Assignment5_AbhishekPandit

Abhishek Pandit

26 November 2019

R Markdown

Load the platforms.csv file containing the 2016 Democratic and Republican party platforms. Note the 2X2 format, where each row is a document, with the party recorded as a separate feature. Also, load the individual party .txt files as a corpus.

```
library(tm)
```

```
## Warning: package 'tm' was built under R version 3.5.3
```

```
## Loading required package: NLP
```

```
## Warning: package 'NLP' was built under R version 3.5.2
```

```
library(grid)
library(wordcloud)
```

```
## Warning: package 'wordcloud' was built under R version 3.5.3
```

```
## Loading required package: RColorBrewer
```

```
## Warning: package 'RColorBrewer' was built under R version 3.5.2
```

```
library(wordcloud2)
```

```
## Warning: package 'wordcloud2' was built under R version 3.5.3
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.5.3
```

```
## -- Attaching packages ---------------------------------------------------------------------------- ti
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
## Warning: package 'tibble' was built under R version 3.5.3
```

```
## Warning: package 'tidyr' was built under R version 3.5.2
```

```
## Warning: package 'readr' was built under R version 3.5.2
```

```
## Warning: package 'purrr' was built under R version 3.5.2
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
## Warning: package 'stringr' was built under R version 3.5.2
```

```
## Warning: package 'forcats' was built under R version 3.5.2
```

```
## -- Conflicts ------------------------------------------------------------------------------------ tidyverse
## x ggplot2::annotate() masks NLP::annotate()
## x dplyr::filter()     masks stats::filter()
## x dplyr::lag()        masks stats::lag()
```

Now loading our data

```
docs <- VCorpus(DirSource("D://Dropbox//Programming//Temp_Data"))
summary(docs)
```

```
##               Length Class             Mode
## d16.txt       2      PlainTextDocument list
## platforms.csv 2      PlainTextDocument list
## r16.txt       2      PlainTextDocument list
```

```
#writeLines(as.character(docs[1])) # Check the corpus... did it work?
#writeLines(as.character(docs[1]))
```

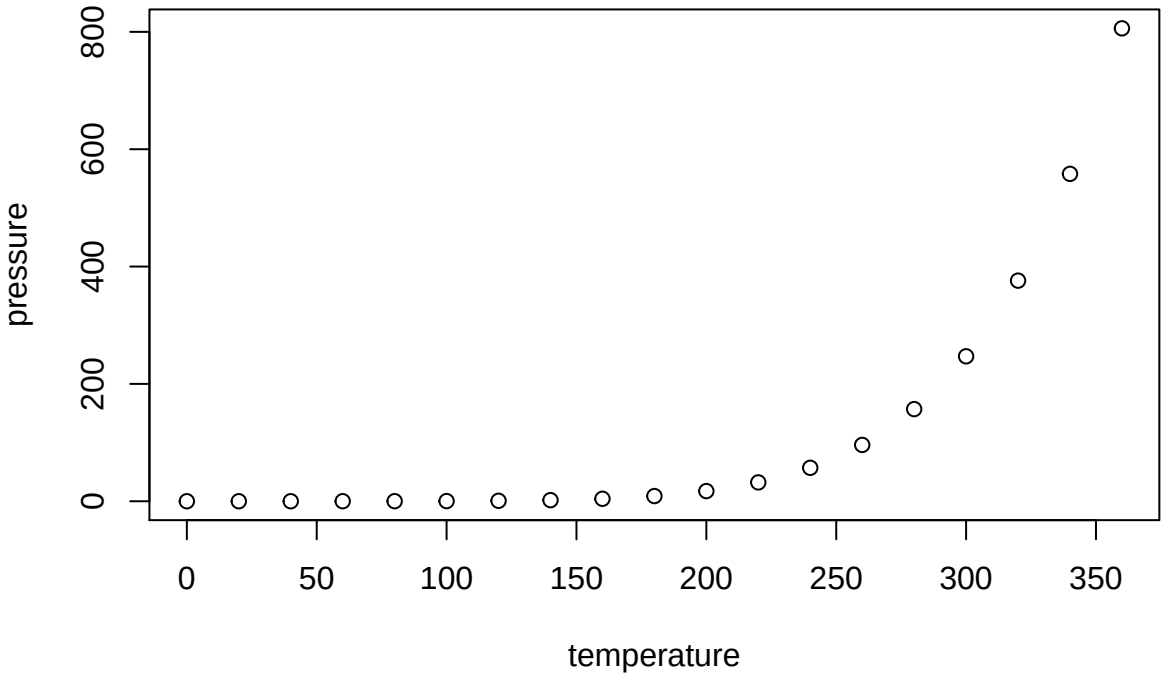
Including Plots

Create a document-term matrix and preprocess the platforms by the following criteria (at a minimum): Convert to lowercase Remove the stopwords Remove the numbers Remove all punctuation Remove the whitespace

```
# Start with punctuation
docs <- tm_map(docs, removePunctuation)
#writeLines(as.character(docs[1]))
```

SENTIMENT ANALYSIS

4. Use the "Bing" and "AFINN" dictionaries to calculate the sentiment of each cleaned party platform. Present the results however you'd like (e.g., visually and/or numerically).
5. Compare and discuss the sentiments of these platforms: which party tends to be more optimistic about the future? Does this comport with your perceptions of the parties?



TOPIC MODELS

6. With a general sense of sentiments of the party platforms (i.e., the tones related to how parties talk about their roles in the political future), now explore the topics they are highlighting in their platforms. This will give a sense of the key policy areas they're most interested in. Fit a topic model for each of the major parties (i.e. two topic models) using the latent Dirichlet allocation algorithm, initialized at `k = 5` topics as a start. Present the results however you'd like (e.g., visually and/or numerically).
7. Describe the general trends in topics that emerge from this stage. Are the parties focusing on similar or different topics, generally?
8. Fit 6 more topic models at the follow levels of `k` for each party: `5, 10, 25`. Present the results however you'd like (e.g., visually and/or numerically).
9. Calculate the perplexity of each model iteration and describe which technically fits best.
10. Building on the previous question, display a barplot of the `k = 10` model for each party, and offer some general inferences as to the main trends that emerge. Are there similar themes between the parties? Do you think `k = 10` likely picks up differences more efficiently? Why or why not?

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.