Overleaf doc for Report:

<https://www.overleaf.com/3339694899nqqnqhtbqpgb>

Zoom Link

<https://ucla.zoom.us/j/3405126815>

Meeting 1/8/21 Notes

* Discussion video
* Maybe watch lecture 3
* Documentation for nltk
* Documentation for sltk.learn.count

Sunday (1/10/2021) @ 10:30 AM

Question 8 Notes

Question 9 Overview

* Document has to be represented by same number of features/dimensions
  + Representation should not be too high
  + Vector dimension <= 300 (Glove matrix 400,000 x 300)
* We cannot use TF-IDF, or any measure that looks at complete dataset
* Need to make each document 300 features at most
  + Use keywords/subject-line to determine features
  + Average the vectors of the keywords+subject-line to get a single vector that represents the whole document
* Convert each document into (# features x # reduced words) matrix after passing it through Gove
* Train classifier on this (# features x # reduced words) matrix
* Count vectorize
* Take every word in subject line and keywords (if they exist) -> take glove representation
* Add glove representations together and normalize

1/15 Meeting Notes

* Email TAs, go to discussion, office hours on Monday
* Edwin: #1,3, 6
* Connor: #2, 4
* Sarah: #5
* Look at question 12
* Meet Saturday 1/16 at 4:00pm

1/16 Meeting notes

* For loop that loops through documents and converts subject line and keywords into a vector of words
* Retrieve glove embedding of each word and add them together, divide by total # of glove embeddings found to get average
* Normalize the found vector by dividing by the max of the glove vector
* Append to a # of documents by 300 matrix
* We can train with this matrix

1/18 Meeting Notes

* Edwin: GLOVE embedding #9
* Connor: GLOVE classifier #10
* Sarah: pipeline (#7), Overleaf