Final Exam: Take Home

Assignment: This exam is unusual. For this exam, you will need to collect data from the internet using the web scrape approach described in class for gathering news and information from websites.

You are not allowed to talk with classmates about this exam. You can use all course material and material you might find online. There will be a Q&A on Monday, Aug. 3rd. Please send any written questions to myself before the Q&A.

Data File: Use Newspaper3K and NewAPI for web scraping articles. You can build your data using one of two option: the 2020 Presidential election new, or the Covid-19 news. Please collect 1,000 or more news articles using both Newspaper3K and NewsAPI. Here are your suggested search words using the default "or" search.

SEARCH WORDS (or):

Option 1: [trump, biden, democrats, republicans]
Option 2: [covid, coronavirus]

Save your articles into excel files. You will need to merge these files. Newspaper3K automatically inserts the agency name into your DataFrame. NewsAPI does not, and it uses different names for the same agencies. Please put the agency names in the data collected from NewsAPI.

Note: you will need to obtain your personal API Key from NewsAPI. Go to their website http://newsapi.org/pricing and register as a developer to obtain your own free key. You only get one, and it is limited to a maximum of 300 searches/day. You may need to collect data over a 2 or 3 day period to obtain a total of 1,000 articles. If you create several files, you will need to merge all of them into one excel file. If you would rather use another filetype, that is fine, but CSV and tab delimited files are problems with news articles that contain commas and tabs.

Once you have collected your data, count the number of search words that appear in each article. For your articles, tally the number of times "Trump", "Biden" and "Fauci" each appear in each article.

Now conduct a sentiment analysis for your articles. Plot the average sentiment versus the number of times each Trump, Biden and Fauci appear in the article. For each agency, report the average sentiment and the average number of times they use the names Trump, Biden and Fauci.

Lastly, conduct a topic analysis for your articles. Be careful to filter stray html terms and punctuation. For each agency, report the proportion of its articles that fall into your topic categories.

After the data are collected, the entire analysis can be done in SAS EM or Python, it's your choice. Please return the following:

1. Describe which data you collected (Covid or Election). Describe how many articles you were able to collect from news3k and from newsapi.
2. The python file you used to collect, merge, preprocess and store data. (preprocess refers to calculating the number of times the 3 actors appear in each paper, and determining the agency for data from NewsAPI. It might include more, but that's your choice.
3. Screen shots of your SAS EM work, if any. This would be a picture of your main diagram, and your property windows for the parse and filter nodes.
4. Please report and describe the average sentiment for each agency, and the scatter plot of the number of times each actor appears in an article versus the average sentiment.
5. Report a word cloud for each topic to help describe and characterize the topics.
6. Lastly report the proportion of articles for each agency that fall into each of your topic categories.
7. A general discussion of your impressions from your analysis. Is there anything surprising that you saw? For example, do your analytics say anything about how the agencies group? Do they say anything about the average sentiment for each agency? This discussion can be written or a video/audio recording.