CS 674 – Data Mining on Multimedia Data

HW 4
Due 4/26/17 at 4:30pm

Download and install a topic modeling package. Some available packages are listed below. Some of the packages will have default values for parameters (α, β) and sampling procedure (# burn in iterations, # data collection iterations). Make sure you pick a package that gives you enough flexibility for the rest of the assignment.

UCI Matlab Topic Modeling toolbox
Mallet (machine learning for language, Java based implementation of topic modeling)
Mahout (Java API that does topic modeling)
Stanford Topic Modeling Toolkit
Python implementation and documentation
R implementation and documentation
C implementation

Part 1. 50 points

Use the NY Times dataset from the UCI Machine Learning Archive (https://archive.ics.uci.edu/ml/datasets/Bag+of+Words), and explore topic modeling with one of the packages above (or you can find your own package). Find a few interpretable topics and present them by showing the highest probability words (10-20) within the topic, and give a label to the topic. Experiment with different parameters.

Part 2. 50 points (up to 75 points)

Next you will experiment applying topic modeling on different types of data. Choose one of the followings, or do both for extra credit.

(1) Image data. Use the MNIST handwritten digits on this website: http://www.cs.nyu.edu/~roweis/data.html. Treat each image as a document, and each pixel as a word. Identify the top "topics" and plot them. Start with the test set only. If you're extra adventurous, you can add in the training set and see if you get better results with larger dataset. Here is a tutorial: http://psiexp.ss.uci.edu/research/programs_data/exampleimages2.html

(2) Time series data. Use the same datasets that you were given, convert them to "bag-of-patterns" using SAX (you'll need to experiment with different SAX parameters), and see if you can identify some meaningful "topics" on the time series.

Write a report describing the package you use, parameter settings and your findings.

Submit: Report, screenshots with output