CS 410 Project Proposal

Captain: Yung Chieh Huang - ych10
Yuriy Kotskyy - ykotsk2
Sudhendu Sharma - sharma78
Nivedita Chatterjee - nc19

For our project, we have chosen Theme 3.4: System extension on the following unlisted Text Mining System approved by the professor and course staff: https://timan102.cs.illinois.edu/tms

The system is related to the class because it is a text mining and topic generation system for social science research that provides an easy way for researchers in social science to mine a large corpus of historical newspaper articles.

Our goals for this project are to enhance this project in several ways with an overall focus on intractability. We will be adding features to visualize existing topic generation framework that uses LDA for topic generation. If time permits we will enhance the system for Seeded-LDA Algorithm for topic generation.

Since the existing code base is largely written in Python and various libraries which extend its functionality, we will simply fork the existing code base and develop our functions and features within this fork. It will take us quite some time to understand such a complex system but once we do it will be quite easy to implement our features. Some of the libraries and tools which we will be using are flask, matplotlib and sqlite.

We have 4 group members so out workload is expected to be 80 hours:

Task	Estimated Time
Understanding the Code Base	20 hours
Topic Generation for user uploaded docs	20 hours
Time Range based Topic Generation for existing documents	20 hours
Visualization improvement (Based on suggestion from Bhavya)	20 hours
Seed Lda Implementation if time permits us(Optional)	20 hours