



SENTIMENT ANALYSIS IN R

Refresher on the text mining workflow

Ted Kwartler
Data Dude

So far ...

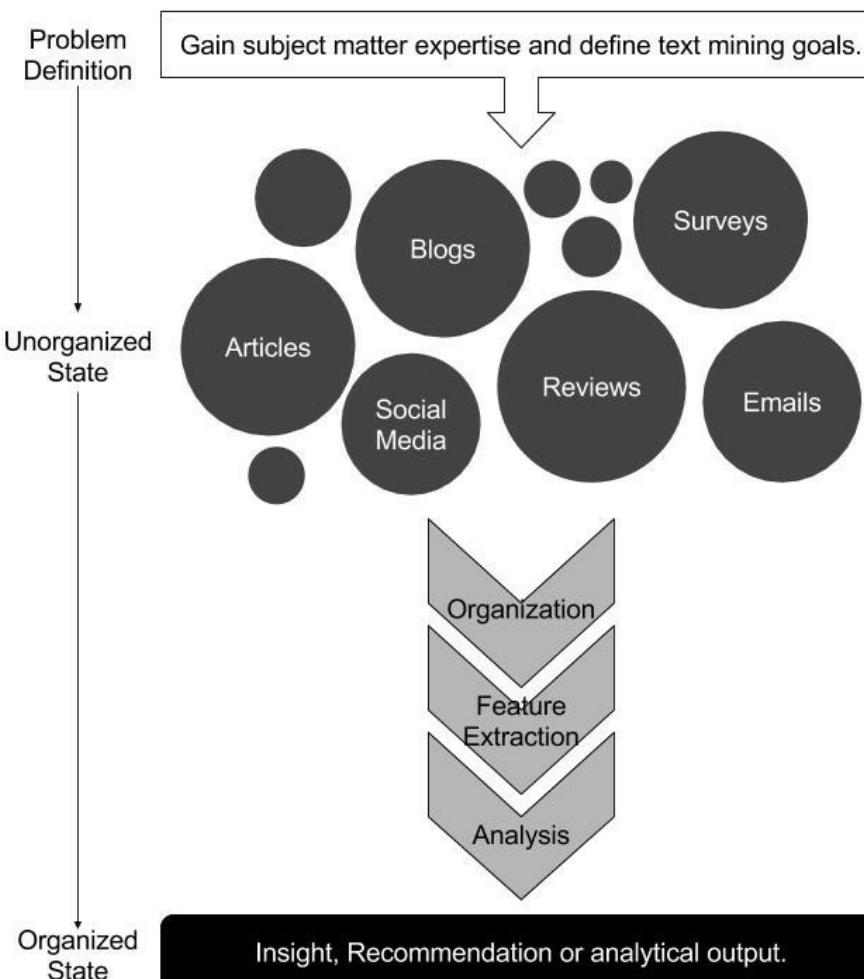
- polarity()
 - Valence shifters
- tidytext, dplyr, tidyr
 - bing, nrc, afinn
- Visualizations

The screenshot shows the Airbnb search results 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", "Filters", "Price", and a user profile for "Jon". On the far right is a green "List Your Space" button. The main area features a map of Boston with red pin icons marking various rental locations. Below the map are four detailed listing cards:

- [6] Guest House Harvard & MIT**
Private room · 17 reviews · Cambridge
\$85
- Back Bay 1BR Apt / Heart of Boston!**
Entire home/apt · 26 reviews · Back Bay, Boston
\$239
- Comfy private queen bed in Brighton**
Private room · 32 reviews · Allston-Brighton, Brighton
\$83
- large 2 bdrm South End by Copley Sq**
Entire home/apt · 3 reviews · South End, Boston
\$275

At the bottom left is a "Language and Currency" 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



SENTIMENT ANALYSIS IN R

Let's practice!



SENTIMENT ANALYSIS IN R

Step 3: Organize (& Clean) the Text

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)
```



SENTIMENT ANALYSIS IN R

Let's practice!

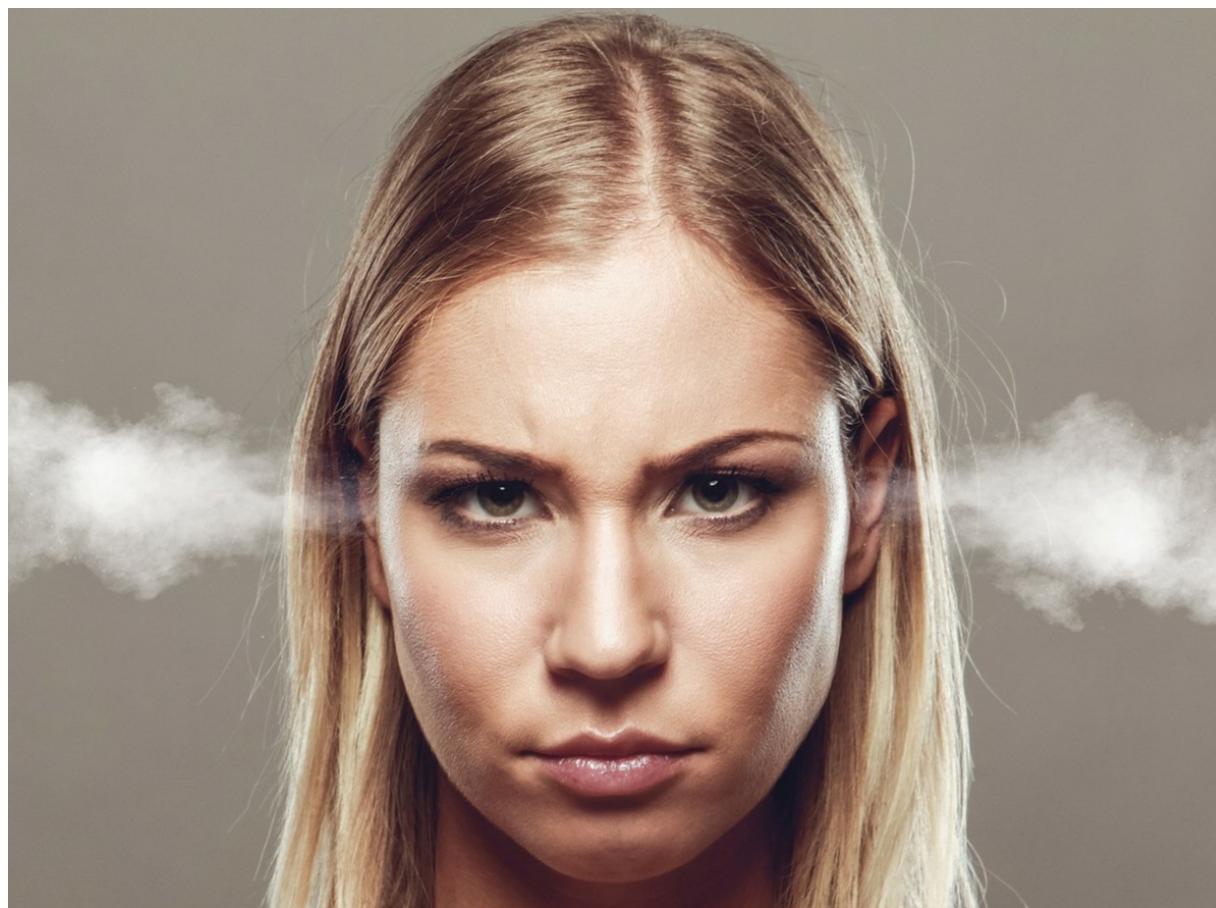


SENTIMENT ANALYSIS IN R

**Step 4: Feature
Extraction & Step 5:
Time for analysis...
almost there!**

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

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

revised SOTU 2011

Always more analysis can be done!





SENTIMENT ANALYSIS IN R

Let's practice!



SENTIMENT ANALYSIS IN R

Step 6: Reach a conclusion

Ted Kwartler
Data Dude

Find the light bulb moments!





SENTIMENT ANALYSIS IN R

Let's practice!



SENTIMENT ANALYSIS IN R

Your turn!

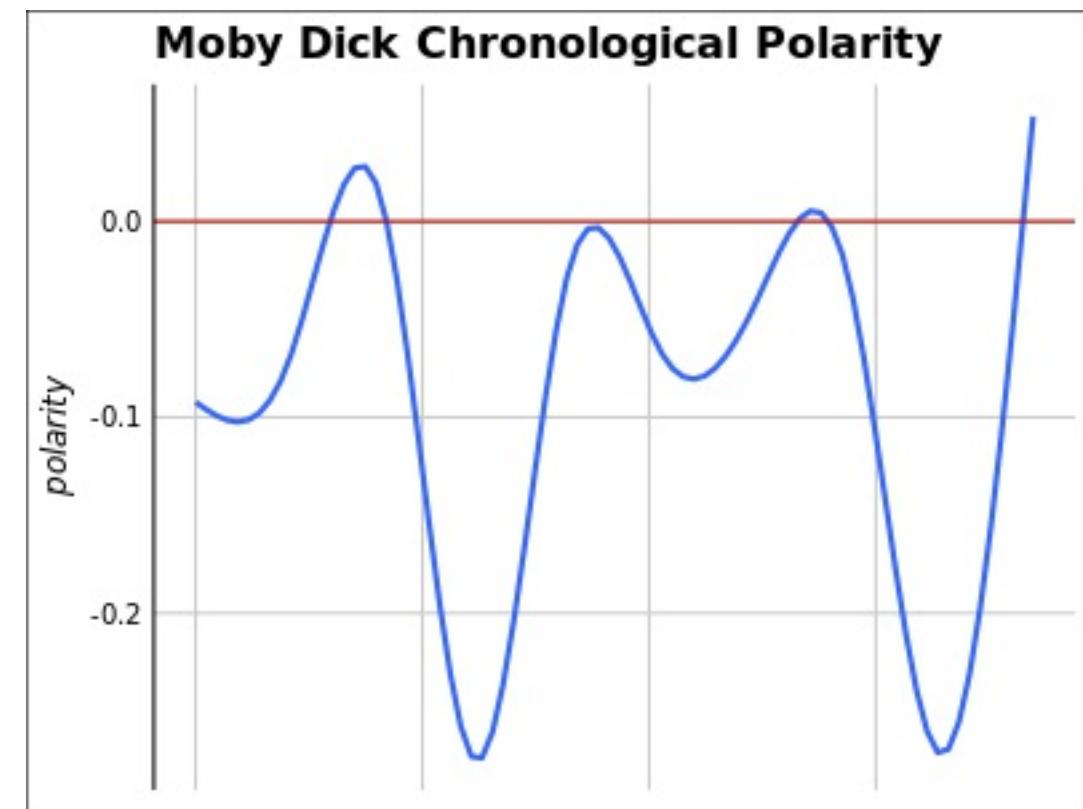
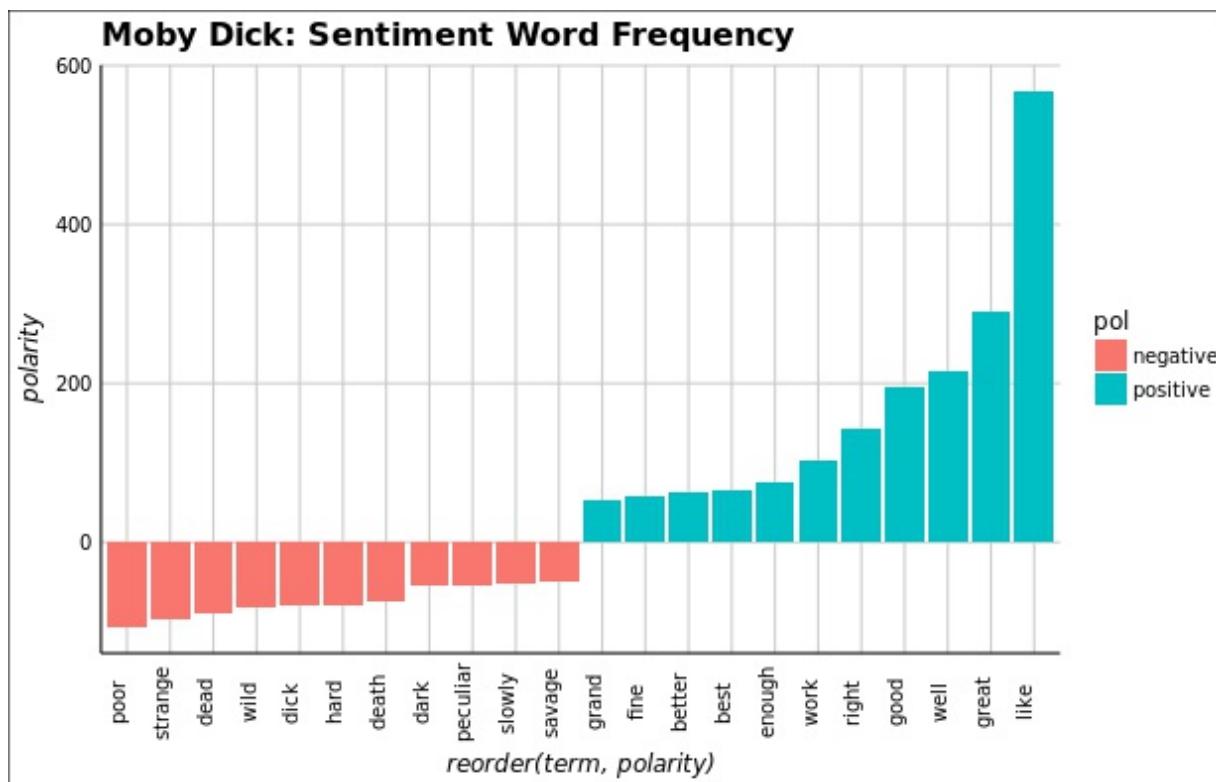
Ted Kwartler
Data Dude

Congratulations!!

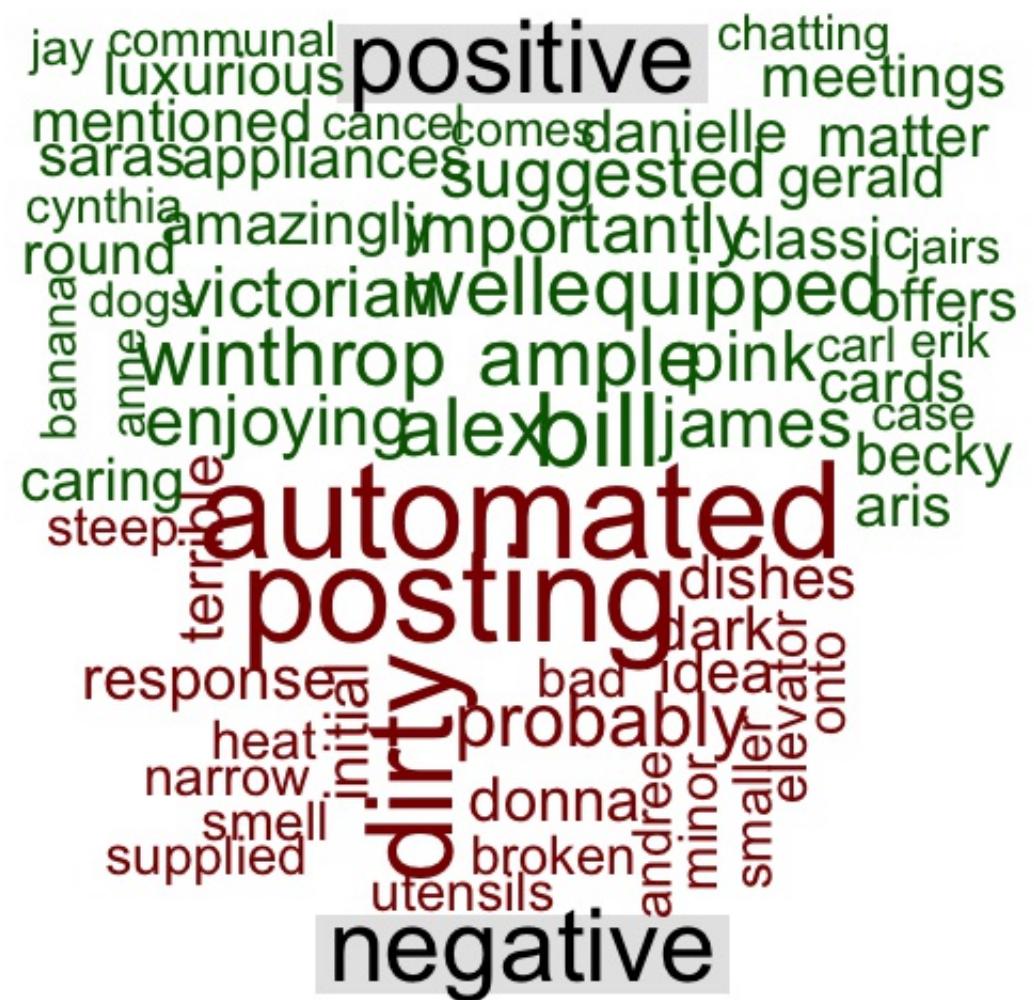
In this course you learned:

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

Congratulations!!



Congratulations!!





SENTIMENT ANALYSIS IN R

Good luck!