CS529– Applied Artificial Intelligence

Instructor : Shashi Shekhar Jha (shashi@iitrpr.ac.in)

Lab Assignment - 2

Due on 13/10/2019 2400 Hrs

Total: 100 Points

Submissions Instructions:

The submission is through google classroom in a single zip file. In case you face any trouble with the submission, please contact the TA:

- Neeru Dubey, neerudubey@iitrpr.ac.in

You can comment in the google classroom with your queries as well which will be beneficial for the whole class.

Your submission must be your original work. **Do not indulge in any kind of plagiarism or copying.** Abide by the honour and integrity code to do your assignment.

Any late submissions after the due date will attract penalties.

You submission must include:

- A legible PDF document with all your answers to the assignment problems/ questions.
- A folder named as 'code' containing the scripts for the assignment along with the other necessary files to run your code.
- A README.txt file explaining how to execute your code.

Naming Convention:

Name the ZIP file submission as follows:

YourName_rollnumber.zip E.g. if your roll number is 2016csx1234 and Abcd,

then you should name the zip file as: Abcd_2016csx1234.zip

Programming Language

As was the case in Lab 1, you will be using **Python** as the chosen programming language. All data wrangling operations should be done using ***Numpy*** and ***Pandas***.

In case you are new to python, you can use the following resource to get you started on Python https://wiki.python.org/moin/BeginnersGuide/Programmers

Part- 1 [Recommender System]

About the Dataset

For this part, you will be using the same dataset on movie reviews as was in Lab 1. Also, use the same train and test sets, that you created for Lab 1.

Questions [Latent Factor Models] [55 points]

Q.1. Implement the Unconstrained Matrix Factorization for the ratings matrix. You can either use the MLE estimate or your own heuristic to fill in the missing entries before factorization. Look into the eigenvalues of the ratings matrix to choose a good value of **K**.

A. Implement the ***batch*** and ***stochastic*** gradient descent methods to get the User and Item factors matrices. Compare the number of iterations it takes to get to convergence using both the methods. State your heuristic for initialisation of the User and Item factors matrices. How are you deciding the step-size α for the gradient descent (use same alpha for both the methods so that they can be compared). [15 points]

B. Whichever method is more practical in (A), use that method to factorize the ratings matrix with regularization parameters λ. Use cross-validation to choose a better value for the parameter lambda λ**. [**5 points]

C. Using your best α**,** λ and gradient descent method, perform unconstrained factorization of the ratings matrix with now including the user and item biases in your model. [5 points]

D. Compare the predicted ratings in A, B and C using RMSE error metric. Which is the factorization having best performance? [5 points]

Q.2. Using the same value of **K**, as in Q.1., perform the SVD factorization of the ratings matrix (you can use the built-in function to perform for SVD

factorization). Use the iterative method mentioned in the class (check slides) to update the missing values. How does the results of the SVD factorization compare with the unconstrained matrix factorization in Q.1? [10 points]

Q.3. Perform the Non-negative Matrix Factorization (NMF) of the ratings matrix using the same regularization parameters λ as in Q.1(C) and your best gradient descent method. The value of **K** remains same as in previous questions. How does NMF perform as compared to the best results in Q.1. and Q.2? The advantage is NMF is better interpretability of the factorization in terms of user and item affinities. Can you discuss your NMF factors highlighting such affinities? [15 points]

Part- 2 [Topic Modelling] [45 points]

About the Dataset

The dataset for this part of the assignment is uploaded alongside in the classroom in a file named - ***nytimes_news_articles.txt.*** The file contains the URLs and news articles that follow. There are about 8800 articles in the dataset. The articles are from New York Times (https://www.nytimes.com) for a period of 2.5 months from mid April to end June 2016.

Questions:

Q.1. You can extract a fair degree of information about the articles by carefully processing the associated URL. The URL contains the date of publication, category such as sports, US (i.e. USA), NY Region, etc. (sometimes, categories are followed by sub-categories and the title of the article. Create a ***processed dataset*** with all the information regarding the article from the URLs alongside the articles itself. Mention the number of categories and sub-categories present in the dataset. [5 points]

Q.2. Perform pre-processing on the articles to remove stop-words (https://www.ranks.nl/stopwords) and stemming. Use the pre-processed articles to create the Vector-Space model with TF.IDF frequencies. What is the size of the Vector-space model? [5 points]

Q.3. Use Latent Semantic Indexing (LSI) to generate the topics discussed in the articles. What criteria did you use to set the latent parameter **Z** (number of topics) in the LSI model**?** Print the top 10 words of each latent topic. [5 points]

Q.4. Implement the Probabilistic Latent Semantic Indexing (PLSI) to generate topics. Use the EM algorithm discussed in the lectures to derive all the parameters (keep the number of topics same as in Q. 2.). The advantage of PLSI approach is that the values in the factor matrices can be interpreted as

probabilities. Print the 10 most probable words for each of the latent topics in the articles. Can you co-relate the most probable words in the latent topics and the categories/sub-categories ? Do the words make sense for the topics? [15 points]

Q. 5. Implement the Latent Dirichlet Allocation (LDA) model to generate the topics. Use the Gibbs sampling based approach discussed in the lecture to estimate the model parameter values. Set the number of topics same as in previous questions. What advantage do you find with the LDA implementation compared to the PLSA implementation? Vary the parameters α and η from the set $\{0.1, 0.5, 1, 1.5, 2\}$. What difference do you find for the different values of α and η? What values are the most suitable? Print the 10 most probable words for each of the latent topics in the articles. Can you compare the your results with that of Q.3. and Q.4.? [15 points]

_____-******_____******_____