CS410 Fall 2022
Cristina Ross
NetID: cb37
Group Name: The Ross Factor

Course Project Proposal

For my project, I will be completing the project alone so I have included my name and NetID in the top left corner, but will include them below again for completeness. In being the single member of the group I will be the captain and will have all administrative duties.

Team Member Name	Team Member Net ID
Cristina Ross	cb37

The topic that I will explore is around Modeling using Healthcare Notes, more specifically I will be using both of the data mining and topic identification parts of the course to explore unstructured note data in the healthcare space to reliably generate features or predictors for use in machine learning applications. This is an important topic as the healthcare space has seen many advancements with the inclusion of unstructured note data in predictive models, improve quality measures for both patient and clinician alike. Most existing software packages generate thousands of features for each note (usually 1 or 2 word phrases), taking up very large amounts of space, resulting in data scientists spending multiple cycles using heuristics to attempt to reduce the feature space, hoping not to exclude an important word of phrase that may hold significance to the outcome of interest.

The goal of my project will be to create a package that can take the unstructured data from a collection of healthcare notes, use an existing or generate a new background model, identify a set of unique topics for each document, and test their usefulness using a simple model built for a known outcome. The approach will be as suggested below:

1. Build collection of healthcare notes in a readable format (i.e. avoid pdfs/image documents) that have a known outcome of interest, ideally the outcome will have at least a 10% prevalence in the collection
2. Build a background model with healthcare notes in general (could be from collection or other notes that don't meet the prevalence requirement or don't have outcome data but are a healthcare note)
3. Build out functionality to generate a new background model based solely on the input collection (needed if we are only looking at one specific type of healthcare note and a general background model is too general)
4. Build out functionality to identify unique topics for document, this step will likely involve existing packages such as scikit-learn's LatentDirichletAllocation decomposition package.
5. Build a simple linear or logistic model using scikit-learn's modeling packages to predict the outcome based on the topics identified for each document

6. Evaluate the ability of the model to accurately predict the outcome based on the topics selected to evaluate overall usefulness of topic selection

This project will be carried out using the Python language exclusively. I have included the time justification below based on the numbered approach items described previously.

1. 3 hours
2. 4 hours
3. 5 hours
4. 5 hours
5. 4 hours
6. 2 hours