In the Name of God, The Compassionate, The Merciful

Statistical Machine Learning H.R. Rabiee
Sharif University of Technology CE 40-956

Due: August 01, 2020 (1399/05/11)

SML Final Project (Spring 2020): Topic Modeling

Project Description

- **Problem Definition**
  Topic modeling is the process of identifying topics in a set of documents. This can be useful for search engines, customer service automation, and any other instance where knowing the topics of documents is important. Lots of unlabeled documents and data are publicly available on the internet, but labeled data are expensive, so the main purpose of this project is to implement automatic unsupervised topic modeling. For more information, you can refer to the useful links part.

- **Solving and Implementation**
  Each group should solve the above problem using both Non-Parametric Bayesian (NPB) and Deep-learning based methods (a hybrid method is preferred). The model selection, implementations, and the whole problem-solving procedure in both kinds of methods can be made by your own group choice. Innovative ideas receive extra bonus points.

- **Evaluation**
  In the last step, each group should compare the results of two kinds of methods using citation graph of papers and discuss the similarity of topics in competing methods (the distance function for calculating similarity and vicinity is optional).

- **Data**
  Three citation datasets are available on the CE course website under resources -> project. each group should perform the implementations on at least one of these datasets. The datasets can also be downloaded from the following links:
  Dataset1: Cora .
  Dataset2: Citeseer .
  Dataset3: cit-HepTh .

- **Useful Links**
  To get acquainted with Topic Modeling and LDA (which is one of the famous solutions for this problem), link1, the original article of LDA or video1 can be useful.

  Some studies have used Deep methods to solve the Topic Modeling problem. One of these articles is available from this link.

- **Project Report and Presentation**
  Each group should prepare a complete report of their results (5 to 20 pages including: Abstract, Introduction, Previous Works, Method description, Experimental results, Conclusions and Future works), and submit it on the day of presentations. Please prepare a 30 slides presentation as well, since you have 30 minutes to present your work.

Good Luck :)