

Een algemene kijk naar toespraken van de verkiezingen in de VN in 2008

Onderzoeksverslag van de datavisualisatie van de toespraken van de verkiezingen
van de VN van 2008

Door Marten Folkertsma

1 Inleiding

In 2008 vonden de verkiezingen voor een nieuwe president in Amerika plaats. Barack Obama won deze verkiezingen en werd de 44ste president van Amerika. Tijdens deze verkiezingen streden verscheidende kandidaten voor het presidentschap. Zij reisden over het gehele land om toespraken te geven en zo de mensen over te halen om voor hun te stemmen. Deze visualisatie probeert een inzicht te geven in een aantal eigenschappen van deze toespraken en hoe deze de populariteit van de kandidaten heeft beïnvloed.

Er zijn een aantal karakteristieken uit deze toespraken gehaald en die worden tegenover de populariteit van de sprekers gezet om te kijken of er mogelijk een correlatie te vinden is. Hier gaat het vooral om de moeilijkheidsgraad van een toespraak en op welke manier het inspeelt op het publiek. Dit is gedaan om een gevoel te krijgen hoe presidenten zich voorbereiden op de verkiezingen en nog meer hoe zij zich proberen aan te passen op hun publiek. Inspiratie en ook een aantal karakteristieken kwam uit een artikel van Ban and Oyabu (2009) die eenzelfde onderzoek naar de moeilijkheidsgraad tijdens de verkiezingen in 2008.

In dit verslag wordt de manier waarop de visualisatie is opgebouwd besproken. Het gaat hier zowel om hoe het technisch in elkaar zit als wat voor designkeuzes er zijn gemaakt. Eerst zal de doelgroep waar de visualisatie op gebaseerd is kort besproken worden. Vervolgens hoe de data verkregen is en hoe deze is omgezet in presenteerbare data. Op het laatst zal de visualisatie zelf besproken worden, hoe deze in elkaar zit en welke keuzes er voor gemaakt zijn.

2 De doelgroep van de visualisatie

Deze visualisatie richt zich op de verkiezingen van de VS, dus het is logisch dat het inspeelt op zij die daarin geïnteresseerd zijn. Specifieker is het gericht op zij die het interessant vinden om te zien hoe presidentskandidaten en specifieker de PR-teams van de kandidaten proberen om de toespraken zo goed mogelijk in te spelen op het publiek waar zij voor spreken. Het idee is om een zo objectief mogelijk beeld te geven van deze toespraken. Dat betekent dat er alleen algemene harde data wordt verkregen, dat de karakteristieken niet te veel een kandidaat bevoordelen.

Ook is het de bedoeling dat iedereen die hierin geïnteresseerd is er iets uit zou moeten kunnen halen. Dit betekent dat er geen achtergrondkennis vereist is van noch textanalyse noch de Amerikaanse verkiezingen. Een leek hoort er naar te kijken en er mogelijk zelf een conclusie uit te kunnen halen. Ik zal dan ook geen eigen conclusie geven over de data die wordt weergegeven.

3 Data Verkrijgen

Om de toespraken te kunnen analyseren zijn ze eerst van het internet gehaald. Vervolgens zijn ze bewerkt met Python zodanig dat het gevisualiseerd kan worden. Bepaalde karakteristieken zijn eruit gehaald, gebruik makende van Python. De toespraken zijn van de universiteit van Californië gehaald (EDU). Hier zijn ze vannaf gescraped met Python en de pattern library. Het is als een lijst van lijsten binnengehaald, waarin elk datapunt bestond uit een lijst. In deze lijst stond de spreker van de toespraak, de datum, de titel en uiteindelijk de speech zelf. Dit is per president van het internet afgehaald en in een csv bestand gezet. Dit heb ik gedaan met een scraper waarbij steeds de basis site veranderde zodat een andere president werd opgepikt.

De data van de verschillende polls zijn ook van het internet gehaald, deze komen van de volgende site(pol). Hiervoor waren in totaal drie verschillende scrapers nodig, die de data van verschillende plekken van de site haalden. Het ging hier om de pre presidential verkiezingen van de democraten en van de republikeinen en om de presidentiële verkiezingen. Deze data was al vrij snel klaar voor gebruik, alleen een paar datums deden raar, maar die zijn met de hand in een editor aangepast.

Nadat de toespraken binnen waren gehaald werden deze bewerkt zodanig dat er grafieken van konden worden gemaakt. Hiervoor werd ook weer Python gebruikt. De vier verschillende karakteristieken die uit de text werden gehaald waren: de gemiddelde zinlengte, de gemiddelde woordlengte, een graad van moeilijkheid en het aantal keer dat het woord "we" voorkwam. Hiervoor is een programma in Python geschreven.

Om de gemiddelde zinlengte te bepalen zijn alle zinnen in een toespraak geteld, dit is vervolgens gedeeld door het totaal aantal woorden. De gemiddelde woordlengte ging redelijk hetzelfde alleen worden daarvoor alle woorden en alle karakters per speech geteld en opgeslagen.

Voor de graad van moeilijkheid is een extrern woordenboek gebruikt. Dit was een woordenboek met 850 basiswoorden van het Engels die elke persoon met Engels als moedertaal zou moeten kennenwoo. Deze woordenlijst werd vergeleken met toespraak en de verhouding van woorden die in de lijst stonden en woorden die niet in de lijst stonden werd als graad van moeilijkheid genomen. Dit is een karakteristiek die Ban and Oyabu (2009) ook hebben gebruikt in hun artikel. Een probleem hiermee was dat werkwoordsvervoegingen en vervoegingen in het algemeen niet in het woordenboek staan. Dit betekent dat er een klein beetje een vertekend beeld wordt gegeven van wat het eigenlijk zou moeten zijn. Aangezien dit echter voor alle spreker gelijk was, geeft deze karakteristiek toch een graad van moeilijkheid weer.

De laatste karakteristiek die gebruikt is, is het percentage van de woorden "we". Dit is gebruikt om te kijken hoe de spreker probeerde aan te geven dat het over iets gezamelijk gaat. Dus dat de sprekers proberen het publiek over te halen vanwegen dat zij het samen met hen president konden worden. De statistiek zelf is betrekkelijk simpel. Het aantal keer dan "weōf" "We" voorkwam werd getelt en dit werd gedeeld door het aantal woorden.

Met deze operaties zijn de texten omgezet van een set woorden tot een aantal getallen bij een bepaalde datum, zodat er grafieken van gemaakt konden worden. De datums die van de site gehaald zijn, zijn ook met python omgezet tot een format makkelijk te lezen voor javascript. Naast dat de data op een bepaalde datum nodig was voor deze visualisatie was ook de locatie van een gegeven toespraak belangrijk. Jammer genoeg stonden er geen locaties bij de toespraken waar deze vandaan kwamen. In de titels van de toespraken stond echter wel vaak een locatie. Om de locatie dan ook uiteindelijk te krijgen zijn de titels vergeleken met de namen van alle staten en de top honderd belangrijkste steden. Als de naam van een staat of de naam van een stad in de titel stond werd de titel hiermee gevonden waar de toespraak gegeven was. De staat waar dit was is opgeslagen in de bewerkte data.

Deze bewerkingen zijn voor alle presidenten apart gedaan en als output werd per president een csv file gegenereerd met deze data erin. In deze csv files stonden regels met de volgende opbouw: de spreker, de datum, de locatie, gemiddelde zinlengte, gemiddelde woordlengte, graad van moeilijkheid, percentage we. Met deze data is vervolgens de visualisatie gegenereerd.

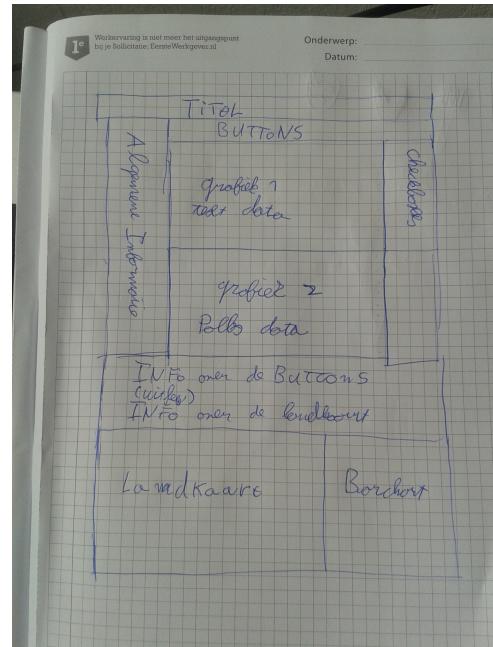
4 De bouw van de visualisatie

Met alle data in csv files klaar om in te laden, kon nu de visualisatie gemaakt worden. In dit stuk zal de opbouw van de visualisatie besproken worden. Hiervoor was als eerst een design document gemaakt. Hierin stond een schets hoe de visualisatie er uiteindelijk ongeveer uit moest zien. Op basis van het design document is vervolgens een HTML pagina gebouwd. Deze HTML pagina werd ondersteund door Javascript en CSS om interactiviteit en juiste plaatsing van onderdelen te regelen.

4.1 design document



(a) Figuur 1: Beginschets van de visualisatie



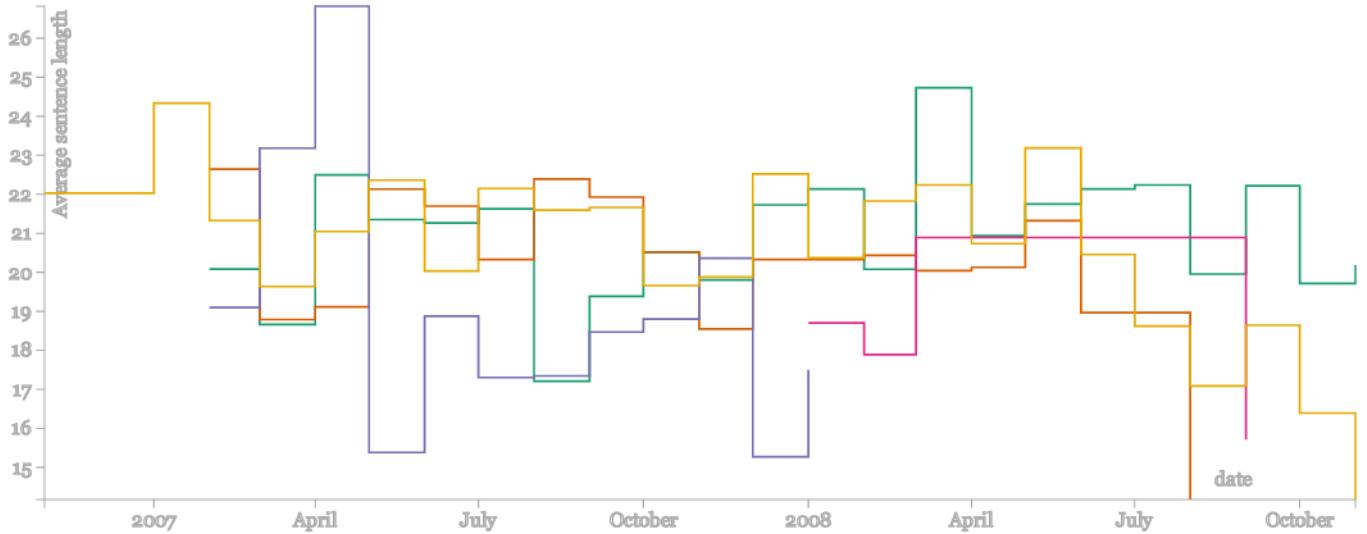
(b) Figuur 2: Schets hoe de visualisatie er uit ziet

Hierboven staat het design document zoals het eerst was voorgesteld in figuur 1 en in figuur 2 staat wat het uiteindelijk geworden is. In deze paragraaf zal kort besproken worden waarom figuur 1 was voorgesteld en welke veranderingen er zijn toegepast naar figuur 2 toe.

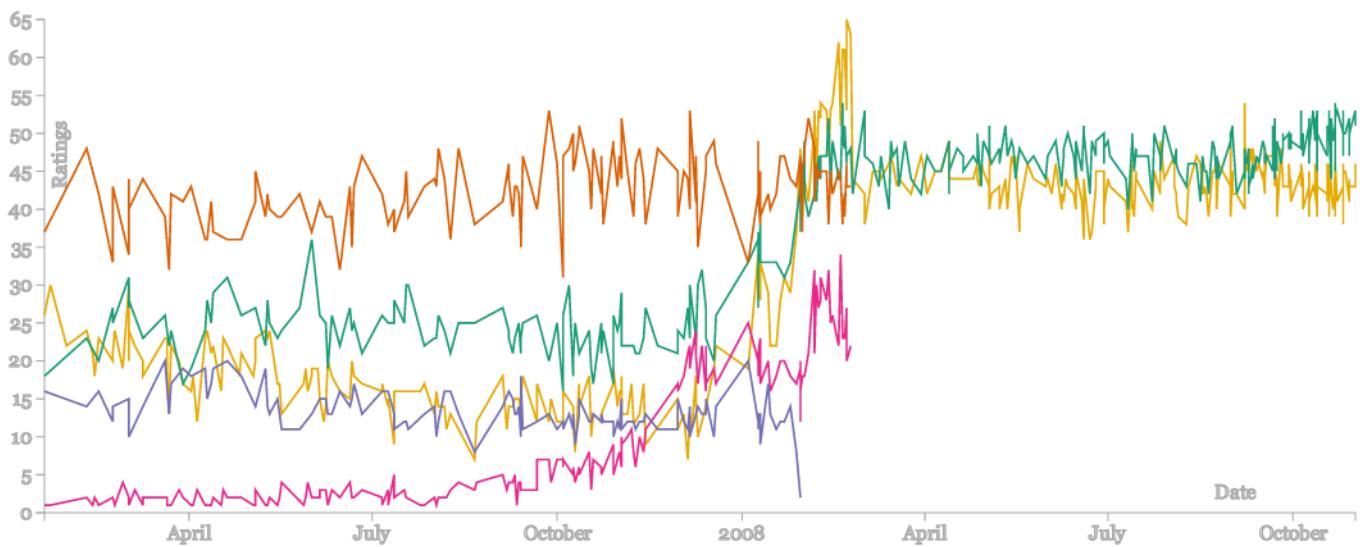
Het idee achter de visualisatie was om te laten zien hoe de sprekers hun toespraken veranderden over de tijd en hoe dat mogelijk invloed had op hun populariteit. Om deze data weer te geven zijn twee grafieken gebruikt met dezelfde assen over de tijd. De één geeft de populariteit weer en de ander de karakteristieke eigenschappen die onderzocht worden. Hiervoor waren de twee grafieken van belang. In het midden stonden in het eerste idee een groep checkboxes. Deze zijn er om verschillende presidenten te benadrukken of juist niet. Zou een box van een president zijn gecheckt dan krijgt de bij de president hoordende lijn een duidelijke kleur, anders is deze grijs.

Uiteindelijk is er voor gekozen om de grafieken onder elkaar te zetten zoals weergegeven in figuur twee. Dit is gedaan zodat een gebruiker makkelijker beide grafieken met elkaar kan vergelijken. Door te zorgen dat de assen gelijk zijn en door de grafieken boven elkaar te zetten kan je makkelijker twee gebeurtenissen in de tijd vergelijken. Natuurlijk staat op de x-as de tijd en op de y-as de waarde van het op dat moment gekozen datatype. Het blok met algemene informatie is ook verschoven naar naast de grafieken om meteen informatieernaast te hebben staan. Zodat er niet hoeft te worden gescrolled om een beetje achtergrond bij de grafieken te kunnen lezen. Een verdere uitleg van de knoppen staat wel onder de grafieken. Zo wordt de gebruiker getrokken door de grafieken en krijgt hij eerst een inleidend stuk te lezen over de visualisatie. Ook zijn de knoppen over het algemeen redelijk voor zichzelf sprekend.

4.1.1 De grafieken



Figuur 2: Grafiek van de gemiddelde zinlengte



Figuur 3: Grafiek van de opiniepeilingen

De grafieken zullen er uit zien als hierboven weergegeven in figuur drie en vier. Voor de opiniepeilingen is gekozen om per datum waarop een opiniepeiling is gedaan in een grafiek te zetten. De peilingen van de pre presidentiële verkiezingen voor zowel democraten als republikeinen is samen met de presidentiële verkiezingen in een grafiek gezet. Hierdoor ontstaat een kleine onrealistische weergave, omdat het lijkt alsof McCain ineens heel veel zakt in de opinies. Dit is toch zodanig in de grafiek gehouden om continuïteit in de weergave te behouden.

In figuur drie wordt de grafiek met de text-data weergegeven. Hier is het een weergave van de grafiek met de gemiddelde zinlengtes. Voor deze grafiek is gekozen om steeds het gemiddelde per maand weer te geven. De y waardes geven dus het gemiddelde per maand voor de desbetreffende president weer en de x waardes de maand waarin dit zo was. Dit is gedaan omdat de grafieken anders nauwelijks af te lezen waren vanwege de chaos er in. Ook is gekozen voor een stapgrafiek, dit ook weer vanwege de gemiddelde per maand.

In beide grafieken is te zien dat alle lijnen verschillende kleuren hebben, dit is voor verbinding met de president. Elke president heeft zijn eigen kleur, die in deze visualisatie telkens wordt gebruikt. Er is bij het uitzoeken van de kleuren op geled dat alle kleuren dezelfde waarde dragen. De ene kleur lijkt niet belangrijker dan een andere

kleur. Ook is er voor gezorgd dat er niet al te politieke kleuren zoals rood en blauw gebruikt zijn. Deze kleuren woorden op de hele pagina gebruikt om de verschillende kandidaten weer te geven.

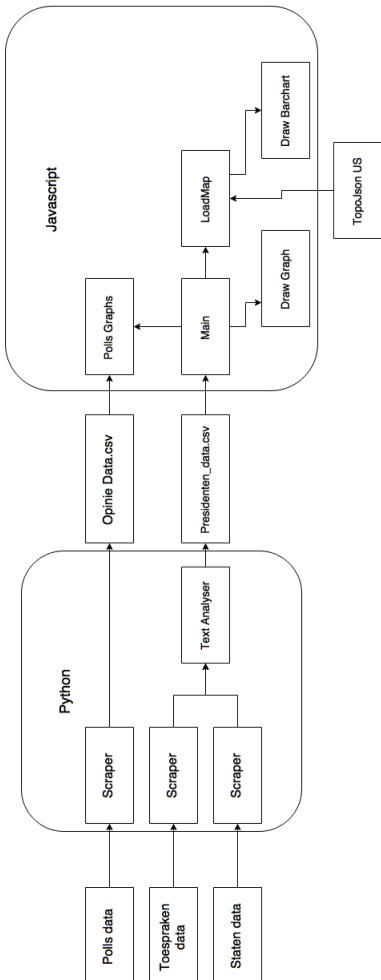
4.1.2 De kaart van Amerika

Het tweede deel van de visualisatie bestaat uit data op locatie. Van een deel van de toespraken was ook uitgezocht waar deze plaats hebben gevonden. Deze data wordt weergegeven met een kaart van Amerika en een barchart. De kaart is er om te kunnen selecteren tussen de verschillende staten. Door te klikken op een staat ontstaat er een barchart met daarin de waardes van een specifieke dataset per president. Het gaat hier om de gemiddelde waardes van de toespraken per president die zijn gehouden in een bepaalde staat. Als een president drie toespraken heeft gehouden in die staat wordt het gemiddelde genomen van die drie toespraken. Als een staat geselecteerd wordt ligt deze blauw op zodat het duidelijk is welke staat geselecteerd is.

Om een goed verschil weer te kunnen geven tussen de presidenten is er voor gekozen om van deze data een barchart te maken. Dit laat in een oogopslag de verschillen tussen de verschillende presidenten zien. Om ook nog het verschil tussen de verschillende staten te kunnen bekijken wordt het bereik van data bepaald op de gehele dataset, dus niet alleen op de data in een geselcteerde staat. Zo is het makkelijk om verschil tussen verschillende staten te kunnen zien. Aangezien er maar één waarde per president hoeft weergegeven te worden was de barchart een handige keuze.

4.2 De bouw van de code

Deze visualisatie is gebouwd met Javascript en in het bijzonder d3, een bibliotheek in Javascript. Naast d3 zijn ook jquery.js queue.js en bootstrap gebruikt om deze visualisatie in elkaar te zetten. Het idee achter de visualisatie was om een soort pipeline te bouwen waarmee de data gegenereerd kon worden en ook in een keer kon worden gebruikt. Het generen van de data is gedaan met Python en hierboven besproken. In de visualisatie zelf is een soort pipeline structuur met vertakkingen gebruikt, waarbij een functie steeds doorverwijst naar een volgende functie. Als eerst wordt de informatie behorende bij de speeches ingeladen. Deze data is overal nodig en wordt dus aan het beginpunt van de pipeline ingeladen. Hiermee wordt de eerste grafiek gegenereerd en een setup gebouwd voor de interactiviteit die met deze grafiek te maken heeft. Daarna wordt een nieuwe functie aangeroepen die de mapdata inlaadt om de kaart van America te generen. In die functie wordt alles behorende bij de landkaart en bij de barchart in elkaar gezet. Vervolgens wordt er een functie aangeroepen om de opiniedata in te laden en te visualiseren. Hiermee wordt alles neergezet op de juiste plekken. Ter verduidelijking staat er in figuur 5 een visualisatie van dit proces.



Figuur 4: Weergave van de pipeline

5 Reflectie op de visualisatie

Een van de grootste problemen die ik tegen ben gekomen was het verzamelen van data. Als ik het project opnieuw zou doen zou ik veel eerder willen beginnen met het uitzoeken en opzoeken van mijn data zodat dat eerder klaar staat. Zodra dat eenmaal goed en klaar is kan de rest van de visualisatie makkelijker in elkaar gezet worden. Een ander groot probleem is hoe je het best de ruimte op het scherm kan gebruiken en daarbij nog belangrijker; hoe dat kan op verschillende schermen. Als iemand anders het opent op een kleiner scherm of dat dan nog goed staat.

De uiteindelijke visualisatie is naar mijn gevoel naar behoren geïmplementeerd, de manier waarop de functies samen werken zit redelijk prima in elkaar. Waar ik een verbetering zou willen zien vooral is de manier waarop ik de data heb geoordend in mijn programma. Er zouden meer functies aan de datastructuur zelf verbonden kunnen worden zodat opzoeken van dingen een stuk makkelijker zou kunnen. Dit en ook transitions in de grafiek zou best leuk zijn geweest, of zeker bij de landen.

Wat wel erg jammer is, is dat er eigenlijk niet genoeg data was voor de locaties. Er is op veel plekken geen data aanwezig. Dit zou verbeterd kunnen worden door te kijken waar presidenten waren via andere bronnen op een bepaald tijdstip. Hier was echter geen tijd voor in de implementatie van deze visualisatie.

Het laatste punt waarop uitgebreid zou kunnen worden is in de data verkregen uit de teksten. Door wat extra woordanalyses toe te voegen aan het Pythonscript zou er makkelijk extra data kunnen worden toegevoegd in de visualisatie.

Referenties

2008 presidential election documents. <http://www.presidency.ucsb.edu/2008election.php>. Accessed : 02 – 06 – 2015.

Ogden's basic english word list. Accessed: 03-06-2015.

Polls data. <http://www.pollster.com>. Accessed: 02-06-2015.

Ban, H. and Oyabu, T. (2009). Metrical analysis of the speeches of 2008 american presidential election candidates. In *Fuzzy Information Processing Society, 2009. NAFIPS 2009. Annual Meeting of the North American*, pages 1–5. IEEE.