Final Project

COEN 240 Winter 2022

All the details and backgrounds of our midterm project could be found in the "Lecture-12_Final Project" slide. A Notebook demo has also been provided in "Camino $->$ Modules $->$ Notebooks $->$ FinalProject" to help you kick off the project. We list the detailed requirements of this project as follows.

1. Preprocess the 20 Newsgroup dataset and visualize its statistical information (*e.g.*, term-frequency distribution). (10 pts)
2. Build and save your vocabulary (termed as `Vocab_v1`) upon Step 1. Learn Bag-of-words (BoW) and TF-IDF model for all the documents. (10 pts)
3. Train a LDA model with `Vocab_v1`. Visualize topics with at least two different methods, and get the topic distribution (as features) for each document. (15 pts)
4. Train Word2Vec and Doc2Vec models upon `Vocab_v1`. Visualize your learned word and document embedding space (*e.g.*, using t-SNE). Collect Doc2Vec representation of each document. (15 pts)
5. Conduct document clustering by K-means with four different doc. representations that you obtained from Step 2-4, including i) BoW; ii) TF-IDF; iii) Topics distribution; and iv) Doc2Vec. Obtain another vocabulary, denoted as `Vocab_v2`, by taking the top 2K words from `Vocab_v1`. Learn the above four doc. representations (*i.e.*, i-iv) with `Vocab_v2` and get the K-means clustering result by NMI for each doc. representation method. To sum up, you should have (1) a table with the size of 2×4 NMI values and (2) a visualization (similar to HW3) result of the best clustering result indicated by NMI. (25 pts)
6. Do experiment analysis from one of the following aspects: 1) Impact of different preprocessing ways (e.g., compare the K-means or LDA visualizations between different vocabs); 2) Impact of different topic numbers; and 3) Impact of different clustering methods on the clustering result. (10 pts)
7. Propose one supervised task on your own. Give the motivation, methodology, and experimental results, accordingly. For example, you could propose a classification task upon the doc. representations we learned so far and develop one classification model (*e.g.*, Softmax regression, SVMs, KNNs, etc) to compare different representations. (15 pts)
8. The bonus could be obtained by designing one doc. representation learning method with a clearly improved clustering performance (around 3% NMI value) over the best performance in Step 5. For example, you may try to learn new doc. representations by using RNNs, temporal CNNs, pre-trained word embeddings/BERT, or averaging the learned word embeddings within one document. (10 pts)

The final submission is required as follows.

- A pdf report (4-8 pages) including 1) a brief introduction to the project and your method; 2) all the necessary results (including figs and tables) and analyses; 3) references for the tools and papers you used in this work (the references can be put into an extra page, containing nothing but references).
- A zip file including all the source codes and experimental results (e.g., the saved figs, numpy files, etc).

Note: The final project will be mainly graded by the results posted in the report. However, the submission without executable codes and complete experimental results will not be graded. The report exceeding the maximum page length will be punished by deducting 10% of the grade. Any Latex/Word/Markdown template could be used. The only requirement is all the fonts, figures, and tables must be legible.
Hint: The NMI values given by TF-IDF and Doc2Vec should around 0.2 or 0.3.