David Gutierrez
CS410: Text Information Systems
Fall 2020
November 29, 2020

Course Project Progress Report

Individual team (NetID: davidmg4)

1) Progress made thus far:

I have successfully applied and been approved for a Twitter developer account. This has enabled me to pull all of the most recent tweets in a search for the previous ten days (instead of the prior year as I had hoped). I have been able to implement code to pull the text data into a CSV file. The resultant data file is then read into a second script that loads it into Pandas data frames and parses the text data using stop-words from NLTK and pulling a sentiment analysis for each tweet using TextBlob. Zero values are dropped and then the remaining sentiment scores are averaged together by state and topic to produce the final values, visualized as percentages. This data is then dropped into an HTML file that features a map of US states as well as a sortable table for each state in addition to a composite score to gauge overall sentiment across the five topics.

2) Remaining tasks:

       a. Re-run the Twitter scraper to get the full text of tweets (currently truncated at 140 characters)

       b. Fully implement dataViewer.html to pull data automatically from the CSV file

       c. Rework the CSS on the same to enable a visual histogram for each topic when hovering over the map for a particular state and color enhancements for numerical values

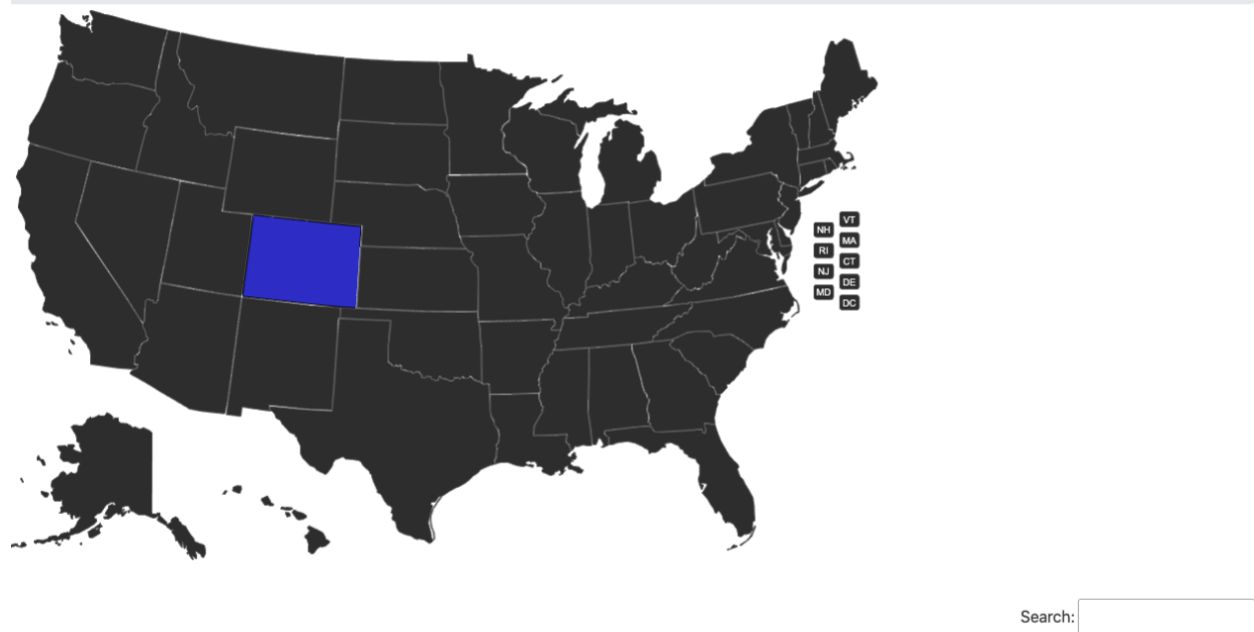   d. Code cleanup and integration into a single workflow

3) Any challenges/issues being faced:

       a. Time series data limited to prior 7-10 days

       b. API limited to 15-minute intervals (takes all day to collect five topic-words for all 50 states)

       c. Spam/bot usage is rampant on Twitter

       d. Overlap in states/topics (including a gross over-weighting of political data due to residual election data)

       e. Location data also severely limited and often inaccurate

       f. Limited data for less populated states

       g. Some topic words are insufficiently measurable (quality of life) or inherently biased (e.g. pollution)

state	lifestyle	quality of life	pollution	government	weather	total
Alabama	-0.038593074	0.103787879	0.33443526	0.06359846	0.14125135	0.120895975
Alaska	0.195849784	0.139393939	0.30078125	0.18654125	0.04625597	0.173764439
Arizona	0.284974166	0.375833333	0.39172619	0.14308267	0.23329426	0.285782124
Arkansas	0.150906385	0.020138889	0.25511815	0.08608419	0.13626968	0.129703459
California	0.279154153	0.171415344	0.14353878	0.06968667	0.17894849	0.168548687
Colorado	0.256762777	0.223778999	0.22362689	0.12952935	0.17125236	0.200990075
Connecticut	0.171590909	0.314141414	-0.0045455	0.13074179	0.11579444	0.145544611
Delaware	0.35952381	0.233333333	0.1	0.01087788	0.35861014	0.212469033
Florida	0.145561274	0.24743266	0.20764209	0.03976945	0.19905182	0.167891459
Georgia	0.150653727	0.277604167	0.2374868	0.20462724	0.13502976	0.201080339
Hawaii	0.271198593	0.35	0.27326389	0.14411322	0.15608345	0.238931831
Idaho	0.3375	nan	nan	0.03511349	0.13303641	0.10112998
Illinois	0.444444444	0.079166667	0.23502747	0.05486355	0.2188336	0.206467146
Indiana	0.102582645	0.209375	0.25738508	0.06051745	0.12351845	0.150675725
Iowa	0.174456727	0.372222222	-0.5	0.09831056	0.11394067	0.051786036
Kansas	0.094491342	-0.025568182	0.36413813	0.00638822	0.13583832	0.115057566
Kentucky	0.251923077	0.025992063	nan	0.05969398	0.15560239	0.098642302
Louisiana	0.13030303	-0.011060606	0.13712121	0.16676449	0.10752152	0.106129929
Maine	0.252247475	0.046428571	0.39014881	0.07940122	0.13205464	0.180056143
Maryland	0.340791126	0.4	-0.2066667	0.05821282	0.23312024	0.165091497
Massachusetts	0.29743266	0.596590909	0.15	0.06211873	0.07782118	0.236792696
Michigan	0.224787415	0.178869048	0.40208333	0.11253065	0.08554799	0.200763687
Minnesota	0.153396104	0.375	0.16982323	0.09022538	0.17124012	0.191936967
Mississippi	0.139015152	0.125	0.01527778	0.07990588	0.18634358	0.109108478
Missouri	0.3	nan	0.12944444	0.08386878	0.06770248	0.11620314
Montana	0.075127324	0.5	0.15	0.18550224	0.15107065	0.212340043
Nebraska	0.233928571	nan	0.46857143	0.04883183	0.11547421	0.173361208
Nevada	0.312317621	0.35	0.05960884	0.08662706	0.21419744	0.204550192
New Hampshire	0.136363636	nan	0.23770643	0.13310231	0.1771875	0.136871975
New Jersey	0.040909091	0.132575758	0.17594066	0.09921351	-0.039263	0.081875204
New Mexico	0.199896694	nan	0.14355374	0.13235972	0.12113809	0.119389649
New York	0.153218622	0.139308428	0.06559096	0.10017699	0.12145422	0.115949844
North Carolina	0.001851852	-0.277777778	0.43105742	0.1058544	0.23483246	0.099163671
North Dakota	0.164393939	nan	0.51041667	-0.0787578	0.08803436	0.136817434
Ohio	0.137540608	0.091666667	0.20897968	0.06535248	0.13966798	0.128641483
Oklahoma	0.157692308	-0.225	-0.1609848	0.09847333	0.06302151	-0.01335953

Sentiment Analysis of American States by Topic
Based on Tweet Data from 11/21/2020 to 11/28/2020

Search:

State ▲	Government	Weather	Pollution	Quality of Life	Lifestyle	Overall Sentiment
Alabama	+6.36%	+14.13%	+33.44%	+10.38%	-3.86%	+12.09%
Alaska	+18.65%	+4.63%	+30.08%	+13.94%	+19.58%	+17.38%