SENTIMENT ANALYSIS IN R: THE TIDY WAY

Analyzing TV news

Julia Silge
Data Scientist at Stack Overflow

Closed captioning text from TV news

climate_text

- station, the TV news station where the text is from
- show, the show on that station where the text was spoken
- show_date, the broadcast date of the spoken text
- text, the actual text spoken on TV

Texts available from the Internet Archive's TV News Archive

Sentiment analysis of TV news

```
> climate_text %>%
+     select(text)

# A tibble: 593 x 1
                                                text
                                               <chr>
 1 the interior positively oozes class raves car..
 2 corporations have withdrawn from the chamber...
 3 he says he was bumped by the greeter but cops..
 4 especially at at time now where the climate....
 5 lots more coming up quite simply here green....
 6 so they`re carrying a lot of water for john....
 7 let me ask you about something else that in....
 8 other important news we`re following including.
 9 let democrats be democrats craig crawford of...
10 you know there are real fights to have over....
# ... with 583 more rows
```

Sentiment analysis of TV news

```
> climate_text %>%
+     unnest_tokens(word, text)

# A tibble: 41,076 x 4
   station             show           show_date       word
     <chr>            <chr>              <dttm>      <chr>
 1   MSNBC Morning Meeting 2009-09-22 13:00:00        the
 2   MSNBC Morning Meeting 2009-09-22 13:00:00   interior
 3   MSNBC Morning Meeting 2009-09-22 13:00:00 positively
 4   MSNBC Morning Meeting 2009-09-22 13:00:00      oozes
 5   MSNBC Morning Meeting 2009-09-22 13:00:00      class
 6   MSNBC Morning Meeting 2009-09-22 13:00:00      raves
 7   MSNBC Morning Meeting 2009-09-22 13:00:00        car
 8   MSNBC Morning Meeting 2009-09-22 13:00:00   magazine
 9   MSNBC Morning Meeting 2009-09-22 13:00:00      slick
10   MSNBC Morning Meeting 2009-09-22 13:00:00        and
# ... with 41,066 more rows
```

SENTIMENT ANALYSIS IN R: THE TIDY WAY

Let's practice!

SENTIMENT ANALYSIS IN R: THE TIDY WAY

Comparing TV stations

Julia Silge
Data Scientist at Stack Overflow

Comparing TV stations

```
> climate_text %>%
+     count(station)

# A tibble: 3 x 2
   station     n
     <chr> <int>
1      CNN   148
2 FOX News   183
3    MSNBC   262
```

Finding totals for each sentiment

```
> tv_sentiment %>%
+     count(station, sentiment, station_total)

# A tibble: 30 x 4
   station    sentiment station_total     n
     <chr>        <chr>         <int> <int>
 1     CNN        anger         10713   187
 2     CNN anticipation         10713   152
 3     CNN      disgust         10713    89
 4     CNN         fear         10713   545
 5     CNN          joy         10713    97
 6     CNN     negative         10713   331
 7     CNN     positive         10713   522
 8     CNN      sadness         10713   139
 9     CNN     surprise         10713   127
10     CNN        trust         10713   368
# ... with 20 more rows
```

Finding proportions for each sentiment

- Define a new column with `mutate()`
- Filter for one sentiment
- Use `arrange()` to order the results

Exploring contributions by word

```
> tv_sentiment %>%
+     count(sentiment, word)

# A tibble: 2,019 x 3
   sentiment         word     n
       <chr>        <chr> <int>
 1     anger   aggressive     7
 2     anger     alienate     1
 3     anger        angry     2
 4     anger   annihilate     1
 5     anger annihilation     1
 6     anger        argue     6
 7     anger     argument    14
 8     anger      assault     1
 9     anger       attack     8
10     anger    attacking     1
# ... with 2,009 more rows
```

SENTIMENT ANALYSIS IN R: THE TIDY WAY

Let's practice!

SENTIMENT ANALYSIS IN R: THE TIDY WAY

Sentiment over time

Julia Silge
Data Scientist at Stack Overflow

Using the lubridate package

For handling dates and times, try the lubridate package

```
floor_date(show_date, unit = "6 months")
```

Using the lubridate package

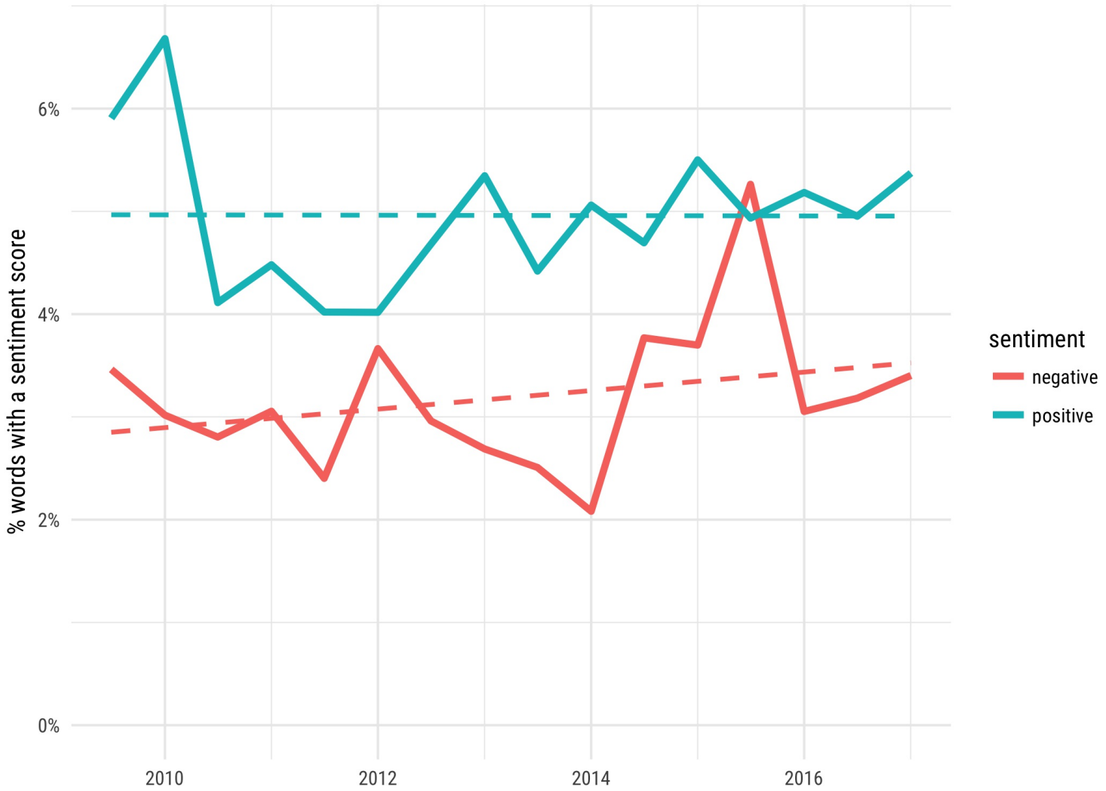
```
> library(lubridate)
>
> floor_date(as.Date("2016-09-27"), unit = "3 months")
[1] "2016-07-01"
```

Tidy date handling

```
> sentiment_by_time %>%
+     filter(sentiment %in% c("positive", "negative")) %>%
+     count(date, sentiment, total_words)

# A tibble: 32 x 4
         date sentiment total_words     n
       <dttm>     <chr>       <int> <int>
 1 2009-07-01  negative         491    17
 2 2009-07-01  positive         491    29
 3 2010-01-01  negative         464    14
 4 2010-01-01  positive         464    31
 5 2010-07-01  negative         535    15
 6 2010-07-01  positive         535    22
 7 2011-01-01  negative         982    30
 8 2011-01-01  positive         982    44
 9 2011-07-01  negative        3955    95
10 2011-07-01  positive        3955   159
# ... with 22 more rows
```

Sentiment over time



SENTIMENT ANALYSIS IN R: THE TIDY WAY

Let's practice!