INFORMATION RETRIEVAL (CS F469) ASSIGNMENT 3

Recommender System

Members

Nikhil Joshi 2015A7PS0179H

Nikhil Madaan 2015A7PS0283H

Bhavesh Gawri 2015A7PS0116H

GETTING DATASET:

The dataset is from <http://shuaizhang.tech/2017/03/15/Datasets-For-Recommender-System/> which contains Jester-Movie Ratings. Dataset contains more than 5000 users and their rating to 100 movies in .xls format. The data is sparse and ratings are between -10 to +10 and and value 99 represents particular user has not rated the movie.

Functionality Implemented:

* Collaborative Filtering (calculating similarity by users and predicting missing ratings)
* Collaborative Filtering using global baseline approach.
* SVD
* SVD with 90% retained energy
* CUR
* CUR with 90% retained energy

MAJOR DATA STRUCTURES USED

* list in python – List in python were used to manipulate data.
* Dictionary in python - Dictionary in python is equivalent to Hash tables

DESIGN ARCHITECTURE:

BRIEF NOTES ON CODE EXECUTION:

* Input from compressed wikipedia articles and extracted using wikiextractor tool
* Above uncompressed file were futher processed to obtainn indivisual articles
* These Individual articles were tokenized, removing of stop words, stemming , punctuation removal and finally case folding were respectively implemented.
* Parallely we calculate and save tf of corresponding term in a document
* generating inverted index was the next process which was further followed by calculation of idf.
* File corresponding to tf-idf and after normalization is obtained into a dictionary object
* Input Query is processed and vector corresponding to query is generated and cosine similarity is generated with respect to all the documents
* Auto correction and wild card queries were implemented using trie.
* Further tf-idf data were used to calculate similarity between given documents and document in corpus.

OUTPUT

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Recommender**  **System**  **Technique** | **Root Mean**  **Square Error**  **(RMSE)** | **Precision on top K** | **Spearman Rank**  **Correlation** | **Time taken for**  **prediction** |
| Collaborative | 0.0282946268 | (top 10 for every user)  0.19934617091 | 0.99999983523 |  |
| Collaborative with  Baseline  approach \* | 0.0302534122 | (top 10 for every user)  0.20330233926 | 0.99999981162 |  |
| SVD |  |  |  |  |
| SVD with 90%  retained energy |  |  |  |  |
| CUR |  |  |  |  |
| CUR with 90%  retained energy |  |  |  |  |