



UNIVERSITY OF GEORGIA

Digital Humanities

# DATA-DRIVEN: THE HUMANITIES GET DIGITAL

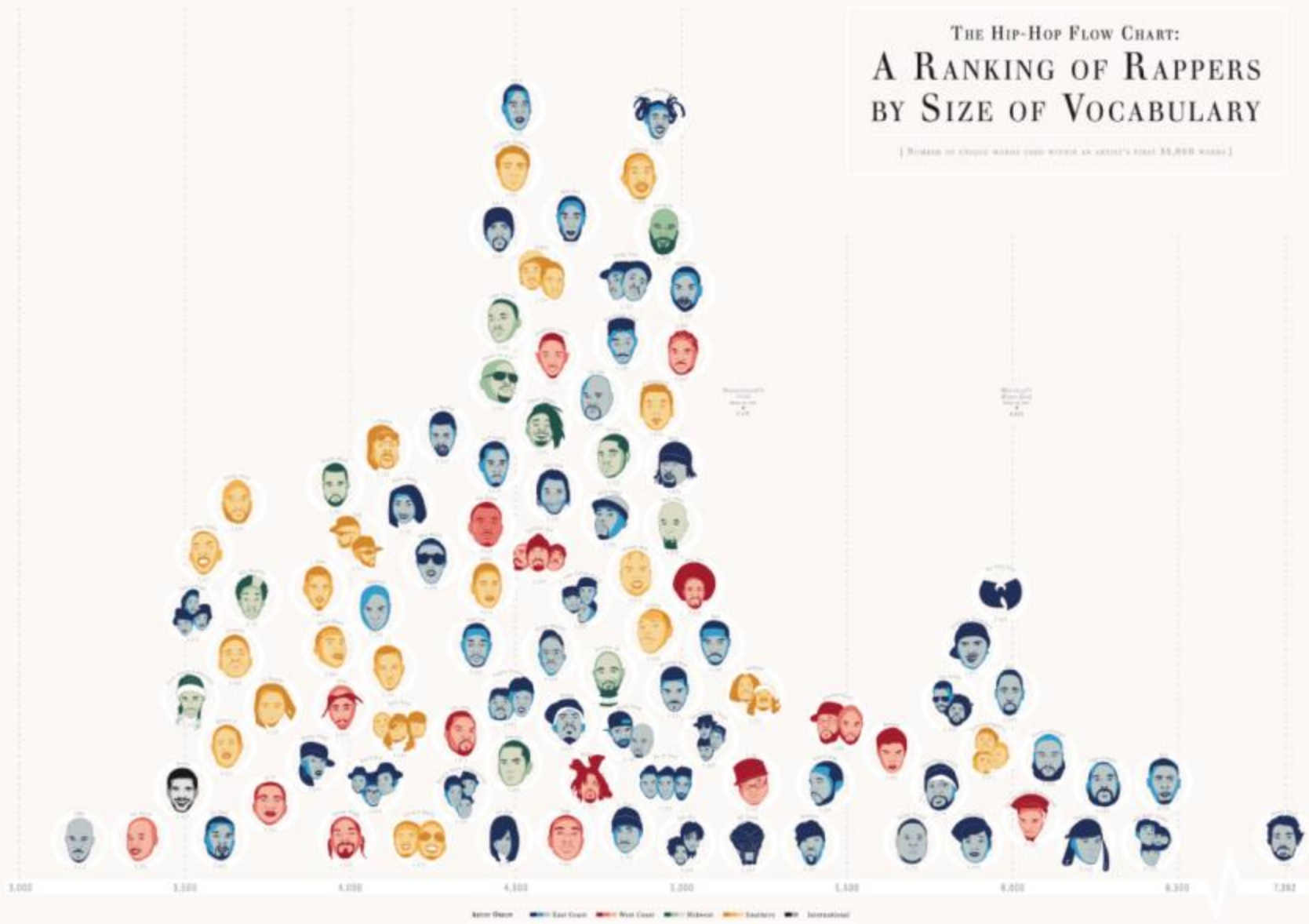
---

Katie Ireland Kuiper  
PhD Candidate, Linguistics

Emily McGinn, PhD  
Head of Digital Humanities

# THE HIP-HOP FLOW CHART: A RANKING OF RAPPERS BY SIZE OF VOCABULARY

[ NUMBER OF UNIQUE WORDS USED WITHIN AN ARTIST'S FIRST 25,000 WORDS ]



# ASKING QUESTIONS OF DATA

---

- What's in your data?
  - What do you have?
  - What's missing?
- What do you want to know?
  - What questions do you have?
  - How would you test these questions with the data you have?

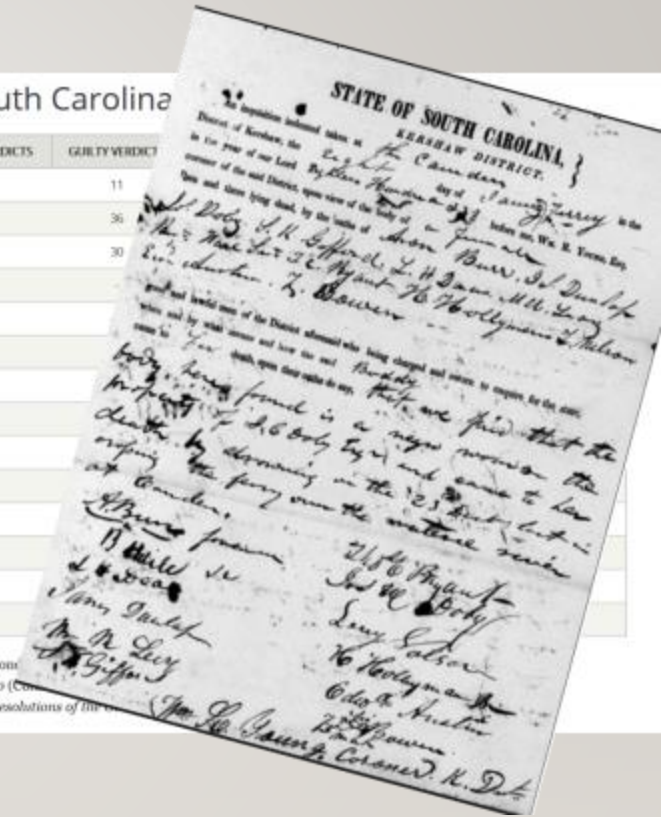
# CSI: DIXIE

Collecting the extant coroners' inquests for South Carolina between 1840 and 1880, "CSI Dixie" provides a glimpse into the sad intimacies inherent in the varied ways people go out of the world.

Murder Cases Tried in South Carolina

YEAR	NUMBER OF HOMICIDES TRIED	NOT GUILTY VERDICTS	GUILTY VERDICT
1887	79	54	11
1888	117	61	36
1889	120	69	30
1890	incomplete returns	-	-
1891	151	76	-
1892	incomplete returns	-	-
1893	incomplete returns	-	-
1894	incomplete returns	-	-
1895	210	112	-
1896	201	110	-
1897	215	120	-
1898	248	105	-
1899	205	83	-
1900	224	127	-

Credit: John Hammond  
Carolina, 1880-1920 (Com  
from Reports and Resolutions of the



# RECONSTRUCTING LOST HISTORY

Emigrants by the ship *Albion* from Dept. of the South from Norfolk 7 June 1842

Providence May 30<sup>th</sup> 1842

The Rev. Mr. M<sup>r</sup> & Son  
Dear Sir

The following is a list of the Black and Colored persons, who it is intended shall take passage in the ship *Albion* for Louisiana and Liberia Bay.

James, an African by birth, has been their commanding as a man in whom confidence may be placed for his honesty is a Brickmaker and is acquainted with plantation work, he is aged about 30 Years.

Henrietta, his wife, aged about 35 Years, Mr. Gage, his son aged 30 Years a Bricklayer, a very little better than our former ones.

Builder of excellent character.

Ellis, his son, a Sugar Maker, understanding the mode of cultivating the cane, and a Bricklayer & Brickmaker aged 22 Years.

Molly Daughter of Henry aged 21 Years

Kenneth do of do do 9 do

Emmanuel Son of do do 7 do

Charity's Daughter of Molly, do 6 do

Anne do of do do 4 do

Willis Son " " 3 "

Shims Gray an excellent man a Carpenter, & a man of all trades, aged about 34 Years, Molly, or Amelia, his wife, aged about 28 Years

Elizabeth, his Daughter aged 9 Years, Lucia, his son 5 Years

Richard, a Minister of the Gospel aged about 30 Years.

Mask, a Carpenter by trade, and a school teacher aged 30 Years also a Brickmaker, a Sugar Maker, a Carpenter and accustomed to plantation work, Bay, plowing, planting, Haying, &c. aged about 35 Years

Thomas, his wife a Tailor, Brickmaker, Miller, &c. aged 30 Years

Charles, his son a Bricklayer, Sugar Maker, and accustomed to plantation work, also a Brickmaker aged 20 Years.

Sarah, their daughter aged 15 Years, Maria, their daughter 12 Years

Sarah, Anne do do 9 do

Peter a first rate Blacksmith and excellent man aged 30 Years

Quana his wife, aged about 35 to 40 Years, Thomas their son a first rate Blacksmith, and a well disposed man, aged about 30 Years

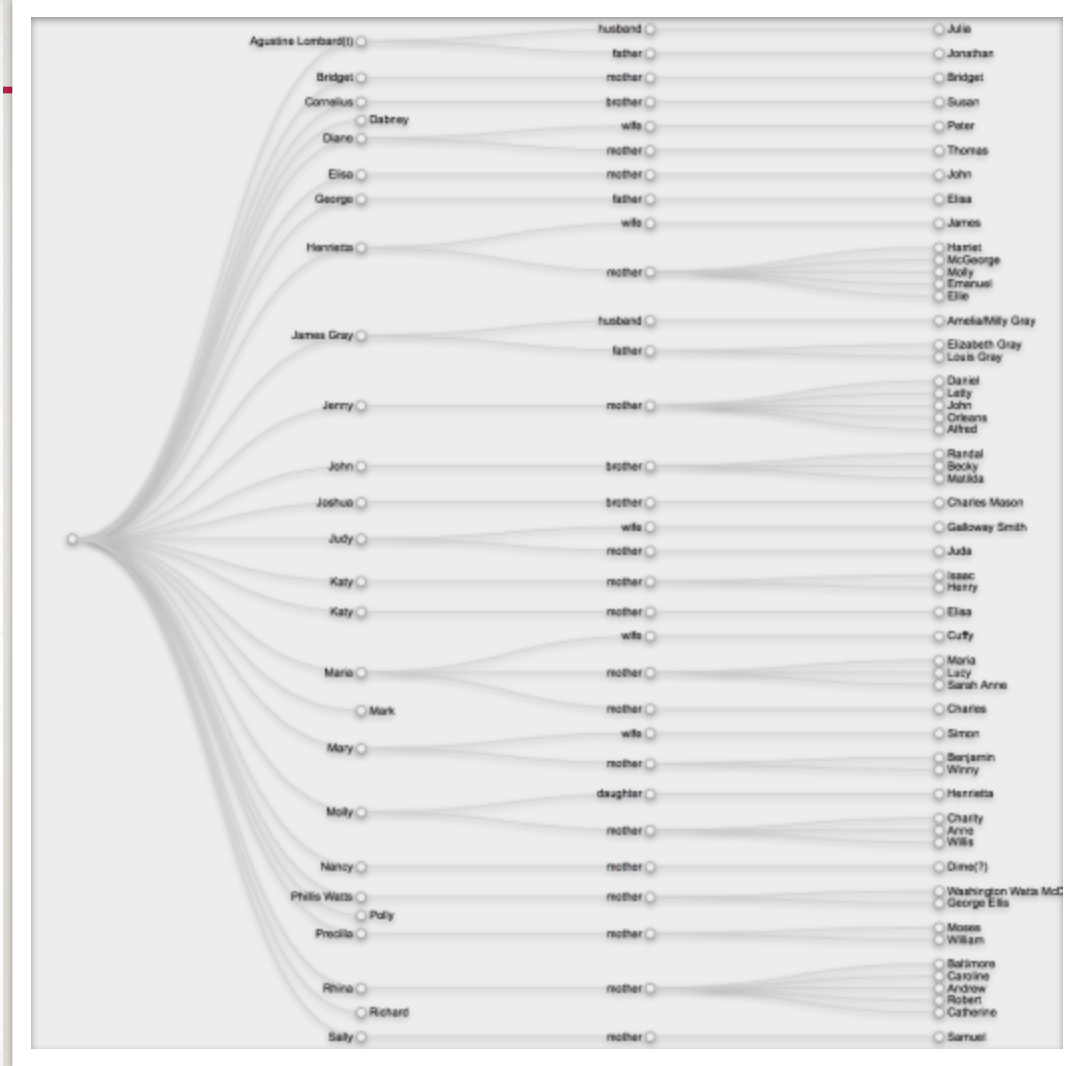
Judy, an excellent woman, a first rate Midwife a Spinner of Cotton and wool, &c. aged about 30 Years, Sally, aged 12 Years

a first rate Carpenter, and Sugar Mill Builder, aged about 32 Years

Judy, Daughter of India aged 19 Years, Jenny, aged 35 Years.

Shim, his son, a Sugar Maker & Carter aged about 19 Years

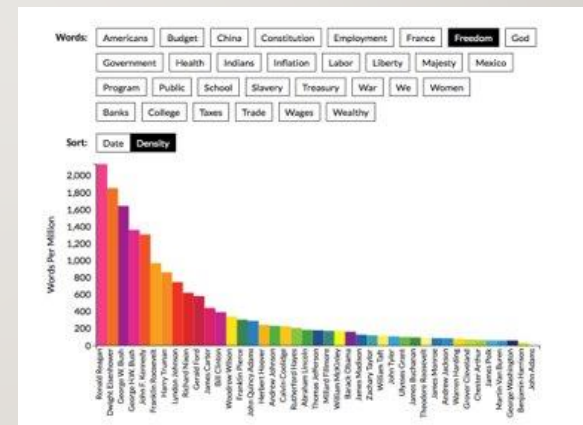
Orlando his son aged 14 Years, Alfred, his son aged 11 Years, Dean, or a Sugar Maker, a Ship Sawyer a Carter, and a Plowman, also a Brickmaker and



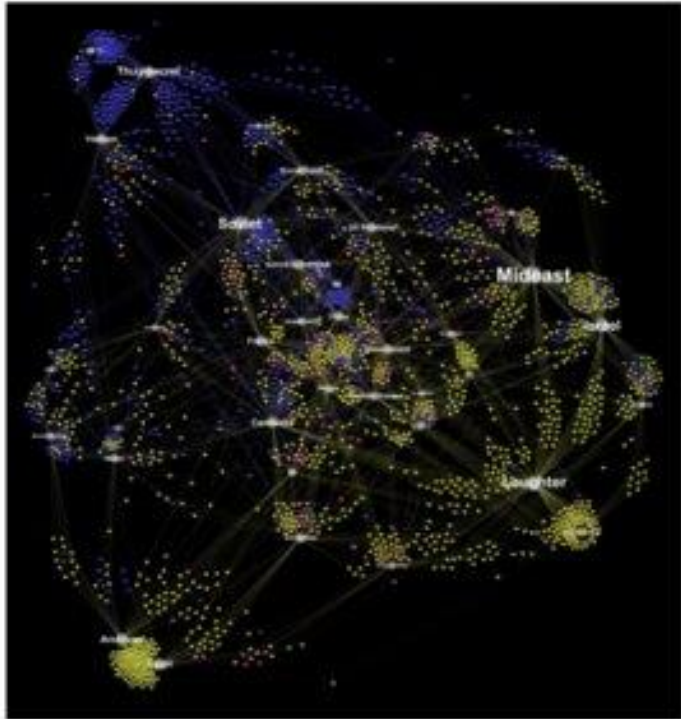
# TEXT ANALYSIS IN DH

---

- Distant Reading
  - Instead of reading one book or a few books, reading 100 or 1000 books at once
- Examples:
  - Mark Algee-Hewitt [The Performance of Character](#)
  - Ben Schmidt and Mitch Frass [The Language of the State of the Union](#)



Memcons: Static Topic Model Force Graph



# TOPIC MODELING

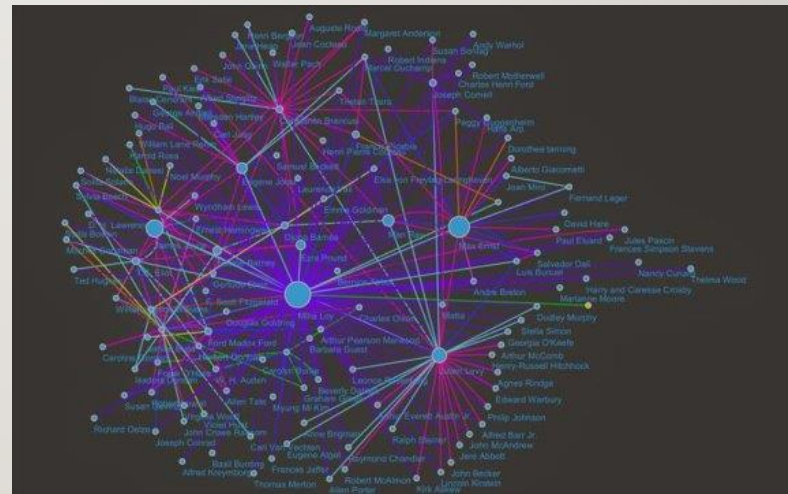
---

- Creating probabilistic models of words that are likely to appear together
- "Reading Tea Leaves" of text analysis
- Example:
  - Micki Kaufman Quantifying Kissinger

# NETWORK ANALYSIS

---

- Graphing connections between groups, people, organizations to find influential actors that may have been erased or forgotten
- Examples:
  - [Six Degrees of Francis Bacon](#)
  - [Mina Loy: Navigating the Avant Garde](#)
    - [Interactive](#)

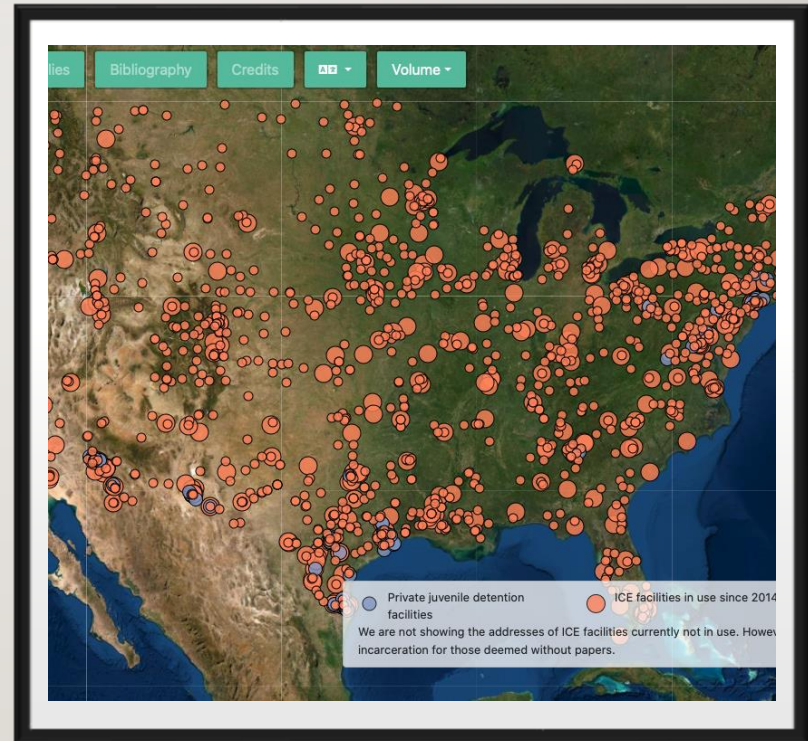




# DH BEYOND ACADEMIA

---

- To be DH argument is necessary
  - Oral History
  - Curated collection
  - Data Journalism
    - Data tells its own narrative
  - Examples:
    - [Covid-19: The Global Crisis – in data](#)
    - [Cliches of ESPN](#)
    - [Torn Apart/Separados](#)



# START SMALL

---

- DH doesn't have to be flashy
- DH doesn't have to be expensive
- DH doesn't have to be hard
  - What can you do today? This week? This month?



# VOYANT TOOLS

---



WEB-BASED ENVIRONMENT



INCLUDES OPTIONS FOR USING  
THEIR DATA OR CREATING  
YOUR OWN CORPUS



ANALYSIS OPTIONS:  
FREQUENCY, NGRAMS,  
COLLOCATIONS, KWIC, TOPICS,  
AND VISUALIZATION(S)

# VOYANT TOOLS AND R

The Project Gutenberg eBook of Emma, by Jane Austen

This eBook is for the use of anyone anywhere at no cost and with almost no restrictions whatsoever. You may copy it, give it away or re-use it under the terms of the Project Gutenberg License included with this eBook or online at [www.gutenberg.org](http://www.gutenberg.org)

Title: Emma

Author: Jane Austen

Release Date: August, 1994 [Etext #158]

Posting Date: January 21, 2010

Last Updated: March 10, 2018

Language: English

Character set encoding: UTF-8

\*\*\* START OF THIS PROJECT GUTENBERG EBOOK EMMA \*\*\*

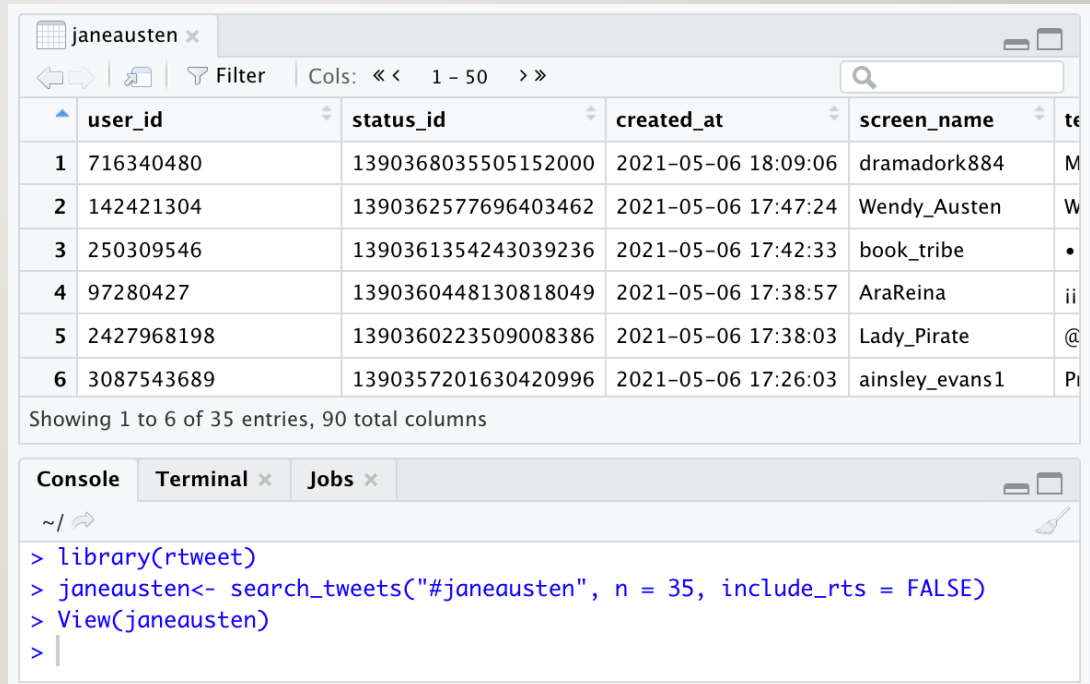
Produced by An Anonymous Volunteer

EMMA

By Jane Austen

VOLUME I

CHAPTER I



The screenshot shows a R console window with a data table and code execution. The table has 6 rows and 5 columns: user\_id, status\_id, created\_at, screen\_name, and te. The code in the console shows the library for rtweet, a search for tweets with #janeausten, and a view of the results.

	user_id	status_id	created_at	screen_name	te
1	716340480	1390368035505152000	2021-05-06 18:09:06	dramadork884	M
2	142421304	1390362577696403462	2021-05-06 17:47:24	Wendy_Austen	W
3	250309546	1390361354243039236	2021-05-06 17:42:33	book_tribe	•
4	97280427	1390360448130818049	2021-05-06 17:38:57	AraReina	ii
5	2427968198	1390360223509008386	2021-05-06 17:38:03	Lady_Pirate	@
6	3087543689	1390357201630420996	2021-05-06 17:26:03	ainsley_evans1	Pr

Showing 1 to 6 of 35 entries, 90 total columns

```
> library(rtweet)
> janeausten<- search_tweets("#janeausten", n = 35, include_rts = FALSE)
> View(janeausten)
> |
```



# ADDITIONAL RESOURCES: DATA

---

- [Linguistic Data Consortium](#)
- 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](#): combines social networks, biological, graph data and tools for analyzing and comparing available datasets

# WEB-BASED RESOURCES

- [iScience Maps](#): focused on 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

# OTHER R PACKAGES

---

- **streamR**: access to Twitter Streaming API
- **twittR**: also useful for getting twitter data in R
- **Rfacebook**: gather data through Facebook API
- **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.
- **xml2** and **rvest** work well together for harvesting web data
- **Rcurl** & **RSelenium**



# PYTHON LIBRARIES

---

- **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**: NLP pipeline for multilingual applications, includes options for preprocessing, analysis of sentiment, morphological features, and more.
- **TextBlob**: includes options for pos-tagging, noun phrase extraction, classification, translation and sentiment analysis

# WEB SCRAPING TOOLS

## Python libraries:

- [Facebook SDK](#): Facebook data scraper
- [Twitter scraper](#): for use with Python 3.6+; can get tweets based on user or other search terms
- [Reddit scraper](#): interacts with Reddit API and PRAW library to obtain Reddit data
- [Tweepy](#) in Python will interact with Twitter API
- [URS](#): Universal Reddit Scraper; command line tool to obtain Reddit data
- [MOZDEH](#): Windows based programming for gathering social media data

## Web scraping:

- [Chrome plugin](#)
- [Beautiful Soup](#): useful python library for webscraping; better for smaller amounts of data
- [Scrapy](#): python library; best for larger datasets
- [Selenium](#): flexible, also beginner friendly library



# THANKS FOR LISTENING!

- [mcginn@uga.edu](mailto:mcginn@uga.edu)
- [katherine.kennedy@uga.edu](mailto:katherine.kennedy@uga.edu)
- <https://digi.uga.edu/bcu/>
- Twitter
  - @DigiLab\_UGA
  - @EmMcGinn
  - @Kannireland

# WORKS CITED

---

- Al-Rfou, Rami. 2015. Polyglot. <https://polyglot.readthedocs.io/en/latest/>
- Barbera, Pablo. 2018. Package 'streamR'. <https://cran.r-project.org/web/packages/streamR/streamR.pdf>
- Barbera, Michael Piccirilli, Andrew Geisler, and Wouter van Atteveldt. 2017. Rfacebook: Access to Facebook API via R. <https://cran.r-project.org/web/packages/Rfacebook/index.html>
- Barbera, Pablo, Tiago Dantas, Jonne Guyt. 2016. Package 'instaR'. <https://cran.r-project.org/web/packages/instaR/instaR.pdf>
- Beckman, Matthew, Stéphane Guernier, Justin Lee, Roberto Molinari, Samuel Orso & Igor Rudnitsky. 2020. An Introduction to Statistical Programming with R. <https://smoc-group.github.io/ds/index.html>
- Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, Akitaka Matsuo, William Lowe. (2018). "qanteda: An R package for the quantitative analysis of textual data." *Journal of Open Source Software*, 3(30), 774. doi: [10.21105/joss.00774](https://doi.org/10.21105/joss.00774), <https://quanteda.io>.
- Bird, Steven, Ewan Klein, and Edward Loper. 2019. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit
- Brown, Simon. 2016. Tips for Computational Text Analysis. <https://matrix.berkeley.edu/research/tips-computational-text-analysis>
- Bussiere, Kirsten. 2018. [Digital Humanities - A Primer](#).
- Cribbin, Timothy, et al. 2021. The Chorus Project: making sense from Twitter. <http://chorusanalytics.co.uk/about/>
- Csardi, G., Nepusz, T. (2006). "The **igraph** software **package** for complex network research." *InterJournal, Complex Systems*, 1695. <https://igraph.org>.
- De Smedt, Tom. 2018. Pattern 3.6 <https://pypi.org/project/Pattern/>
- Digital Footprints Research Group. 2021. Digital Footprints. <https://digitalfootprints.dk/about> Aarhus University.
- Evert, Stefan. 2007. Corpora and collocations. [http://www.stefan-evert.de/PUB/Evert2007HSK\\_extended\\_manuscript.pdf](http://www.stefan-evert.de/PUB/Evert2007HSK_extended_manuscript.pdf)
- Feinerer, Ingo. 2020. Introduction to the tm Package: Text Mining in R. <https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf>
- Gentry, Jeff. 2015. twitterR: R Based Twitter Client. <https://cran.r-project.org/web/packages/twitterR/index.html>
- Grün, Bettina & Kurt Hornik. topicmodels: An R Package for Fitting Topic Models. <https://cran.r-project.org/web/packages/topicmodels/vignettes/topicmodels.pdf>
- Gruzd, Anatoly, & Philip Mai. Netlytic: Making sense of public discourse online. <https://netlytic.org/home/>
- Han, Na-Rae. Python 3 tutorials. <http://www.pit.edu/~nraehan/python3/>.
- Harrison, John & Ju Yeong Kim. 2020. RSelenium. <https://cran.r-project.org/web/packages/RSelenium/RSelenium.pdf>
- Honnibal, Matthew and Montani, Iles and Van Landeghem, Sofie and Boyd, Adriane. 2020. [spaCy: Industrial-strength Natural Language Processing in Python.] (<https://doi.org/10.5281/zenodo.1212203>)
- Kearney, Matthew. 2018. R: Collecting and Analyzing Twitter Data: featuring (tweet). Ni CAR 2018. [http://mkearney.github.io/nicar\\_workshop/#/](http://mkearney.github.io/nicar_workshop/#/)
- Kearney, Matthew, Andrew Heiss, and Francois Briatte. 2020. Package 'rtweet'. <https://cran.r-project.org/web/packages/rtweet/rtweet.pdf>

# WORKS CITED

---

- Kross, Sean et al. 2020. swirl: Learn R, in R, <https://cran.r-project.org/web/packages/swirl/index.html>
- Kuiper, Katie Ireland. 2021. Text Analysis Glossary. *DigiLab*. <http://316.wul.mnif02clvg22bwb9l-wpengine.netdna-ssl.com/wp-content/uploads/sites/9/2021/03/Text-Analysis-Glossary.pdf>
- Leskovec, Jure, and Krevl Andrej. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. <http://snap.stanford.edu/data>
- Loria, Steven. 2020. [TextBlob: Simplified Text Processing.](<https://textblob.readthedocs.io/en/dev/>)
- Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. [The Stanford CoreNLP Natural Language Processing Toolkit](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* pp. 55-60.
- Pederson, Thomas. 2021. ggraph: an implementation of grammar of graphics for graphs and networks. <https://cran.r-project.org/web/packages/ggraph/index.html>
- Radim, R. & P. Sojka. 2010. Software Framework for Topic Modeling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. 45-50.
- Reips, U.-D., & Garaizar, P. 2011. Mining Twitter: Microblogging as a source for psychological wisdom of the crowds. *Behavior Research Methods*, 43, 635-642. doi <http://dx.doi.org/10.3758/s13428-011-0116-6>
- Rivera, Ian. 2019. package RedditExtractoR. <https://cran.r-project.org/web/packages/RedditExtractoR/RedditExtractoR.pdf>
- Roberts et al. 2019. stm package for structural topic modeling. <https://www.jstatsoft.org/article/view/v09i02>
- Rüdiger, Sophia, and Daria Dayter. 2020. Corpus Approaches to Social Media. In *Studies in Corpus Linguistics*.
- Schrading, N. 2015. Intro to NLP with spaCy. <https://nicschrading.com/project/intro-to-nlp-with-spacy/>
- Shah, Chirag. [infeextractor.infoextractor.org](http://infeextractor.infoextractor.org)
- Silge, Julia, and David Robinson. 2017. Text Mining with R: A Tidy Approach. <https://www.tidytextmining.com/>
- Thelwall, M., & Sud, P. 2012. Webometric research with the Bing Search API 2.0. *Journal of Informetrics*, 6(1), 44-52.
- Toret, Matthieu. 2021. Naoyun- Visualize Live Twitter Activity. <http://matthieu-toret.fr/Koumin/tod/s/naoyun/>
- van Gompel, Maarten. 2016. PyNLPi. <https://pynpli.readthedocs.io/en/latest/index.html>
- Wasser, Leah, and Carson Farmer. 2020. Twitter Data in R Using Rtweet: Analyze and Download Twitter Data. *Earth Data Science*. <https://www.earthdatascience.org/courses/earth-analytics/get-data-using-apis/use-twitter-api-r/>
- Watanabe, Kohei. 2021. Example: social media analysis <https://quantda.io/articles/pkgdown/examples/twitter.html> quantda package examples.
- Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, <https://ggplot2.tidyverse.org>.
- Wickham et al., 2019. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>
- Wickam et al. 2020. xml2: Parse XML. <https://cran.r-project.org/web/packages/xml2/index.html>
- Zonin, A. 2015. SocioViz: A Free Social Network Analysis Tool for Twitter [Software]. Available from <http://socioviz.net/>