Pyladies Prep:

Rachel & Sanjana -> talking about ML

Blaize -> Async

Machine Learning & Data Science:

Focusing on Enron email corpus.

Steps:

1. Reading the data from the csv files
2. Cleaning the data.
   1. Show example of the cleaner data
3. Storing the emails in the database
   1. Show structure of the database: sender, sender email, domain, etc
4. Filtering the content (removing stopwords/other unwanted stuff)
   1. Show example of cleaned content
5. Generate LDA Model
   1. Show a finished LDA Model
   2. Go over how there are different topics
   3. Show the pyLDAvis graph. Hover over the topics so they can see some of the example words
6. Yield Document Topic Vectors for each document and store it in a separate database
   1. Show document topic vector database structure
7. Find similar documents using nmslib
   1. Creating the different indices, some to speed up the process, or the other required index
8. Do search features: key word search or by topics.
   1. If doing keyword search then have input words and just find the intersection of the documents in which the words occur
   2. If doing topic search then have input words, find the topics for those words, and return documents that contain that word and relate to the topics found
9. Construct the visual. D3.js is what I’m working with. Should show my graph of it.

Future areas for analysis

* Seeing which nouns are capitalized in the raw data (besides names)
* See what topics this person is involved with (sender or receiver)