Homework

Recommender Systems

11/06/2015

Due: Sunday, November 15.

In this homework, you will analyze a part of Amazon product review data collected by Julian McAuley http://jmcauley.ucsd.edu/data/amazon/

The file *videoGames.json.gz* contains data. The starter script will load data for you. There are 100,000 reviews. For each review we have the following information:

- **itemID** The ID of the item. This is a hashed product identifier from Amazon.
- **reviewerID** The ID of the reviewer. This is a hashed user identifier from Amazon.
- **rating** The rating that the reviewer gave to the item.
- **helpful** Helpfulness votes for the review. This is a list with two subfields, 'nHelpful' and 'outOf'. The latter is the total number of votes this review received, the former is the number of those that considered the review to be helpful.
- **reviewText** The text of the review.
- **summary** Summary of the review.
- **unixReviewTime** Time of the review in seconds since 1970.
- **reviewTime** Plain-text representation of the review time.
- **category** Category labels of the product being reviewed.

The starter script will create a rating matrix from this information. As you will see, the starter script keeps only users that have rated more than 2 video games and video games rated by more than 3 users. This is primarily done for computational reasons, so that you do not have to wait too long before getting an answer.

Your tasks:

- Find the user that has rated the most amount video games.
- Which video games has been rated by the most amount of users?
- Find the user that is most similar to "U141954350".
- Recommend a video game to the user "U141954350". For example, you may find a video game that the user "U141954350" has not bought yet and you expect that he would rate it high. Explain the recommendation strategy you used.