Amazon product data

Julian McAuley, UCSD

New - Q/A data!

See our newly-released Q/A data (described in our WWW 2016 paper)!

Description

This dataset contains product reviews and metadata from Amazon, including 142.8 million reviews spanning May 1996 - July 2014.

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).

Files

"Small" subsets for experimentation

If you're using this data for a class project (or similar) ***please*** consider using one of these smaller datasets below before requesting the larger files. To obtain the larger files you will need to contact me to obtain access.

K-cores (i.e., dense subsets): These data have been reduced to extract the k-core, such that each of the remaining users and items have k reviews each.

Ratings only: These datasets include no metadata or reviews, but only (user,item,rating,timestamp) tuples. Thus they are suitable for use with mymedialite (or similar) packages.

Books	5-core (8,898,041 reviews)	ratings only (22,507,155 ratings)
Electronics	5-core (1,689,188 reviews)	ratings only (7,824,482 ratings)
Movies and TV	5-core (1,697,533 reviews)	ratings only (4,607,047 ratings)
CDs and Vinyl	5-core (1,097,592 reviews)	ratings only (3,749,004 ratings)
Clothing, Shoes and Jewelry	5-core (278,677 reviews)	ratings only (5,748,920 ratings)
Home and Kitchen	5-core (551,682 reviews)	ratings only (4,253,926 ratings)
Kindle Store	5-core (982,619 reviews)	ratings only (3,205,467 ratings)
Sports and Outdoors	5-core (296,337 reviews)	ratings only (3,268,695 ratings)
Cell Phones and Accessories	5-core (194,439 reviews)	ratings only (3,447,249 ratings)
Health and Personal Care	5-core (346,355 reviews)	ratings only (2,982,326 ratings)
Toys and Games	5-core (167,597 reviews)	ratings only (2,252,771 ratings)
Video Games	5-core (231,780 reviews)	ratings only (1,324,753 ratings)
Tools and Home Improvement	5-core (134,476 reviews)	ratings only (1,926,047 ratings)
Beauty	5-core (198,502 reviews)	ratings only (2,023,070 ratings)
Apps for Android	5-core (752,937 reviews)	ratings only (2,638,172 ratings)
Office Products	5-core (53,258 reviews)	ratings only (1,243,186 ratings)
Pet Supplies	5-core (157,836 reviews)	ratings only (1,235,316 ratings)
Automotive	5-core (20,473 reviews)	ratings only (1,373,768 ratings)
Grocery and Gourmet Food	5-core (151,254 reviews)	ratings only (1,297,156 ratings)
Patio, Lawn and Garden	5-core (13,272 reviews)	ratings only (993,490 ratings)
Baby	5-core (160,792 reviews)	ratings only (915,446 ratings)
Digital Music	5-core (64,706 reviews)	ratings only (836,006 ratings)
Musical Instruments	5-core (10,261 reviews)	ratings only (500,176 ratings)
Amazon Instant Video	5-core (37,126 reviews)	ratings only (583,933 ratings)

Complete review data

Please see the **per-category** files below, and only download these (large!) files if you really need them:

raw review data (20gb) - all 142.8 million reviews

The above file contains some duplicate reviews, mainly due to near-identical products whose reviews Amazon merges, e.g. VHS and DVD versions of the same movie. These duplicates have been removed in the files below:

user review data (18gb) - duplicate items removed (83.68 million reviews), sorted by user

product review data (18gb) - duplicate items removed, sorted by product

ratings only (3.2gb) - same as above, in csv form without reviews or metadata

5-core (9.9gb) - subset of the data in which all users and items have at least 5 reviews (41.13 million reviews)

Finally, the following file removes duplicates more aggressively, removing duplicates even if they are written by different users. This accounts for users with multiple accounts or plagiarized reviews. Such duplicates account for less than 1 percent of reviews, though this dataset is probably preferable for sentiment analysis type tasks:

aggressively deduplicated data (18gb) - no duplicates whatsoever (82.83 million reviews)

Format is one-review-per-line in (loose) json. See examples below for further help reading the data.

Sample review:


```
{
  "reviewerID": "A2SUAM1J3GNN3B",
  "asin": "0000013714",
  "reviewerName": "J. McDonald",
  "helpful": [2, 3],
  "reviewText": "I bought this for my husband who plays the
piano.  He is having a wonderful time playing these old hymns.
The music  is at times hard to read because we think the book
was published for singing from more than playing from.  Great
purchase though!",
  "overall": 5.0,
  "summary": "Heavenly Highway Hymns",
  "unixReviewTime": 1252800000,
  "reviewTime": "09 13, 2009"
}
```


where

- `reviewerID` - ID of the reviewer, e.g. A2SUAM1J3GNN3B
- `asin` - ID of the product, e.g. 0000013714
- `reviewerName` - name of the reviewer
- `helpful` - helpfulness rating of the review, e.g. 2/3
- `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)

Metadata

Metadata includes descriptions, price, sales-rank, brand info, and co-purchasing links:

metadata (3.1gb) - metadata for 9.4 million products

Sample metadata:


```
{
  "asin": "0000031852",
  "title": "Girls Ballet Tutu Zebra Hot Pink",
  "price": 3.17,
  "imUrl": "http://ecx.images-
amazon.com/images/I/51fAmVkTbyL._SY300_.jpg",
  "related":
  {
    "also_bought": ["B00JHONN1S", "B002BZX8Z6", "B00D2K1M3O",
"0000031909", "B00613WDTQ", "B00D0WDS9A", "B00D0GCI8S",
"0000031895", "B003AVKOP2", "B003AVEU6G", "B003IEDM9Q",
"B002R0FA24", "B00D23MC6W", "B00D2K0PA0", "B00538F5OK",
"B00CEV86I6", "B002R0FABA", "B00D10CLVW", "B003AVNY6I",
"B002GZGI4E", "B001T9NUFS", "B002R0F7FE", "B00E1YRI4C",
"B008UBQZKU", "B00D103F8U", "B007R2RM8W"],
    "also_viewed": ["B002BZX8Z6", "B00JHONN1S", "B008F0SU0Y",
"B00D23MC6W", "B00AFDOPDA", "B00E1YRI4C", "B002GZGI4E",
"B003AVKOP2", "B00D9C1WBM", "B00CEV8366", "B00CEUX0D8",
"B0079ME3KU", "B00CEUWY8K", "B004FOEEHC", "0000031895",
"B00BC4GY9Y", "B003XRKA7A", "B00K18LKX2", "B00EM7KAG6",
"B00AMQ17JA", "B00D9C32NI", "B002C3Y6WG", "B00JLL4L5Y",
"B003AVNY6I", "B008UBQZKU", "B00D0WDS9A", "B00613WDTQ",
"B00538F5OK", "B005C4Y4F6", "B004LHZ1NY", "B00CPHX76U",
"B00CEUWUZC", "B00IJVASUE", "B00GOR07RE", "B00J2GTM0W",
"B00JHNSNSM", "B003IEDM9Q", "B00CYBU84G", "B008VV8NSQ",
```

```
"B00CYBULSO", "B00I2UHSZA", "B005F50FXC", "B007LCQI3S",
"B00DP68AVW", "B009RXWNSI", "B003AVEU6G", "B00HSOJB9M",
"B00EHAGZNA", "B0046W9T8C", "B00E79VW6Q", "B00D10CLVW",
"B00B0AVO54", "B00E95LC8Q", "B00GOR92SO", "B007ZN5Y56",
"B00AL2569W", "B00B608000", "B008F0SMUC", "B00BFXLZ8M"],
    "bought_together": ["B002BZX8Z6"]
  },
  "salesRank": {"Toys & Games": 211836},
  "brand": "Coxlures",
  "categories": [["Sports & Outdoors", "Other Sports",
"Dance"]]
}
```


where

- `asin` - ID of the product, e.g. 0000031852
- `title` - name of the product
- `price` - price in US dollars (at time of crawl)
- `imUrl` - url of the product image
- `related` - related products (also bought, also viewed, bought together, buy after viewing)
- `salesRank` - sales rank information
- `brand` - brand name
- `categories` - list of categories the product belongs to

Visual Features

We extracted visual features from each product image using a deep CNN (see citation below). Image features are stored in a binary format, which consists of 10 characters (the product ID), followed by 4096 floats (repeated for every product). See files below for further help reading the data.

visual features (141gb) - visual features for all products

The images themselves can be extracted from the `imUrl` field in the metadata files.

Per-category files

Below are files for individual product categories, which have already had duplicate item reviews removed.

Books	reviews (22,507,155 reviews)	metadata (2,370,585 products)	image features
Electronics	reviews (7,824,482 reviews)	metadata (498,196 products)	image features
Movies and TV	reviews (4,607,047 reviews)	metadata (208,321 products)	image features
CDs and Vinyl	reviews (3,749,004 reviews)	metadata (492,799 products)	image features
Clothing, Shoes and Jewelry	reviews (5,748,920 reviews)	metadata (1,503,384 products)	image features
Home and Kitchen	reviews (4,253,926 reviews)	metadata (436,988 products)	image features
Kindle Store	reviews (3,205,467 reviews)	metadata (434,702 products)	image features
Sports and Outdoors	reviews (3,268,695 reviews)	metadata (532,197 products)	image features
Cell Phones and Accessories	reviews (3,447,249 reviews)	metadata (346,793 products)	image features
Health and Personal Care	reviews (2,982,326 reviews)	metadata (263,032 products)	image features
Toys and Games	reviews (2,252,771 reviews)	metadata (336,072 products)	image features
Video Games	reviews (1,324,753 reviews)	metadata (50,953 products)	image features
Tools and Home Improvement	reviews (1,926,047 reviews)	metadata (269,120 products)	image features
Beauty	reviews (2,023,070 reviews)	metadata (259,204 products)	image features
Apps for Android	reviews (2,638,173 reviews)	metadata (61,551 products)	image features
Office Products	reviews (1,243,186 reviews)	metadata (134,838 products)	image features
Pet Supplies	reviews (1,235,316 reviews)	metadata (110,707 products)	image features
Automotive	reviews (1,373,768 reviews)	metadata (331,090 products)	image features
Grocery and Gourmet Food	reviews (1,297,156 reviews)	metadata (171,760 products)	image features
Patio, Lawn and Garden	reviews (993,490 reviews)	metadata (109,094 products)	image features
Baby	reviews (915,446 reviews)	metadata (71,317 products)	image features
Digital Music	reviews (836,006 reviews)	metadata (279,899 products)	image features
Musical Instruments	reviews (500,176 reviews)	metadata (84,901 products)	image features
Amazon Instant Video	reviews (583,933 reviews)	metadata (30,648 products)	image features

Citation

Please cite one or both of the following if you use the data in any way:

Image-based recommendations on styles and substitutes
J. McAuley, C. Targett, J. Shi, A. van den Hengel
SIGIR, 2015
pdf

Inferring networks of substitutable and complementary products
J. McAuley, R. Pandey, J. Leskovec
Knowledge Discovery and Data Mining, 2015
pdf

Code

Reading the data

Data can be treated as python dictionary objects. A simple script to read any of the above the data is as follows:

```
def parse(path):
  g = gzip.open(path, 'r')
  for l in g:
    yield eval(l)
```

Convert to 'strict' json

The above data can be read with python 'eval', but is not strict json. If you'd like to use some language other than python, you can convert the data to strict json as follows:

```
import json
import gzip

def parse(path):
  g = gzip.open(path, 'r')
  for l in g:
    yield json.dumps(eval(l))

f = open("output.strict", 'w')
for l in parse("reviews_Video_Games.json.gz"):
  f.write(l + '\n')
```

Pandas data frame

This code reads the data into a pandas data frame:

```
import pandas as pd
import gzip

def parse(path):
  g = gzip.open(path, 'rb')
  for l in g:
    yield eval(l)

def getDF(path):
  i = 0
  df = {}
  for d in parse(path):
    df[i] = d
    i += 1
  return pd.DataFrame.from_dict(df, orient='index')

df = getDF('reviews_Video_Games.json.gz')
```

Read image features

```
import struct

def readImageFeatures(path):
  f = open(path, 'rb')
  while True:
    asin = f.read(10)
    if asin == '': break
    feature = []
    for i in range(4096):
      feature.append(struct.unpack('f', f.read(4)))
    yield asin, feature
```

Example: compute average rating

```
ratings = []
```

```
for review in parse("reviews_Video_Games.json.gz"):
  ratings.append(review['overall'])

print sum(ratings) / len(ratings)
```

Example: latent-factor model in mymedialite

Predicts ratings from a rating-only CSV file

```
./rating_prediction --recommender=BiasedMatrixFactorization --
training-file=ratings_Video_Games.csv --test-ratio=0.1
```