Practice Sentiment analysis

Emily Maloney

February 10, 2019

Trying out some methods for sentiment analysis of subreddits.

Libraries

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(bigrquery)
library(tidyr)
library(sentimentr)
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

Data

just getting comment histories of two subreddits: conservative and The Donald

```
#set project name
project <- 'conversion-narratives'

#standardSQL
sql <- "SELECT
          author,
          created_utc,
          url,
          title,
          selftext,
          id,
          num_comments,
          ups,
          downs,
          score
        FROM
          `fh-bigquery.reddit_posts.201*`
        WHERE
          subreddit = 'The_Donald'
"

#get post data
df_reddit_TD <- query_exec(sql, project = project, use_legacy_sql = FALSE)
```

```
## Auto-refreshing stale OAuth token.
```

```
## 118.7 gigabytes processed
```

```
## Warning: Only first 10 pages of size 10000 retrieved. Use max_pages = Inf
## to retrieve all.
```

```
#standardSQL
sql <- "SELECT
          author,
          created_utc,
          url,
          title,
          selftext,
          id,
          num_comments,
          ups,
          downs,
          score
        FROM
          `fh-bigquery.reddit_posts.201*`
        WHERE
          subreddit = 'Conservative'
"

#get post data
df_reddit_cons <- query_exec(sql, project = project, use_legacy_sql = FALSE)
```

```
## 118.7 gigabytes processed
```

```
## Warning: Only first 10 pages of size 10000 retrieved. Use max_pages = Inf
## to retrieve all.
```

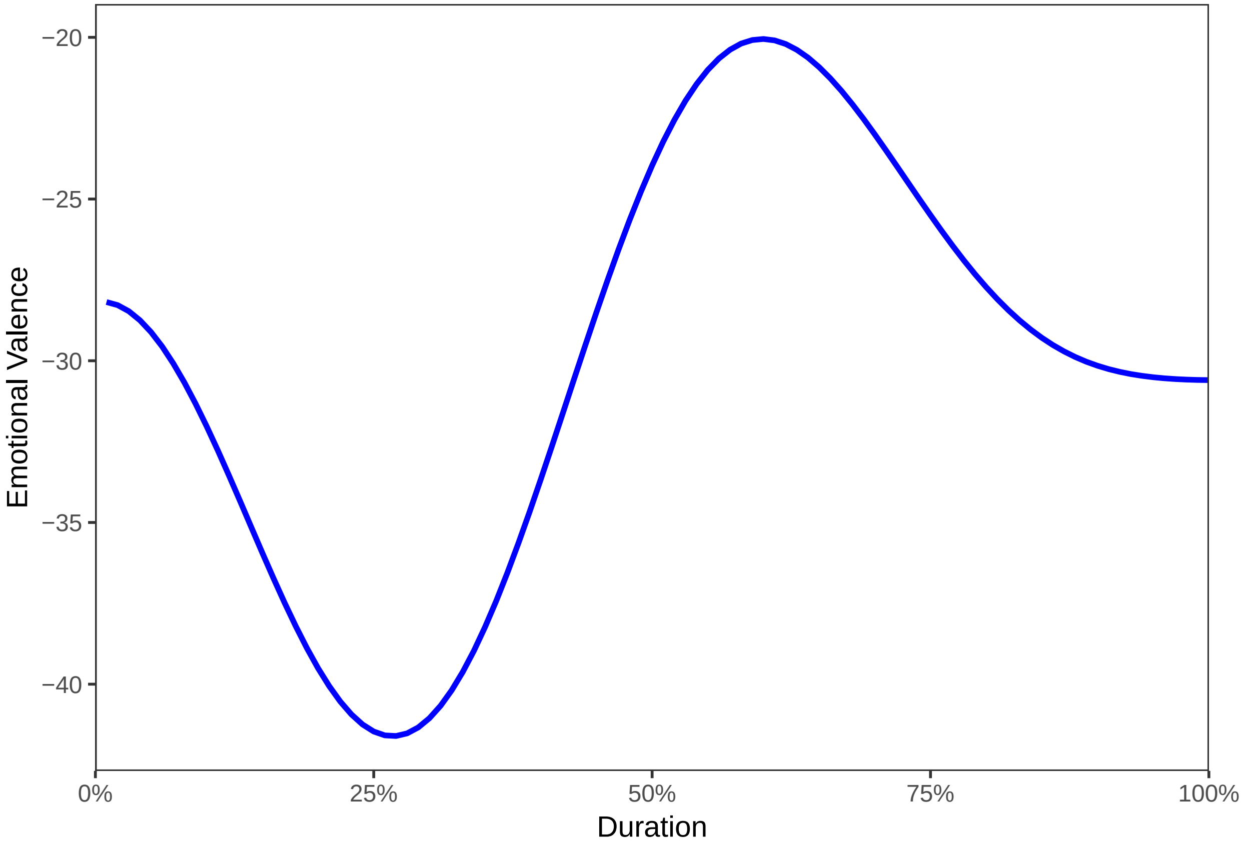
```
df_TD_sent <- df_reddit_TD %>%
              mutate(post_split = get_sentences(title)) %$%
              sentiment_by(post_split)

df_TD_sent %>% summarise(mean = mean(ave_sentiment))
```

```
##          mean
## 1 -0.02264612
```

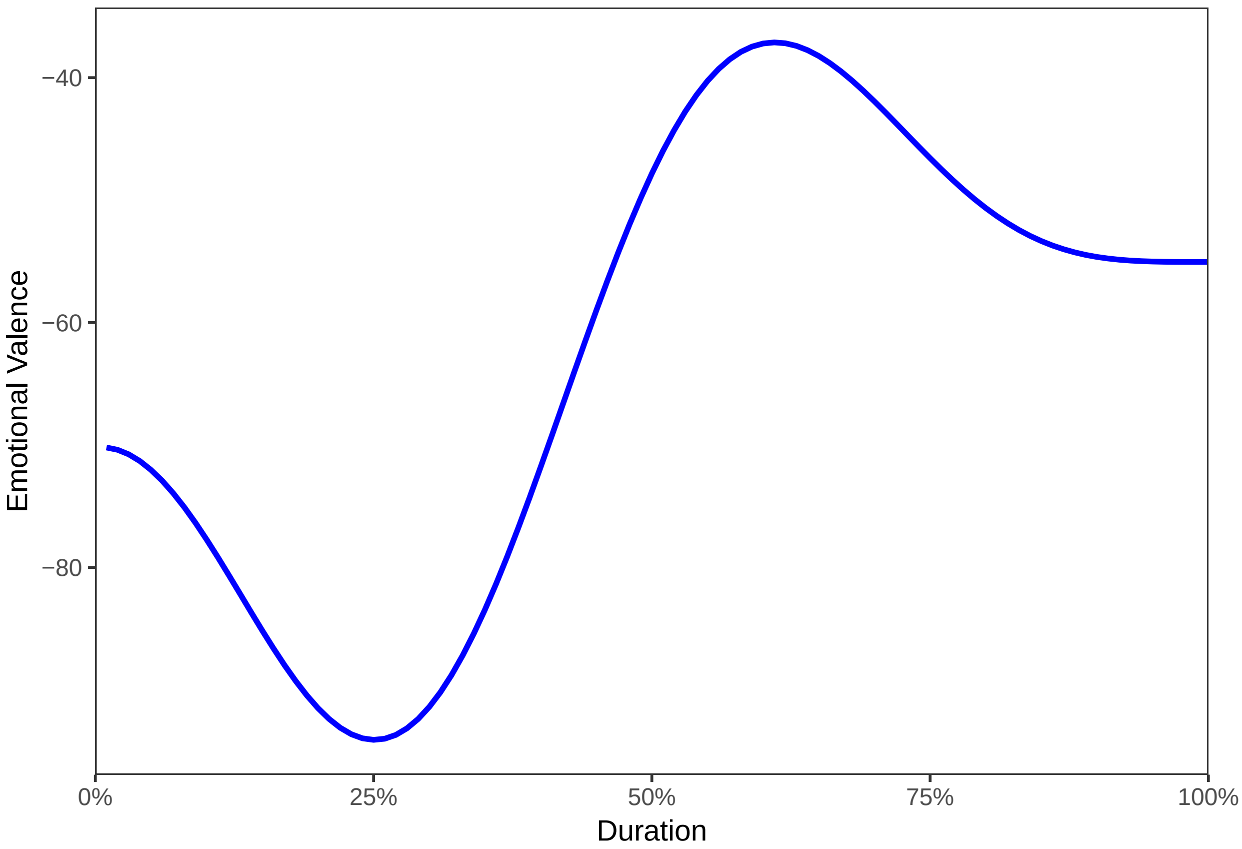
```
df_C_sent <- df_reddit_cons %>%
              mutate(post_split = get_sentences(title)) %$%
              sentiment_by(post_split)

df_C_sent %>% summarise(mean = mean(ave_sentiment))
```

```
##         mean
## 1 -0.0581569
```

```
plot(uncombine(df_TD_sent))
```



```
plot(uncombine(df_C_sent))
```


−40
−60
−80
Emotional Valence
0%
25%
50%
75%
100%
Duration