Northeastern University

EECE 5642 Data Visualization

Midterm Project
Instructor: Mr. Zhiqiang Tao
Submission Due Date: **11:59 pm Mar. 18**
Presentation Date: Mar. 11
Submission: Blackboard

All the details and backgrounds for our midterm project could be found in the "Lecture- Midterm Project" slide. We list the requirements for this project as follows.

1. Preprocess 20 Newsgroup dataset as corpus and visualize its statistical information. (10')
2. Build two different vocabularies upon different preprocessing ways; Learn Bag-of-words (BoW) and TF-IDF model with each vocabulary accordingly. (10')
3. Train two LDA models upon the vocabularies in Step 2; Visualize topics with four different methods; and eventually get the topic distribution (as feature) for each document. (20')
4. Train two Doc2Vec models upon the vocabularies in Step 2; Visualize your leaned word and document embedding space; Collect Doc2Vec representation of each document. (20')
5. Conduct document clustering by K-means with four different doc. representations: 1) BoW; 2) TF-IDF; 3) Topics distribution; and 4) Doc2Vec. Compare different results by Normalized Mutual Information (NMI) and visualize the clustering results. (20')
6. Do experiment analysis from the following aspects: 1) Impact of different preprocessing ways (*e.g.*, how to filter vocabulary; using n-gram model); 2) Impact of different topic numbers; and 3) Different training methods for Doc2Vec; 4) What's the key factor for doc. visualization? (20')
7. Learn document representation beyond the above ones. For example, how to use temporal context in a document? (Bonus)

Every group is required to give a presentation with slides in our class. The presentation time is about 5-10 minutes. Each talk should include the following contents.

1) Introduction to your group members and team assignments.
2) A clear illustration for your project workflow.
3) All the experimental results you have obtained with some necessary experimental analyses.
4) A live demo for your visualization result. For example, a demo for visualizing different topics, or displaying your word/doc embedding space. (Jupyter Notebook is recommended.)

The final submission is required as follows.

1) A two-page pdf report including 1) a brief introduction to the project and your method; 2) all the necessary results and analyses; 3) references for the tools and papers you used in this work.
2) A package file including all your source codes and visualization results.

Hint: We do not require the format (e.g., single-column or double-column, font size and line space) for the final report. However, you need to make sure it is neat and readable. Some good and highly recommended (Word, Latex) templates could be found from IEEE Transactions or ACM conference.