DIGILAB WORKSHOP SERIES:
TEXT ANALYSIS 101

SOCIAL MEDIA APPLICATIONS

KATIE IRELAND KUIPER
22 APRIL 2021

UNIVERSITY OF GEORGIA
IN PREPARATION

I JUST LOVE DATA.

DATA'S MY FAVORITE.
APPLICATIONS OF SOCIAL MEDIA
METHODS: CORPUS ANALYSIS

- Gather data
- Organize metadata & dataset(s)
- Annotate and format
- Analyze
• **tidytext**: helpful for data formatting and visualization; works well with other packages in the Tidyverse (Silge & Robinson 2016)

• **quanteda**: incredibly useful package; includes preprocessing abilities, dtm function, as well as statistical analyses options like document classification and topic modeling

• **tokenizers**: useful options for text analytics, including tokenization and stemming
  
  • **ggplot2**
  
  • **ggraph**
  
  • **igraph**

• **Sentiment analysis** using quanteda and the AFINN lexicon
DATA SOURCES

- **Linguistic Data Consortium**
- The World Wide Web
- Kaggle.com: many pages of social media datasets, including tweets, and others: example: disaster tweets dataset, Instagram data, emojis, reddit, and many many others.
- **Stanford SNAP:** large network dataset collection, including data from amazon, social media, Wikipedia and others
- **Network Repository:** including social networks, biological, graph data and tools for analyzing and comparing available datasets
iScience Maps: web-based option for getting Twitter data, with options for sorting and analyzing the data

Naoyun: software for connecting Twitter data with Gephi, with options for visualizing “live Twitter activity”

Netlytic: uses APIs to collect public data from Twitter, YouTube, and RSS feeds. Includes free and paid user options, with network and text analytics

Socioviz: get and analyze Twitter data in this web-based environment

The Chorus Project: free web-based option for analyzing and obtaining Twitter data; based out of the UK

Webometric Analyst: free Windows-based program for gathering data, including Social media, from the Statistical Cybermetrics Research

Digital Footprints: obtain and analyze Facebook data; web-based service available for researchers, based out of Aarhus University

InfoExtractor: no longer maintained, but offers options for getting data from different URLs

Snoopreport: free for researchers; focus on obtaining Instagram data
• **streamR**: Access to Twitter Streaming API via R

• **twittR**: also useful for getting twitter data in R

• **Rfacebook**: Rfacebook: Access to Facebook API via R

• **instaR**: access Instagram data via the Instagram API; an approved developer account is required

• **RedditExtractor**: utilizes Reddit API to obtain posts, comments, and subreddit information

• **Rtweet**: useful package for getting Twitter data, with options for accessing followers, retweets, geolocation, and additional metadata.
• **spaCy**: pos tagging, tokenization, dependency parsing, etc. Check out this [tutorial](#) for more about NLP with spaCy

• **CoreNLP**: lemmatization, pos tagging, tokenization, named entity recognition

• **NLTK**: Natural Language ToolKit; contains over 50 corpora, includes options for tokenization, tagging, parsing, document classification

• **Gensim**: useful for various types of topic modeling

• **PyNLPI**: open-source NLP library; great for of tasks ranging from building simplistic models and extraction of n-grams and frequency lists, with support for complex data types and algorithms

• **Pattern**: useful for web-crawling (webscraping) for creating your own corpora; includes options for tokenizing, pos tagging, etc

• **Polyglot**: very useful library for other languages than English

• **TextBlob**: includes options for pos-tagging, noun phrase extraction, classification, translation and sentiment analysis
ADDITIONAL TOOLS

- **AntConc**: A free corpus analysis toolkit for concordancing and corpus-based methods
- **Voyant Tools**: web-based text reading and analysis environment
- **Google Books Ngram Viewer**: online search engine that charts the frequencies of any set of comma-delimited search strings
- **Wordseer**: text analysis environment that combines visualization, information retrieval, and nlp methods
LEARN TO TELL THE DIFFERENCE

GOOD DATA

BAD DATA
• Silge and Robinson’s Text Mining with R
• Social Media Research using R
• Rswirl package
• Python 3 tutorials
• Social Media Analytics: helpful overview of options and types of analyses
• Text Analysis Glossary
• Corpus Approaches to Social Media (Rüdiger & Dayter 2020)
• DigiLab tutorials & Linked-in Learning Tutorials
• Twitter and Tear Gas by Z. Tufekci
• UGA CQP Server: new social media corpus BLM, added courtesy of Jordan Graham and Dr. Hale!
COURSES AT UGA

• Maymester 2021:
  • #TheDigitalLifeofLanguage ROML 4120/6120 & LING 4910

• This Fall:
  • Natural Language Processing: LING 4570/6570
  • Style: ENGL/LING 4826/6826
  • Text and Corpus: ENGL/LING 4886/6886

• Note: These all count toward the Digital Humanities Undergraduate certificate!
THANKS FOR LISTENING!

KATHERINE.KUIPER25@UGA.EDU

PLEASE FILL OUT THIS SURVEY.


• Bird, Steven, Ewan Klein, and Edward Loper. 2019. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit


• Han, Na-Rae. Python 3 tutorials. http://www.pitt.edu/~naraehan/python3/


• Lincoln, Mullen, 2018. https://cran.r-project.org/web/packages/tokenizers/vignettes/introduction-to-tokenizers.html
• Machlis, Sharon. 2020. How to search Twitter with rtweet and R. infoworld.com


• https://monkeylearn.com/text-analysis/

• Millot, Thomas. Photo. Unsplash


• Rivera, Ian. 2019. package RedditExtractoR. https://cran.r-project.org/web/packages/RedditExtractoR/RedditExtractoR.pdf


• Wiedemann, Gregor & Niekler, Andreas. 2017. Hands-on: A five day text mining course for humanists and social scientists in R. Proceedings of the 1st Workshop on Teaching NLP for Digital Humanities (Teach4DH@GSCL 2017), Berlin.