UCLA CS 145 Homework #6

DUE DATE: Sunday, 12/09/2018 11:59 PM

Note

- HW6 is optional. We will select your highest 5 homework grades to calculate your final homework grade (up to 25% based on the grading policy).
- You are expected to submit both a report and code. The submission format is specified on CCLE under HW6 description.
- Copying and sharing of homework are NOT allowed. But you can discuss general challenges and ideas with others. Suspicious cases will be reported.
- "# ======== YOUR CODE HERE ========" is used where input from you is needed in the code file.

In Homework 6, you will complete two basic text mining model in lectures: Naive Bayes and Topic modeling (pLSA). Python 3.6 is recommended in HW6 and you may need to install additional packages such as `sklearn` and jieba (`numpy`, `pandas` and `matplotlib` are also needed.) Besides, you are allowed to slightly change the code or add new functions other than "YOUR CODE HERE" as long as you keep the structure and do not use external model toolkits.

You need to submit all your python files, report and output files (including figures and text files). Do NOT include original dataset files.

1 Naive Bayes for Text (50 points)

Naive Bayers is one generative model for text classification. In the problem, you are given a document in `dataset` folder. The original data comes from "20 newsgroups". You can use the provided data files to save efforts on preprocessing.

(a) Complete the implementation of Naive Bayes model for text classification in `nbm.py`. After that, run `nbm_sklearn.py`, which uses `sklearn` to implement naive bayes model for text classification (Note that the dataset is slightly different).

(b) Report your classification accuracy on train and test documents. Also report your classification matrix. Show one example document that Naive Bayes classifies incorrectly ((you can fill in the following table).

(c) Question: Is Naive Bayes a generative model or discriminative model and Why? What is the difference between Naive Bayes classifier and Logistic Regression? What are the pros and cons of Naive Bayes for text classification task?

Table 1: Report accuracy for Naive Bayes Model

	Train set accuracy	Test set accuracy
sklearn implementation		
your implementation		

Table 2: Incorrect Examples

Words (count) in the example document	Predicted label	Truth label
For example, student(4), education(2), etc	Class A	Class B

(d) Question: Can you apply Naive Bayes model to identify spam emails from normal ones? Briefly explain your method.

2 Topic Modeling: Probabilistic Latent Semantic Analysis (pLSA) (50 points)

In this section, you will implement Probabilistic Latent Semantic Analysis (pLSA) by EM algorithm.

(a) Complete the implementation of pLSA in `plsa.py`. You need to finish the E step, M step and likelihood function.

(b) Choose different K (number of topics) in `plsa.py`. What is your option for a reasonable K in `dataset1.txt` and `dataset2.txt`? Give your results of 10 words under each topic by filling in the following table (suppose you set $K = 4$).

Table 3: Topic words

Dataset 1			
Topic 1	Topic 2	Topic 3	Topic 4
Dataset 2			
Topic 1	Topic 2	Topic 3	Topic 4

(c) Question: Are there any similarities between pLSA and GMM model? Briefly explain your thoughts.

(d) Question: What are the disadvantages of pLSA? Consider its generalizing ability to new unseen document and its parameter complexity, etc.