BITS - Pilani,Hyderabad Campus

CS F469 IR Assignment - 3

Deadline: 21/10/2017

This assignment is aimed at implementing and comparing various techniques for building a Recommender System taught in the class.

Kindly continue with the same team for this assignment.

Programming Languages:

The assignment can be implemented in any programming language of your choice. STL's and inbuilt packages can be used for tasks like sparse representation of the matrix, matrix multiplication and finding out eigenvalues and vectors. You are expected to code the core functionality without using any packages.

Task:

The task is to compare various techniques used in implementing Recommender Systems on the basis of their errors using Root Mean Square Error, Precision on top K and Spearman Rank Correlation. Also compare their overall running time and prediction time.

Expectations:

1. Successful implementation of the techniques on a reasonably sized dataset.
2. Ensuring generous raters and strict raters are handled appropriately.
3. Explaining the results in the design document and clearly stating all the formulation.
4. Mention how many rows or columns were used in case of CUR, how many neighbors were considered in case of Collaborative Filtering and other essential details in the design document.
5. Filling the following table:

Recommender System Technique	Root Mean Square Error (RMSE)	Precision on top K	Spearman Rank Correlation	Time taken for prediction
Collaborative				
Collaborative along with Baseline approach*				
SVD				
SVD with 90% retained energy**				
CUR				
CUR with 90% retained energy**				

* Modeling Local and Global Effects

** Retain few largest singular values while doing Dimensionality Reduction

How Many Singular Values Should We Retain?

A useful rule of thumb is to retain enough singular values to make up 90% of the *energy* in Σ. That is, the sum of the squares of the retained singular values should be at least 90% of the sum of the squares of all the singular values. In Example 11.10, the total energy is $(12.4)^2 + (9.5)^2 + (1.3)^2 = 245.70$, while the retained energy is $(12.4)^2 + (9.5)^2 = 244.01$. Thus, we have retained over 99% of the energy. However, were we to eliminate the second singular value, 9.5, the retained energy would be only $(12.4)^2/245.70$ or about 63%.

Additional Resources:

1. Mining Massive Datasets: Module 8 & 9
2. Datasets:
    a. https://www.quora.com/Where-can-I-find-dataset-for-a-recommender-system
    b. Datasets for Recommender Systems

c. Movielens
d. 9 datasets for recommender systems.

Instructions:

1. Take data of reasonable size. Make sure the data can be loaded entirely into the RAM at once. As the algorithms involve matrix multiplications and multiple iterations ensure that the data isn't too large.
2. Preferably choose datasets which contain only ratings. Datasets containing both reviews and ratings will have to undergo the process of removal of reviews. Only datasets which contains more than 1000 products and 1000 users should be used.

Deliverables:

The final submission must contain the following documents:

1. **Design Document** –This document should contain all the formulas and packages used along with a brief description. All the assumptions, pros & cons of your model, running time etc. should be well documented. Make sure names and ID numbers of all the team members are mentioned on the first page itself.
2. **Code** – The code should be well commented.
3. **Documentation** – All the classes, functions and modules of the code must be documented. Software that automatically generate such documents can be used – pydoc for Python, Eclipse for Java etc.
4. **README** – The README file should describe the procedure to compile and run your code for various datasets.

Submission Guidelines:

All the deliverables must be zipped and submitted to **bphc.ir@gmail.com** latest by **deadline.**

You are expected to demo your application and present your results as per the schedule that will be made available.

Evaluation Criteria for Task :

S.No.	Task	Marks
1.	Implementation of the Algorithms	20
2.	Design Document	5
3.	Filling The Table	15
4.	Handling generous raters and strict raters	5
5.	Viva	5
	Total	**50**

It should be noted that all the assignments would be run through a plagiarism detector and based on the results, the marks would be altered. The final decision lies in the hand of the instructor and only one submission per group would be allowed for one assignment.