Data Science 311

Lab 3 (10 points)
Due at 10am on Oct. 19, 2022

Read all of the instructions. Late work will not be accepted.

Overview

In this lab you will use the Spacy NLP package to analyze the fake and real news datasets from a Kaggle competition that we have been working with in class. You can use any of the data processing code presented in class for this lab.

Collaboration

For this lab, you are encouraged to spend the lab period working together with a partner. Together means synchronously and collaboratively: no divide an conquer. After the lab period ends, you will work independently and submit your own solution, though you may continue to collaborate the same partner if you wish. Your submission must acknowledge which person you worked with, if any, and for what parts of the lab (this should be included as a statement at the top of your notebook).

Details

Data

The data you will be working with are contained in two files Fake.csv and True.csv. You can download them from here:
`https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset`

Tasks

In this lab you will get aquainted with using Spacy to uncover meaningful insights from the two text corpora contained in True.csv and Fake.csv. In a single notebook titled `lab3.ipynb`, perform the following analyses on ***both*** the True.csv data and the Fake.csv data.

1. Using news article titles, provide a list of all the countries that are mentioned in the month of March 2017. Briefly comment on your findings.
2. Using the news article titles, calculate the top five people mentioned on election day November 8th, 2016? Briefly comment on your findings.
3. Create histograms of real and fake article sentiment mentioning the top person (from the previous analysis). Briefly comment on your findings.
4. Plot histograms for real and fake news article sentiment (by title) for a 2 month period of your choosing. Briefly comment on your findings.
5. Perform 1 additional analysis of your choosing on some date range/date of your choosing. Justify your analysis choice, and comment on your findings.

6. Time running a Spacy language model on 5 full text news articles. Use back of the napkin math to calculate how long it would take to run the language model on the full text of one week of articles. Briefly comment on your findings.
7. Perform analyses 2-5 on the full text of a single week's real and fake news articles. Are there any differences between the two corpora discernable from these analyses? Discuss any limitations of these analyses to uncover differences between the two corpora. Suggest additional analyses that might uncover meaningful differences, between the two corpora and salient qualities of each corpora.

Submitting Your Work

Please zip your lone ipynb file (using the naming conventions stated above, where spelling, spacing and capitalization matter) and upload the zip via Canvas.

Grading

Analyses will be graded on the following:

- 50% of total grade will be on analyses 1-6.
- 50% of total grade will be on analysis 7.

For individual analyses:

- 80% based on the quality of the analyses, including both correctness of and readability of the code to conduct the analyses.
- 20% based on the clarity and completeness of presentation of findings from the analyses.