Homework 5

Yu-Ru Lin

University of Pittsburgh
INFSCI 2160: Data Mining

yurulin@pitt.edu

2017-03-15

Homework 5 I

You will use the Newsgroups dataset, the movie rating dataset, and the book rating dataset for this assignment.

- Submit your report in PDF, and your code in *.R, via courseweb.
  - If you use "R Markdown," make sure your code is reproducible, and you provide clear answers (elaborate description) to the questions.
  - Follow the homework guideline posted on Piazza.
- Due: 2017-03-28 11.59am
- You can reference to the sample R code

Homework 5 II

Task 1 (Text Mining): analyze the topical clusters from text data

Dataset & description: http://www.cs.umb.edu/~smimarog/textmining/datasets/
Data csv: http://www.yurulin.com/class/spring2017_datamining/data/Newsgroup.csv

1. Plot the histogram of number of documents per topic. Find and list the four most popular topics in terms of number of documents.
2. Extract contents in these top 4 topics as your corpus. Run pre-processing on this corpus and use terms that appear at least four times in the corpus to create a term-document matrix. Use the term-document matrix to generate an MDS plot where each node represents a document with color indicating its topic.

Homework 5 III

3. Apply TFIDF weighting, latent semantic analysis (LSA) and non-negative matrix factorization (NMF) on the term-document matrix. Generate MDS plots corresponding to these matrices (TFIDF weighted matrix, LSA approximated matrix, and NMF approximated matrix). ***Hint***: see the sample code from class08.
4. Write down your observation based on these plots.

Task 2 (Network Analysis): create a movie-movie network and identify the community structure and central nodes

Download the movie rating dataset (MovieLens 100k) from:

`http://grouplens.org/datasets/movielens/`

Read the dataset description:

`http://files.grouplens.org/datasets/movielens/ml-100k-README.txt`

Homework 5 IV

1. Create a movie-to-movie co-rating network. Load u.data dataset, and extract data where the ratings are generated after timestamp 03/20/1998 00:00:00 and equal to rating 5. Extract the top 30 most frequently rated movies as the nodes. Generate a network with edge weights $>= 10$, i.e., two movies have a link if they are rated by at least 10 common users. Load u.item to replace movieID with movies titles. List the names of the top 10 movies and their number of ratings. ***Hint***: see the hw5 sample code on how to extract popular movies.
2. Identify the community structure in the network by using the modularity-based community detection algorithm. Plot the network with the detected community structure (use 'plot') and the dendrogram (use 'dendPlot').

Homework 5 V

3. Identify the most central nodes in the network based on different centrality measures, degree centrality, closeness centrality, betweenness centrality, and PageRank. Plot different networks where the nodes are sized based on the centrality measures. Highlight the top 5 nodes with the highest centrality measures in each network.
   Hint: see the sample code from class10.
4. Write down your observation based on these plots.

Task 3 (Recommendation): create a recommender system based on the book rating data

Homework 5 VI

Download the book rating dataset from:

http://www2.informatik.uni-freiburg.de/~cziegler/BX/

In the dataset, load the BX-Book-Ratings.csv & BX-Books.csv data. Extract books published from 1998 to present. Create a book rating matrix from these books, and in this matrix, only consider books that were rated by at least 10 users, and users that rated at least 10 books. ***Hint***: see the hw5 sample code on how to extract the book-rating matrix.

1. Run and test a recommender system built with different recommendation methods, including random, popular, user-based collaborative filtering, item-based collaborative filtering. Evaluate the different methods by using *k*-fold cross-validation (*k*=4). Generate a performance table in terms of performance measures MAE, MSE and RMSE as follows.

Homework 5 VII

	Random	Popular	UBCF	IBCF
MAE				
MSE				
RMSE				

Hint: Use the ’recommenderlab’ package. The documentation can be found:

http://cran.r-project.org/web/packages/recommenderlab/vignettes/recommenderlab.pdf

Hint: see the sample code from class11.

2. Write down your observation based on the performance table.