Amazon Customer Reviews Dataset

Amazon Customer Reviews (a.k.a. Product Reviews) is one of Amazon's iconic products. In a period of over two decades since the first review in 1995, millions of Amazon customers have contributed over a hundred million reviews to express opinions and describe their experiences regarding products on the Amazon.com website. This makes Amazon Customer Reviews a rich source of information for academic researchers in the fields of Natural Language Processing (NLP), Information Retrieval (IR), and Machine Learning (ML), amongst others. Accordingly, we are releasing this data to further research in multiple disciplines related to understanding customer product experiences. Specifically, this dataset was constructed to represent a sample of customer evaluations and opinions, variation in the perception of a product across geographical regions, and promotional intent or bias in reviews.

Accessing the Amazon Customer Reviews Dataset

Over 130+ million customer reviews are available to researchers as part of this release. The data is available in TSV files in the `amazon-reviews-pds` S3 bucket in AWS US East Region. Each line in the data files corresponds to an individual review (tab delimited, with no quote and escape characters). Samples of the data are available in English and French; more details on the information in each column can be found here.

If you use the AWS Command Line Interface, you can list data in the bucket with the "ls" command:

```
aws s3 ls s3://amazon-reviews-pds/tsv/
```

To download data using the AWS Command Line Interface, you can use the "cp" command. For instance, the following command will copy the file named amazon_reviews_us_Camera_v1_00.tsv.gz to your local directory:

```
aws s3 cp s3://amazon-reviews-pds/tsv/amazon_reviews_us_Camera_v1_00.tsv.gz .
```

You may list the available files using the CLI or check this index file which includes URLs for all available files.

About the Data

The dataset contains the customer review text with accompanying metadata, consisting of three major components:

1. A collection of reviews written in the Amazon.com marketplace and associated metadata from 1995 until 2015. This is intended to facilitate study into the properties (and the evolution) of customer reviews potentially including how people evaluate and express their experiences with respect to products at scale. (130M+ customer reviews)
2. A collection of reviews about products in multiple languages from different Amazon marketplaces, intended to facilitate analysis of customers' perception of the same products and wider consumer preferences across languages and countries. (200K+ customer reviews in 5 countries)
3. A collection of reviews that have been identified as non-compliant with respect to Amazon policies. This is intended to provide a reference dataset for research on detecting promotional or biased reviews. (several thousand customer reviews). This part of the dataset is distributed separately and is available upon request – please contact the email address below if you are interested in obtaining this dataset.

Data Formats

The dataset is currently available in two file formats.

1. Tab separated value (TSV), a text format - s3://amazon-reviews-pds/tsv/
2. Parquet, an optimized columnar binary format - s3://amazon-reviews-pds/parquet/

To further improve query performance the Parquet dataset is partitioned (divided into subfolders) on S3 by **product_category**. This allows for queries using a **WHERE** clause on **product_category** to only read data specific to that category.

For example:

```
SELECT product_title, star_rating FROM table_name WHERE product_category = 'Books'
```

The above SQL query will only read data from s3://amazon-reviews-pds/parquet/product_category=Books/ improving performance and reducing query costs.

To quickly get started with the dataset, in regions where AWS Glue is available you can use a nice feature called the crawler to automatically discover the data and create the required tables you will later query.

Alternatively you can head over to the Amazon Athena console and manually create a table as follows:

```
CREATE EXTERNAL TABLE amazon_reviews_parquet(
  marketplace string,
  customer_id string,
  review_id string,
  product_id string,
  product_parent string,
  product_title string,
  star_rating int,
  helpful_votes int,
  total_votes int,
  vine string,
  verified_purchase string,
  review_headline string,
  review_body string,
  review_date bigint,
```

```
  year int)
PARTITIONED BY (product_category string)
ROW FORMAT SERDE
  'org.apache.hadoop.hive.ql.io.parquet.serde.ParquetHiveSerDe'
STORED AS INPUTFORMAT
  'org.apache.hadoop.hive.ql.io.parquet.MapredParquetInputFormat'
OUTPUTFORMAT
  'org.apache.hadoop.hive.ql.io.parquet.MapredParquetOutputFormat'
LOCATION
  's3://amazon-reviews-pds/parquet/'
```

Once the table is created execute the following in the Athena console only once:

```
MSCK REPAIR TABLE amazon_reviews_parquet
```

Contact

If you have questions about the data, please email us at customer-review-dataset@amazon.com.

License

By accessing the Amazon Customer Reviews Library (“Reviews Library”), you agree that the Reviews Library is an Amazon Service subject to the Amazon.com Conditions of Use and you agree to be bound by them, with the following additional conditions:

In addition to the license rights granted under the Conditions of Use, Amazon or its content providers grant you a limited, non-exclusive, non-transferable, non-sublicensable, revocable license to access and use the Reviews Library for purposes of academic research. You may not resell, republish, or make any commercial use of the Reviews Library or its contents, including use of the Reviews Library for commercial research, such as research related to a funding or consultancy contract, internship, or other relationship in which the results are provided for a fee or delivered to a for-profit organization. You may not (a) link or associate content in the Reviews Library with any personal information (including Amazon customer accounts), or (b) attempt to determine the identity of the author of any content in the Reviews Library. If you violate any of the foregoing conditions, your license to access and use the Reviews Library will automatically terminate without prejudice to any of the other rights or remedies Amazon may have.

This license language is also available at https://s3.amazonaws.com/amazon-reviews-pds/license.txt

---

AWS Public Datasets