Group 4 Capstone Project Design Document

GOALS:

Week 1 Goals (Week of 7/11/2022)

Long Term Goals: Finish Part One of google & Ethical Considerations Worksheet

Intermediate Goals: Complete Upwards of 40% of the powerpoint slides. Have upwards of 50% of the Readme.md file completed. Complete Phase 1 & 2 of the decision log.

Week 2 Goals (Week of 7/18/2022)

Long Term Goals: Completion of the entire capstone project (Phase 3 & 4)

Intermediate Goals: Completion of the Readme.md file & decision log. Complete any revisions that were suggested from the TA's.

ROLES:

Tobi – Project Lead, Colab Notebook File, & Design Document, Decision Log

Laila – Ethical Considerations Worksheet (Completed on 7/15/2022)

Jaden – Read.md File

Brianna – PowerPoint Presentation

PROBLEM SPACE

We have 2 challenges we need to complete. The first challenge is us isolating 3 newsgroup datasets of the original 20, and then proceed to display the top 100 most frequent words within those 3 newsgroup datasets. Our second challenge is omitting out certain words of the 3 newsgroup datasets and applying lemmatization to the datasets, and plot the optimal K. This is going to require a lot of skills, mainly the use of pandas and scikit learn commands. When we do plot the optimal K, will there be another more efficient way to effectively plot the optimal K? will there be a better method to have a more effective cluster?

Data Acquisition and Preparation

Google Colab Part 1 -

1. Acquiring the 20 news data sets with sickit learn.
2. Exploratory Data Analysis of the 20 news datasets
3. Isolate the three topics needed for the word frequency
4. Create a code to set the 3 isolated news datasets as targets.
5. Create the code to list the top 100 frequencies between the 3 news datasets.
6. Print the code for #5

Google Colab Part 2

1. Import all functions needed.
2. Test if a word would count as a token.
3. Strip these words out the corpus for the given topics and apply lemmatization.
4. Find the optimal K

5. Plot the optimal K.
6. Test with the optimal K again and display the top 3 sets of topics associated with each cluster.

Motivated Questions

1. Where in our model will the slope “fall off”?
2. Will our methodology create bias and how can we prevent it?
3. How will our findings help improve future research?