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

1. Have you completed what you have planned?
   From the proposal, there were initially 6 steps proposed consisting of: (1) Building a collection of healthcare notes (2) Build a background model (3) Build out functionality to generate a new background model (4) Build out functionality to identify unique topics (5) Build a model to predict an outcome and (6) Evaluate the model. Of these 6 steps, step 1 proved to be very difficult and I was only able to procure a small subset of healthcare documents, some only partial notes, that clearly identified the diagnosis of the patient. With just a small subset the following steps 2-6 were able to be completed, but not with the accuracy that I was striving for. Steps 2, 3, and 6 were completed as proposed. Steps 4 and 5 were combined to develop a model that used word distributions from labelled note sets to predict whether a new note would be similar or not. With some noted changes, all parts were completed.

2. Have you got the expected outcome?
   The model that was trained with over 100 notes each for a septic and non-septic outcome performed extremely well on a small set of notes of the same type, which got me very excited! However, these notes are not publicly available and when using this same model on septic and non septic notes from publicly available notes the performance was terrible. Additionally, when using this model on the gout data set, which consisted of a quick blurb of nurses's evaluations to decide whether or not a patient had a final diagnosis of gout performed equally terrible. This was likely due to a very small word distribution for each "note", at less than 20-30 words each, there was more noise than signal.