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.

**OR If you cannot work with Mallet – you can use my code for LDA which also generates a great interactive vis.**

**Here is my folder filled with many LDA and API examples:**

[**https://drive.google.com/drive/folders/1\_QMxLIffDshlY8U2Nn\_yJlAY\_5ToF\_0e?usp=sharing**](https://drive.google.com/drive/folders/1_QMxLIffDshlY8U2Nn_yJlAY_5ToF_0e?usp=sharing)

**Again – you do not need to use Mallet – you may if you wish. You may also use both or multiple methods if you want to go deep.**

There are a few other parameters you can tune, such as ngram (for Mallet only). 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.

**ALSO – do not start with this HUGE dataset. First, create a small and balanced sample of data or even a different dataset that you find or make that will work. Make sure your code runs on that smaller data first. Then, when everything work – try it on the large dataset. If you still cannot – cut the large dataset down until you can.**

NOTE from other Professor:

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.