Information Retrieval: Homework 2.

Project Proposal: Classification of Online User Review and Five Stars Scaling

I. Motivation

As a part of my interest of recommendation system, the classification of online user reviews is also an important tool for some recommendation systems. In this project, I want to explore several methods that can not only classify a review as a good one or bad one, but also scale it into 5 stars rating scale.
The project will first investigate approaches using in the research of Bo Pang and Lillian Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," and then comparing between Pang, and Lee's approaches to the approach using in this project.

II. Methods

In other to work on this projects, several tools will be using such as Support Vector Machines, Weka Toolkit, and the main programming language is Java.
Some of information extraction may be applied to preprocess the database; however, I still haven't understood how this can be used.

III. Pilot Results

Database

The database using for the pilot results was the same one as Pang and Lee did in their research. The scaleset database can be found here: http://www.cs.cornell.edu/people/pabo/movie-review-data/

First approach

The first approach using to analyze the review in database was base on the #good-review-keywords divided by #bad-review-keywords. This ratio will then map to the 5 stars rating scale.

A review is:

5 stars if ratio >= 2
4 stars if 1.3 <= ratio < 2
3 stars if 0.9 <= ratio < 1.3
2 stars if 0.5 <= ratio < 0.9
1 star if ratio < 0.5

The numbers are chosen based on the analysis of many reviews in the database.

Result

The method using is very simple, but turn out it works very well on scaling into 5 stars rating. Particularly, it roughly the same as the result of Pang and Lee did on their research.
Surprisingly, it classifies the bad and good review very well.

Plans on final project

In the final project, I will try to extends the use of keywords by the group of keywords, and sentences in order to determine if the review is good or bad, and then using the ratio of #good-review-keywords-sentences/#bad-review-keywords-sentences to scale the review to 5 stars rating scale.
Currently, the good and bad review keywords are defined as a static arrays, but the methods of information extraction will also be researched in order to extract the good keywords from the review.