Data Science Interview Challenge

For this exercise, you will analyze a dataset from Amazon. The data format and a sample entry are shown on the next page.

A. (Suggested duration: 90 mins)
With the given data for 548552 products, perform exploratory analysis and make suggestions for further analysis on the following aspects.

1. Trustworthiness of ratings
Ratings are susceptible to manipulation, bias etc. What can you say (quantitatively speaking) about the ratings in this dataset?

2. Category bloat
Consider the product group named 'Books'. Each product in this group is associated with categories. Naturally, with categorization, there are tradeoffs between how broad or specific the categories must be.

For this dataset, quantify the following:
a. Is there redundancy in the categorization? How can it be identified/removed?
b. Is is possible to reduce the number of categories drastically (say to 10% of existing categories) by sacrificing relatively few category entries (say close to 10%)?

B. (Suggested duration: 30 mins)
Give the number crunching a rest! Just think about these problems.

1. Algorithm thinking
How would build the product categorization from scratch, using similar/co-purchased information?

2. Product thinking
Now, put on your 'product thinking' hat.
a. Is it a good idea to show users the categorization hierarchy for items?
b. Is it a good idea to show users similar/co-purchased items?
c. Is it a good idea to show users reviews and ratings for items?
d. For each of the above, why? How will you establish the same?

Data entry format:

- **Id:** Product id (number 0, ..., 548551)
- **ASIN:** Amazon Standard Identification Number
- **title:** Name/title of the product
- **group:** Product group (Book, DVD, Video or Music)
- **salesrank:** Amazon Salesrank
- **similar:** ASINs of co-purchased products (people who buy X also buy Y)
- **categories:** Location in product category hierarchy to which the product belongs (separated by |, category id in [])
- **reviews:** Product review information: time, user id, rating, total number of votes on the review, total number of helpfulness votes (how many people found the review to be helpful)

Sample data entry:

Id: 15
ASIN: 1559362022
title: Wake Up and Smell the Coffee
group: Book
salesrank: 518927
similar: 5 1559360968 1559361247 1559360828 1559361018 0743214552
categories: 3
|Books[283155]|Subjects[1000]|Literature & Fiction[17]|Drama[2159]|United States[2160]
|Books[283155]|Subjects[1000]|Arts & Photography[1]|Performing Arts[521000]|Theater[2154]|General[2218]
|Books[283155]|Subjects[1000]|Literature & Fiction[17]|Authors, A-Z[70021]|(B)[70023]|Bogosian, Eric[70116]
reviews: total: 8 downloaded: 8 avg rating: 4
2002-5-13 customer: A2IGOA66Y6O8TQ rating: 5 votes: 3 helpful: 2
2002-6-17 customer: A2OIN4AUH84KNE rating: 5 votes: 2 helpful: 1
2003-1-2 customer: A2HN382JNT1CIU rating: 1 votes: 6 helpful: 1
2003-6-7 customer: A2FDJ79LDU4O18 rating: 4 votes: 1 helpful: 1
2003-6-27 customer: A39QMV9ZKRJXO5 rating: 4 votes: 1 helpful: 1
2004-2-17 customer: AUUVMSTQ1TXDI rating: 1 votes: 2 helpful: 0
2004-2-24 customer: A2C5K0QTLL9UAT rating: 5 votes: 2 helpful: 2
2004-10-13 customer: A5XYF0Z3UH4HB rating: 5 votes: 1 helpful: 1

Data source: http://snap.stanford.edu/data/amazon-meta.html