September 22, 2015 Folder ReadMe:

This folder contains unsupervised (LDA) topic lists, R environments, and bar graphs for three different corpora: comedies, tragedies, and a combination of the two found within the EEBO-TCP corpus.

These unsupervised topics were generated from unsegmented text files, which had their TEI encoding removed.

The MALLET+.r script was used to create each model. The script can be found here.

The 'eccostoplist' was used to remove common words, stage directions, and names. The stoplist can be found here.

The purpose of this set of topic models was to see what differences can be found and how the models can be compared among different sets of corpora. Each model was generated using the same parameters in order to make sure the models are comparable and that there is limited influence from other factors beyond the corpus choice for each model.

Folder Contents:

Each of the three folders contains the same types of documents for each of the models.

.RData files: These files are the R environments. Open these in R studio to obtain all the information about the script and results of each iteration. This can be used if one wants to either create further visualizations from this iteration of data or compare the results of the data at an algorithmic level.

topic-labels.csv: These files list the topics of each corpus. Each model generated 75 topics.

Topic-docs.csv: These files list the weight of each topic (the columns, labeled as V#), with each text (the rows). This file is not labeled with document names.

Means.csv: .csv files that contain the same information contained in the topic-docs.csv files, but with the file names attached so as to connect a value to a specific text. The means.csv files were used to create the bar graph pdfs.

Bar graph .pdfs: These files were generated from the means.csv using R's basic visualization function: plot(x). These specific graphs show the weights of each text on one topic per graph, going in sequential order; graph 1 represents topics 1-25, graph 2, 26-50, and graph 3 showing 51-75. The script can be found in the MALLET+.R script and R environments.