Group Project Part 2

Krishna Goel, Toby Han, Elijah Lipkin

October 3, 2018

Setup:

```
setwd("C://Users//kgoel//Desktop//Y3 Sem1//Stat 405//Overall Project")

library(dplyr)
library(stringi)
library(ggplot2)

word_frequencies <- read.csv("data.csv")
fixed_words <- mutate(word_frequencies, word = stri_sub(word, 3, -4))
head(fixed_words)
```

```
##     word total_upvotes
## 1     aa          3410
## 2    aah           586
## 3  aahed           586
## 4 aahing           586
## 5   aahs           586
## 6    aal            11
```

Mutating because the strings we got were misformatted and the API had kicked us out for too many requests. We have a bigger data set downloaded, but we just wanted to demonstrate we could apply dsplyr to the dataset for now. The full csv was too big for R to open. (This data set is a approxiamtly 1000 words startin in "a" that were processed in python to get the total upvotes for each word as that doesn't come by default from the API)

Processing intermediate data in dplyr

Part 1

Arranged data to see range of upvotes

```
arranged <- arrange(fixed_words, desc(total_upvotes))
head(arranged)
```

```
##       word total_upvotes
## 1    abort         15813
## 2  aborted         15813
## 3  aborter         15813
## 4 aborters         15813
## 5 aborting         15813
## 6 abortion         15813
```

Part 2

Got various subsets of the data based on upvote range, some metrics to see what the data looks like an how many rows it contains

```r
filtered_top <- filter(arranged, total_upvotes > 10000)
head(filtered_top)
```

```
##       word total_upvotes
## 1    abort         15813
## 2  aborted         15813
## 3  aborter         15813
## 4 aborters         15813
## 5 aborting         15813
## 6 abortion         15813
```

```r
nrow(filtered_top)
```

```
## [1] 15
```

```r
filtered_mid <- filter(arranged, total_upvotes > 6000, total_upvotes < 10000)
head(filtered_mid)
```

```
##             word total_upvotes
## 1        academe          9498
## 2       academes          9498
## 3       academic          9498
## 4     academical          9498
## 5  academicalism          9498
## 6 academicalisms          9498
```

```r
nrow(filtered_mid)
```

```
## [1] 11
```

```r
filtered_mid_low <- filter(arranged, total_upvotes > 2000, total_upvotes < 3000)
head(filtered_mid_low)
```

```
##        word total_upvotes
## 1   ableism          2677
## 2  ableisms          2677
## 3  aardvark          2570
## 4 aardvarks          2570
## 5  abusable          2179
## 6     abuse          2179
```

```r
nrow(filtered_mid_low)
```

```
## [1] 17
```

```r
filtered_low <- filter(arranged, total_upvotes > 500, total_upvotes < 1000)
head(filtered_low)
```

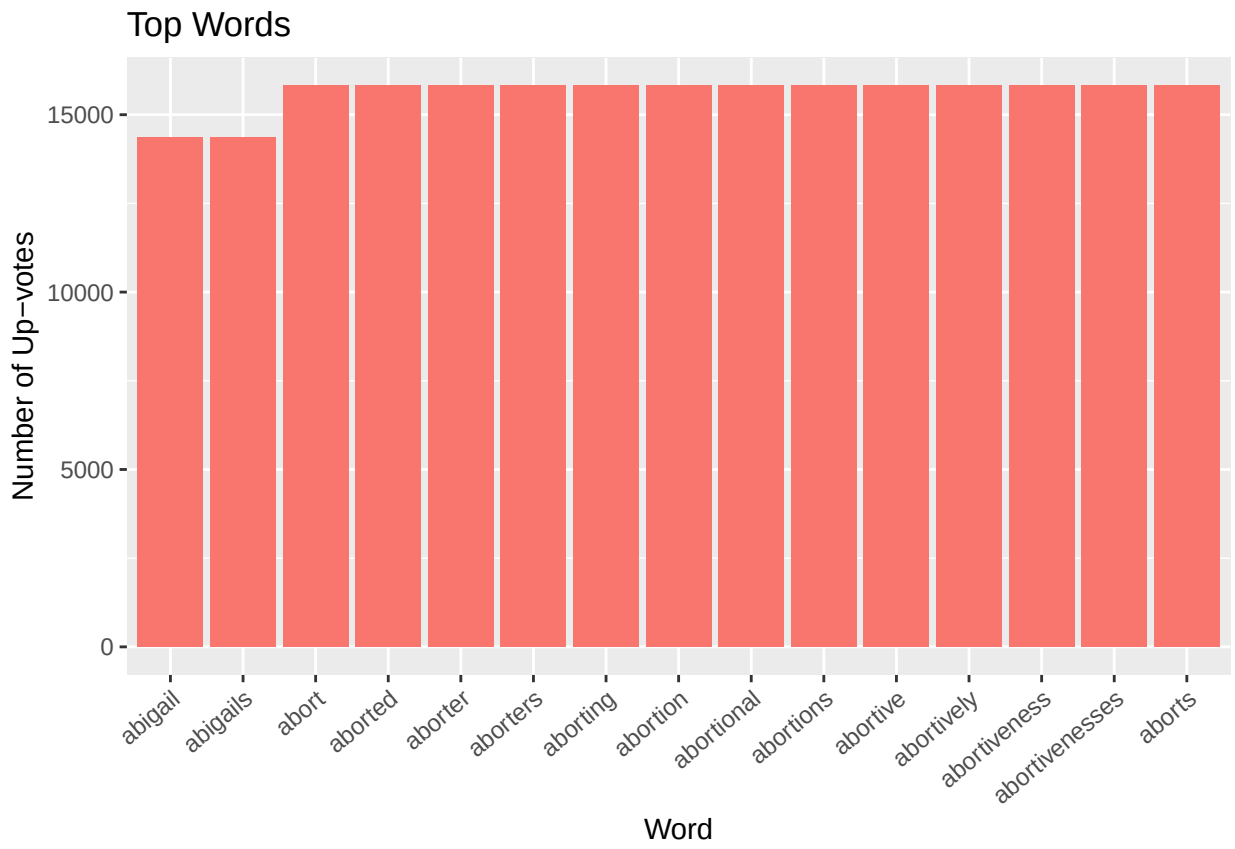
```
##           word total_upvotes
## 1        aargh           868
## 2        aarti           801
## 3       aartis           801
## 4  abracadabra           720
## 5 abracadabras           720
## 6     absolute           658
```

```r
nrow(filtered_low)
```

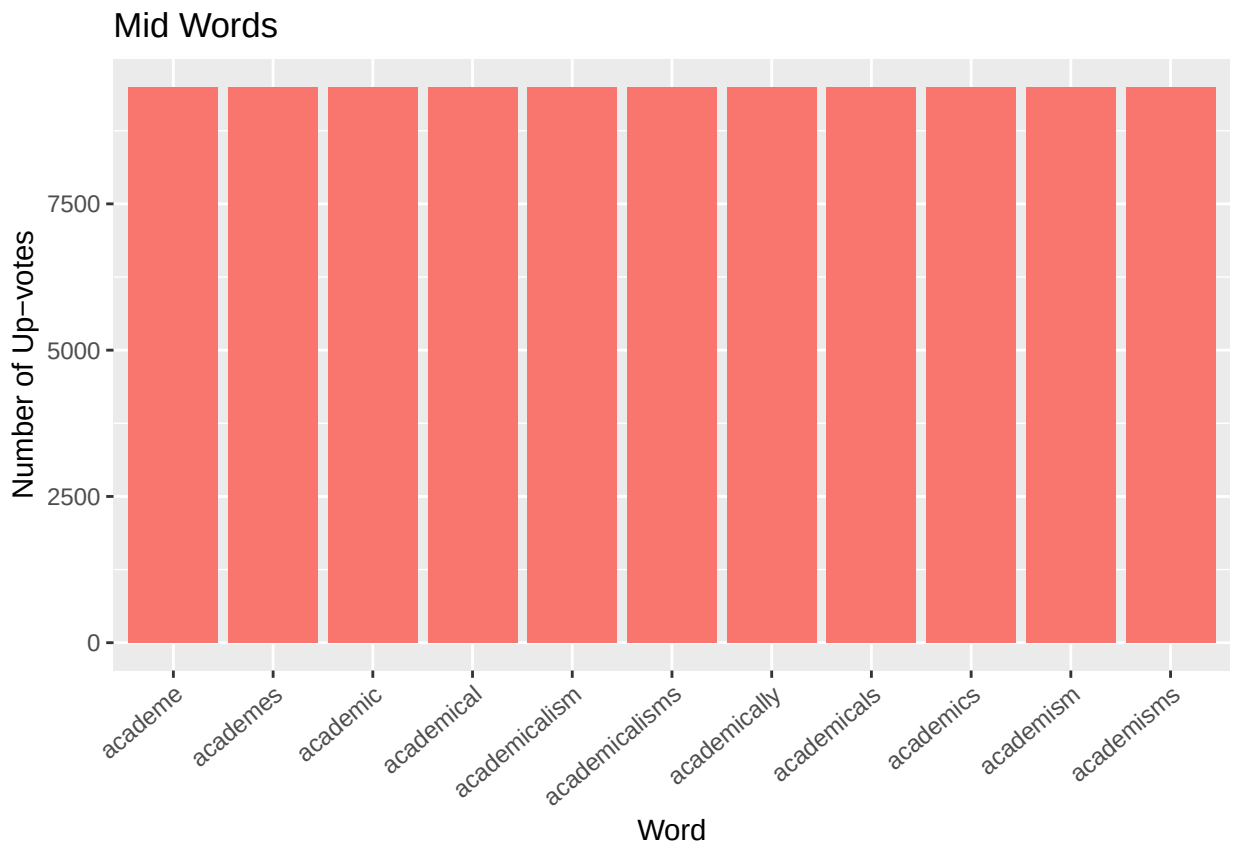
```
## [1] 54
```

Plotting the data to get an idea of what types of words appear

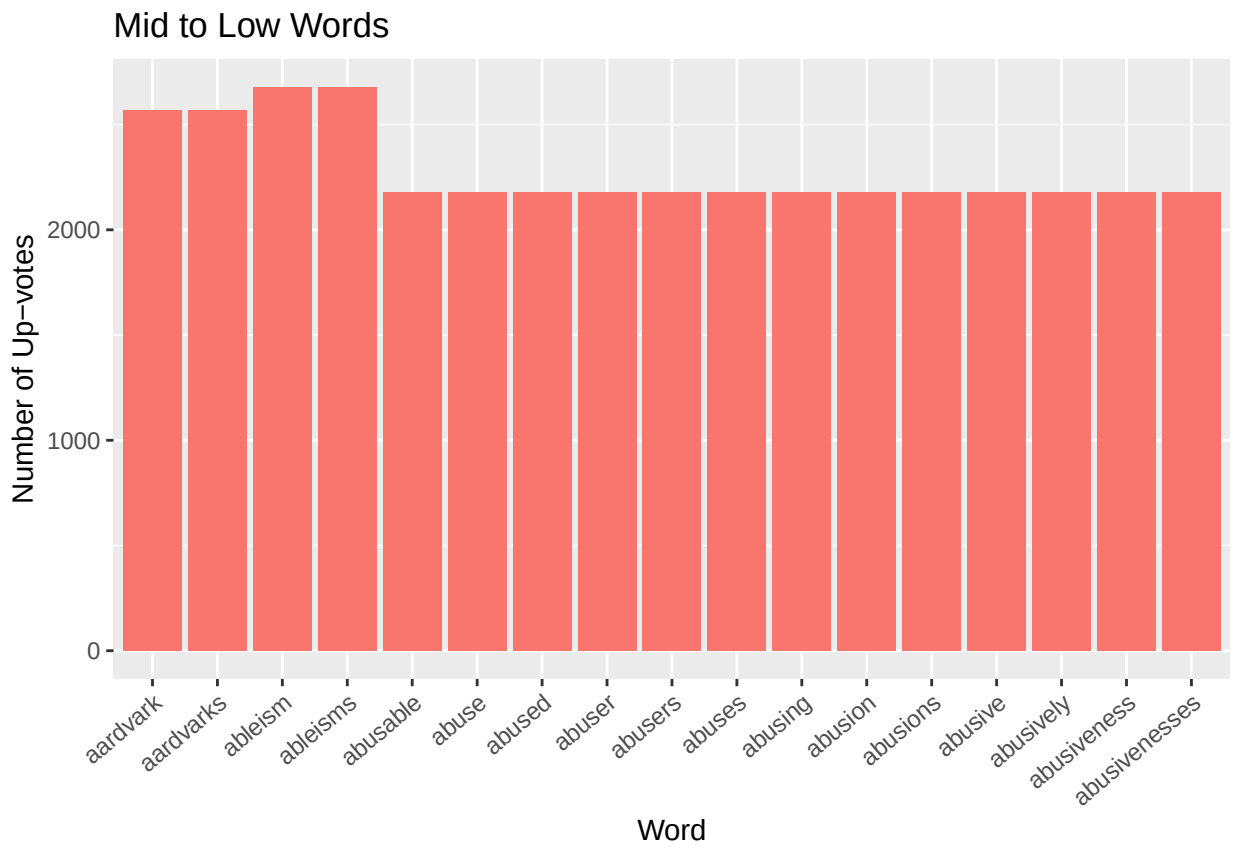
```
ggplot(data = filtered_top) +
  aes(factor(word), total_upvotes, fill = "#2e7f1e") +
  geom_col(position = "identity") +
  labs(x = "Word",
       y = "Number of Up-votes",
       title = "Top Words") +
  theme(axis.text.x = element_text(angle = 40, hjust = 1)) +
  theme(legend.position="none")
```



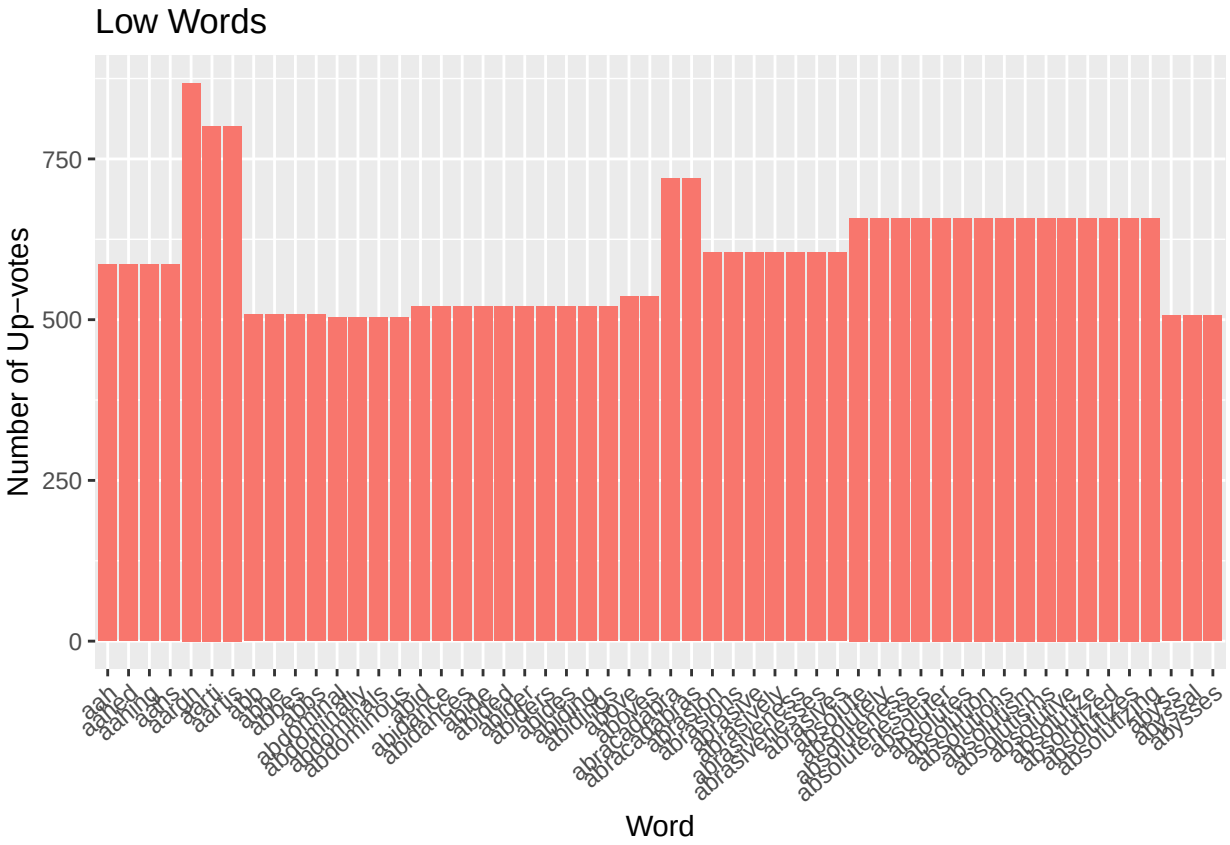

```
ggplot(data = filtered_mid) +
  aes(factor(word), total_upvotes, fill = "#1a2e89") +
  geom_col(position = "identity") +
  labs(x = "Word",
       y = "Number of Up-votes",
       title = "Mid Words") +
  theme(axis.text.x = element_text(angle = 40, hjust = 1)) +
  theme(legend.position="none")
```




```
ggplot(data = filtered_mid_low) +
  aes(factor(word), total_upvotes, fill = "#1a2e89") +
  geom_col(position = "identity") +
  labs(x = "Word",
       y = "Number of Up-votes",
       title = "Mid to Low Words") +
  theme(axis.text.x = element_text(angle = 40, hjust = 1)) +
  theme(legend.position="none")
```




```
ggplot(data = filtered_low) +
  aes(factor(word), total_upvotes, fill = "#1a2e89") +
  geom_col(position = "identity") +
  labs(x = "Word",
       y = "Number of Up-votes",
       title = "Low Words") +
  theme(axis.text.x = element_text(angle = 40, hjust = 1)) +
  theme(legend.position="none")
```


Low Words
Number of Up−votes
750
500
250
0
Word