Title of the thesis

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

0.1 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 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 wat de belangrijke onderwerpen in een speech waren en welke thema's behandelt werden over de jaren.
De State of the Union is niet de enige speech die al lange tijd wordt gehouden. Zo ook de troonredes van ons eigen Nederland. Deze troonredes worden sinds 1818 jaarlijks door de koning/koningin gegeven. 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 update over de staat van het land. De staat van het land wordt ook wel besproken, maar meer in de context van de wereld, namelijk 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?

RQ1 We beantwoorden deze vraag met behulp van de volgende deelvragen:

1. Wat is "word co-occurence" en hoe worden deze uit teksten verzameld? Evaluatie sectie 0.4.
2. Hoe kan betekenis worden gegeven aan co-occurences?
3. Hoe wordt de context van de tekst in acht gehouden?
4. Wat voor invloed heeft de verandering van taal(gebruik) over de jaren op de analyse technieken?

Overview of thesis Hier geef je even kort weer wat in elke sectie staat.

0.2 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.

0.2.1 RQ1

0.2.2 RQ2

0.3 Methodology

0.3.1 Description of the data

De dataset die gebruikt gaat worden om deze vragen te beantwoorden is een verzameling van troonredes sinds 1814. Deze troonredes zijn terug te vinden op www.troonredes.nl [Herko Coomans, 2015] . Om een duidelijker beeld te geven van wat de troonrede nu precies is eerst een kleine introductie van wat de troonredes nu precies zijn.
De troonrede wordt jaarlijks door de koning(in) uitgesproken op Prinsjesdag. De eerste troonrede werd in 1818 als een algehele toespraak voor de Staten-generaal gehouden. De troonredes worden vooral gebruikt om wets- en beleidsveranderingen door te geven en 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. Sinds 1848 worden de troonredes geschreven door ministers en is het kabinet verantwoordelijk voor de uitspraken. Dit zorgt ervoor dat de troonredes een beeld geven van wat de Nederlandse regering op dat moment belangrijk vindt.

0.3.2 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 1818 wordt gerepresenteerd door een troonrede. Omdat de inhoud van de troonredes worden 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.

0.3.3 Quality of the data

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

0.3.4 Wat plotjes en tabelletjes

Zie het IPython Notebook voor de code om vanuit pandas een poltje op te slaan en een dataframe als tabel op te slaan. Het werkt ideaal!

De interrupties van Wilders staan beschreven in Figure **??** en Tabel **??**.

0.3.5 Methods

Het gehele corpus bestaat uit 117 verschillende troonredes, welke in totaal uit 143875 woorden bestaan. Om uit deze data te kunnen achterhalen of er overkoepelende thema's/onderwerpen zijn en wat deze mogelijk zijn wordt er gebruik gemaakt van tekstanalyse technieken. Specifiek wordt er gebruik gemaakt van co-occurence aanpakken. Callon et al. [1991] Hierbij worden categorin geinduceerd door te kijken naar het samen voorkomen van termen in afzonderlijke stukken tekst. Hierbij is het doel om uit te vinden welke termen relevant zijn en hoe deze zich over tijd met andere termen associren. Met behulp van deze informatie zouden er uitspraken gedaan kunnen worden over de onderwerpen die termen impliceren. Hiermee kunnen dan uitspraken worden gedaan over de inhoud van een specifieke troonrede aan de hand van de termen in de tekst.
Voordat deze uitspraken echter kunnen worden gedaan moet de data eerst worden verwerkt zodat er analyse op kan worden gedaan. Hiervoor worden de teksten van de troonredes eerst gelemmatiseerd. Hierdoor worden de woorden herleid tot hun lemma(stam), waardoor ze kunnen worden geanalyseerd als n term. Met behulp van deze gelemmatiseerde termen wordt een co-occurence matrix opgesteld. Deze matrix omvat voor alle combinaties van termen de frequentie dat ze samen in een paragraaf voorkomen.
Voor alle combinaties uit deze matrix wordt een nabijheidsscore berekend via het volgende algoritme.

$$S(W_1, W_2) = \frac{\sum_{c \epsilon W\{W_1,W_2\}, I(W_1,c)>0} min(I(W_1, c), I(W_2, c))}{\sum_{c \epsilon W\{W_1,W_2\}, I(W_1,c)>0} I(W_1, c)}$$

Dit algoritme maakt gebruik van de woorden uit het gehele corpus om context te bepalen. Hiervoor wordt voor elke combinatie woorden (W1,W2) gekeken naar de woorden waarmee ze samen in een paragraaf voorkomen. De som van I(W,C) voor alle woorden (C) uit het corpus waarvoor geldt dat I(W,C) ¿ 0 wordt voor beide woorden berekend. Als hieruit volgt dat I(W,C) voor beide woorden gelijk is kan men stellen dat als W1 in een paragraaf voorkomt W2 ook voorkomt en vice versa. Een verwantschapsscore 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. Afhankelijk van de termen binnen deze clusters kunnen deze clusters gezien worden als in-

dicatief 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.
Om tot het gewogen netwerk te komen dat gebruikt kan worden voor de clustering wordt de dataset van woorden binnen het gehele corpus eerst gereduceerd tot de meest relevante data. Allereerst wordt de dataset gelemmatiseerd, waarna de 1000 meest voorkomende termen gebruikt worden om de co-occurence matrix op te stellen. De nabijheidsscore voor de termen uit deze matrix worden berekend en om ervoor te zorgen dat enkel de meest relevante termen worden meegegeven aan het community detectie algoritme wordt een drempel bepaald. Deze drempel geeft aan vanaf welke waarde van 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 telkens op te hogen. De drempelwaarde is bereikt als er binnen het netwerk een component van meer dan 2 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 wordt bepaald aan de hand van de termen die er onder vallen of er een onderwerp mee geassocieerd kan worden.

RQ1

RQ2

0.4 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.

0.5 Conclusions

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

0.5.1 Acknowledgements

Hier kan je bedanken wie je maar wilt.

Bibliography

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.

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`.

Alix Rule, Jean-Philippe Cointet, and Peter S. Bearman. Lexical shifts, substantive changes, and continuity in state of the union discourse, 17902014. *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`.

.1 Slides