Homework 3

Summary of Hidden Markov Model for Stock Trading:

What seems to be the most important idea of this paper is that HMM is a very strong prediction model that outperforms other traditional models in predicting and trading stock. After describing what the problems are for traders (when do I buy/sell stock?) and providing some introduction for HMM, the paper discusses how HMM can be used as a solution. Since HMM is so flexible with its observable data and hidden states and HMM does not need a lot of states for accurate prediction, it offers a better alternative to previously used models. This is seen in the results of the paper, where HMM tends to be more efficient than HAR and HMM tends to outperform HAR and the buy/hold method for profits as well. It surprises me that, with such a simplistic model, HMM performs so well. It looks like a very strong option whenever a predictive model is needed, especially when there are hidden factors at play. The downside may be the complexity of the algorithms used, although I do not really have much information on what other models used, so I cannot really make an accurate comparison there.

Summary of Gene finding and the Hidden Markov models:

The main problem that this paper addresses is that gene finding in eukaryotes is much harder than in prokaryotes. This is largely due to eukaryote genes having introns and exons. A more popular technique used to solve this problem is HMM. HMM are used to identify segments of genes to find their borders and to find introns/exons. HMM seem to work better than traditional techniques because they combine basic models and offer a more flexible predictive model for biological sequencing. Similar to the last paper, this discusses what a HMM consists of and its main tasks (1. finding the emission and transition matrices given observable and hidden sequences and 2. finding the hidden sequence given an observable sequence, the emission matrix, and the transition matrix). Additionally, it seems that HMM is powered by breaking down the larger problem into subproblems and then connecting the subproblems together to pose a potential solution. HMM need to be trained for them to be predictive (in this case to find those special segments that have certain genetic properties).

Summary of Project Source:

The source that I will be summarizing from the ones listed in the personal project proposal (down below) is the dataset. Our dataset is from kaggle.com and is prepared by the White House and a coalition of other research groups. It contains 52,000 scholarly articles, which all pertain to the COVID-19 outbreak. It is provided in the hopes of generating more insights that can help society combat the disease. The dataset is periodically updated as well. It is a larger dataset, so I do not know how much of it we will use, but hopefully it should be fine for the purposes of filtering and performing PCA analysis on it.

Project Proposal

Team: Alex Bishka, Eve Kazarian

Research question: What is known about a vaccine for COVID-19?

Methods: We plan on looking through the literature on COVID-19, filtering for documents pertaining to vaccination. Once the documents that are irrelevant to our research are filtered out, we plan on performing PCA on the data.

After using PCA to separate the dataset of research projects into topics, we will choose the number of dimensions of our PCA to be three and get a visualization of the topic distribution in three dimensions. We can then analyze the articles in the topics to vaguely discern each topic's category. If we find a cluster about vaccines, we will look specifically into that cluster to get the latest research in that area.

We are building off an already-existing literature clustering project on Kaggle but modifying it to plot in 3D, changing our choice of stop words and number of clusters, and where exactly in the clusters we do our analysis.

Sources:

- Dataset
- Article clustering kernel on Kaggle
- 3D PCA on Kaggle