IITM-CS6730 : Probabilistic Reasoning in AI — Given on: Apr 20, 5pm

Programming Assignment #4 — Due on : Apr 30, 5pm

- The goal of this assignment is to gain an understanding of LDA.
- This is an individual assignment. Collaborations and discussions with others regarding the problems and solutions are strictly prohibited.
- You have to turn in your report and code in the prescribed format in Moodle. Report should be typesetted in Latex only.
- You can use only Python.

---

1. The objective of this programming assignment is to perform LDA on the given document collection and use it to cluster the documents. You have been provided with a dataset with the following description.

   - The dataset contains 2761 documents each belonging to one of the following 5 categories: Medicine, Hockey, Baseball, Windows, Religion.
   - All the documents are provided in a single file "doc.txt". Each line in the file corresponds to a document. The first word in the line is the document id and the rest being the bag of words in the document.
   - The documents are already pre-processed and stop words are removed.

   Implement Gibbs Sampling for LDA and learn a topic model for this dataset. Refer http://www.uoguelph.ca/∼wdarling/research/papers/TM.pdf for implementation guidelines. Input to your code is the file "doc.txt". Output should be the documents represented as $k$ dimensional real vector where $k$ is the number of topics.

   Now perform K-means on this LDA representation. You have been provided with the ground truth in "truth.txt". Calculate cluster purity (http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html) and report the same.

   1. Vary the number of topics and perform clustering by fixing $k$ in k-means to be 5. Report the cluster purity.
   2. What is the effect of varying the $k$ in K-means, provided the number of topics is fixed?
   3. Report cluster purity when number of topics is 5 and number of clusters is 5.
   4. Report cluster purity when number of topics is 4 and number of clusters is 5.
   5. For the case of 5 topics, submit a data file containing the word distribution in each topic. Also report 10 most probable words in each topic. Analyze these words qualitatively.

# Submission Instructions

Submit a single tarball/zip containing the following files in the specified directory structure. Use the following naming convention. 'rollno.tar.gz' with all capital letters. Ex: CS12S043.tar.gz

rollno
- Report
  - Report.pdf
- Code
  - all your code files.
- Data
  - all the data that are needed for your code to run.