IST736 Text Mining
HW8

# Topic Modeling

LDA is an algorithm that can "summarize" the main topics of a text collection, now you are asked to use this algorithm to analyze the main topics in the floor debate of the 110th Congress (House only). According to political scientists, there are usually 40-50 common topics going on in each Congress. Tune the number of topics and see if LDA can get you the common topics, such as defense, education, healthcare, economy, etc.

The data set "110" consists of four subfolders. For the subfolder names, "m" means "male", "f" means "female", "d" means "democrat", "r" means "republican". You can merge all of them into one folder to run Mallet LDA.

There are a few other parameters you can tune, such as ngram. You can decide what parameters to use and explain your decision in the report.

Interpreting topic clustering results is very difficult. See if this article "Reading Tea Leaves" may help you. http://www.umiacs.umd.edu/~jbg/docs/nips2009-rtl.pdf. The recommended readings are also great examples to demonstrate how to articulate topic modeling results.

This is a fairly large data set (100M pure text, more than 400 files). Please start working on it early because it may take a long time to run. To prevent your program from being interrupted, run it as a backend process by adding "&" to the end of your command (for Linux system). Or you can use one subset of the data to build a topic model and explain what topics you have discovered from the data.