**Building a Text Dataset**

**Motivation**

Of all of the hashtags being used today, one of the most interesting to me is #okboomer. While this response is likely one of the least aggressive possible, it is full of passive aggressive dismissal. This (supposedly) millennial-driven come-back to being called “snowflake” is prohibitive to meaningful discussion about modern barriers to success, but it is also super entertaining. My goal in choosing this dataset is to see what attributes tweets containing this hashtag might carry.

**About This Dataset**

The sample of tweets I pulled from the Twitter API contains 2,780 tweets from 2,507 different authors that all include the hashtag “okboomer.” These tweets were created between 5:05 am on 12/3/19 and 12:50 am on 12/8/19. Table 1 shows that the largest number (43% of total tweets) were located in the USA, the second largest group (41% of total tweets) was created without a specific location, and the rest (16% of total tweets) were listed as in another country. The attached appendix contains a table of the 50 most common tokens from the gathered tweets. Figure 1 below shows that the word count decreases greatly as you go down the list for the top 10 most common words, and then levels out from there . Of interest are the words old, generation, age, and millennials. These tweets combined have a lexical diversity of 3.39.

Table : Tweet Count by Location

|  |  |
| --- | --- |
| **Location** | **Count of Tweets** |
| N/A | 1146 |
| Other | 448 |
| USA | 1186 |
| **Grand Total** | **2780** |

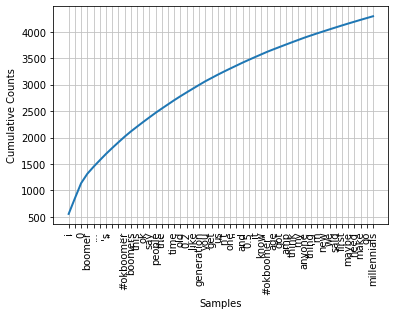


Figure : Token frequency cumulative counts

**Reflection**

As with all of my classes this semester, having a more solid understanding of Python would have greatly helped my process. While I am able to Google things until the it works, it would be great to be able to write at least half of the code from scratch myself.

Extending the previous point, if I were to continue this project in the future, I would like to refine the cleaning segment of code as it feeds into the file writing segment. In general, having a better understanding of the language would help me be more efficient, and understand how to optimize my files.

Appendix

|  |  |
| --- | --- |
| **Item** | **Count** |
| i | 556 |
| ' | 294 |
| 0 | 286 |
| boomer | 178 |
| . . . | 132 |
| " | 123 |
| 's | 120 |
| " | 112 |
| … | 108 |
| #okboomer | 106 |
| boomers | 98 |
| this | 91 |
| ok | 90 |
| say | 87 |
| people | 86 |
| the | 81 |
| `` | 79 |
| time | 79 |
| old | 75 |
| 0.2 | 72 |
| like | 71 |
| generation | 70 |
| you | 69 |
| get | 62 |
| us | 62 |
| n't | 59 |
| one | 57 |
| '' | 56 |
| and | 56 |
| 0.5 | 53 |
| item | 52 |
| know | 51 |
| #okboomer" | 51 |
| age | 46 |
| got | 45 |
| amp | 45 |
| think | 44 |
| my | 44 |
| anyone | 43 |
| thing | 30 |
| 'm | 39 |
| new | 38 |
| we | 38 |
| said | 37 |
| first | 37 |
| maybe | 36 |
| need | 35 |
| make | 35 |
| go | 34 |
| millenials | 34 |