A Computational Linguistic Analysis of Resting Twitter Data by Shawn Gardy

# Introduction

Twitter is an exciting social media platform that allows users to freely participate in global discourse from wherever they are around the world, providing they abide by the firms agreed upon terms of service. Their firm is kind enough to make most of the content generated by its users open to the public to access analysis within a variety of different contexts.

Within the field of academia, users of this data are particularly interested in the language applications this presents within the fields of Computational Linguistics. Using Natural Language Processing techniques such as n-gram analysis and some commonly knows statistical calculations it would be interesting to take a snapshot of this data in order to identify if any new phrases are being used with respect to various cities within North America.

Also, taking into consideration that users can choose to access this platform on a mobile app or web browser, further attention will be paid to identify as to what extent the frequency of use of text abbreviation or short-message-service acronyms are being used.

# Literature Review

Deitel, Harvey Deitel, Paul J. Python for Programmers. [Addison-Wesley Professional](https://learning-oreilly-com.ezproxy.torontopubliclibrary.ca/library/publisher/addison-wesley-professional/),

March 2019

This is a fabulous resource with which I gleaned a tremendous amount of information about using Python to extract Twitter data. An entire chapter is devoted to Twitter API’s with specific reference to the easy to use Tweepy API. Links are provided to Github repositories that are updated and useful.

An introduction to n-grams and the context with which they are used is also discussed. The most valuable part of this book is the discussion of statistical calculations and how they can be used in comparing different corpus of text.

Bensalem1, Imene Chikhi, Salime Rosso Paolo “*On The Use Of Character n-grams As The Only Intrinsic Evidence Of Plagiarism*”, Lang Resources & Evaluation (2019) 53:363–396 Published online: 31 January 2019 Springer Nature B.V. 2019 <https://doi.org/10.1007/s10579-019-09444-w>

One of the main uses of n-grams is for author authentication. Assuming there is some prevalence to the term Fake News, without trying to quantify to what extent it exists, it would be relevant to consider Twitter within this context. This paper attempts to identify the use of n-grams within the context of plagiarism.

Twitter has a Retweet option which allows people to piggyback off of what other people are discussing. The data cleaning employed in my analysis omits these entries. Time is spent looking at lexical diversity so it is easy to analyze if people are retweeting the same things and taking ownership of what they are saying as opposed to simply Retweeting. Also, the detection of Twitter Bots tweeting out the same information can also be considered.

Held, Leonhard, Schwab, Simon “Improving The Reproducibility Of Science”, Significance, Feb 2020 Volume 17 Issue 1

This is a great article that discusses the collection and processing best practices of good data science and statistical research. What was of particular interest for me from this article were the ways to avoid unreliable results through “Repeated Selection Of Data” and “Selective Reporting.” There were key things that I wanted to avoid when collecting the data for my experiment.

Rashel Rana, S.M. “Location Based Popularity Analysis Off Twitter Data”, Ryerson University, 2015. <https://digital.library.ryerson.ca/islandora/object/RULA%3A3661>

This is a masters thesis written to examine location based twitter data with respect to sentiment and phrase popularity. Numerous different applications are created and combined in an attempt to achieve this task drawn from different platforms, including Google Maps.

One of the key takeaways from this work were the accuracy with which it produced results in congruent with the final Toronto Municipal Election Results of 2014. Another very important factor gleaned from this paper was that, despite the authors valiant attempts to link twitter data with geographic location they still ran into the fact that location based twitter data is very difficult to obtain. This was insightful information to proceed with.

Sidorov, Grigori. *Syntactic n-Grams in Computational Linguistics*. Springer, 2019.

This was a fabulous place to start about learning about n-grams. It takes the reader on a journey through the main theories and applications of this lexical tool. A primary focus is placed on the application of Authorship Attribution with respect to Syllables and Punctuation. Further attention is paid to Types of Syntactic n-grams according to their components.

What this book did is help break down the technical jargon that makes up the components of an n-gram to identify what we are looking for. Lexical elements, for example “words” are what we are looking for in they hypothesis of the different regions use different words separately. Tags of Syntactic Relations and the use of Characters help test the the main hypothesis with respect to these linguistic components as well.

[Wasser](https://www.earthdatascience.org/authors/leah-wasser), Leah, [Morrissey](https://www.earthdatascience.org/authors/martha-morrissey), Martha “Use Twitter Social Media Data in Python - An Introduction”

[Earth Lab at University of Colorado, Boulder](http://www.colorado.edu/earthlab). September 11, 2020

<https://www.earthdatascience.org/courses/use-data-open-source-python/intro-to-apis/social-media-text-mining-python>

This is a fabulous resource that provided many inspirational tips on indexing and working with twitter data. It provided many easy to follow examples with respect to cleaning and transforming data within its many forms. A tremendously valuable free resource, even though it did not make use of the Tweepy API directly.

# Dataset

The tweets were individually extracted using the Tweepy API after permission keys were provided by Twitter Inc. using the Anaconda distribution of Python.

The search terms with a correction for retweets was conducted us the words “Toronto” and “Vancouver”. Because “New York City” and “San Francisco” are composed of multiple syllables, extra precautions were taken so that parts of these names were not double counted.

Each query generated a series of at least four thousand tweets, with the exception of Toronto. In this case more then twelve thousand tweets were scraped. They were exported and saved into excel files and then imported back into Python for further analysis. The time of the tweets was also recorded to prove they were within the same time period, which was literally seconds in most cases.

# Approach

* **Step 1:**  **Extract Data using Tweepy API And Exported It Into Excel**

After gaining proper permissions from the Twitter Corporation to have access to the developer keys, basic search queries were used taking into consideration factors already mentioned early with respect to retweets. The data was visually inspected to see if it logically made sense.

* **Step 2: Returned Data Back To Python For Further Cleaning**

Using some of the tools mentioned in the bibliography, the time stamp column was removed as all the tweets were within a couple of minutes of each other and it was no longer relevant for analysis. This was broken down into the following series of steps.

1. Using the re(regular expression) function the special characters, symbols, emoticons and urls were removed.
2. Ii. Then, the data was transformed into a list. All the letters were adjusted to lowercasing and the words for each tweet were individually split up.
3. All Stop Words were removed.
4. All Collection Words were removed

As part of the cleaning process the individual numbers were left alone and some people them as part of short form communication. Eg b 4 u leave = Before you leave

* **Step 3: Basic Word Frequency Methods Used**

The first 1000 most frequently used terms were examined to see if there was any new individual words being used. Furthermore, the data was transformed into bar charts to see how the effects of removing the Stop Words and Collection Words affected the most frequently used terms in the entire collection. As these changes were taking place, simple calculations of how much of the corpus was actually not duplicate(ie. Unique) were preformed for observation as well.

* **Step 3: Basic Advanced NLP Packages For Deeper Analysis**

At this stage the nltk library was imported and the bigram and trigram features were employed to further examine the corput for commonly occurring two and three word phrases.

A second Anaconda notebook for each of the four queries was created and the data the same files were originally reimported. Any word over 100 characters was removed and the file was

transformed into a text file so that it could be read by the Textatistic package for further analysis and testing**.**

* **Step 5: Final Comparison Analysis Done Addressing Hypothesized Questions**

Some of the key test results and calculations were aggregated and placed into tables for referential examination across the regions. Some of the lesser knows results from Textastistic package were further researched for their meaning and relevance according to topics being addressed.

A very quick internet google search was preformed based on some of the results that appeared out of the ordinary in order to come up with a quick explanation about what may have occurred.

# Results

The following graph shows the readability scores when comparing three of the four data files. The reason why Toronto was omitted from the graph was because it had about three times the amount of tweets and the other three cities. Given that the Vancouver tweets have slightly higher Gunningfog and Smog Scores, more multi syllables words a relatively lower Flesh Score when looking for new phrases or words, that would be the first data set to look at for new words of phrases.

The table below provides a full overview of all the scores that the Textastistic provided.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table1: Testastistic Library Results** | | | | |
| **Characteristic** | **San Francisco** | **New York** | **Vancouver** | **Toronto** |
| Char Count | 103583 | 83425 | 113363 | 280964 |
| Word Count | 17998 | 16417 | 19125 | 48303 |
| Sent Count | 1768 | 1483 | 1852 | 5073 |
| Sylb Count | 26609 | 21828 | 29859 | 68246 |
| Notdalechall Count | 8555 | 6319 | 8637 | 20110 |
| Polysyblword Count | 1834 | 1156 | 2919 | 4564 |
| Flesch Score | 71.426 | 83.115 | 64.271 | 77.642 |
| Fleschkincaid Score | 5.826 | 4.417 | 6.860 | 4.795 |
| Gunningfog Score | 8.148 | 7.245 | 10.236 | 7.588 |
| Smog Score, | 8.947 | 8.173 | 10.301 | 8.548 |
| Dalechall Score | 11.647 | 10.263 | 11.280 | 10.683 |

If one further looks at the table below, it becomes evident that the San Francisco and New York tweets were gather at the same time. Furthermore, that Vancouver And Toronto tweets were collected at the same time as well. It is also interesting to note that the three twitter files that are roughly the same size have approximately the same proportion of words that are unique and it is much higher then the Toronto corpus which is much larger. Another important thing to consider is that the multi syllable cities can often come in the form of abbreviations, that could possibly cause reference words to them in their tweets to skew the final results.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table2: Basic Tweet Metric Aggregates** | | | | |
| Characteristic | San Francisco | New York | Vancouver | Toronto |
| Number of Tweets | 4415 | 4406 | 4416 | 12540 |
| Collect Start Time | 2021-06-09 23:59 | 2021-06-09 23:59 | 2021-06-10 23:59 | 2021-06-10 23:59 |
| Collect End Time | 2021-06-09 0:00 | 2021-06-09 0:00 | 2021-06-10 0:00 | 2021-06-10 0:00 |
| Length Full Tweet | 101330 | 96866 | 99569 | 286437 |
| Unique Words Full Tweet | 15816 | 14003 | 16270 | 32805 |
| % Full Tweet Unique | 5.6% | 14.45% | 16.34% | 11.45% |
| Length Tweets Minus  Stop&Collect Word | 61313 | 53246 | 61224 | 175428 |
| Unique Tweet Words  Minus Stop&Collect | 15684 | 13869 | 16134 | 32667 |
| % Tweets Unique  Minus Stop&Collect | 25.58% | 26% | 26.35% | 18.62% |

# Conclusions

Hypothesis One: Using Natural Language Processing techniques such as n-gram analysis it would be interesting to identify of new phrases and idioms are being used in each region.

The construction of n-grams(bigram or trigrams) will provide a response to the first hypothesis of whether or not new phrases or idioms are showing up associated with each selected city. Code block 57 provides the frequency of the top 500 bigrams across the four workbooks. Nothing seems to appear out of the ordinary except for the most popular bigram in the Toronto coding of

'gospel + chh'. This appears 200 times. Upon further inspection it appears that this is being picked up as part of the ntlk package and it is included in a hashtag and the bigrams package is not editing out.

To further search for any other unusual linguistic terms the trigram(block 66)

code needs to be examined. It is interesting to note that the bigram term in the prior paragraph is not to be found. Everything appears to be quite normal except for the top five trigrams in the Toronto book that are different combinations of the words Toronto, Russia, Embassy. On that day there was an announcement of the reunification of diplomatic relations between the United States and Russia according to the second link below. [[1]](#footnote-1) As a result there may have been an increase in spam bot activity promoting this. Aside from this there does not appear to be any other terms or phrases naturally using the query search methodology employed.

Hypothesis Two: Taking into consideration that users can choose to access this platform on a mobile app or web browser, further attention will be paid to identify to what extent the significance of use of text abbreviation or short-message-service acronyms are being used.

The information needed to determine the accuracy of this second hypothesis can be found by looking at code block 22 across the four python workbooks. This contains the frequency of the most common 1000 phrases in the tweets that were scraped, Everything appears to be quite grammatically fluid as even the less frequently used words are spelt in full. There could be several reasons for this. The most obvious one is that twitter is not a direct messaging app such as Whatsapp, Skype or Facebook Messenger. Another reason could be that some of these terms were eliminated when cleaning the data using the regular expression function.

Also, more people may be accessing the twitter platform at home due to the lockdowns. With less concerts and sporting events there is fewer opportunities for people to take photos and selfies and use the mobile app. Also, if more people are staying at home then they may be discussing matters of social and political discourse that require more thought and thus may want to refrain from using words that may bring into question my their credibility. At this point we don’t have data indicating whether a tweet was sent using the app or a website so these inferences mentioned previously in this paragraph represent my own attempts at logical reasoning.

There was clearly a problem with the Toronto search results. They were much larger then the rest of the search files. A bigram analysis indicated that the most popular bigram was linked to an unremoved hashtag hashtag that may have been related to bot retweets. A new Toronto search query was not conducted for two reasons. The first reason is that in the trigram analysis using the same cleaning formulas and stops it did not appear. The second reason is for fear of over mining and compromising the objectivity of the research.

Going forward, in the future streaming APIs could be tried to get more real time results. Other APIs aside from Tweepy could also be employed to see if more can be done to narrow down the geographic location of where the tweets are being sent from as well to see if a more accurate search proxy for location can be used.

1. https://www.cnn.com/2021/06/10/politics/biden-putin-ambassadors/index.html [↑](#footnote-ref-1)