O'REILLY®
oscon

oscon.com
#oscon

YouR Feelings

How To Conduct A Sentiment Analysis Using R Programming

Pierre DeBois
July 19th, 2018

Overview

- Cultural and Business Trends That Brought Our Feelings Online
- Explain Sentiment Analysis
- 3 steps to develop a model based on Twitter data
  - Create Corpus and Invoke Libraries
  - Token The Text
  - Apply sentiment models
- Keep In Minds (KIMs)

Communication with media has evolved

1960s - Our devices (TV, radio) and media showed real time events that generated limited responses

Omar Salha
@o_salha

Over 600 people attended RTP's #OpenIftar in Portland, USA. Another strong message of unity in the face of fear #Muslims4Portland

5:24 AM - May 28,

234 103 p

#5 Oh Don't Mind Me. I'm Just Pipetting While Being #distractinglysexy

Meg Massa
@MegMassa

Oh don't mind me. I'm just pipetting while being #distractinglysexy #TimHunt #WomenInSTEM #womeninscience

4:33 PM - 11 Jun 2015

226 329

David Sheen @davidsheen · May 3

1000's of Jewish Africans & their allies now occupying Tel Aviv highway in #BlackLivesMatter protest (via @Yomgashum)

291 128

Barney Dellar @branaby · Apr 20

Wow. I guess it's time we all stopped using @eventbrite. They claim the right to attend your event, film it, and own the copyright. eventbrite.com/support/artic

166 2.2K 2.1K

Eventbrite
@eventbrite

Following

Replying to @branaby

Hi Barney - the terms were designed to help create promotional content with our creators, but the language we used was broader than necessary. We have not recorded any footage at events and have now removed the clause entirely. Apologies for any concern this caused.

5:00 PM - 22 Apr 2018

2018 - We research real-time events with our devices (smartphones) and media (social) for multichannel widespread responses

The Clapback Age

- **Confluence of our media interactions with brands, institutions, and other people creates a mirror of what we feel in a moment**
- Our online conversations reflect real world influences…
- The spark of those conversations has scaled with nuanced emotions and expressions…
- The technology for examining those conversations are beginning with statistical prowess

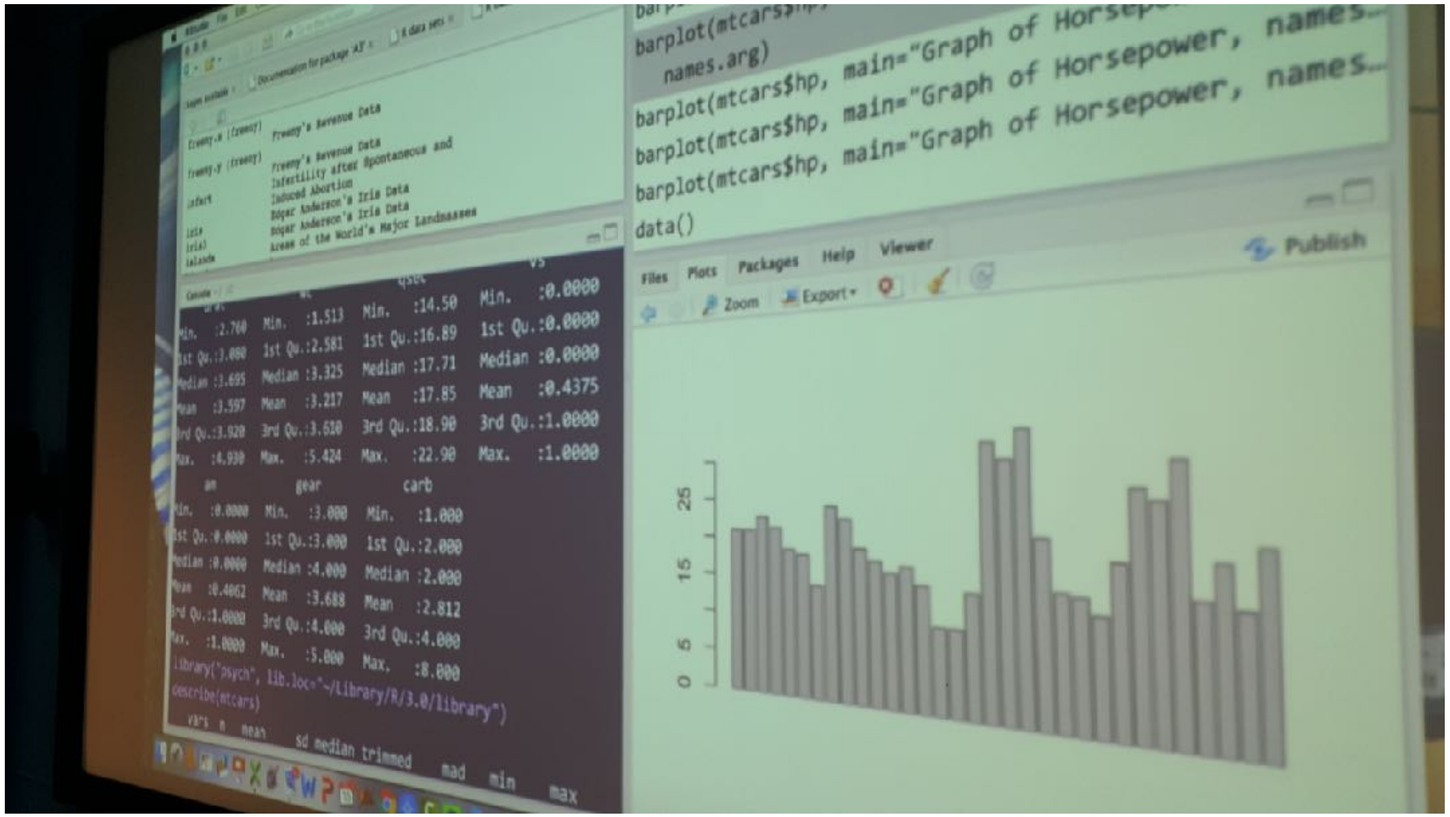
Look for Digital Behaviors Online To Develop An Idea

- US Adults spend 5.9 hrs/day on digital media (3.3 - mobile) - drives mobile payment & eCommerce activity*
- Ethical expectations from brands influences customer purchase decisions**
- People seek news online, generate conversations
  - Pew survey shows 50% now seek info online; 7% difference from television vs. 19% difference in early 2016***
  - Twitter leads Facebook in the percentage of users who look for news (74% vs 68%)***
- African-American, Hispanics demographic trends online are also visible due to smartphone access***

* source: 2018 Internet Trends Report, Mary Meeker, Partner - Kleiner Perkins Caufield & Byers, May 30th
** source: eMarketer 2015
*** source: Pew Institute

Sentiment Analysis / R Programming

- Natural Language Processing techniques that classifies text in a document (Corpus)
- To analyze, the corpus is reduced into a token - a “bag of words”
- High interest in using R and Python to create statistical models
  - R was developed for statistics modeling and analysis
  - Attracts data scientists with skills and insights from other industries



1. Start with A Corpus and Libraries

- Invoke libraries (packages) - programs that contain functions
  - Search for packages at cran.r-project.org or search within R-Studio (Files-Plot-Package Pane)
  - Each library has a document to explain functions and parameters
  - Some libraries connect to databases or API
- Put a collection of text in a data frame - a data table object.

```
#Sentiment_Analysis_IHOP
#
#Thursday July 19th 2018
#
#Call libraries - twitteR (requires "ROAuth" for access
#Twitter account and "devtools").
#Corpus steps based on Chapter 10 of the book
#R and Data Mining: Examples and Case Studies by Yanchang Zhao
#and tidytext sentiment, developed by Julia Silge and David Robinson
#
library(twitteR)
library(devtools)
library(ROAuth)
```

Package 'twitteR'

August 29, 2016

Title R Based Twitter Client

Description Provides an interface to the Twitter web API.

Version 1.1.9

Author Jeff Gentry <geoffjentry@gmail.com>

Maintainer Jeff Gentry <geoffjentry@gmail.com>

Depends R (>= 2.12.0)

Imports methods, bit64, rjson, DBI (>= 0.3.1), httr (>= 1.0.0)

Suggests RSQLite, RMySQL

License Artistic-2.0

LazyData yes

URL http://lists.hexdump.org/listinfo.cgi/twitter-users-hexdump.org

Collate allGenerics.R base.R account.R statuses.R users.R trends.R
s4methods.R convert.R dm.R cauth.R ccmm.R followers.R search.R

Why And How To Use Twitter As A Corpus

- People post frequently and in real time - statistical opportunity
- Public acceptance for tweeting an immediate thought and attracting response
  - Get 4 API code from apps.twitter.com (consumer key, consumer secret key, access token, access secret token)
  - Download and invoke TwitteR library
  - Use setup_TwitterOAuth function from TwitteR library to access Twitter parameters
  - Use searchTwitter function to return tweets containing keyword or hashtag

```
11 library(twitteR)
12 library(devtools)
13 library(ROAuth)
14 #call Twitter with OAuth via ROAuth
15 #obtain keys from dev.twitter.com - a Twitter account is required
16 setup_twitter_oauth("SSAEGOWJ2OI5LT7tDUPjeP96v","bpyHAXEQu943fHuqHEKz1XsDh
17 #call a Timeline if you want to see an account and verify.
18 userTimeline("IHOP")
19 #create an object and call searchTwitter to see what is assocaited with a
20 #list to be used for the sentiment analysis.
21 IHOBb4 <- searchTwitter("IHOB", since = '2018-05-11')
22 summary(IHOBb4)
```

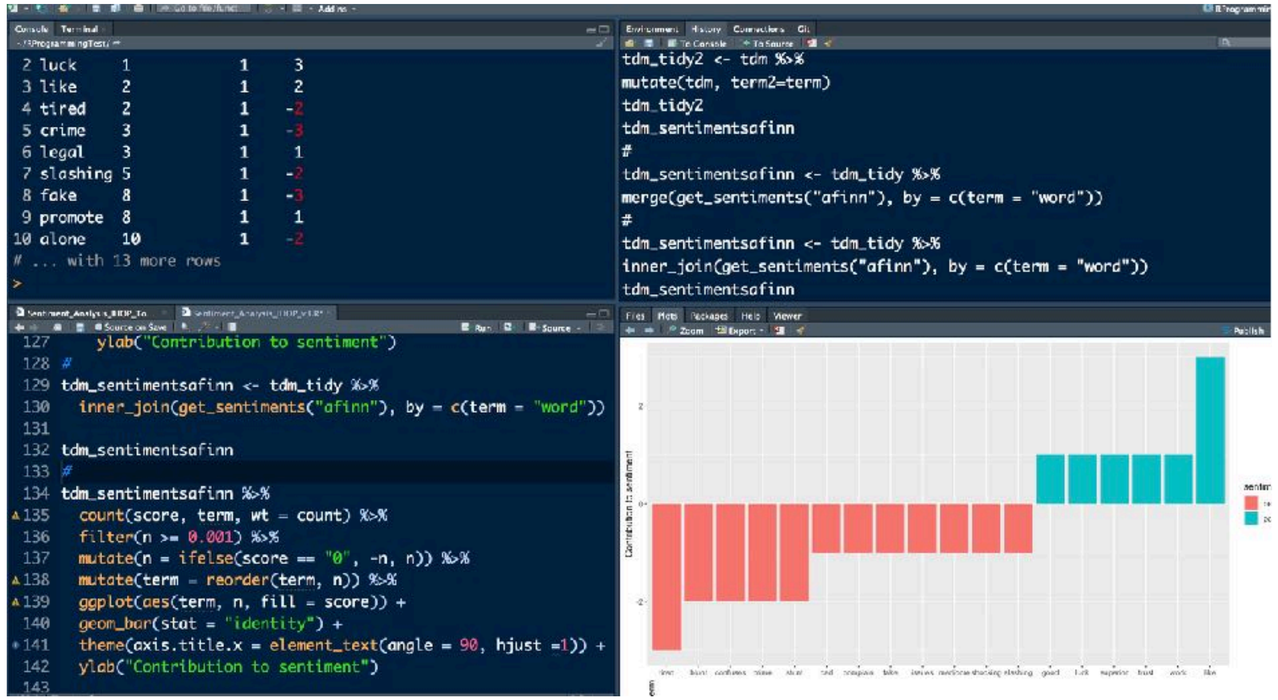
2. Token Your Text

- Tokenizing - The reduction of a corpus into units
- Remove punctuation, special characters, and capital letters
- Use library tm to change data frame into a corpus
- Apply functions for stopwords - words that repeats in an already expected manner and really don’t advance a narrative
  - prepositions
  - pronouns
- use tm_map at each step to token the corpus

```
#from the corpus
myStopwords <- c(setdiff(stopwords('english'), c("r", "big")),
                 "the", "a", "to")
myCorpus <- tm_map(myCorpus, removeWords, myStopwords)
#the following line is for removing white space
myCorpus <- tm_map(myCorpus, stripWhitespace)
#
#remove punctuation
myCorpus <- tm_map(myCorpus, removePunctuation)
#remove numbers
myCorpus <- tm_map(myCorpus, removeNumbers)
#
```

3. Apply Statistical Sentiment, Then Visuals

- Objective - Visualize which words match a lexicon or how frequently it appears
- Basic lexicons via get_sentiment function
  - AFINN - assigns words with a score between -5 to 5
  - Bing - assigns positive or negative
  - NRC - categorizes words as yes or no for several sentiments (positive, negative, anger anticipation, fear, joy, sadness, surprise, and trust)
- Bar chart (lexicons)
- Histogram (word frequency)
- Wordcloud



Topic Modeling

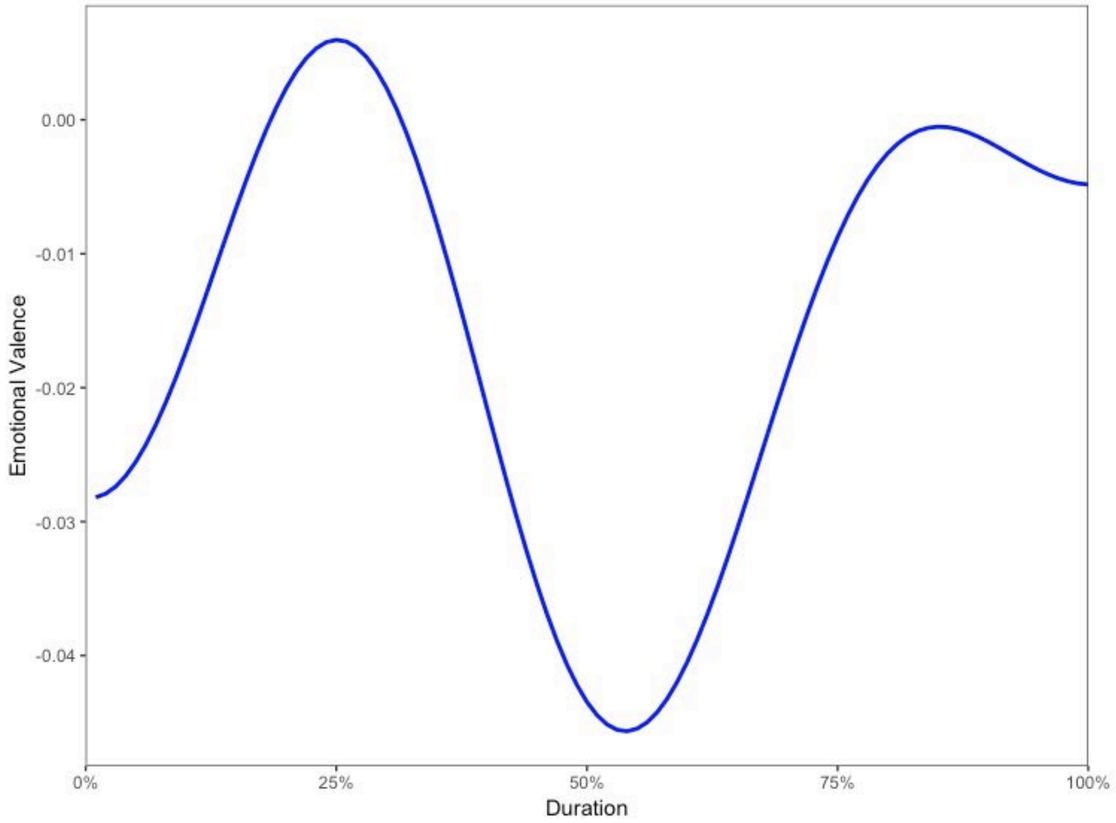
- Examine multiple word or phrase association in multiple documents
- Uses Term Document Matrix - table with terms in a row, documents in columns (library tm required)
- Metric: tf-idf (Term Frequency-Inverse Document Frequency) - weight to determine the importance of a word to a given document
- tidytext includes a bind_tf_idf function - calculates and bind the term frequency and inverse document frequency of a tidy text dataset

$$tf\,idf = \left(\frac{x}{y} \right) \left(\log \frac{N_1}{N_2} \right)$$

x = number of times a term appears
y = number of terms in a given document
N_1 = number of documents
N_2 = number of documents containing the term x

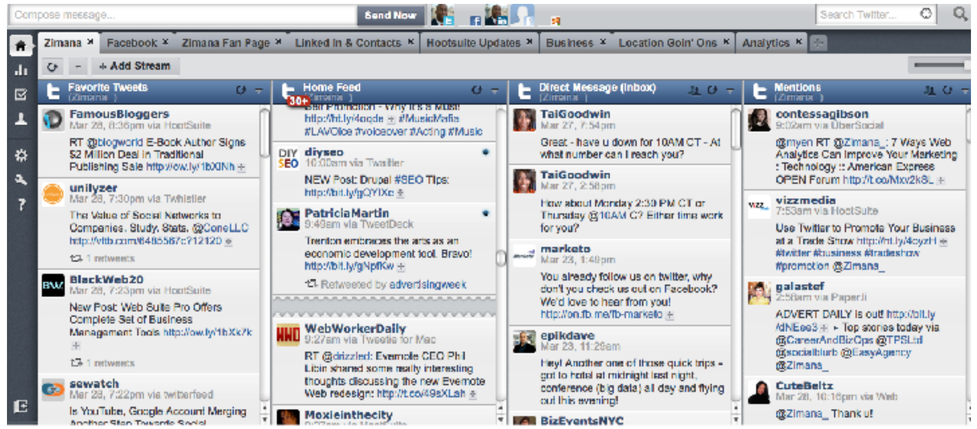
Sentimentr

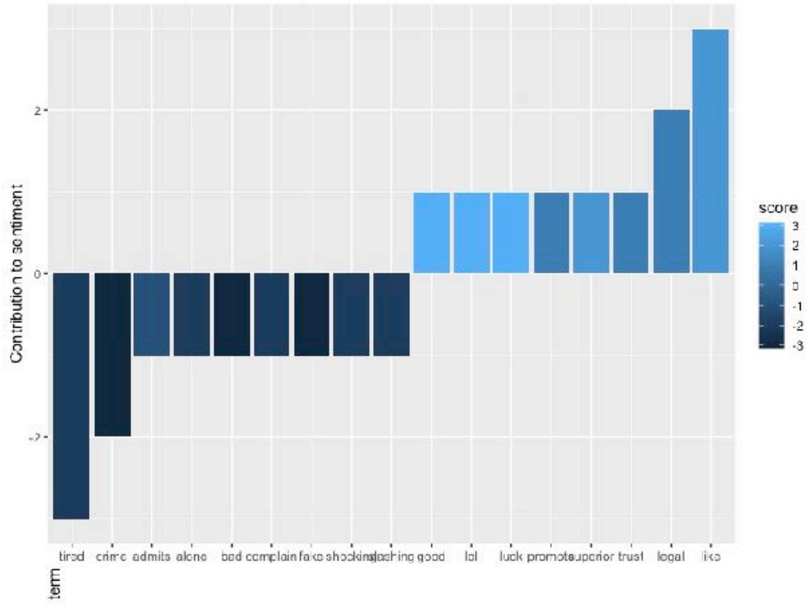
- Different sentiment R programming library (Tyler Rinker)
- Analyzes a word set within a corpus rather than singular words
- get_sentences - splits text into sentences
- sentiment_by() - outputs a polarity score; Can plot by duration
- Includes practice data (presidential_debates_2012, hotel_reviews dataset 2011, trip advisor, new york times articles, canon_reviews)



Keep In Minds (KIMs)

- Keep a sensibility of the timeline when examining social media data
- Monitor a Hootsuite or Tweetdeck channel for conversations around a hashtag or word
- Measuring sentiment on an influencer stream can be a hit or miss
- Recognize data restrictions with APIs
- Recognize when data is being combined that leads to Personal Identifiable Information
- Be ready for social data to continue growing while providing continual sentiment lessons for study





To Summarize Your Steps In Sentiment Analysis

- Review Digital Trends - Learn What Are People Doing and Imagine Your Data
- Start with A Corpus (and Libraries) in R Programmable
- Tokenize (Remove punctuation, adjust stopwords)
- Apply Statistical Sentiment (lexicon) and Visualization

Thank You!

- Twitter: @zimanaanalytics
- LinkedIn: Pierre DeBois
- Facebook Pages: /ZimanaAnalytics and /pierredeboisbiz
- code available at https://github.com/zimana/OSCON

Appendix

Resources

- R programming - 3.5 latest version cran.r-project.org
  - Updating R (linked in post by) https://www.linkedin.com/pulse/3-methods-update-r-rstudio-windows-mac-woratana-ngarmtrakulchol/
  - Use UpdateR library (Mac - required devtools library) or installr (Windows)
- R-Studio (IDE for running R programming)
- Libraries
  - tm
  - tidytext (contains lexicons AFINN, bing, NRC lexicons)
  - twitteR (there is also an alternative library Rtwitter)
  - ROAuth (for connecting R to an OAuth)
  - ggplot (visualization)
  - dplyr (for joining data frames, tables)

Resources

- Libraries (continued)
  - syuzhet package (contains NRC lexicon)
  - devtools
  - wordcloud (optional)
- A list of Data joins (http://stat545.com/bit001_dplyr-cheatsheet.html#full_joinsuperheroes-publishers)
- Optional: Twitter search engine (Socialbearing) https://socialbearing.com/ for comparing results in a data range, although range is limited in this application
- Term Document Matrix - Julia Silge and Davide Robinson (https://cran.r-project.org/web/packages/tidytext/vignettes/tidying_casting.html)
- tf-idf basics http://www.tfidf.com/

Tidy Text Resources

- Libraries
  - tidyverse
  - tidytext - Gabriela De Queiroz, Julia Silge and David Robinson
- Book: Text Mining With R - Julia Silge and David Robinson (O'Reilly)
- Tidy Text principles (https://cran.r-project.org/web/packages/tidytext/readme/README.html)
- Book: R and Data Mining: Examples and Case Studies by Yanchang Zhao (http://www2.rdatamining.com/uploads/5/7/1/3/57136767/rdatamining-book.pdf)

Images Sources

- Reporter at Vietnam War - Television Museum
- Civil Rights Meme - Southern Poverty Law Center
- Tweets - Twitter via @zimanaanalytics
- Special Thanks to Mendy Butler of Mendy Butler Virtual Business Support for background assistance with verifying Twitter resources online

Other Useful Libraries

- tm - text mining
- SnowballC - stemming (reducing words to a common stem)