**Corpus Building Project: Perspectives on COVID-19**

**Project Report**

**Background**

The purpose of this project is to build a series of corpora that will enable research on a **distinct perspective on** **COVID-19** using the computer assisted text analysis tools we have been taught in this course. Keeping this in mind, I elected to build my corpus focusing on discussion happening within three political subreddits on the website Reddit, focusing on discussions happening during the period of the COVID-19 pandemic. My primary corpus was built entirely from data scraped from the r/conservative subreddit. I chose r/conservative (a predominantly right leaning subreddit) as my primary source as I was particularly interested in analysing conservative discussion surrounding the coronavirus pandemic and government response, including how that discussion shifted during key political events. To supplement this primary corpus, I built 2 comparable subcorpora, collected from the r/libertarian and r/socialism subreddits respectively. I chose these two subreddits as I believe they provide two distinct perspectives from those presented on r/conservative. I believe all three of these subreddits to be appropriate sources to draw upon as all three have large membership bases with frequent user contributions. r/conservative is the largest with over 850k members, r/libertarian the second largest with 490k members, and r/socialism the smallest with 350k members.

I believe my primary corpus, in conjunction with my two reference corpora can be used by researchers to analyse shifting public sentiment within the r/conservative community, in comparison to other political subreddits. I also built my corpus in such a way that it is possible to make these comparisons in specific time periods where events of interest occurred. The memberships of these subreddits (and reddit as a whole) has a strong American presence. As such, I believe my corpora will be most suitable for research surrounding American political discussion. It is important to realise that any research performed using these corpora will not be without their limitations. I will expand on these in a later section.

**Corpus**

My primary corpus (built from data scraped from the r/conservative subreddit) contains 50 documents, 11152 word types, and 132708 word tokens. My first subcorpus (built from data scraped from the r/libertarian subreddit) contains 25 documents, 9315 word types, and 101308 word tokens. My second subcorpus (built from data scraped from the r/socialism subreddit) contains 25 documents, 10631 word types, and 107594 word tokens. I believe the size of these corpora to be of an appropriate size to perform the type of research detailed in the background section.

All of my corpora were built using **comments only**. This was a deliberate design choice on my part as, even though posts on these communities occasionally contain sizeable text data to scrape, more often the posts were non-text based and provided an inconsistent source of suitable corpus data.

To build all of my corpora, I laid out a strict timeframe from which all my data would be drawn from. The timeline begins on the 21st of December 2019 and ends on the 21st of August 2021. This gave me a total of 20 months to draw data from and work with, from which I gathered 2 documents each containing approximately 2.5 word tokens. In addition to this, I elected to draw more data from certain months of interest. The months of interest, and their justifications are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Month | Dates | Documents | Events of Interest |
| 2 | 21/1/2020  - 21/2/2020 | X4 | First U.S coronavirus vase reported **January 20th** |
| 3 | 21/2/2020 –  21/3/2020 | X4 | COVID-19 declared pandemic **March 11th** , U.S state of emergency declared **March 13th** |
| 6 | 21/5/2020 –  21/6/2020 | X4 | Operation *Warp Speed* announced **May 21st** , U.S passes 2 million cases **June 10th** |
| 11 | 21/10/2020 -21/11/2020 | X4 | 100k cases on a single day **November 4th** , U.S presidential election **November 5th** |
| 20 | 21/7/2021 –  21/8/2021 | X4 | 5th Coronavirus wave in the U.S, Biden’s first presidential crisis (U.S withdrawal of Afghanistan) |

This was the timeline used for building my primary corpus (r/conservative). The dimensions for my two subcorpora were halved, i.e one document gathered per month apart from the 5 months of interest where two documents were gathered.

There are several word tokens present within all these corpora that I believe could be of particular use to any researchers working with this data. These include *covid*, *pandemic*, *China*, *Trump,* and *lockdown*. By performing collocation analysis on these search terms, researchers could accrue some level of sentiment towards these terms by analysing the words that appeared alongside these word tokens frequently. Researchers could also use keyness analysis to determine which word tokens were more predominant on the r/conservative subreddit.

**Method**

The main resource I had at my disposal was the Pushshift API, more details of which can be found [here](https://github.com/pushshift/api/blob/master/README.md). Reddit does provide its own API to use for data gathering purposes but, after experimenting with this and finding it less than ideal, I settled on Pushshift. Pushshift provides enhanced functionality and search capabilities for data scientists collect text data from reddit. The script I wrote to run Pushshift was rather simple and can be found as a separate file within this submission (titled ‘pushshift\_scraping\_script’). The script takes on the parameters start date, end date, subreddit, and then returns the desired number of comments. I simply changed these parameters in line with the dates outlined in my timeline and to gather from my desired subreddits.

Figuring out the appropriate number of comments for each document required some experimentation. First, I tried gathering 1000 comments per document. This returned on average 22k word tokens, which I thought would be too large for a corpus of 50 documents. After some more experimentation, I eventually settled on gathering 100 comments per document which returned approximately 2.2k word tokens.

Considering the number of documents that I would be producing (100 total), I deemed it necessary to be very strict with my naming conventions. I settled on the template ‘r\_subreddit\_month\_version’, so the third 100 comment document collected for the month of 21st of January 2020 – 21st of February 2020 from the r/conservative subreddit is named ‘r\_conservative\_2\_3’.

I also experimented with the exact date parameters for my Pushshift API script. In the timeline listed above, I detail the months of particular interest to my research, and how many copies I wanted to gather from each month. Initially, I set the date parameters to cover the entire month (21st -21st) and then gathered 4 copies using these dates. However, upon data inspection, I realised that this method produced significant duplication in corpus data. To counter this, I compartmentalised the timeline even further, breaking the months down into roughly week by week sections. This means that in the months of particular interest I would gather four documents, using the following time periods: 21st – 28th, 28th – 5th, 5th-13th, and 13th – 21st. In the other months I would use two time periods: 21st – 5th, and 5th -21st. These dates would change to compensate for the different lengths of months. This method fixed my issue of duplication and had the added benefit of allowing even more specific time periods for any research performed on this data.

**Limitations**

As mentioned earlier, there are some significant limitations inherent in this dataset that must be taken into account by any prospective researchers. First and foremost, while the data in my primary corpus and reference corpora may be gathered from specific political communities, there is no way to verify that the commenters on these subreddits actually subscribe to the political ideology of said community. In my research into these three subreddits, I saw significant evidence of ‘brigading’. This term refers to commenters who clearly do not subscribe to the political ideology of the subreddit they are commenting in but are commenting anyway in the interests of discussion and/or argument. Another limitation to this data that must be addressed is the demographics of reddit as a whole. A 2019 survey found that 64% of users were between the ages of 18 and 29, and another 29% were between the ages of 30 and 49. In addition to this, users were found to be as high as 69% male[[1]](#footnote-1). Clearly, then, any data gathered from reddit will not be drawn from a representative sample. The issues of low demographic diversity alongside the issue of brigading mentioned earlier does somewhat reduce the veracity of the data I have gathered here. Despite this, I do still believe that the corpora I have gathered do still possess some utility for researchers who wish to analyse discourse across political subreddits during the period of COVID-19. I am looking forward to doing my own research on these corpora during the next phase of this assignment.

1. https://www.alphr.com/demographics-reddit/ [↑](#footnote-ref-1)