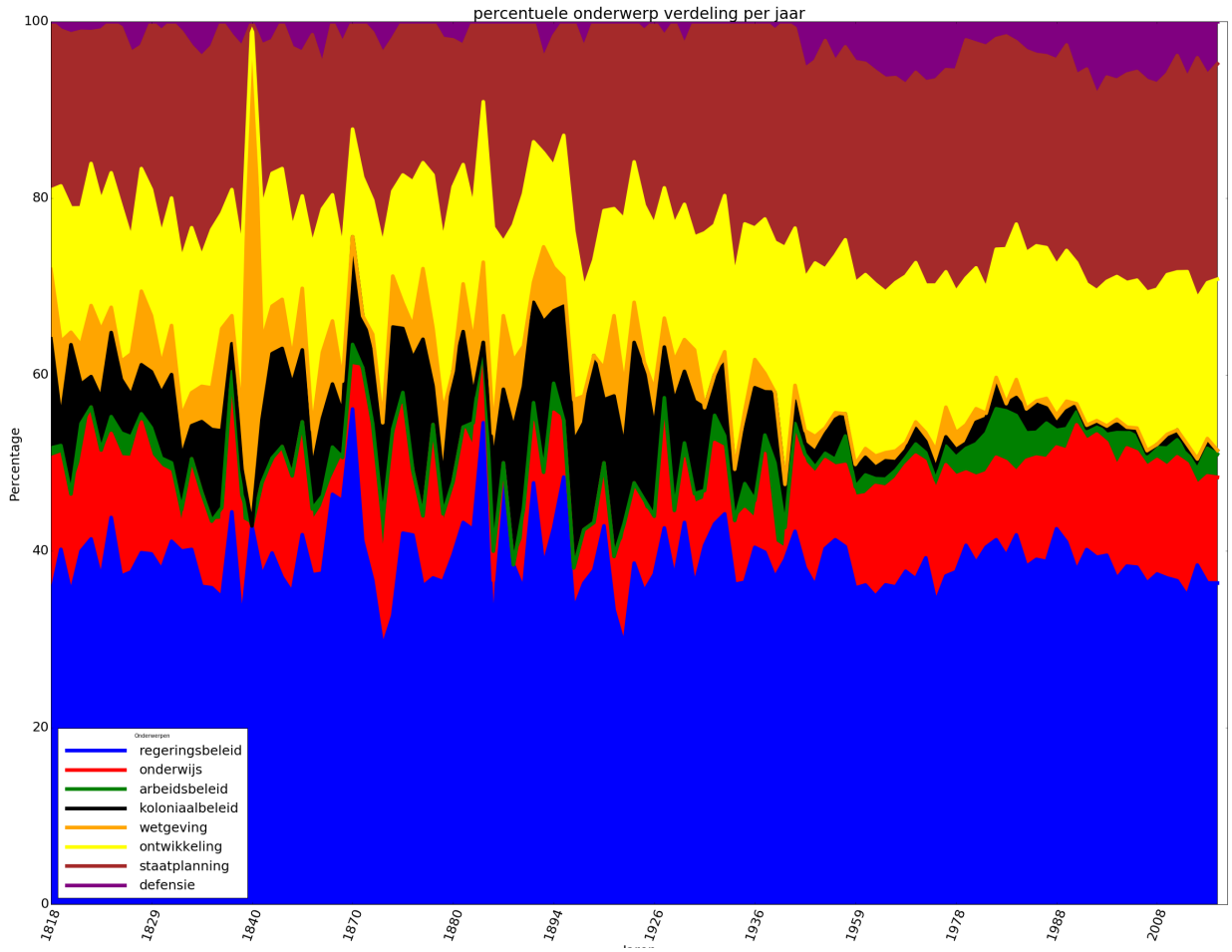
De verkregen resultaten zijn de geïnduceerde onderwerpen en het daarmee verkregen overzicht van onderwerp verschuivingen. Hiervoor was het belangrijkste dat er vanuit het semantische netwerk door middel van het community clustering algoritme duidelijke clusters gevormd werden. Vanuit de termen binnen een cluster kan een achterliggend onderwerp worden bepaald dat door deze termen wordt geïmpliceerd. Het is dus noodzakelijk dat de clusters duidelijk verschillend van elkaar zijn en de relatie tussen de termen in een cluster zo min mogelijk verschillende onderwerpen impliceren. Hoe duidelijker de relatie is tussen de termen in een cluster hoe makkelijker het is om er een onderwerp aan te koppelen. Dit is het makkelijkst voor clusters die gecentreerd zijn rondom één term, zoals "onderwijs" of "ontwikkeling".

De onderwerpen die gevonden zijn staan in tabel 2:

Onderwerpen
regeringsbeleid
onderwijs
arbeidsbeleid
koloniaalbeleid
wetgeving
ontwikkeling
staatplanning
defensie

tabel 2: Geïmpliceerde onderwerpen

Voor elk onderwerp is er een lijst met termen waarvoor geldt dat als deze voorkomen in een tekst ze dat onderwerp impliceren. Aan de hand van deze lijst is een grafiek opgesteld met daarin voor elk jaar de hoeveelheid dat een onderwerp besproken werd. Dit is gedaan door te kijken naar de volledige tekst van dat jaar en voor elk woord na te gaan bij welk onderwerp het hoort. Hierna is door gebruik te maken van een normaal verdeling een grafiek gevormd die per onderwerp procentueel aangeeft hoeveel van de troonrede besteed was aan de behandeling van dat onderwerp. Hieruit volgt de grafiek in figuur 2:



figuur 2: Verschuiving van inhoud van troonredes over de jaren aan de hand van de gevonden onderwerpen.

Vanuit deze grafiek is duidelijk af te lezen welke onderwerpen het belangrijkst waren in een jaar. Zo is bijvoorbeeld duidelijk te zien dat er rond 1839-40 het onderwerp wetgeving sterk naar voren komt, wat terug te leiden is naar de werkelijkheid, omdat in 1840 de officiële scheiding plaatsvond tussen Nederland en België ondertekend in het verdrag van Londen. Door dit verdrag is een groot deel van de Nederlandse grondwet aangepast. [Schroeder, 1996] De resultaten zijn dus zichtbaar terug te leiden op de werkelijkheid.

5 Conclusions

Hierin beantwoord je jouw hoofdvraag op basis van het eerder vergaarde bewijs.

5.1 Acknowledgements

Hier kan je bedanken wie je maar wilt.

References

Charu C Aggarwal and ChengXiang Zhai. A survey of text clustering algorithms. In *Mining Text Data*, pages 77–128. Springer, 2012.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10): P10008, 2008.

Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40, 2009.

Michel Callon, Jean Pierre Courtial, and Françoise Laville. Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. *Scientometrics*, 22(1):155–205, 1991.

Pieter C. Lagas. Herko Coomans. troonredes.nl, 2015. URL `www.troonredes.nl`.

Yutaka Matsuo and Mitsuru Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(01): 157–169, 2004.

Juan Ramos. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, 2003.

Rijksoverheid. rijksoverheid, 2016. URL `https://www.rijksoverheid.nl/onderwerpen/koninklijk-huis/inhoud/positie-en-rol-staatshoofd/troonrede`.

Alix Rule, Jean-Philippe Cointet, and Peter S. Bearman. Lexical shifts, substantive changes, and continuity in state of the union discourse, 1790–2014. *Proceedings of the National Academy of Sciences*, 112 (35):10837–10844, 2015. doi: 10.1073/pnas.1512221112. URL `http://www.pnas.org/content/112/35/10837.abstract`.

Paul W Schroeder. *The transformation of European politics, 1763-1848*. Oxford University Press, 1996.

Sayori Shimohata, Toshiyuki Sugio, and Junji Nagata. Retrieving collocations by co-occurrences and word order constraints. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 476–481. Association for Computational Linguistics, 1997.

A Slides