

# Refresher on the text mining workflow

SENTIMENT ANALYSIS IN R



Ted Kwartler

Data Dude

# So far ...

- `polarity()`
  - Valence shifters
- `tidytext, dplyr, tidyr`
  - `bing, nrc, afinn`
- Visualizations

The screenshot shows the Airbnb search interface for Boston, MA, United States. At the top left is the Airbnb logo and a search bar with the location "Boston, MA, United States". To the right are buttons for "Browse" and "List Your Space". The main area features a map of Boston with red pins indicating rental locations across various neighborhoods like Somerville, Cambridge, Allston-Brighton, and South End. Below the map are three listing cards:

- (6) Guest House Harvard & MIT**  
Private room · 17 reviews · Cambridge  
\$85   
 A bedroom with two beds covered in red and white patterned bedding.
- Back Bay 1BR Apt / Heart of Boston!**  
Entire home/apt · 26 reviews · Back Bay, Boston  
\$239   
 An exterior view of a brick apartment building with a gold-colored entrance door.
- Comfy private queen bed in Brighton**  
Private room · 32 reviews · Allston-Brighton, Brighton  
\$83   
 An interior view of a bright room with a queen-size bed, a wooden dining table, and bookshelves.

At the bottom left are "Language and Currency" settings, and at the top right is a user profile for "Jon" with a "Help" button and a "List Your Space" button.

# The text mining workflow



# 6 defined steps

1. Define the problem & specific goals
2. Identify the text
3. Organize the text
4. Extract features
5. Analyze
6. Draw a conclusion/reach an insight

# Step 1: Define your problem

Tips:

- Be precise
- Avoid a "scope creep"
- Iterate and try new methods and/or subjectivity lexicons to ensure some consistency

# Step 2: ID your text

Tips:

- Find appropriate sources (e.g. searching Wikipedia for stock prices may make less sense than examining a stock forum)
- Follow the terms of service for a site, be mindful of web scraping
- Text sources affect the language used...become familiar with the source's tone and nuances

# Let's practice!

SENTIMENT ANALYSIS IN R

# Step 3: Organize (& clean) the text

SENTIMENT ANALYSIS IN R



Ted Kwartler  
Data Dude

# Get to it!

Initial goal: Use the `polarity()` function to define subsections of the text for examination.

```
pos_comments <- subset(bos_reviews$comments,  
                      bos_reviews$polarity > 0)  
neg_comments <- subset(bos_reviews$comments,  
                      bos_reviews$polarity < 0)  
  
pos_terms <- paste(pos_comments, collapse = " ")  
neg_terms <- paste(neg_comments, collapse = " ")
```

# More organization

Goal: Use the tidy rental reviews to create the tidy formatted polarity scoring.

```
library(tidytext)
library(dplyr)

tidy_reviews <- bos_reviews %>%
  unnest_tokens(word, comments)

tidy_reviews <- tidy_reviews %>%
  group_by(id) %>%
  mutate(original_word_order = seq_along(word))
```

# Tidy text polarity scoring

Recall the "bing" lexicon in `sentiments` has words categorized either as positive or negative.

```
library(tidytext)
library(tidyr)
library(dplyr)

bing <- sentiments %>%
  filter(lexicon == "bing")

pos_neg <- tidy_reviews %>%
  inner_join(bing) %>%
  count(sentiment) %>%
  spread(sentiment, n, fill = 0) %>%
  mutate(polarity = positive - negative)
```

# Let's practice!

SENTIMENT ANALYSIS IN R

# Revising the comparison cloud

SENTIMENT ANALYSIS IN R



Ted Kwartler  
Data Dude

# Author effort



# Comparisons



SOTU 2010  
values year took  
act bill families  
americans  
race new can just  
world future best  
SOTU 2011

# Revising the comparison cloud



revised SOTU 2010

office  
values  
billyear  
act  
families  
best future  
now race  
make want  
world years

revised SOTU 2011

# Always more analysis can be done!



# Let's practice!

SENTIMENT ANALYSIS IN R

# Step 6: Reach a conclusion

SENTIMENT ANALYSIS IN R



Ted Kwartler  
Data Dude

# Find the light bulb moments!



# Let's practice!

SENTIMENT ANALYSIS IN R

# Your turn!

## SENTIMENT ANALYSIS IN R



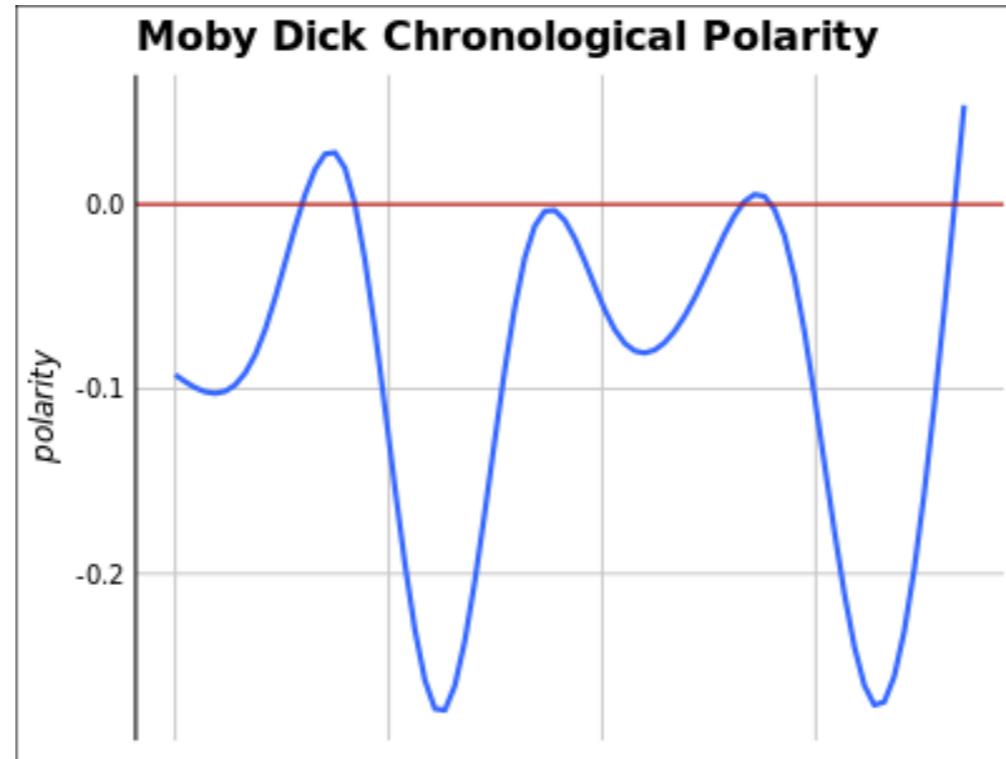
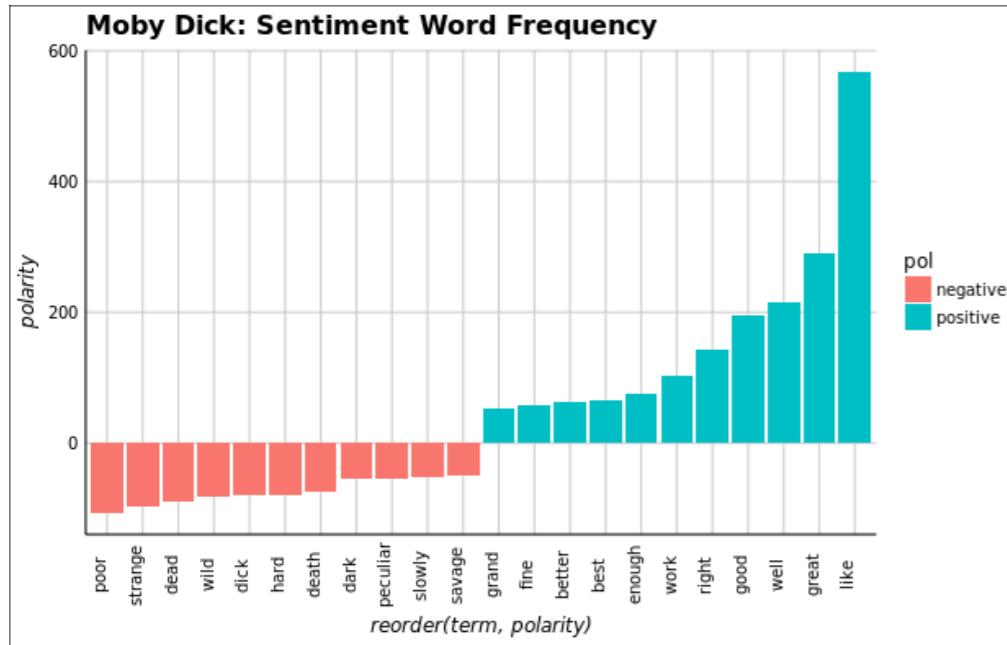
Ted Kwartler  
Data Dude

# Congratulations!!

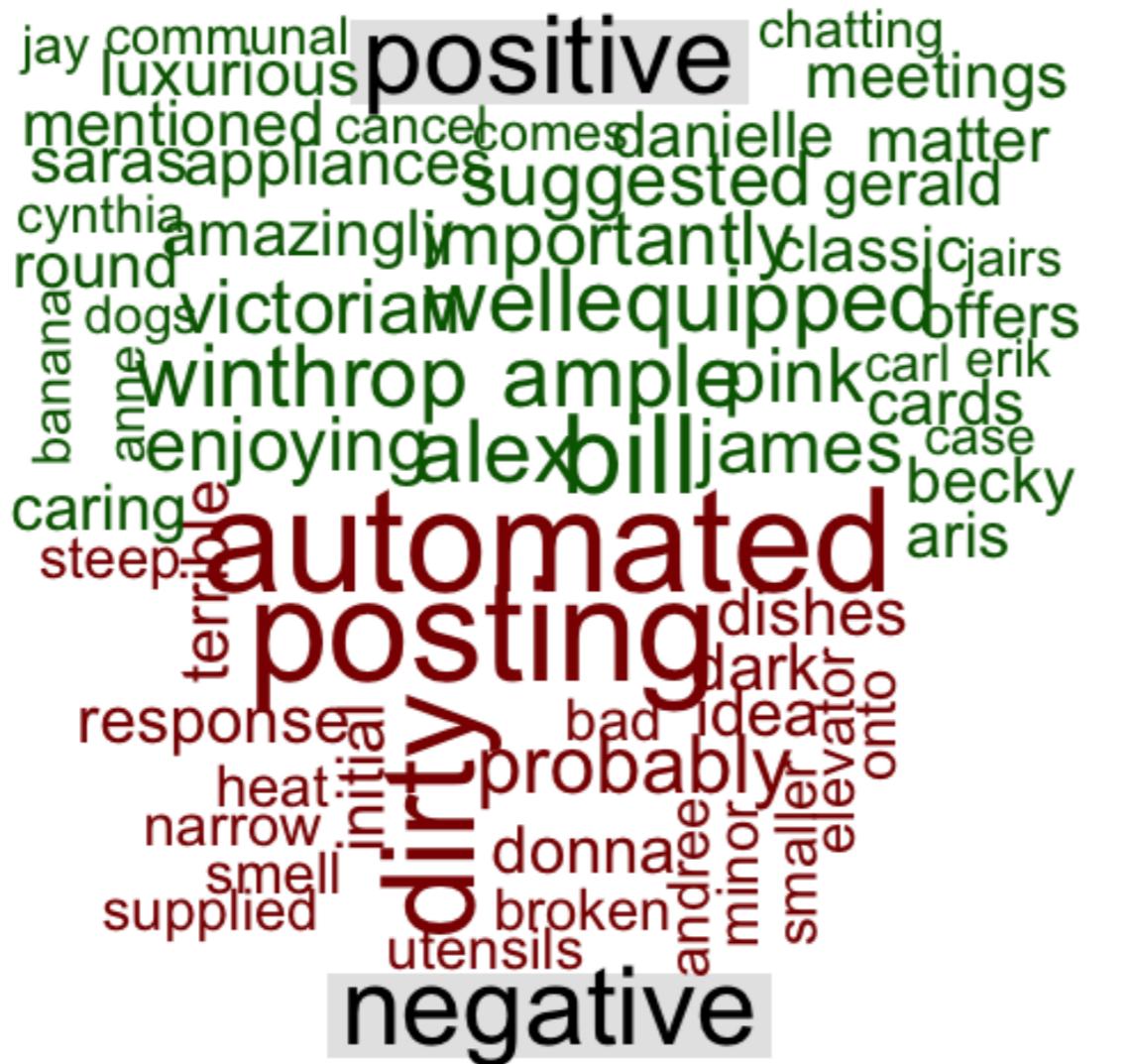
In this course you learned:

- `qdap`'s `polarity()` function
- `tidytext` data formats and `tidy` data functions
- `inner_join` with subjectivity lexicons

# Congratulations!!



# Congratulations!!



# Good luck!

SENTIMENT ANALYSIS IN R