IST736 Text Mining
HW8

Document Clustering and Topic Clustering

Task 1: Document clustering

The data file data-science-course-desc.csv includes 20 courses that are currently listed as required or elective courses in the CAS in Data Science. Use k-means to explore this data set and see whether you can find meaningful clusters of courses.

If you do find meaningful clusters, explain what courses are in each cluster and why they form meaningful clusters. This finding can be helpful to help students choosing courses in relevant topics to build a strong profile for job interviews.

If you do not find meaningful clusters, explain what might be the factors that affect the clustering algorithm performance. For example, if highly relevant courses were separated into different clusters, what words in their descriptions made them look different to the clustering algorithms? At the same time, if not-so-relevant courses were grouped into one cluster, what words made them look similar?

Task 2: Topic clustering

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,” and “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 n-gram. 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.

A sample solution that used the Princeton LDA-C tool is provided to demonstrate how to make sense of a topic clustering result. However, you can be creative and think of innovative approaches to tackle the problem of interpreting the topic clustering results.

-- Optional: if you are interested in comparing the topic clustering results from two LDA implementations, the Mallet and the Princeton LDA-C, use the following tools to preprocess the data to LDA-C format.

This open-source Python script can read txt files in a folder and convert them to the data format that LDA requires:
https://github.com/JoKnopp/text2ldac

Usually removing stopwords would help topic clustering. You can use this stoplist for data preprocessing:
http://jmlr.org/papers/volume5/lewis04a/a11-smart-stop-list/english.stop.

The Python preprocessing command should look like this:

```
python text2ldac.py 110 –o output –e .txt –stopwords english.stop.txt
```

110 is the input folder.
You will create a new folder “output” to store the output.
-e specifies the extension name of the files that you want to read in.
--stopwords specifies the name of the stoplist file.

-- end of optional

This is a fairly large data set (100M pure text, more than 400 files). Please start working on it early because it may take hours 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). For example, your command is "lda est 1.0 2 settings.txt mydata.dat random model". Now add that "&" to the end; the command is now " lda est 1.0 2 settings.txt mydata.dat random model &". Your command will be run uninterrupted until it is done. Better run it before you go to bed.