Phase 1 – Group 03

Dataset

Topic

We have chosen a dataset that compiles user's product reviews of the Amazon website that originally included 142.8 million reviews spanning May 1996 - October 2018.

This dataset was aggregated by researchers who made it available for non-commercial use.

The dataset was very large, so we have decided to reduce the volume and we used a random number generator to create a distribution of 0's and 1's where there is a 2% probability of generating a 1. We applied this mask so that only 2% of the entries are retained. The dataset file size was reduced from14.3GB to 1,8GB.

Last update

October 2018

Size

1,8GB

Link to the file

https://nijianmo.github.io/amazon/

File Type

JSONL

Information contained

This dataset includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs).

Identification of Business Capabilities

What our startup will do is to manage product reviews for e-commerce websites.

To demonstrate the capabilities we use the amazon dataset.

Our application needs to have the following:

- GET, POST, PUT, DELETE;
- GET for all data fields:

- reviewerID - ID of the reviewer, e.g. A2SUAM1J3GNN3B
- asin - ID of the product, e.g. 0000013714
- reviewerName - name of the reviewer
- vote - helpful votes of the review
- style - a disctionary of the product metadata, e.g., "Format" is "Hardcover"
- reviewText - text of the review
- overall - rating of the product
- summary - summary of the review
- unixReviewTime - time of the review (unix time)
- reviewTime - time of the review (raw)
- image - images that users post after they have received the product.