SHADES OF #MAGA

Final report for visual analysis project exploring the use of the “Make America Great Again” slogan on Twitter

Devon Mordell

ARST 575H: Information Visualization and Visual Analytics

November 30, 2017

Table of Contents

Executive Summary

This report describes an exploratory visual analysis conducted on Twitter data containing the #MAGA ("Make America Great Again") hashtag. It discusses key findings from the analytic process; steps involved with data collection and cleaning; the analytic tools used, and tasks performed with them; and an evaluation of the project. The primary project outcome was a website that featured static, stylized visualizations exported from Gephi and Voyant, and a set of embedded dynamic visualization tools from Voyant.

Overview of Analytic Tasks

At the outset of the project, my objective was to investigate the context and meaning(s) of the #MAGA hashtag for the purposes of better understanding how the slogan is discursively constructed on Twitter.[1] My argument was that its lack of referentiality – i.e. the elusiveness of its precise meaning – allows #MAGA to accumulate meanings through its usage, which could be elucidated through a visual analysis of Twitter data. I determined that the most suitable approach to meeting my objective was to undertake:

1. A social network analysis, to visualize the topology of its use, and
2. A textual analysis, to visualize its semantic dimensions.

In my project plan, the overarching analytic tasks I initially proposed were to make preliminary observations in Tableau using tabular data (i.e. word counts), to graph the relationships between users of the #MAGA hashtag in Gephi, and to compute sentiments and cluster tweets using Jigsaw.

[1] The acronym is emblematic of Donald Trump's "Make America Great Again" campaign slogan.

Data Provenance

The dataset was comprised of Twitter data harvested from a five-day period between 12:00AM on November 6 and 11:59PM on November 10, 2017. The primary fields used in the analysis from the original Twitter data were the tweet's text and sender, and the sender's number of followers. Other data (e.g. @ mentions) was generated from the tweet text using OpenRefine and Microsoft Excel.

Harvesting the Data

The proposed extent of the project was to collect data for three seven-day periods surrounding significant events in Trump's presidency (his election, the white nationalist rally in Charlottesville, and a "control" week of non-controversial events).[2] An initial test run of the data collection process, however, revealed that the planned dataset would be unwieldy; in one day alone, Twitter users sent over 100,000 tweets containing the #MAGA hashtag.[3] Furthermore, the Twitter API limits the retrieval of tweets to the previous nine days.[4] I modified the scope of the data collection accordingly, to five days around one significant event: the one-year anniversary of the 2016 election – November 8, 2017.

To harvest the data, I used a Python script (Galea, 2016/2017). My initial intent was to use Martin Hawksey's TAGS 6.1 spreadsheet (2014) but it became evident that the tool was only returning a small fraction of the requested tweets. The Python script was able to retrieve all

[2] Part of the project proposal was to explore links between the use of the #MAGA hashtag and the emboldening of white supremacists and white nationalists, hence the reason for including the week surrounding the events in Virginia on August 11 & 12, 2017.
[3] Twitter users in US and Canada, that is – the data gathered by the Python script was bound by geographical coordinates.
[4] I explored options for harvesting older tweets through web scraping techniques in the meantime, but ultimately decided to focus instead on cleaning the large corpus of data returned by the Python script.

tweets from each day, but as I was cleaning the data, I discovered that only the first 140 characters were captured.[5] I decided, however, that a full set of incomplete tweets was preferable to a proportionally insignificant number of complete tweets.

Cleaning the Data

I set the script to create separate files for each day, and for the #makeamericagreatagain and #MAGA hashtags respectively; the data was retrieved in JSON format. While my approach made it possible to get more granular insight into the data, it also entailed repeating pre-processing tasks on five to ten datasets and introducing more opportunities for error.[6] To clean and prepare the data for analysis, it was necessary to:

1. Convert the JSON data to CSV using a Python script (Dolinar, 2015/2017)
2. Remove URLs and unescape HTML entities, e.g. &amp; using a Python script[7]
3. Replace Unicode code point characters, e.g. \u2026 using Notepad++[8]
4. Create an edge file for Gephi with weights and directions using OpenRefine & Excel[9]

[5] Regrettably, Twitter increased the maximum number of characters allowed in a tweet to 280 on November 7 – and, not being a frequent user of Twitter, I was oblivious to the development. The Python script I used accessed the API through a compatibility mode that was meant to work with legacy applications, which truncated longer tweets to 140 characters with an ellipsis to indicate more content. While it would have been possible to update the script to capture the entire tweet, it would have required additional testing and re-scraping the entire dataset when I was supposed to be moving on to data pre-processing according to my project timeline.

[6] Pre-processing the harvested data *en masse* would not have been possible given the size of the full dataset, regardless; I encountered enough 'application not responding' messages with the per-day datasets. Nor could the visualization tools accommodate datasets of that magnitude. I also eventually omitted the #makeamericagreatagain datasets from the textual and social network analyses because – after accounting for duplicate tweets in which senders had included both hashtags – there was a negligible number of tweets that only contained the #makeamericagreatagain hashtag.

[7] Script written by the author – see Appendix B.

[8] Replacing the Unicode code points was the most labour-intensive aspect of data cleaning by far: the steps involved are detailed in Appendix B, which could likely be automated by a more competent Python programmer.

[9] The edge files were created using the instructions from Lab 4; the maximum number of mentions included was capped at 5.

After the initial set of analytic tasks, I prepared additional subsets of data in OpenRefine and Excel to isolate 1) retweet mentions, to discover who was being retweeted most frequently and 2) tweets containing the keyword "white," to take up a question from my project plan about associations between the #MAGA hashtag, national identity and whiteness.

Analytic Process

My analytic process followed two separate trajectories: one for social network analysis, the other for textual analysis. I undertook the textual analysis first, hypothesizing that the most interesting insights would emerge from the lexical and sentiment analyses. I then performed the social network analysis. I had intended for subsequent visual analyses to reflect a sense-making loop (Keim et al. 2008, pp. 164-165) that integrated the two strategies to provide a richer representation of the #MAGA hashtag's use, but chose in the end to focus my further analytical iterations on more detailed aspects of each approach.

Textual Analysis

For the textual analysis component of the project, I sought to determine what words were most frequently used in conjunction with the #MAGA hashtag (lexical analysis) and in what sense they were meant (sentiment analysis). The most notable departure from the project plan was the use of Voyant instead of Jigsaw as the primary analytic tool for the text corpus. The rationale was, for the most part, pragmatic: Jigsaw was unable to handle large datasets that Voyant could easily manage. Voyant tokenized the tweets as tabular data and presented the word tokens through multiple visual encoding idioms at once on its dashboard, allowing me to quickly identify the most frequently used terms in the corpus; for example, the prevalence of "RT" throughout all five days of data indicated that many tweets were not original but retweets from other users. Noting that the five most frequently mentioned terms tended to remain consistent over the five days of analysis, I created a new dataset from the top 100 terms for each day and weighted them according to their cumulative frequency to create the static bubble plot and cirrus graph visualizations shown on the project website.[10] I also used the aggregate dataset of tweets containing the keyword “white” to create a cirrus graph (figure 1).

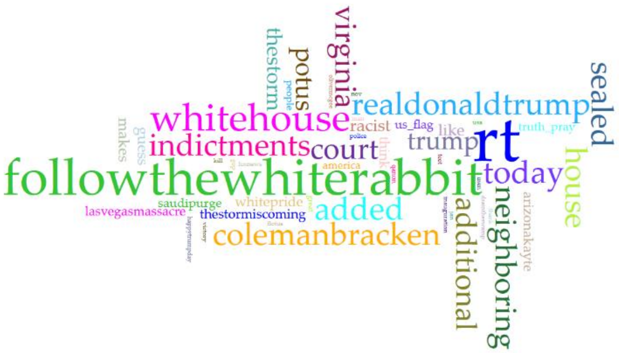

Figure 1: Cirrus graph of tokens from #MAGA tweets containing the keyword "white." A larger, interactive version is available on the project website.

For visual analysis, Voyant offered an impressive range of visual encoding and interaction idioms; many of them, however, fell short of communicating the desired abstraction – that is, what words appeared most frequently in tweets containing the #MAGA hashtag (compare the streamgraph in figure 2 with the pre-attentively legible bubble plots in figure 3). Also, because the size and colour encodings of various idioms appeared to be arbitrary and/or not proportional at times, it was difficult to trust the accuracy of the visualizations thus lowering their efficacy (figure 3).

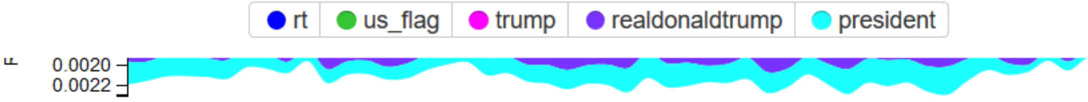

Figure 2: Streamgraph in Voyant showing changes in frequency of term use over time, from the end of the day (11:59PM) to the beginning (12:00AM); the streamgraph places an additional cognitive load on the prospective viewer to process an unfamiliar idiom and infer what it is attempting to communicate.

[10] “Cirrus graph” is Voyant’s term for word clouds.

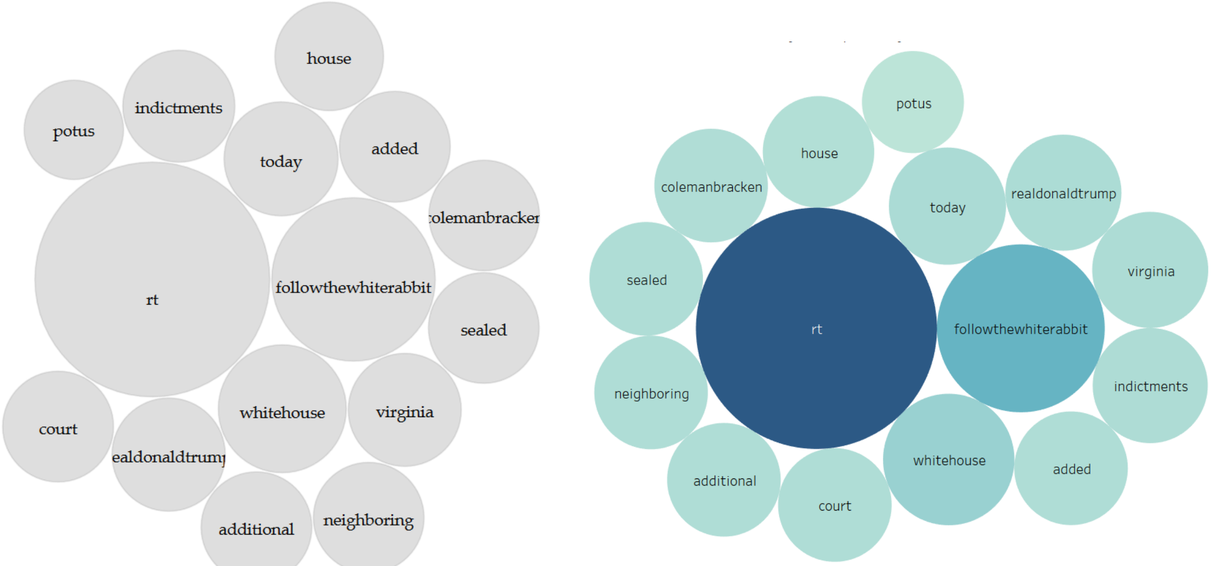

Figure 3: Validating a bubble plot in Voyant (left) using a representation of the same data in Tableau (right) for the corpus of tweets containing the word "white." Note that the proportional size of the bubbles in Voyant is only accurate when the "scale" slide is at its maximum (5), as it is in the visualization above.

Moreover, the reliance upon lexical analysis for most of the idioms in Voyant meant that the insights gleaned – though helpful – were relatively superficial. Knowing that the term "Trump" appears frequently is less telling without also knowing the context of its use: does it belong to an expression of support or a caustic critique? In my project plan, I had indicated the possibility of rudimentarily coding the tweets by sentiment but the size of the dataset meant relying upon automated techniques. The Voyant dashboard includes a sentiment analysis tool, but it is strictly textual rather than visual in its display. Later in the analytical process, I was able to work with Jigsaw using the smaller "white" keyword dataset. Leveraging Jigsaw's semantic analysis algorithms through visual encoding idioms like its document cluster (figure 4) and document grid views hinted at the possibility of what I could do in the future to gain a more nuanced understanding of the corpus.

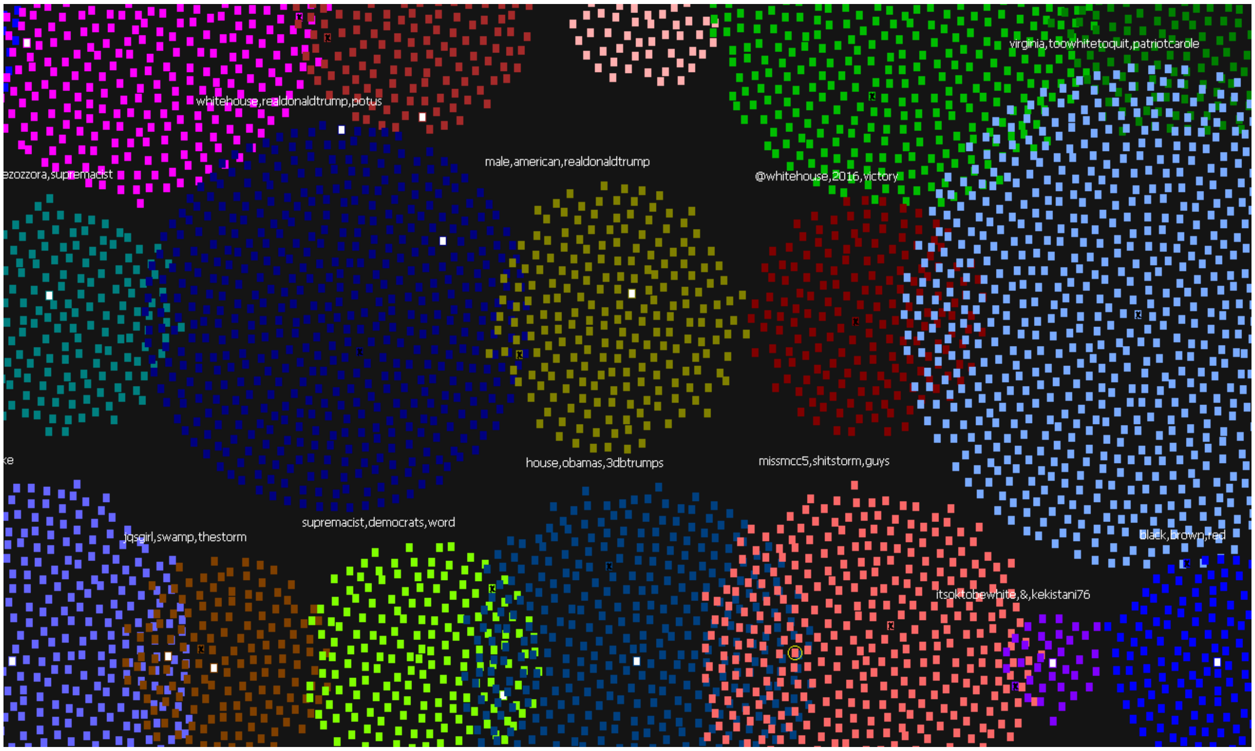
virginia,toowhitetoquit,patriotcarole
whitehouse,realdonaldtrump,potus
male,american,realdonaldtrump
ezozzora,supremacist
@whitehouse,2016,victory
ke
house,obamas,3dbtrumps
missmcc5,shitstorm,guys
supremacist,democrats,word
jqsgirl,swamp,thestorm
black,brown,red
itsoktobewhite,&,kekistani76

Figure 4: Document cluster view in Jigsaw from "white" keyword dataset.

Social Network Analysis

With the social network analysis component of the project, my intent was to visualize the network topology of #MAGA: to locate the Twitter users who most frequently mentioned and/or were mentioned in relation to the #MAGA hashtag, and the connections – if any – between them. I used Gephi to make sense of the complex network data generated through the #MAGA hashtag's use in the intuitive form of a force-directed placement idiom.

As Munzner notes, force-directed placement diagrams can quickly degenerate into a "hairball" where the link density is such that nodes are occluded by edges (2014, p. 210); I filtered the visible nodes to a more manageable number (~300) by excluding nodes with a degree – i.e. number of connecting edges – lower than 50 (figure 5). I then arranged the remaining nodes with Gephi's layout algorithms (Fruchterman Reingold & Force Atlas) to reveal the underlying structure of the network, and manually dragged some of the larger nodes out to the periphery to evaluate their connectedness. In order to establish who was doing the mentioning and who was being mentioned, I manipulated the colour hue and size channels of the nodes: first, mapping the hue to in-degree to create a visualization for the targets of the tweets, and then mapping the hue to out-degree to highlight the sources (the mapping for the size of the nodes remained constant, ranked based on in-degree).

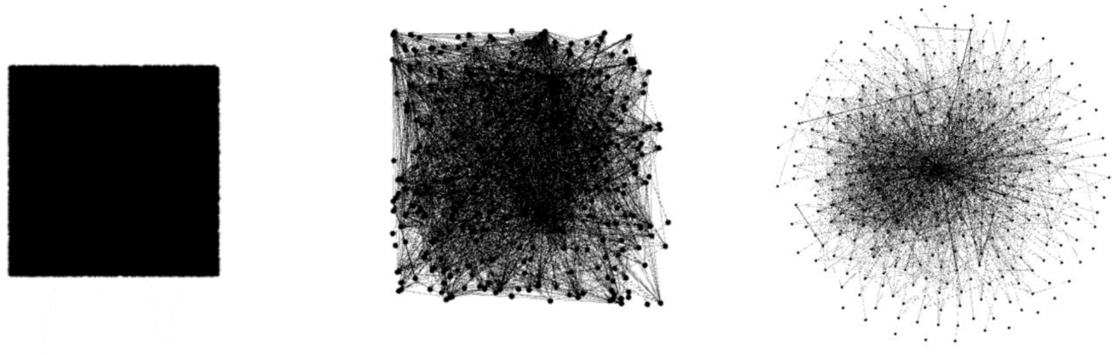

Figure 5: Graph data representing users associated (i.e. mentioning and/or being mentioned in tweets) with the #MAGA hashtag on November 8, 2017 as first imported into Gephi (left), with a filter based on degree applied (center) and following the application of a Fruchterman Reingold layout algorithm (right).

Although the visualizations created in Gephi did offer some analytical insights, there were multiple issues with the tool's stochasticity: one significant weakness of Gephi – and of the force-directed placement idiom more generally – is its non-deterministic layouts, meaning that subsequent runs of the algorithm do not result in predictable outcomes nor can layouts be undone. Likewise, modifications to the channels of the visualization exemplified what Munzner refers to as "brittleness" (2014, p. 206), or a context-sensitivity to parameters, making it difficult to trust the visualization without a sophisticated grasp of the mathematical principles behind them. From a technical standpoint, Gephi's inability to accommodate more than 100,000 nodes at a time required limiting the dataset to one day's worth of tweets, which can hardly be viewed as a representative sample.[11] The option to export a filtered data table from Gephi, however, may offer a viable approach to visualizing the larger topology of #MAGA's usage.

Analysis Highlights

Given the sizeable dataset to work with and the richness of its contents, both the social network and textual analyses yielded numerous insights, from the expected to the surprising. The first was the prevalence of retweets (i.e. the term "RT"), by far the most frequently used term across all five days. It suggests that the #MAGA hashtag circulates within a populist echo chamber of sorts, rather than being a site of new knowledge creation that one might associate with a research topic hashtag (e.g. #blockchain). At the centre of the text and context of #MAGA is, of course, the polarizing figure of Donald Trump. Considering the pervasive collocation of #MAGA and @realDonaldTrump, the hashtag may be imagined as opening a direct line of communication to the president – for praise and lambastes alike. Indeed, if #MAGA can be said to mean anything, it is a reflection of Trump's public persona: big on self-referentiality, hazy on specifics. It would, for example, be nearly impossible to determine Trump's political values or even party affiliation from the bubble plots.[12]

More surprising, as illustrated by the force-directed placement graph in which colour hue encodes the out-degree attribute, is how little Trump himself participates in the ongoing discursive construction of #MAGA – even on the anniversary of his election. In fact, the @realDonaldTrump handle does not appear once in the 'screen_name' field of the dataset

[11] Force-directed placement diagrams are inherently not scalable beyond a certain threshold, as evidenced by figure 5, but the full dataset could ideally be filtered for a more accurate representation.
[12] See project website.

(that is, as a sender of a #MAGA tweet) during the entire five days of data collection. It would seem that #MAGA behaves less like a political slogan than as an extension of Trump's brand; its visibility – like a designer's name on clothing – due mostly to the free promotion by others.

With respect to the more direct question of examining #MAGA's uptake in the discourses of white nationalists and white supremacists, as investigated through the visual analysis of the "white" keyword subset, the results were inconclusive but worth further study as predicted. Hashtags symptomatic of white backlash to anti-racist critiques like #itsokaytobewhite, #whitelivesmatter, #whitepride and #whitegenocide have a small but not insignificant presence within the corpus. An unintended consequence of the "white" keyword analysis was the surfacing of conspiracy theory-oriented hashtags like #followthewhiterabbit, #quanon and #thestormiscoming, a facet of #MAGA's use that – by definition, as a fringe opinion – would not be discernible at the overview level of analysis. I was impressed by the capacity of visual analysis to unveil the semantic landscape of #MAGA's use, from its centre down to its margins; it would be an invaluable means for researchers to discover what might not be evident from tracking the #MAGA hashtag natively in Twitter.

Discussion

In addition to the insights gathered from the data itself, the project also revealed much about the larger process of visual analysis. The most notable – and heartbreaking – lesson was the importance of reliable and robust data collection and cleaning methods. Although there was a sufficient amount of data to make the analysis meaningful, the incompleteness of the tweets owing to their truncation by the Twitter API would undermine the validity of the results if I wished to use them for purposes beyond my own edification. Moreover, in working with such a large dataset, it was difficult to fully observe and grasp the destructiveness of some of the pre-processing steps; that is, the potential for another layer of information to be sloughed off the dataset. My concerns about the dataset's integrity also reinforced the necessity of documenting the project's data and analytic provenance so that others may trust its findings.

Another dimension of the project for further consideration is the privacy of the users in visual analysis projects that work with Twitter data. Lemieux and Rossert highlight the ethical quandaries in the secondary use of Twitter data, particularly in relation to tweets that intersect with political commentary (2016, para. 2 & 3); in the case of the "white" keyword dataset especially, there is the potential to expose or mischaracterize the participation of certain individuals who may believe they are only communicating with a small audience. To respect the "privacy in public" of the users, I took measures like encoding the number of followers attribute to the label size in Gephi so that the label would – in theory – be too small to read if a user had fewer followers.[13] For a visualization that I created in Tableau to determine which users were most frequently retweeted (figure 6, not used in the project website), I de-identified users with fewer than 25,000 followers – the rationale being that a Twitter user with more than that number of followers may be assumed to have a reduced expectation of privacy when tweeting. The question of how to ethically represent Twitter users in textual or social network analyses, however, remains to be contemplated.[14]

[13] In the end, this proved to be a poor design choice, as users with large numbers of followers were disproportionately visible in the graph in some cases. The number of followers attribute may have been more effectively represented through a mapping to node size. Rendered as an SVG, the visualization could also be enlarged to view the names of smaller labels. An alternative strategy like the one used in the Tableau visualization from figure 6 could be achieved by exporting the data table from Gephi and performing a find and replace function to de-identify users based on their number of followers.

[14] The data in Voyant, which has not been de-identified, runs the risk of violating the privacy of the implicated users, although by not hosting the project website publicly I am relying upon the imperfect measures of "security by obscurity" and Voyant's own privacy policy.

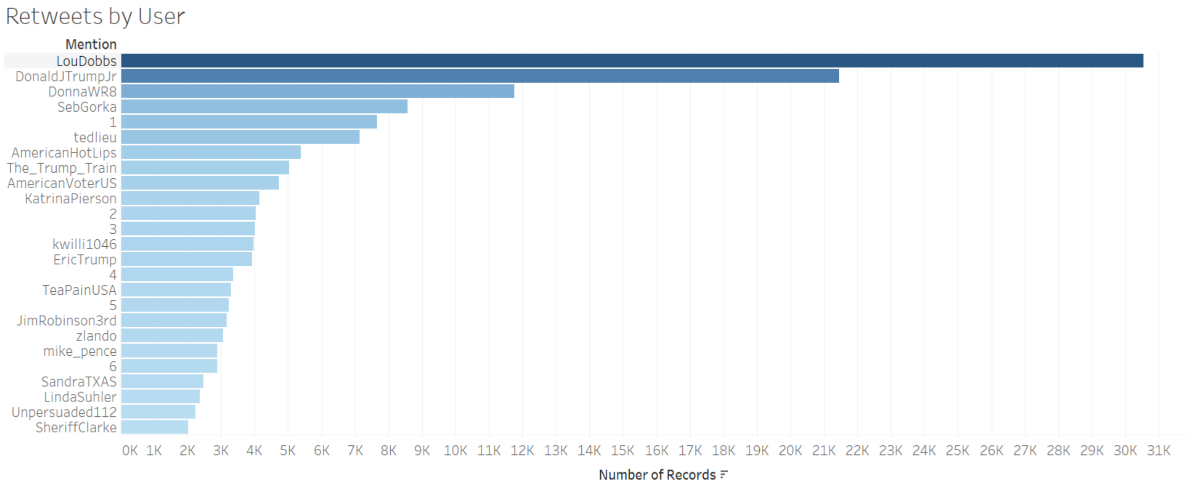

Figure 6: A horizonal bar graph showing the most frequently retweeted users of the #MAGA hashtag, calculated cumulatively for the five days of data collection. Users with fewer than 25,000 followers have been de-identified.

Lastly, the project was admittedly ambitious in scope. Although I feel that the visual analysis methodology I used was well-suited to realizing my research objective, the size of the dataset – and the repercussions for pre-processing and visualizing it – prevented more than a superficial understanding of the findings in the available time. Ultimately, the project became an elaborate prototype or feasibility study that spoke more to process than an end product; any further analyses or renderings of the results would have been pointless without gathering a new and more complete dataset. But I am encouraged by my results to refine my approach and undertake the study again with a greater awareness of the time demands involved. It will also require me to further develop my skills in Gephi, Jigsaw and Voyant in order to interpret their visualizations on a more sophisticated level, or adopt a pair analytics approach in which I work with someone who already has that expertise (Arias-Hernandez et al. 2011). At the end of the project, I have an appreciation of how much I have left to learn – about the tools, and about the dataset – and a desire to see my research questions properly addressed.

References

Arias-Hernandez, R., Kaastra, L. T., Green, T. M., & Fisher, B. (2011). Pair Analytics: Capturing Reasoning Processes in Collaborative Visual Analytics. In *2011 44th Hawaii International Conference on System Sciences* (pp. 1–10). https://doi.org/10.1109/HICSS.2011.339

Dolinar, S. (2017). *twitter_JSONParser_ASCII.py*. Python. Retrieved from https://github.com/seandolinar/Twitter-Tutorial (Original work published 2015)

Galea, A. (2017). *twitter_search: High level script for finding tweets using Python 3 and Tweepy*. Python. Retrieved from https://github.com/agalea91/twitter_search (Original work published 2016)

Hawksey, M. (2014, September 22). Get TAGS. Retrieved November 27, 2017, from https://tags.hawksey.info/get-tags/

Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., & Melançon, G. (2008). Visual Analytics: Definition, Process, and Challenges. In *Information Visualization* (pp. 154–175). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-70956-5_7

Lemieux, V. L., Gormly, B., & Rowledge, L. (2014). Meeting Big Data challenges with visual analytics: The role of records management. *Records Management Journal*, *24*(2), 122–141. https://doi.org/10.1108/RMJ-01-2014-0009

Lemieux, V., & Rossert, B. A. (2016). When Public Data Shouldn't Be Public: A Case Study on the Ethics of using Twitter Data in Big Data Research. https://www.researchgate.net/publication/311986390_When_Public_Data_Shouldn%27t_Be_Public_A_Case_Study_on_the_Ethics_of_using_Twitter_Data_in_Big_Data_Research

Munzner, T. (2014). *Visualization Analysis and Design*. Boca Raton, FL; London; New York: A K Peters/CRC Press. https://doi.org/10.1201/b17511-3

Appendix: Detailed Description of Data Cleaning Tasks

Data pre-processing steps for Gephi

As indicated in the report, data pre-processing steps were carried out according to the directions for Lab 4. To automate the assigning of directed/undirected values in the edges spreadsheet, I used the a function in Excel(=IF(C2="","undirected","directed")).

The one difference was the addition of a "followers" column in the nodes spreadsheet. It was also necessary to clean the data to avoid the duplication of nodes in Gephi; I used search and replace to strip out any errant punctuation marks and the =LOWER() function in Excel to convert all letters to lowercase (Gephi is case sensitive).

Data pre-processing steps for Voyant & Jigsaw

I initially used OpenRefine to tokenize the data and then clustered for similarity to avoid having multiple spellings of the same word. The large dataset (>1,500,000 words per day) overburdened my computer's resources, however, and stripped the words of their context within a tweet. Performing find and replace in Excel threw up errors (forum searches suggest that it was due to cells starting with @), so I used Notepad++ for most of the pre-processing.

A complete list of steps taken in Notepad++ to clean the textual data is listed below:

1. Replace \n with whitespace
2. Replace \U0001f1fa\U0001f1f8 with [US_flag]
3. Replace truncating ellipsis (\u2026) and the text directly preceding it – to avoid partial words – with whitespace (regex: \b(@|#|[\w])+([[:word:]]\\u2026))
4. Replace \u2026 with NULL

5. Replace \\u2018|\\u2019 with ‘ (type in quotation mark; regex)

6. Replace \s+(@\,|#\,) with “,” (regex)

7. Replace \s+(@\”\,|#\”\,) with “”,” (regex, replace quotes)

8. Replace (\s)+\, with “,” (regex) and (\s)+”\, with “”,”

9. Replace “.” with “. ” (NOT regex!!!!)

10. Use find & replace to locate unicode chars

11. Replace “ [ ]+” with “ ” (regex)

Python Script for Cleaning HTML Entities & URLs

```
import html

import re


with open('input.csv', 'r') as inputCSV, open('output.csv', 'w', newline='') as output:

    content = html.unescape(inputCSV.read())

    noURL = re.sub(r"(https?\://)[^\s\"|,]+", "", content)

    output.write(noURL)
```