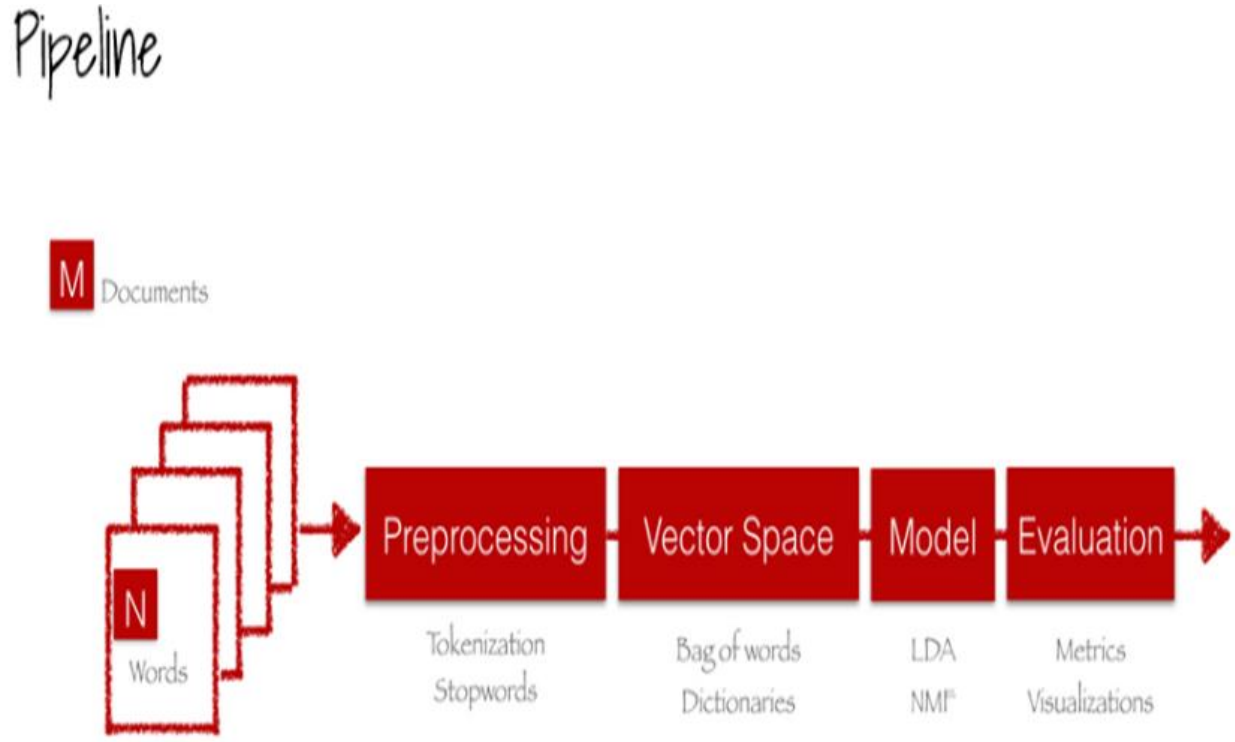
The last project is all about Topic Modeling analysis and being able to know when to use which tool to get the job done.

Below is the pipeline to use when you are using Topic Model analysis on a corpus. Based on the size of the corpus, one can easily justify when to use NMF or Latent Dirichlet Allocation (LDA).



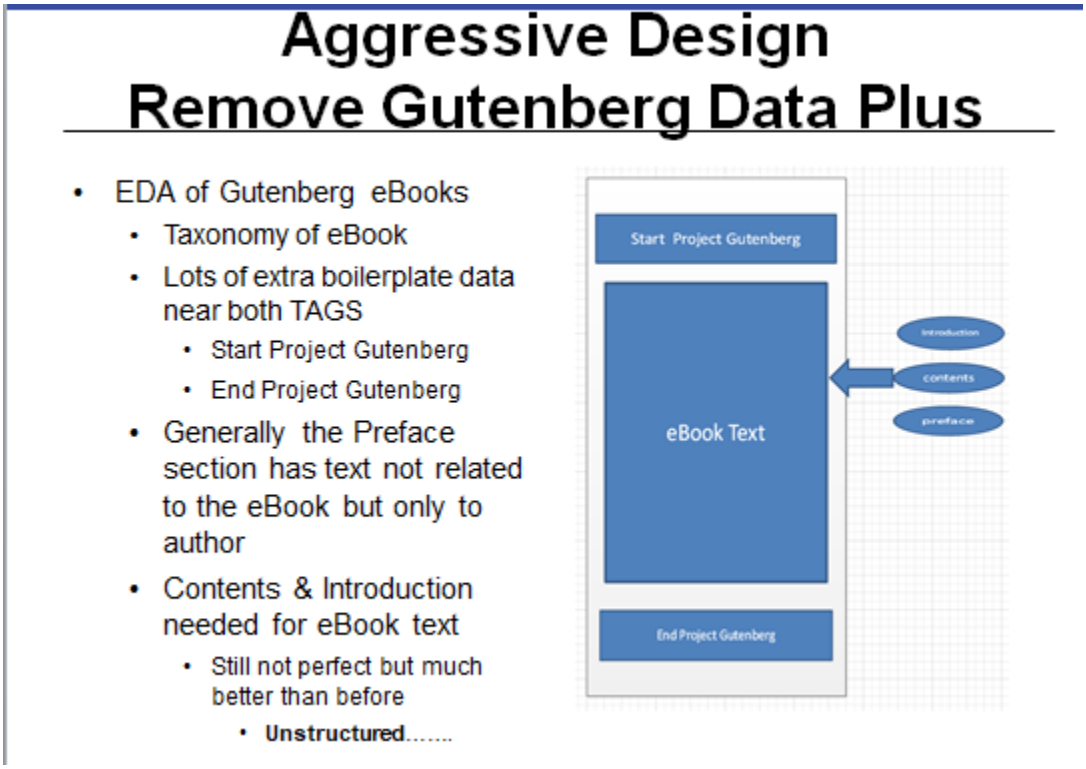
The task for the final project is twofold. First, we want everyone to get good hands-on experience working with NMF and/or LDA and the necessary Python libs to analyze and visualize your results. Second, when your analysis is complete, you will need to create a power point presentation of your hypothesis, results, summary, and conclusion.

Below are the datasets you will choose from to use as corpora for the Topic Modeling Final Project.

Choose either:
1) IMDB Movie Reviews analysis of Good and Bad Movie Reviews from homework 5 and 7
or
2) The Gutenberg eBooks from homework 1

If you choose the Gutenberg option, remember that it is important to clean your Gutenberg eBooks dataset before analysis. For this exercise, it is important to understand the Gutenberg eBook document model. You want to capture only the content of the book, not any other special Gutenberg eBook headers. There is a Python Gutenberg utility but it still leaves some unwanted data in the eBooks.

Below is the Gutenberg eBook taxonomy which is considered to be an UNSTRUCTURED dataset.



If you **DON'T** clean your data properly, you WILL obtain a lot of false negatives and false positives indicating your dataset was not cleaned properly. For instance, below, one thought they cleaned their dataset of the Gutenberg Start and Stop Tags but as you can see they did not.

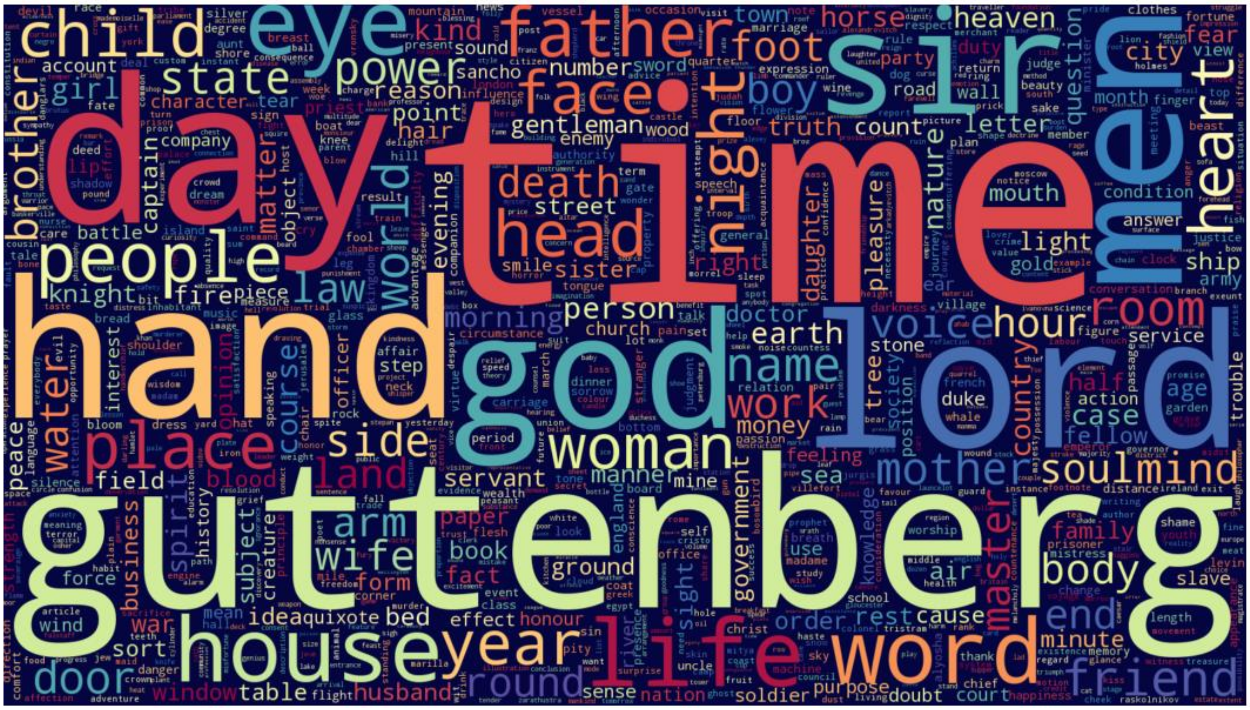


Figure 1 WORD Cloud based Gutenberg book after the book was cleaned

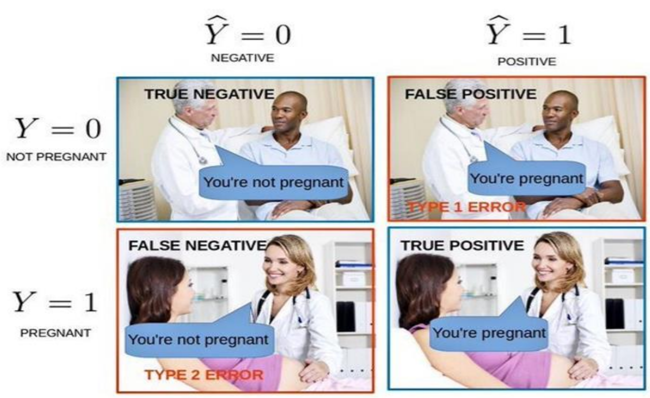


Figure 2 Confusion Matrix →False Positives and False Negatives

The word Guttenberg should not be in the cleaned data set!!! You will lose points according based upon this error.

Please perform Topic Modeling on either one of the items below

1) IMDB Good and Bad Movie Reviews from homework 5 and 7, or
2) Gutenberg eBooks from homework 1

Please be prepared to justify using either NMF or LDA for your Topic Modeling analysis. It is strongly encouraged that you create some data visualization as evidence to prove or disprove your hypothesis or use-case.

I strongly recommend using pyLDAvis,"Stanford Termite – HCI Interactive Data Visualizations", plotly, bokeh, or matplotlib for 2d or 3d data visualizations.

Topic Models Analysis
- pyLDAvis
- Stanford Termite - HCI Interactive Data Visualizations (supports both LDA and NMF)

When the time comes in the semester when we discuss LDA and Topic Models, it would be appropriate how to install pyLDAvis and Stanford Termite.

pyLDAvis Installation
- pip install pyldavis OR
- conda install -c conda-forge pyldavis

Stanford Termite - HC Interactive Data Visualizations
- https://github.com/StanfordHCI/termite
- WORKS ONLY in Python 2.7 but don't let that stop you!!!

Both of the above custom Topic Model interactive data visualization use D3 and JavaScript. It is imperative to save the files with the python code in *.html format.

In either case, depending on the project that you complete, item (1) or item (2), you are expected to do a 5 minute presentation on your results in Week 15 Live session. We strongly advise rehearsing a few times to make sure it fits in five minutes (time yourself!)

Items to submit:

1. Jupyter iPython notebook in *.ipynb and *.html format
2. Power Point Presentation consisting of about 5 or 6 slides
3. Python source code if needed

Items are due the week of 4/16/2019

Note:

You must submit your work in a timely fashion so that Dr. Slater/Ben Brock have time to review your work.