CSC110 Project Proposal: An Investigation of the Political Satisfaction of the General Public over the COVID-19 Pandemic

Aviraj Newatia, Mishaal Kandapath, Rudraksh Monga, Taylor Whatley

Friday, November 5, 2021

Problem Description and Research Question

TODO

Dataset Description

The dataset we will be using was sourced from the online machine learning data repository Kaggle, a subsidiary of Google LLC, which serves as an online community of data scientists and machine learning practitioners.

It is a csv (comma-separated values) file of 17,777,472 unique comments related to covid that were collated from the online forum platform Reddit. The dataset is called The Reddit COVID dataset [https://www.kaggle.com/pavellexyr/the-reddit-covid-dataset].

We transformed this dataset by filtering by the subreddit the comment came from, building a sub-dataset of comments only from political subreddits such as r/uspolitics.

The dataset was too large to load into Excel, so we filtered the first 2 million comments using pandas in Python to yield 32,335 comments.

The dataset has 10 columns, of which we extracted and will be using 2 which are the actual text of the comment ("body") and the timestamp of creation/posting ("created_utc").

	type	id	subreddit.i	subreddit.i	subreddit.i	created_utc	permalink	body
59	comment	hi1verl	2cneq	politics	FALSE	1635206220	https://old.r(	Once my kid can get the
172	comment	hi1uqe2	2cneq	politics	FALSE	1635205901	https://old.r(	Here's two decent little
267	comment	hi1u8wg	2cneq	politics	FALSE	1635205676	https://old.r(	Because everyone else go
336	comment	hi1tslf	2cneq	politics	FALSE	1635205465	https://old.r(	I see a lot of covid republ
374	comment	hi1tmv6	2cneq	politics	FALSE	1635205392	https://old.r(	Because his tone changed

Computational Plan

TODO

References

TODO