MATH 189Z Homework 2:

Topic Modeling of COVID-19 Tweets Over Time (and PCA)

In this homework, you will be using topic modeling to identify topics in tweets about COVID-19 collected from February through April. You will use PCA to visualize the results of your topic modeling as well as investigate the semantic meaning of the topics you have found and the prevalence of these topics over time. The provided Jupyter Notebook will guide you through all of the assignment but, like last time, your deliverable will be a document. The requirements for each task are outlined below. When you have completed your assignment, upload your files to your GitHub repository in a folder clearly labeled Homework 2.

Deliverables:

Task 1: Your list of stop words

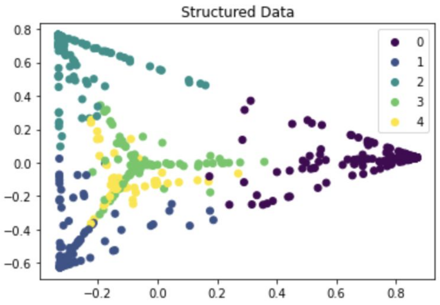
Tasks 2 and 3: An image of your PCA results on the tweet topics (not the example for Structured vs. Unstructured data). In addition to the image, discuss what this graph shows us about the topics (are they more structured or unstructured and what does this mean?). Remember not to use any built in PCA functions! If you are unsure if your PCA is functioning properly, check out the provided graphs at the bottom of this document to compare.

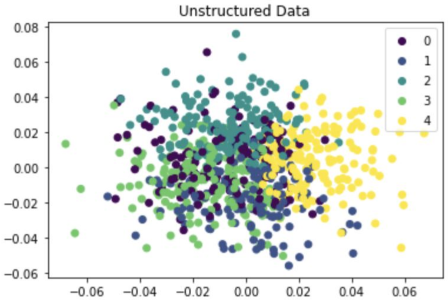
Task 4: Image of 'Monthly Tweet Counts by Topic' bar graph

Task 5: Discussion of results (what you need to discuss is explained in Jupyter Notebook)

PCA Example:

This image is so you can check to make sure your PCA is working properly. If your image does not look similar to the ones provided, there's probably something wrong :)


Structured Data


Unstructured Data