**Revive: Recommender System for Books and Movies**

**Summary:**

On any time movies, books make you revive from the daily work. So in the deluge of options provided by the social networking websites and entertainment websites it often leads to a situation of choosing the right option of movies. So our application provides a list of recommended movies and books based on the genre. Here we will be able to track the user history and store them in a file to recommend his/her well opted genre of books and movies in order avoid the monotony of searching mechanisms.

**Design:**

In order to suggest the suitable genre movies and books we have included the concepts of clustering and recommendation. Here clustering will be done on the basis of genre and recommendation will be done on the basis of ratings and number of people viewed the books.

**Data Model:**

The data models which we have considered are set of csv files which will be used for the recommender system to evaluate the genre based movies and books. There are mainly two sets of data. They are:

* Books and
* Movies

The Books dataset consists of a series of reviews and the name of each book and the identification of each user who has given the review. The Movies data set consists of the movies and their genres. The csv files can be merged in the Hadoop dataset and are being clustered. The datasets can eb viewed in the following github URL:

* https://github.com/SaiKishoreBandaru/MidTerm/tree/master/src/Data

**Integration Model and Algorithm:**

The integration model primarily consists of two part. They are:

* The clustering of the movies data should be done on basis of genre and should be pushed to the Solr
* The books data will run through a recommendation algorithm and the result of these algorithm should be integrated with genre specified by the user and should be recommended as books for his search specification.
* The history of user will be tracked and stored in separate files based on his logging credentials, in order make a content personalization recommendation system.

The following diagram depicts the brief overview of the integration data model of our application:

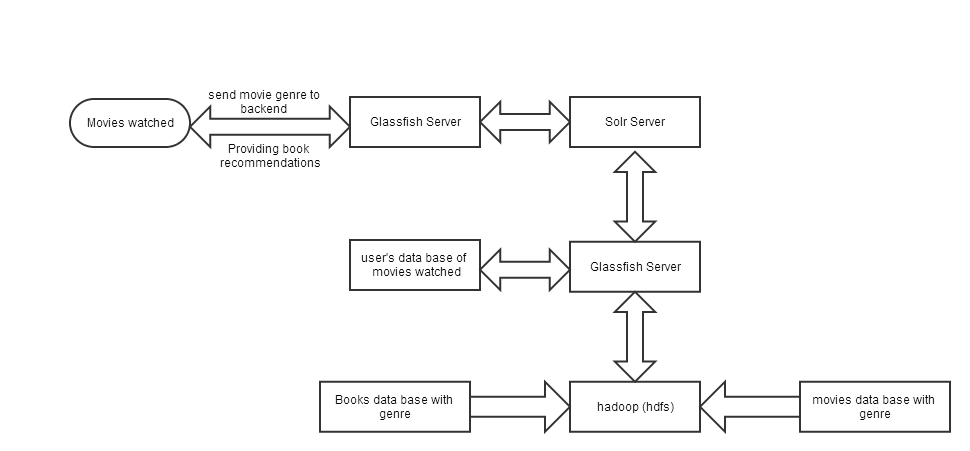


Figure 1Data Model

The algorithms used for the application are basically the clustering algorithm and recommendation algorithm. Initially we thought of using the Apriori algorithm for the data clustering. We have tried these algorithm using the R technology for machine learning, but due to the datasets issues we couldn’t move forward on this algorithm using R.

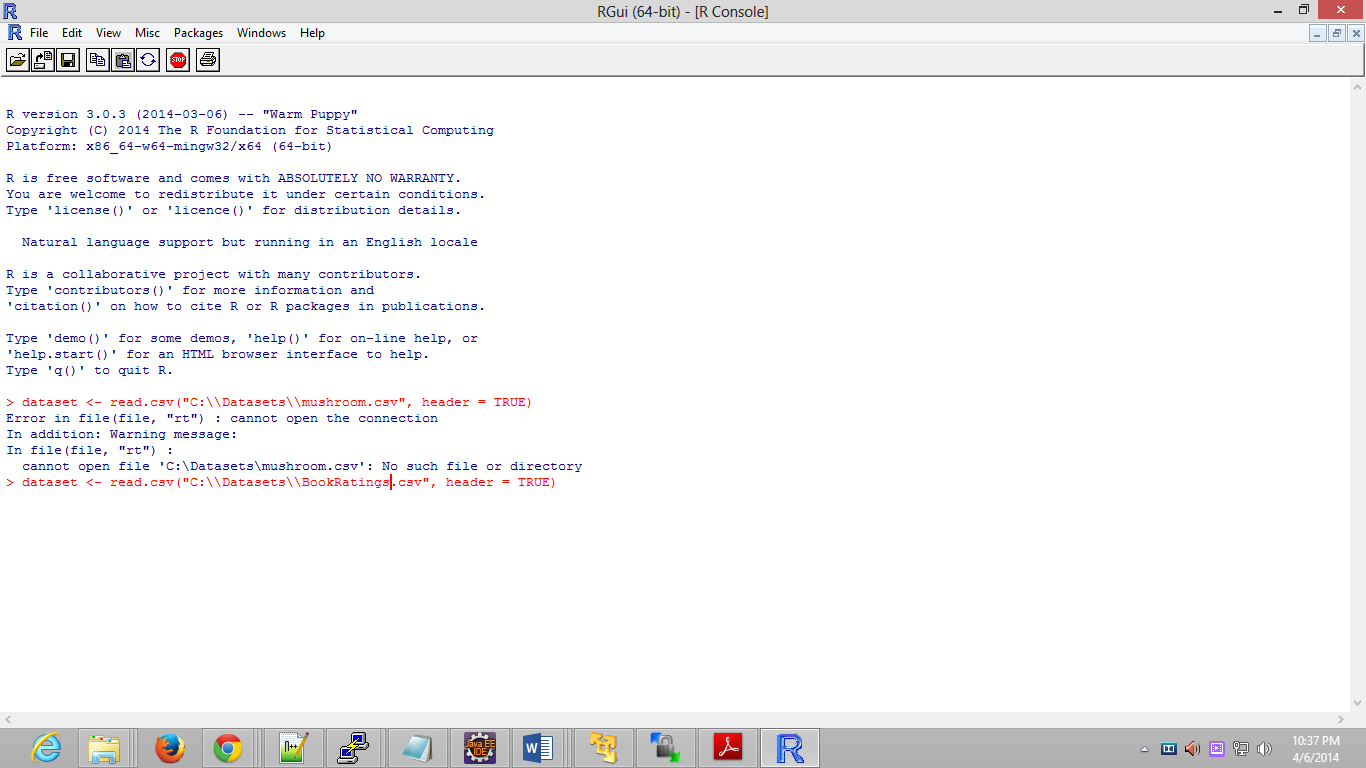


Figure 2R GUI for Apriori

So on the basis of apriori algorithm and K-means clustering algorithm we have developed our new algorithm for finding the movies based on the genre. The algorithm precisely clusters the movies based on the genre and this will be stored in the form of JSON format and stored in the Solr. The recommendation algorithm will be classifying the books based on the user ratings and the number of readers who have approved that book for genuine good ratings.

**Recommendation Algorithm:**

The recommendation algorithm will be classifying the books based on the user ratings and the number of readers who have approved that book for genuine good ratings. We have also developed a personalization mechanism in order to track down the history of users based on the credentials and their frequency viewing movies of specific genre. Content based recommendation is the technique which is being followed in this application.

**Selection of Datasets:**

The datasets are being selected based on the algorithms which have been mentioned in the above mentioned sections of recommendation and clustering algorithms. The datasets comprises a size of 30MB respectively and an application based dataset also developed for tracking the user based on their credentials. The user details will be used for the content based recommendations. This content will be purely restricted to the history of operations but not to any of the geographical or demographic features.

**Application Design:**

The application basically consists of a web page where the user will be selecting a genre for retrieving the results of movies to be watched. And based on the genre selected and also on the User ID’s we will be suggesting the books, which have been mined through the recommendation algorithms. The selection of each genre of the respective user will be tracked so as for further recommendations.

**Implementation:**

**Integration/ Clustering algorithms:**

As mentioned in the design section the algorithm was made to use the k-means and apriori algorithms for the integration and clustering algorithm. But however due to the data availability we have designed our own algorithm for the clustering of data based on the k-means methodology. The algorithm has the below mentioned steps:

* Retrieving of data from the stored data in Hadoop.
* Clustering of data based on the genre and creating a list of movies
* Converting the data from the retrieved format to the JSON format in order to push the output to Solr using the curl commands

We were successfully able to retrieve the data and store it in the form of JSON and push the output into the Solr.

**Recommendation Algorithms:**

The recommendation algorithm basically consists of recommending the books to the users who have read the books of same genre but nit the others based on the ratings. We were able to recommend the required books to the users but somehow unable to track the books based on the ratings of each book of the genre. The following steps define the workflow phases in recommendation:

* Load the data from the Hadoop which is being stored in HDFS
* Quantify the book ratings
* Iterate through the ratings suggested by each user and specify the books to each user of those genre.
* Store the output in the JSON format and push the data to Solr.

**Solr Indexing:**

Solr indexing for the movies clustering was primarily based on the id and the title. The id consists of the genre and title consists of movies which belong to that genre.

The indexing for the books will be made on the recommendations where the id will be the user ID and the title will be the Book ISBN of the recommendation that is being made.

**Application Interface:**

The GUI is being developed in HTML5 and the internal operations such as retrieval of data from Solr and storing of user specifications are done through the java script. However we were unable to prepare the user credentials, for each user and track their history. The application interface will take the input as genre from the user and displays the suggested books and movies of the specified genre. The books will be recommended not only using the genre but also by the ratings.

**Outputs:**

The below snapshot depicts the output of clustering of movies which will be in the JSON format:

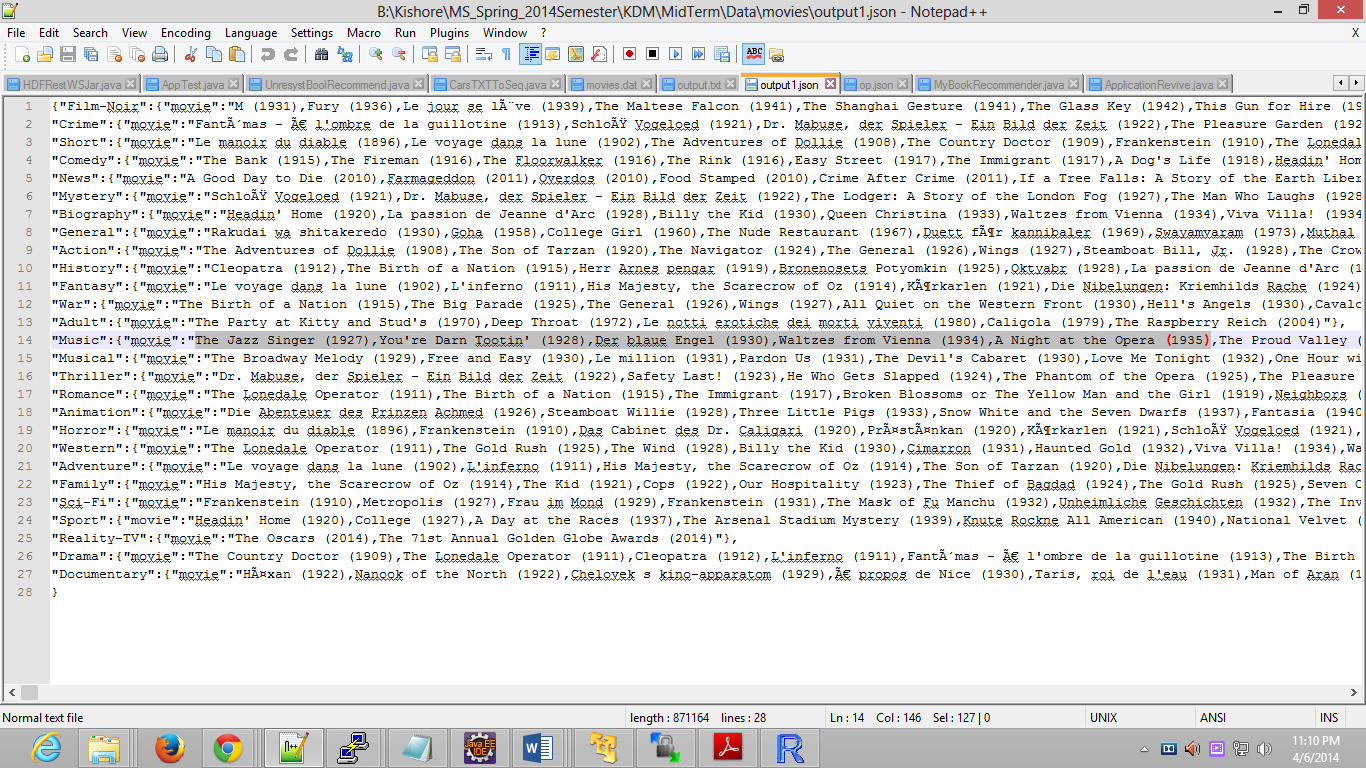


Figure 3Clustering data for movies

The below snapshot shows the pushing of the data retrieved from the clustering mechanism using curl command:

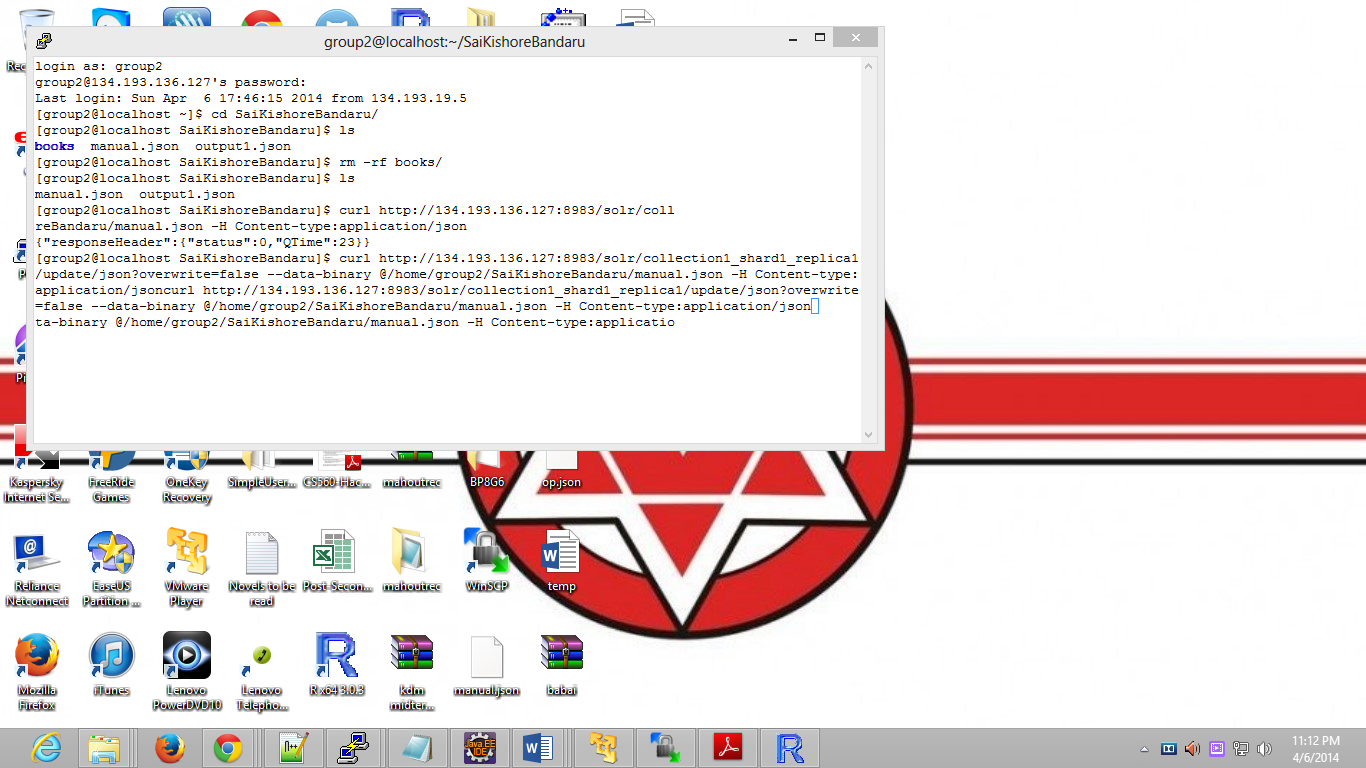


Figure 4Curl command execution for pushing data to Solr

The below snapshot depicts the data that is being retrieved from Solr which is being stored using curl command for the Comedy genre:

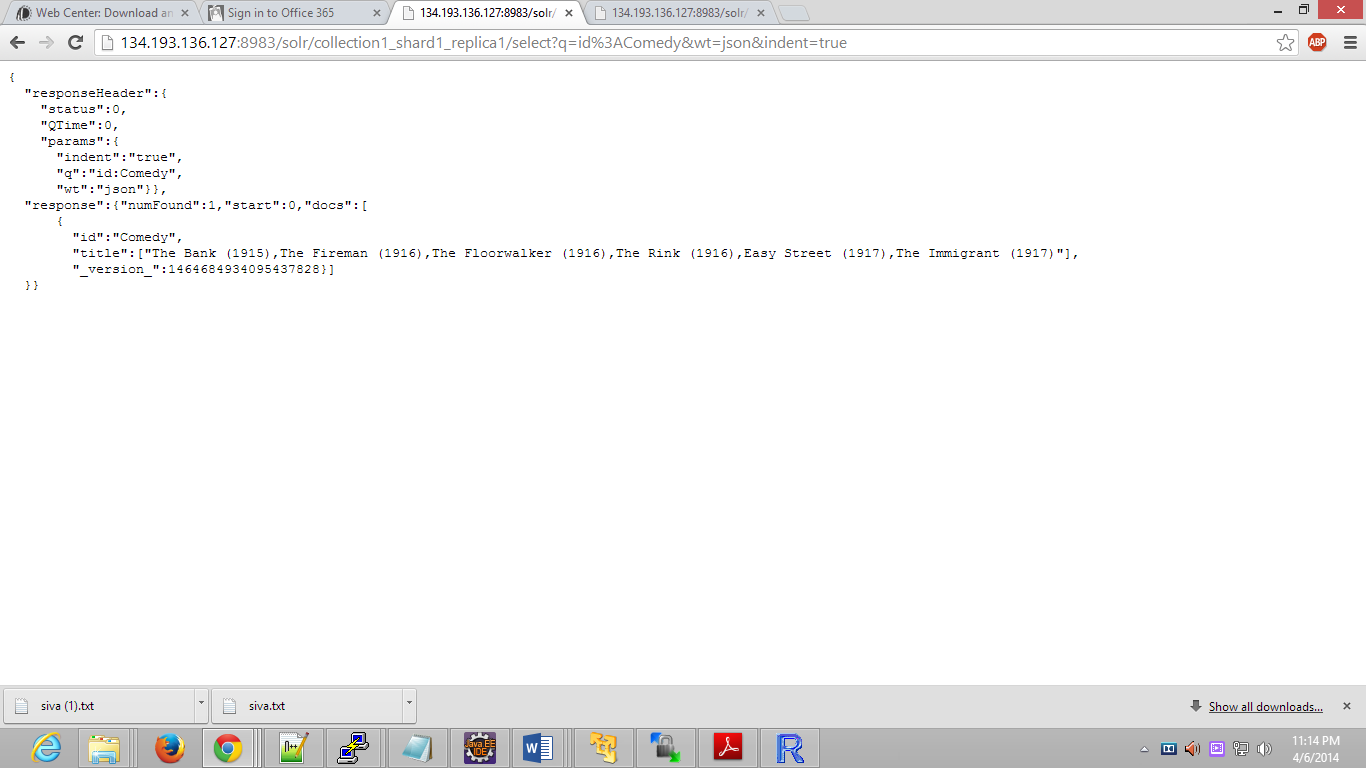


Figure 5Output from Solr for Comedy genre

The below snapshot shows the recommendations that are being for made for each user after running the recommendation algorithm:

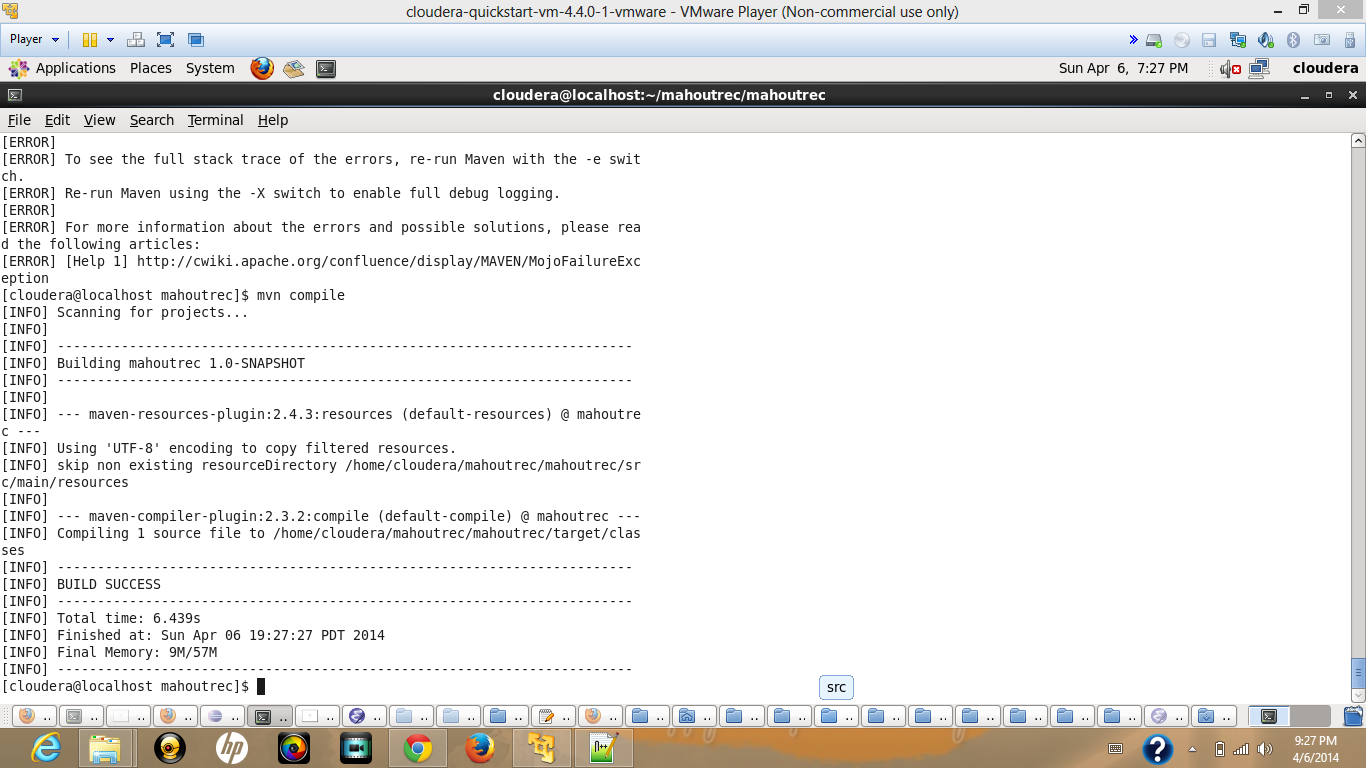


Figure 6Compliation and build success for recommendation code

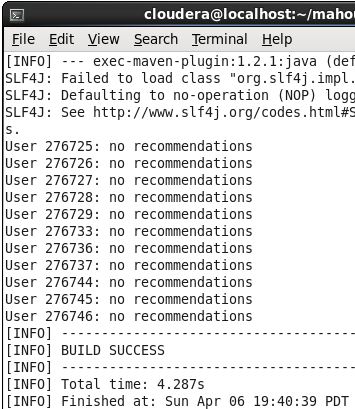


Figure 7Recommendations that are being made

The below snapshot shows the usage of application. The user needs to first select the genre that he wants for the viewing of movies and also for the books suggestions:

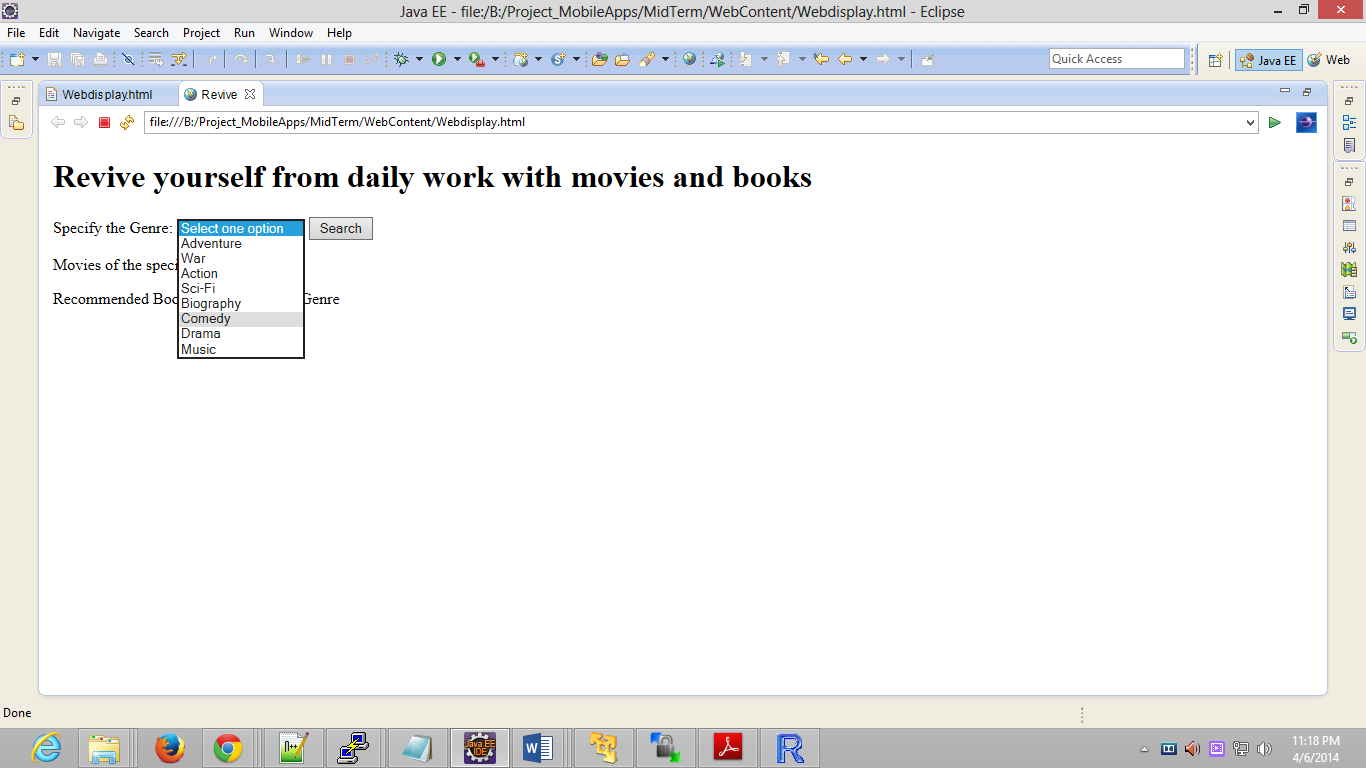


Figure 8GUI interface

The below snapshot reveals the results for books and movies based on the specified genre:

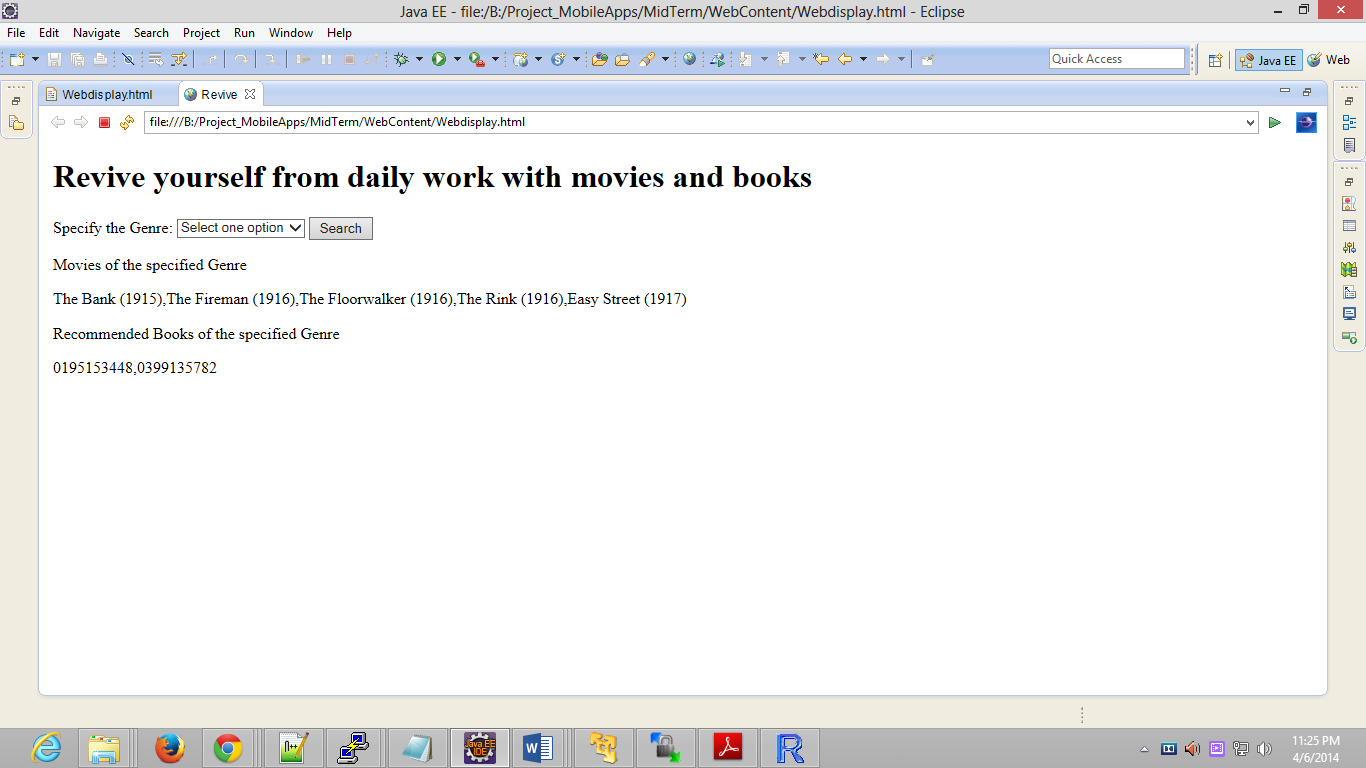


Figure 9 List of movies and books suggested(Books with ISBN numbers)

**GitHub URL:**

The following is the github link for entire code implementation:

* <https://github.com/SaiKishoreBandaru/MidTerm>

**Limitations:**

The limitations of this project are we couldn’t include the user validation and user logins. The data between the recommendation and clustering couldn’t be matched through the basis of user ids as the user validations are not maintained. So the book recommendations are primarily based on the genre not on the ratings and user history.

**Enhancements:**

* Additions include the user validations and maintain content personalization
* Recommendation of books with title names instead ISBN number.

**References:**

* Datasets - <https://github.com/sidooms/MovieTweetings>
* R machine learning - <http://ocw.mit.edu/courses/sloan-school-of-management/15-097-prediction-machine-learning-and-statistics-spring-2012/lecture-notes/MIT15_097S12_lec02.pdf>
* Clustering - <https://tomcat.apache.org/tomcat-6.0-doc/cluster-howto.html>
* Recommendation Algorithm - <http://chimpler.wordpress.com/2013/02/20/playing-with-the-mahout-recommendation-engine-on-a-hadoop-cluster/>