Final Project: Text-mining US Election Debates

*Christoffer Mondrup Kramer*

*08-12-2020*

*Student nr.: 201704767*

*Cultural Data Science*

*Final exam*

*Aarhus University*

***Words****: 1.910*

*(Front-page, table of contents, and code-chunks excluded).*

**Practical Information**

* *Repository:* <https://github.com/Digital-Methods-HASS/au590388_Christoffer_Kramer/tree/master/final_project>
* *Figures:* Also provided in the appendix.
* *Data source:* <https://www.debates.org/voter-education/debate-transcripts/>
* *License*:
  + Transcripts: In the public domain.
  + My repository: Licensed under *creative commons*

**Contents**

[Introduction 3](#_Toc58358229)

[Web Scraping 3](#_Toc58358230)

[Data Cleaning and Data Wrangling 5](#_Toc58358231)

[Cleaning Dates 5](#_Toc58358232)

[Who is Speaking? 6](#_Toc58358233)

[Text-mining: Preparation 8](#_Toc58358234)

[Text-mining: Stop Word Filtering 9](#_Toc58358235)

[Plotting Stop Word Filtering 9](#_Toc58358236)

[Text-mining: Term Frequency - Inverse Document Frequency 10](#_Toc58358237)

[Text-mining: Sentiment Analysis 12](#_Toc58358238)

[Plotting sentiment 13](#_Toc58358239)

[Conclusion 15](#_Toc58358240)

## Introduction

library(tidyverse) #Cleaning data  
library(tidytext) #Text-mining  
library(ggplot2) #Plotting  
library(ggwordcloud) #Plotting word clouds  
library(rvest) #Web scraping  
library(lubridate) #Cleaning Dates  
library(zoo) #Dealing with na's

## Text-mining: Preparation

Since I’m gonna be making a lot of plots, I will make a function for saving my plots. This will make writing the code easier, and it will make it easier to combine all pdf files into one big appendix file, since they all have the same size.

save\_plot\_pdf <- function(name, path\_destination = "../plots/", width\_value = 8.26, height\_value = 11.69){  
 dev.copy(pdf, paste(path\_destination, name, ".pdf", sep = ""), width= width\_value, height= height\_value) #Units are inches  
 dev.off()  
}

## Combining Sentiment and wordclouds

Let’s get an overview of which words the candidates use and their sentiment. This can be achieved with a word cloud, where the size of each word is determined by the count and the color is determined by the sentiment (red = negative, green = postive). Since I wan’t a gradient scale, I’m using the afinn lexicon rather than bing:

sentiment\_debate\_words %>%  
 filter(grepl("[DR]", party)) %>%   
 count(year, word, party, value, sort = TRUE) %>%  
 group\_by(year, party) %>%   
 slice\_max(order\_by = n, n= 50) %>%   
 plot\_wordclouds(color\_value = value) +  
 scale\_color\_gradient(low = "red", high = "green") +  
 facet\_wrap(~year + party) +  
 labs(title = "Fig. 7: Democrats' sentiment and word count year")

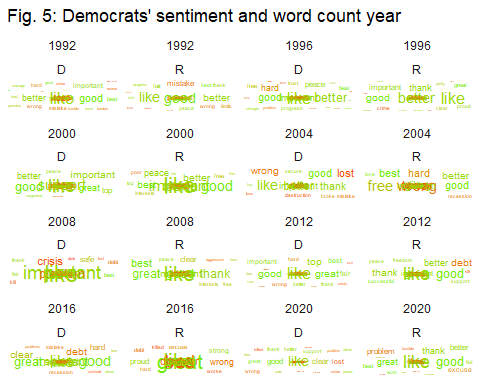


Fig. 5: Afinn Word cloud. The size of each word is the count, the color is the sentiment

save\_plot\_pdf("fig\_7")

Interestingly the same words tends to be used across all debates. Words such as “Like”, “important”, “safe” and “better” are often used as positive words. The word “like” is especially problematic here, since it can be an adjective, a verb, a preposition and an adverb, which changes its sentiment, this suggest that the debates are more negative than our previous analysis showed. What is surprising is how much overlap there actually is between candidates from both parties. They tend to use the same words a lot, and the big difference is in their use of negative words with less common usage such as “terror”, which is primarily used by republicans" and “crisis”, which is more commonly used by democrats.

However, as figure 5 shows, candidates from both parties share a surprisingly similar vocabulary, since words such as “debt”, “wrong”, “problem” and “crisis” appear consistenly across all debates from each party.

**Litterature:**

https://stackoverflow.com/questions/47209828/r-tidytext-stop-words-are-not-filtering-consistently-from-gutenbergr-downloads

https://www.pewresearch.org/topics/political-polarization/).

- Pewreseach.org (n.d.): ”Political Polarization”. *Pew Research Center*. URL: <https://www.pewresearch.org/topics/political-polarization/> .

- Wickham H et al. (2019). “Welcome to the tidyverse.” *Journal of Open Source Software*, 4(43), 1686. doi: 10.21105/joss.01686

- Grolemund, G. & Wickham, H. (2011). ”Dates and Times Made Easy with lubridate”. *Journal of Statistical Software*, 40(3), 1-25. URL http://www.jstatsoft.org/v40/i03/

- Zeileis, A. & Grothendieck, G. (2005). “Zoo: S3 Infrastructure for Regular and Irregular Time Series.” *Journal of Statistical Software*, 14(6), 1–27. doi: 10.18637/jss.v014.i06

- Broman, W. K., & Woo, H. K. (2018). ”Data Organization in Spreadsheets”. *The American Statistician*, 72:1, 2-10, DOI: 10.1080/00031305.2017.1375989

- Silge, J. & Robinson, D. (2017). “Text Mining with R: A Tidy Approach”. *O’reilly Media, Inc.* URL: <https://www.tidytextmining.com/>

- Silge J, Robinson D (2016). “Tidytext: Text Mining and Analysis Using Tidy Data Principles in R.” *JOSS*, 1(3). doi: 10.21105/joss.00037, <http://dx.doi.org/10.21105/joss.00037>

- Wilke, C. O. (2019). “Chapter 10: Visualizing proportions” in *Fundamentals of Data Visualization*. *O’Reilly Media Inc.*, URL: <https://clauswilke.com/dataviz/visualizing-proportions.html>

(Wickham, 2014) (datacarpentry.org, n.d.): https://datacarpentry.org/spreadsheets-socialsci/03-dates-as-data/index.html

“Text Mining with R: A Tidy Approach” (2020).

Wilke, C. O. (XXXX - FIXME). “Chapter 10: Visualizing proportions” in Wilke, C. O. Fundamentals of Data Visualization, O’Reilly Media, Inc, URL: https://clauswilke.com/dataviz/visualizing-proportions.html

Packages

Zeileis, A. & Grothendieck, G. (2005). “Zoo: S3 Infrastructure for Regular and Irregular Time Series.” *Journal of Statistical Software*, 14(6), 1–27. doi: 10.18637/jss.v014.i06

Grolemund, G. & Wickham, H. (2011). ”Dates and Times Made Easy with lubridate”. *Journal of Statistical Software*, 40(3), 1-25. URL http://www.jstatsoft.org/v40/i03/

Karl W. Broman & Kara H. Woo (2018) Data Organization in Spreadsheets, The American Statistician, 72:1, 2-10, DOI: 10.1080/00031305.2017.1375989

Silge J, Robinson D (2016). “tidytext: Text Mining and Analysis Using Tidy Data Principles in R.” JOSS, 1(3). doi: 10.21105/joss.00037, http://dx.doi.org/10.21105/joss.00037

Wickham H, et al. (2019). “Welcome to the tidyverse.” Journal of Open Source Software, 4(43), 1686. doi: 10.21105/joss.01686

Wickham H (2007). “Reshaping Data with the reshape Package.” Journal of Statistical Software, 21(12), 1–20. http://www.jstatsoft.org/v21/i12/

Fellows, Ian (n.d.). “wordcloud: Word Clouds”, located at CRAN here: https://cran.r-project.org/web/packages/wordcloud/index.html