1. **Procedure**: Load the data into a pandas data frame and use the NLTK library for data preprocessing

Experimental setup:

1. Loaded data in a pandas data frame
2. Iterated through each cell in the data frame
3. Removed all data that weren't characters
4. Tokenized the strings using the NLTK library
5. Remove token if it is a stop word, which utilizes the NLTK stop words library
6. Change each token to be lower case
7. Lemmatize each token using the NLTK WordLemmatizer library
8. Append token to the list for the current row of data
9. Repeat for each row of data

2. **Procedure:** create multiple LSA and graph their coherence values and find the best topic model

Experimental Setup:

1. Using the cleaned data, generate a vocabulary for the corpus using the Gensim corpora library
2. Create a bag of words representation of the data using the Gensim library
3. Create the Tf-Idf representation using the bag of words model
4. Iterate through a predefined number of topics (2-11 topics)
5. During each iteration, create an LSA model with the current number of topics
6. Append the model to the list of models
7. Calculate the coherence score for the model and append it to the list of coherence scores
8. Once all iterations have concluded, plot the coherence values of each model using the Matplotlib library
9. Create an LDA model with the optimal number of topics determined by the previous process
10. For LDA models, calculate and print the log likelihood of each model
11. Print the top 10 topics for the LSA and LDA models along with the top 20 words from each topic

Questions:

1. It seems that the Tf-Idf representation works better for the LSA model, while the bag of words representation is optimal for the LDA model
2. Alpha and eta represent the correlation between words, topics, and documents in an LDA model. Alpha determines how many topics are in each document, a lower alpha value indicates that each document will be composed of a fewer number of topics. Eta on the other hand, dictates the amount of words per topic. A high eta value will create topics which contain a greater number of words. For my experimental setup, I chose

for both alpha and eta to be “auto”, so that the model will learn the values as it is being computed.