Liubov Ivashov

Final Project Prospectus: Amazon Fine Food reviews sentiment analysis

Context

This dataset consists of reviews of fine foods from amazon. The data span a period of more than 10 years, including all ~500,000 reviews up to October 2012. Reviews include product and user information, ratings, and a plain text review. It also includes reviews from all other Amazon categories. The data source: Kaggle

Attribute Information:

- Id
- ProductId — unique identifier for the product
- UserId — unique identifier for the user
- ProfileName
- Helpfulness Numerator — number of users who found the review helpful
- HelpfullnessDenominator — number of users who indicated whether they found the review helpful or not
- Score — rating between 1 and 5
- Time — timestamp for the review
- Summary — brief summary of the review
- Text — text of the review

The questions to be investigated

- Classify a review, determine whether the review is positive (Rating of 4 or 5) or negative (rating of 1 or 2).
- Build Simple recommender system

The methodological approach

- Data cleaning + EDA
- Tokenization

- Sentiment analysis using tidytext package using sentiment lexicons AFINN, bing and nrc
- Visualize the results of the sentiments analysis
- Evaluate 8 different models (Item Based Collaborative Filtering with Cosine (IBCF_cosine), Item Based Collaborative Filtering Pearson (IBCF_Pearson), Item Based Collaborative Filtering Pearson Euclidean Distance (IBCF_Euclidean) for similarity score; Single Value Decomposition (SVD); Alternating Least Squares (ALS); Popular; Random) to get ourselves the best working recommender system using **recommenderlab** -> R package, to help us create and evaluate the models we need.

The potential challenges

- Big amount of data -> will take some time to run and evaluate the models for recommendation system. Pre-processing will also take some time.
- We can only suggest three items for the recommendation system. However, we have to make sure that highly relevant products get selected.