Semester Project

Assignment:
You are the new junior member of a data mining consulting firm. A new client has come to the firm with a new project. They have acquired Goodreads data and would like you to build a simple four book recommender system based off the data they have provided.

Your boss would like you to use three different data mining modeling techniques to verify the recommendations are “the best recommendations possible”, and assess which technique is the best. He would like you to use at least one modeling technique that you have researched on your own and not just modeling techniques the company already uses (ones we have discussed in class).

If you really wanted to impress the boss, you might try pulling in data from Amazon, Google Books, WorldCat, or the Open Library to try and augment the data used to recommend books

The client has provided the data in three files:

goodreads_books.csv
book_tags.csv
tags.csv

- goodreads_books.csv contains metadata about each book.
- book_tags.csv contains a list of each book, a tag id for a keyword the book has been tagged with, and the number of people who have tagged the book with that keyword.
- tags.csv contains a list of tag ids and their corresponding keyword tags.
- goodreads_books.csv and book_tags.csv are tied together through goodreads_book_id .
- book_tags.csv and tags.csv are tied together through tag_id .

The details:
The data was collected from community entered data, and as such is not the cleanest. There is duplication in the tags due to inconsistencies in spelling or different representation of the same tags. Before you build the recommender system, you will need to clean the data up and settle on an official representation.

The client was unable to determine what the metadata attributes in books.csv stand for. You will need to explore those on your own, determine what the level of measurement for each is, and make your own assessment on if the attribute can be used to build the recommender system.

Create a simple recommender system using three different modeling techniques. No more than two of those modeling techniques should be ones discussed in class. At least one of those modeling techniques should be a technique you research on your own. Techniques we discuss in class are 1R Classifier, SLR, MLS, K-Means, K-Nearest Neighbor, ID3, Logistic Regression,

Perceptron Regression, DBScan, Agglomerative Hierarchical Clustering, SVM, Gaussian Mixtures, and ANN.

Notes:
You may work together in teams of up to 4 people. Supply me with a list of names before you start to work together. You will be required to anonymously evaluate the contributions of your team members, and anyone not contributing will be given a zero percent on the semester project.

To Evaluate the performance of the different models you will need to make a determination on which model is better and explain the criteria you used to make the determination.

For your Results you must demonstrate the books recommended by the recommender system for someone who has read Stardust by Neil Gaiman. Then demonstrate the books recommended by the recommender system for one other book of your choice.

Check your code into git-classes.mst.edu and grant developer access to koobp and the grader.

Post a pdf of your report on the models evaluated and used. Please make sure to conform to the submission template in Canvas.

When you have completed a model, prepare a PowerPoint presentation of your semester project to show your boss and the client. Make sure to highlight what is in the data, which attributes you used, difficulties you had making the recommender system, the modeling techniques you used, the technique you settled on, and finally, demonstrate your working model. You will have a separate assignment for you to submit your PowerPoint slides.

Once the slides are completed, record a presentation and post it to the Semester Project discussion post for other to review and post comments on.