\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

•What is the motivation of your project?

• Which part of the course topics is most related to your project?

• What are the deliverables you plan to submit by the end of this project?

• Which tentative dataset do you use in your project? Show the basic statistics of the dataset.

• What techniques/algorithms will you apply?

• How will you demonstrate the main function or analysis results?

• What is the existing work that is related to your project, if applicable?

• A very rough timeline to show your project milestone. (The timeline doesn't have to be accurate.)

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Dataset:

<https://www.kaggle.com/snap/amazon-fine-food-reviews>

This dataset consists of a single CSV file, Reviews.csv, and a corresponding SQLite table named Reviews in database.sqlite. The columns in the table are:

* Id
* ProductId - unique identifier for the product
* UserId - unqiue identifier for the user
* ProfileName - name of the user
* HelpfulnessNumerator - number of users who found the review helpful
* HelpfulnessDenominator - number of users who indicated whether they found the review helpful
* Score - rating between 1 and 5, rating of the product
* Time - timestamp for the review (unix time)
* Summary - brief summary of the review
* Text - text of the review

recommender systems, expertise, user modeling

perceptions of products to change

* The age of the product
* The age of the user
* The community the user belongs to

We model such ‘personal development’ through the lens of user experience, or expertise. Starting with a simple definition, experience is some quality that users gain over time, as they consume, rate, and review additional products.

Latent parameters:

‘Experience’ and ‘expertise’

In particular, our goal is not to say whether experienced/expert users are ‘better’ or ‘more accurate’ at rating products. We simply model the fact that users with the same level of experience/expertise rate products in a similar way.

We also note that while we obtain significant benefits on Amazon data, the mean-squared-errors for this dataset are by far the highest.

We use two schemes to build our test sets: our first scheme consists of selecting a random sample of reviews from each user. This is the standard way of selecting test data for ‘flat’ models that do not model temporal dynamics.

The second scheme we use to build our test set is to **consider the final reviews for each user**

To predict rating

One reason is that Amazon users use a full spectrum of ratings from 1 to 5 stars, whereas CellarTracker users (for example) rate wines on a smaller spectrum (after normalizing the ratings, most wines have scores above 4.25); this naturally leads to higher MSEs.

Another reason is that our Amazon data has many products and users with only a few reviews, so that we cannot do much better than simply modeling their bias terms. As we see in Section 5, bias terms differ significantly between beginners and experts, so that modeling expertise proves extremely beneficial on such data.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

1. **Motivation**

**[Point form to collect all thoughts]**

* Find out most and least popular items
  + For item on sales / recommended item
  + Frequent itemset? SVD, CUR, PMF
* To predict the recommendation’s (1) score, OR (2) positive or negative, OR (3) helpfulness
  + (1) score
  + (2) positive or negative, more or less the same as (1), but > 3 be positive and <3 be negative
  + (3) Helpfulness, for example Numerator / Denominator %, say, >80% means “helpful”
  + It depends on how deep we gonna do: SVD, CUR, SVD++ matrix completion (Netflix problem), modeling user expertise (utilize with “time”)
* To predict the above three things only based on “Summary” and “Text”
  + Text mining
  + Frequent itemset on words
  + MapReduce to process words

**[Paragraph form]**

To determine the popularity of a product, a straight forward way is to collect ratings and comments from users. However, the number of responses would not always be sufficiently large enough for reference. In other words, we have a sparse dataset with users versus ratings. The methodology of matrix completing helps to extend the sparse data and have an estimated complete data matrix for further study. In the Amazon Fine Food Reviews, there are user’s scores on different products and our goal is to investigate the popularity of the products. This will benefit online store, such as Amazon, to perform further actions to improve user experience and yield more revenue, for example, to place popular products under “Recommended Items” to drive traffic; or to group less popular products for sale and promotions.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**2 Topics related and deliverables**

Our project is related to collaborative filtering and we would use Probabilistic Matrix Factorization for analysis. We expect to come up with the most and least popularity products.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**3 Algorithm, dataset and demonstration method**

The tentative dataset is the Amazon Fine Food Reviews data. It consists of 568,454 food reviews from Amazon users during Oct 1999 – Oct 2012. There are 256,059 users and 74,258 products.

We would use matrix factorization techniques to perform Probabilistic Matrix Factorization and followed by optimization steps.

***[Anything we can use mapreduce or distributed systems?]***

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**4 Related work**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*