Project for CSG230: Description & Rubrics

Description:
In this project your task is to **design two models**(algorithms) that work well on the movie recommendation task, which comes from the famous competition hold by Netflix(http://www.netflixprize.com/). Search for collaborative filtering on Google or Google Scholar. The paper from http://ccs.mit.edu/papers/CCSWP165.html may provide a good starting point since it presents a system that is relatively simple.

You will get the data set later through email with description on details, which is different from the one from Netflix. In it you can find the data of ranking movies from users, and also the data of the contents of the movies. **The first model will not use the data of contents of movies(no-content based), while the second one will(content based)**.

Of course you can test your code using the original data from Netflix(you need to sign up as a team on Netflix.com for the data), which may need more work on how to deal with large volume data set.

Your models should not be trivial, where a trivial model is one that simply predicts averages. You need to **implement your algorithms** and **tune** them. You can obtain a model from literature, but you should understand and implement it instead of running the source code without any efforts of yourself.

Prepare a **report** that describes what you did. The report will preferably be organized in the following way:

- **Information about your team**. List all the members of the team with NUID(last 4 digits) and their contributions to the project.
- **Introduction**: Describe the problem of collaborative filtering. Describe some existing approaches. Successful unification and classification of existing approaches will be favored in the grading.
- **Theoretical Discussion**: Describe sufficiently well two methods (either your own or from literature) for collaborative filtering.
- Try to provide some theoretical justification why you expect those methods to work. Discuss any modeling assumptions as well as practical “hacks” you made in your model.
- **Results:** Report the root mean squared error (RMSE) on the tuning and test sets for both methods. Report the values of the main parameters of your models that you used. If you have studied how the RMSE depends on a particular parameter you may provide extra tables and plots. Investigation of parameter effects will be favored in grading.
- **Discussion and error analysis**: Try to interpret the results of your models.
- **Appendices and References:** Add necessary appendices and references to support your contents.

Your report should **not exceed 15 pages**. Only content but not length of report will be considered in grading. The report should be typed. The report should be submitted in hard copy(or emailed) on the due date.

There are no restrictions on the use of languages (i.e.Perl, C++, etc.) and tools (i.e. Matlab, WEKA, etc.) as well as operating systems (Windows, Linux, MAC, etc.).

In the course of doing the project you may find yourself implementing and experimenting with more than one method. You can report those results only if they are interesting, i.e. the methods are substantially different than other methods on which you report *or* produce good results. Remember you should have at least one algorithm in the two categories mentioned (using and not using content information from the movies). Extra credit may be given if you experiment with more methods.

Deadline:

23:59 EST, Dec 3rd, 2008
Notice: There will be no extension on this deadline!!!

Submission Checklist:

- Two files with results
- Report
- Source code
- Readme file that tells us how to run your code

Grading:

The grading will be based **solely on the report** you provide. **Try to make the report representative of your efforts.** Even if you cannot obtain good results in terms of RMSE you should report what you did. Points will be assigned as follows:

Report Section	Max. points
Introduction	**10**
Theoretical discussion of method 1 (non-content based) method)	**20**
Theoretical discussion of method 2 (content based method)	**20**
Results on method 1	**20**
Results on method 2	**20**
Discussion	**30**
Appendices and references	**10**
Total	**130**

Extra credit: (up to 20% from the regular credit) may be given for especially good sections of the report or experimentation with more algorithms than required**.**

Notice: If this rubric is different from the project details you got from website, please refer to this rubric.

You can either do the projects alone, or be part of a group of no more than three people. **Doing the project alone will not be weighted more in grading.** Always every group member will receive exactly the same project score regardless of how much contribution each individual has, **unless there are unfair situations described in the contribution information in your report**.

Please use one compressed file named src.* for all your source code of you projects, and your report with name doc.* in another compressed file separately. Use zip or gzip to compress your files(so * above will be zip or gz). Send your submission with **the mail title** in the following format:

[CSG230]project_**xxxx**

Here **xxxx** is the last 4 digits of NUID of the person who submits it.

Please submit to Jian Wen<jarodwen@ccs.neu.edu>.

Notice: Too large attachment or improper extensions(.exe, .rar) will cause transmission failure in the email system in our department. Please make an appointment with me if your project are larger than 8MB.