Hailey Hultquist
Maria Fabiano
Geetali Tyagi

CS 391D Data Mining
Project Proposal

Statement of problem: We aim to use word embeddings created by transformers on electronic health records (EHRs) in conjunction with autoencoders for representation learning. Then, utilize the representation to predict the length of stay of a patient.

What you intend to do:

- Use logistic regression to create classification baseline results
- Fine-tune pre-trained embeddings from BERT with EHR data
- Use autoencoder to get a latent representation of EHR data
- Use latent representation from autoencoder in classification task

What data you will use:

- The MIMIC-III dataset contains over 60,000 rows for patients who stayed in the intensive care unit (ICU). The information is de-identified.
- The National Inpatient Sample (NIS) contains data on more than 7 million hospital stays (documentation is here: https://www.hcup-us.ahrq.gov/db/nation/nis/nisdbdocumentation.jsp).

Tentative timeline:

- March 13: Preprocess and clean data.
- March 27: Establish classification baseline with cleaned data. Learn BERT embeddings on EHR data.
- April 17: Have autoencoder representations.
- April 24: Use representations to begin classification task.
- May 1: Tune BERT embeddings and classification models/parameters.