

Recommender Systems and Personalization Datasets

Julian McAuley, UCSD

Description

This page contains a collection of datasets that have been collected for research by our lab. Datasets contain the following features:

- user/item interactions
- star ratings
- timestamps
- product reviews
- social networks
- item-to-item relationships (e.g. copurchases, compatibility)
- product images
- price, brand, and category information
- GPS data
- heart-rate sequences
- other metadata

Please cite the appropriate reference if you use any of the datasets below.

Datasets are in (loose) json format unless specified otherwise, meaning they can be treated as python dictionary objects. A simple script to read json-formatted data is as follows:

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

Directory by Dataset

[Twitch](#) live-streaming interactions
[NPR](#) interview dialog data
[This American Life](#) podcast transcripts
[Recipes](#) and interactions from food.com
[Paired Recipes](#) from food.com
[EndoMondo](#) fitness tracking data
[Amazon](#) product reviews and metadata
[Amazon](#) question/answer data
[Amazon](#) marketing bias data
[Google Local](#) business reviews and metadata
[Steam](#) video game reviews and bundles
[Goodreads](#) book reviews
[Goodreads](#) spoilers
[Pinterest](#) fashion compatibility data
[ModCloth](#) clothing fit feedback
[ModCloth](#) marketing bias data
[RentTheRunway](#) clothing fit feedback
[Tradesy](#) bartering data
[RateBeer](#) bartering data
[Gameswap](#) bartering data
[Behance](#) community art reviews and image features
[Librarything](#) reviews and social data
[Epinions](#) reviews and social data
[Cant understanding data](#)
[Dance Dance Revolution](#) step charts
[NES](#) song data
[BeerAdvocate](#) multi-aspect beer reviews
[RateBeer](#) multi-aspect beer reviews
[Facebook](#) social circles data

[Twitter](#) social circles data
[Google+](#) social circles data
[Reddit](#) submission popularity and metadata

Directory by Metadata Type

The datasets below can be roughly organized in terms of the types of metadata they contain:

Review text: see [Amazon](#), [BeerAdvocate](#), [RateBeer](#), [Google Local](#), [Google Restaurants](#)
Image data: [Amazon](#), [Behance](#), [Pinterest](#), [Google Restaurants](#)
Item-to-item relationships: [Amazon](#)
Q/A data: [Amazon Q/A](#)
Geographical data: [Google Local](#), [Google Restaurants](#), [EndoMondo](#)
Heart-Rate data: [EndoMondo](#)
Bundle data: [Steam](#)
Peer-to-peer trades: [Tradesy](#), [RateBeer](#), [Gameswap](#)
Social connections: [Librarything](#), [Epinions](#)
Fit feedback: [Modcloth](#), [Renttherunway](#)
Multiple aspects: [BeerAdvocate](#), [RateBeer](#)

Twitch

Description

This is a dataset of users consuming streaming content on Twitch. We retrieved all streamers, and all users connected in their respective chats, every 10 minutes during 43 days.

Basic statistics

100k full		
Users:	100k	15.5M
Streamers (items):	162.6k	465k
Interactions:	3M	124M
Time steps:	6148	6148

Metadata

Start and stop times are provided as integers and represent periods of 10 minutes. Stream ID could be used to retrieve a single broadcast segment from a streamer (not used in our work).

- User ID (anonymized)
- Stream ID
- Streamer username
- Time start
- Time stop

Example

```
1,34347669376,grimmax,5415,5419
1,34391109664,jtgtv,5869,5870
1,34395247264,towshun,5898,5899
1,34405646144,mithrain,6024,6025
2,33848559952,chfhdtgus1,206,207
2,33881429664,sal_gu,519,524
2,33921292016,chfhdtgus1,922,924
```

Download link

See our [data folder](#) containing all Twitch files. The file [full_a.csv.gz](#) contains the full dataset while [100k.csv](#) is a subset of 100k users for benchmark purposes. The code is available in our [Github repository](#).

Citation

Please cite the following if you use the data:

Interview: NPR Media Dialog Data

Description

This dataset contains interview transcripts from [National Public Radio \(NPR\)](#). Data includes full interview transcripts and news article headlines.

Basic statistics

NPR	
Speakers:	185K
Episodes (Interviews):	106K
Utterances:	3.20M
Words:	126.7M

Metadata

- Episode Date and Title
- Speaker Names
- Speaker Utterances
- News Article Headlines

Example

```
episode: 79679
program: Talk of the Nation
title: Forecasting the Future of the Internet
date: 2006-05-26
episode_order: 48
speaker: Professor LARRY PETERSON (Princeton University)
utterance: And this is almost like the neutrality aspect of the issue, that there are places you just can't get to and the universal connectivity of the original Internet is deteriorating. Because of a lack of security built into the Internet your only recourse is to throw up all sorts of protections that are extremely suspicious of every bit of traffic that happens to fly by.
```

Download link

See the [Interview Dataset Page](#) for download information.

Citation

Please cite the following if you use the data:

Interview: Large-scale Modeling of Media Dialog with Discourse Patterns and Knowledge Grounding

Bodhisattwa Prasad Majumder*, Shuyang Li*, Jianmo Ni, Julian McAuley

EMNLP, 2020

[pdf](#)

This American Life Podcast Transcripts

Description

This dataset contains program transcripts from [This American Life](#). Data includes full program transcripts and associated audio.

Basic statistics

This American Life	
Speakers:	6,608

Episodes: 663
Utterances: 163,808
Words: 7,390,793

Metadata

- Episode Act
- Speaker Names
- Speaker Utterances
- Utterance Lengths
- Episode Audio

Example

```
episode: ep-1
act: prologue
utterance_start: 39.96
utterance_end: 54.89
duration: 14.93
speaker: ira glass
utterance: Well, one great thing about starting a new show is
utter anonymity. Nobody really knows what to expect from you. This
interviewee did not know us from Adam.
```

Download link

See the [This American Life Dataset Page](#) for download information.

Citation

Please cite the following if you use the data:

Speech Recognition and Multi-Speaker Diarization of Long Conversations
Huanru Henry Mao, Shuyang Li, Julian McAuley, Garrison W. Cottrell
INTERSPEECH, 2020
[pdf](#)

Food.com Recipe & Review Data

Description

These datasets contain recipe details and reviews from [Food.com](#) (formerly GeniusKitchen). Data includes cooking recipes and review texts.

Basic statistics

Food.com
Number of recipes: 231,637
Number of users: 226,570
Number of reviews: 1,132,367

Metadata

- Ratings and Reviews
- Recipe Name, Description, Ingredients, and Directions
- Recipe Categories (Tags)
- Recipe Nutrition Information

Example

Recipe:

```
name: beer mac n cheese soup
id: 499490
minutes: 45
contributor_id: 560491
submitted: 2013-04-27
tags: 60-minutes-or-less
      time-to-make
      preparation
```

nutrition:	678.8 70.0 20.0 46.0 61.0 134.0 11.0
n_steps:	7
steps:	cook the bacon in a pan over medium heat and set aside on paper towels to drain , reserving 2 tablespoons of the grease in the pan add the onion , carrot , celery and jalapeno and cook until tender , about 10-15 minutes add the garlic and cook until fragrant , about a minute mix in the flour and let it cook for 2-3 minutes add the broth , beer , nutmeg , bacon and macaroni and let cook until the macaroni is al-dente , about 7-8 minutes add the cream , mustard , worcestershire sauce and cheese and cook until the cheese has melted without bringing it back to a boil season with cayenne , salt and pepper to taste
description:	all of the flavors of mac n' cheese in the form of a hot bowl of soup! submitted by kevin lynch
ingredients:	bacon onion carrots celery jalapeno pepper garlic cloves flour chicken broth beer nutmeg elbow macaroni heavy cream dijon mustard worcestershire sauce cheddar cheese cayenne salt and pepper
n_ingredients:	17

Review:

user_id:	8937
recipe_id:	44394
date:	2002-12-01
rating:	4
review:	This worked very well and is EASY. I used not quite a whole package (10oz) of white chips. Great!

Download link

See the [Food.com Dataset Page](#) for download information.

Citation

Please cite the following if you use the data:

Generating Personalized Recipes from Historical User Preferences
 Bodhisattwa Prasad Majumder*, Shuyang Li*, Jianmo Ni, Julian McAuley
EMNLP, 2019
[pdf](#)

Recipe Pairs data

Description

This is a collection recipes paired with variants, e.g. a recipe matched with a vegan version of the same recipe.

Basic statistics

Food.com

Number of recipes: 83,000
Number of base recipes: 36,000
Number of target recipes: 60,000

Metadata

- Ratings and Reviews
- Recipe Name, Description, Ingredients, and Directions
- Recipe Categories (Tags)
- Recipe Nutrition Information

Download link

See the [Recipe Pairs Dataset Page](#) for download information.

Citation

Please cite the following if you use the data:

SHARE: a System for Hierarchical Assistive Recipe Editing

Shuyang Li, Yufei Li, Jianmo Ni, Julian McAuley

EMNLP, 2022

[pdf](#)

EndoMondo Fitness Tracking Data

Description

This is a collection of workout logs from users of [EndoMondo](#). Data includes multiple sources of sequential sensor data such as heart rate logs, speed, GPS, as well as sport type, gender and weather conditions.

Basic statistics

Users: 1,104
Workouts: 253,020

Metadata

- User Identifier
- Gender
- Sport type
- Latitude/Longitude/Altitude sequences (with timestamps)
- Heart rates
- Various derived sequences

Example

```
userId: 10921915
gender: male
sport: bike
id: 396826535
longitude: [24.64977040886879, 24.65014273300767, 24.650910682976246,
24.650668865069747, 24.649145286530256, ...]
latitude: [60.173348765820265, 60.173239801079035, 60.17298021353781,
60.172477969899774, 60.17186114564538, ...]
altitude: [-1.8044666444624418, -1.8190453555595787,
-1.8190453555595787, -1.8511185199732794, -1.871528715509271, ...]
timestamp: [1408898746, 1408898754, 1408898765, 1408898778,
1408898794, ...]
time_elapsed: [-0.12256752559145224, -0.12221090169596584,
-0.12172054383967204, -0.12114103000950663, -0.12042778221853381,
...]
heart_rate: [-8.197369036801112, -5.867841701016304,
-3.961864789919643, -4.173640002263717, -3.961864789919643, ...]
derived_speed: [-7.0829444390064396, -2.8061928357004815,
-0.3976286593020398, -0.7571073884764162, 2.6415189187026646, ...]
distance: [-4.372303649217691, -2.374952819539426,
-0.07926348591212737, 0.4284751220389811, 4.710835498111755, ...]
tar_heart_rate: [100, 111, 120, 119, 120, ...]
```

```

tar_derived_speed: [0, 10.751376415573548, 16.806294372816662,
15.902596545765366, 24.446443398153843, ...]
since_begin: [1378478.8892184314, 1378478.8892184314,
1378478.8892184314, 1378478.8892184314, 1378478.8892184314, ...]
since_last: [2158.84607810351, 2158.84607810351, 2158.84607810351,
2158.84607810351, 2158.84607810351, ...]

```

Download link

See the [FitRec Dataset Page](#) for download information.

Citation

Please cite the following if you use the data:

Modeling heart rate and activity data for personalized fitness recommendation
Jianmo Ni, Larry Muhistein, Julian McAuley
WWW, 2019
[pdf](#)

Amazon Product Reviews

Description

This is a large crawl of product reviews from Amazon. This dataset contains 82.83 million unique reviews, from around 20 million users.

Basic statistics

Ratings:	82.83 million
Users:	20.98 million
Items:	9.35 million
Timespan:	May 1996 - July 2014

Metadata

- reviews and ratings
- item-to-item relationships (e.g. "people who bought X also bought Y")
- timestamps
- helpfulness votes
- product image (and CNN features)
- price
- category
- salesRank

Example

```

{
  "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"
}

```

Download link

See the [Amazon Dataset Page](#) for download information.

The 2014 version of this dataset is [also available](#).

Citation

Please cite the following if you use the data:

Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering
R. He, J. McAuley

WWW, 2016

[pdf](#)

Image-based recommendations on styles and substitutes

J. McAuley, C. Targett, J. Shi, A. van den Hengel

SIGIR, 2015

[pdf](#)

Amazon Question and Answer Data

Description

These datasets contain questions and answers about products from the Amazon dataset above.

Basic statistics

Questions:	1.48 million
Answers:	4,019,744
Labeled yes/no questions:	309,419
Number of unique products with questions:	191,185

Metadata

- question and answer text
- is the question binary (yes/no), and if so does it have a yes/no answer?
- timestamps
- product ID (to reference the review dataset)

Example

```
{  
  "asin": "B000050B6Z",  
  "questionType": "yes/no",  
  "answerType": "Y",  
  "answerTime": "Aug 8, 2014",  
  "unixTime": 1407481200,  
  "question": "Can you use this unit with GEL shaving cans?",  
  "answer": "Yes. If the can fits in the machine it will dispense hot  
gel lather. I've been using my machine for both , gel and traditional  
lather for over 10 years."  
}
```

Download link

See the [Amazon Q/A Page](#) for download information.

Citation

Please cite the following if you use the data:

Modeling ambiguity, subjectivity, and diverging viewpoints in opinion question answering systems

Mengting Wan, Julian McAuley

International Conference on Data Mining (ICDM), 2016

[pdf](#)

Addressing complex and subjective product-related queries with customer reviews

Julian McAuley, Alex Yang

World Wide Web (WWW), 2016

[pdf](#)

Marketing Bias data

Description

These datasets contain attributes about products sold on ModCloth and Amazon which may be sources of bias in recommendations (in particular, attributes about how the products are marketed). Data also includes user/item interactions for recommendation.

Basic statistics

ModCloth Amazon Electronics	
Reviews:	99,893
Items:	1,020
Users:	44,783
Bias type: body shape gender	

Metadata

- ratings
- product images
- user identities
- item sizes, user genders

Example (ModCloth)

```
item_id,user_id,rating,timestamp,size,fit,user_attr,model_attr,c...
7443,Alex,4,2010-01-21 08:00:00+00:00,,,Small,Small,Dresses,,2012,0
7443,carolyn.agan,3,2010-01-27 08:00:00+00:00,,,Small,Dresses,....
7443,Robyn,4,2010-01-29 08:00:00+00:00,,,Small,Small,Dresses,,20...
7443,De,4,2010-02-13 08:00:00+00:00,,,Small,Dresses,,2012,0
7443,tasha,4,2010-02-18 08:00:00+00:00,,,Small,Small,Dresses,,20...
7443,gina.chihos,5,2010-02-25 08:00:00+00:00,,,Small,Dresses,,2...
7443,Kim,2,2010-02-26 08:00:00+00:00,,,Small,Small,Dresses,,2012,0
7443,jess.betcher,5,2010-03-26 07:00:00+00:00,,,Small,Dresses,....
```

Download links

See our [project page](#) for download links.

Citation

Please cite the following if you use the data:

Addressing Marketing Bias in Product Recommendations

Mengting Wan, Jianmo Ni, Rishabh Misra, Julian McAuley

WSDM, 2020

[pdf](#)

Google Local Reviews (2021)

Description

This dataset contains review information from Google Maps (ratings, text, images, etc.), business metadata (address, geographic info, descriptions, category information, price, open hours, etc.), and links (related businesses) up to Sep 2021 in the United States.

See also two variants of this dataset below, including a 2021 version, and a version containing item images.

Basic statistics

Reviews:	666,324,103
Users:	113,643,107
Businesses:	4,963,111

Review

```
{
  'user_id': '101463350189962023774',
  'name': 'Jordan Adams',
  'time': 1627750414677,
  'rating': 5,
  'text': 'Cool place, great people, awesome dentist!',
  'pics': [
    {
      'url':
      ['https://lh5.googleusercontent.com/p/AF1QipNq2nZC5TH4_M7h5xRAd
       61hoTgvYlo9lozABguI=w150-h150-k-no-p']
    }
  ],
  'resp': {
```

```

        'time': 1628455067818,
        'text': 'Thank you for your five-star review! -Dr. Blake'
    },
    'gmap_id': '0x87ec2394c2cd9d2d:0xd1119cfbee0da6f3'
}

```

- **user_id** - ID of the reviewer
- **name** - name of the reviewer
- **time** - time of the review (unix time)
- **rating** - rating of the business
- **text** - text of the review
- **pics** - pictures of the review
- **resp** - business response to the review including unix time and text of the response
- **gmap_id** - ID of the business

Metadata

```

{
    'name': 'Walgreens Pharmacy',
    'address': 'Walgreens Pharmacy, 124 E North St, Kendallville, IN 46755',
    'gmap_id': '0x881614ce7c13acbb:0x5c7b18bbf6ec4f7e',
    'description': 'Department of the Walgreens chain providing prescription medications & other health-related items.',
    'latitude': 41.451859999999996,
    'longitude': -85.2666757,
    'category': ['Pharmacy'],
    'avg_rating': 4.2,
    'num_of_reviews': 5,
    'price': '$$',
    'hours': ['Thursday', '8AM-1:30PM'], ['Friday', '8AM-1:30PM'],
    ['Saturday', '9AM-1:30PM'], ['Sunday', '10AM-1:30PM'], ['Monday',
    '8AM-1:30PM'], ['Tuesday', '8AM-1:30PM'], ['Wednesday', '8AM-1:30PM']],
    'MISC': [
        'Service options': ['Curbside pickup', 'Drive-through', 'In-store pickup', 'In-store shopping'],
        'Health & safety': ['Mask required', 'Staff wear masks'],
        'Staff get temperature checks'],
        'Accessibility': ['Wheelchair accessible entrance',
        'Wheelchair accessible parking lot'],
        'Planning': ['Quick visit'],
        'Payments': ['Checks', 'Debit cards']
    ],
    'state': 'Closes soon · 1:30PM · Reopens 2PM',
    'relative_results': ['0x881614cd49e4fa33:0x2d507c24ff4f1c74',
    '0x8816145bf5141c89:0x535c1d605109f94b',
    '0x881614cda24cc591:0xca426e3a9b826432',
    '0x88162894d98b9lef:0xd139b34de70d3e03',
    '0x881615400b5e57f9:0xc56d17dbe420a67f'],
    'url':
    'https://www.google.com/maps/place//data=!4m2!3m1!1s0x881614ce7c13acb
    b:0x5c7b18bbf6ec4f7e?authuser=-1&hl=en&gl=us'
}

```

- **name** - name of the business
- **address** - address of the business
- **gmap_id** - ID of the business
- **description** - description of the business
- **latitude** - latitude of the business
- **longitude** - longitude of the business
- **category** - category of the business
- **avg_rating** - average rating of the business
- **num_of_reviews** - number of reviews
- **price** - price of the business
- **hours** - open hours
- **MISC** - MISC information
- **state** - the current status of the business (e.g., permanently closed)
- **relative_results** - relative businesses recommended by Google
- **url** - URL of the business

Download links

See the [Google Local Dataset Page](#) for download information.

Citation

Please cite the following if you use the data:

UCTopic: Unsupervised Contrastive Learning for Phrase Representations and Topic Mining

Jiacheng Li, Jingbo Shang, Julian McAuley

Annual Meeting of the Association for Computational Linguistics (ACL), 2022

[pdf](#)

Personalized Showcases: Generating Multi-Modal Explanations for Recommendations

An Yan, Zhankui He, Jiacheng Li, Tianyang Zhang, Julian McAuley

arXiv:2207.00422, 2022

[pdf](#)

Google Local Reviews (2018)

Description

These datasets contain reviews about businesses from Google Local (Google Maps). Data includes geographic information for each business as well as reviews.

Basic statistics

Reviews:	11,453,845
Users:	4,567,431
Businesses:	3,116,785

Metadata

- reviews and ratings
- GPS coordinates and address
- User information (places lived, jobs)
- timestamps
- business category, opening hours, etc.

Example (review)

```
{  
    'rating': 3.0,  
    'reviewerName': u'an lam',  
    'reviewText': u'Ch\u1eae5t l\u01b0\u1ee3ng t\u1ea1m \u1ed5n',  
    'categories': [u'Gi\u1ea3i Tr\u0111\xed - Caf\xec9'],  
    'gPlusPlaceId': u'108103314380004200232',  
    'unixReviewTime': 1372686659,  
    'reviewTime': u'Jul 1, 2013',  
    'gPlusUserId': u'100000010817154263736'  
}
```

Example (business)

```
{  
    'name': u'Diamond Valley Lake Marina',  
    'price': None,  
    'address': [u'2615 Angler Ave', u'Hemet, CA 92545'],  
    'hours': [[u'Monday', [[u'6:30 am--4:15 pm']]], [u'Tuesday',  
              [[u'6:30 am--4:15 pm']]], [u'Wednesday', [[u'6:30 am--4:15 pm']], 1],  
              [u'Thursday', [[u'6:30 am--4:15 pm']]], [u'Friday', [[u'6:30 am--4:15  
pm']]], [u'Saturday', [[u'6:30 am--4:15 pm']]], [u'Sunday', [[u'6:30  
am--4:15 pm']]]],  
    'phone': u'(951) 926-7201',  
    'closed': False,  
    'gPlusPlaceId': '104699454385822125632',  
    'gps': [33.703804, -117.003209]  
}
```

Download links

[Places Data \(276mb\)](#)

[User Data \(178mb\)](#)

[Review Data \(1.4gb\)](#)

Citation

Please cite the following if you use the data:

Translation-based factorization machines for sequential recommendation

Rajiv Pasricha, Julian McAuley

RecSys, 2018

[pdf](#)

Translation-based recommendation

Ruining He, Wang-Cheng Kang, Julian McAuley

RecSys, 2017

[pdf](#)

Google Restaurants

Description

This is a multi-modal dataset of restaurants from Google Local (Google Maps). Data includes images and reviews posted by users, as well as other metadata for each restaurant.

Basic statistics

subset full	
Restaurants:	30K 65K
Users:	37K 1.01M
Reviews:	108K 1.77M
Images:	203K 4.43M

Metadata

- Geographical location and address
- Reviews, ratings and images
- Timestamps
- Business category, opening status, price, etc.

Example

```
"name":"The Fish Spot",
"address":"5101 W Pico Blvd, Los Angeles, CA 90019",
"Description":null,
"Latitude":34.0481627,
"Longitude":-118.3494339,
"category":["Seafood restaurant"],
"gmap_url":"https://www.google.com/maps/place/The+Fish+Spot/",
"Avg_rating":4.3,
"Num_of_reviews":80,
"price": "$$",

"Reviews":
[
  {"user_id": "111210125124533240892",
   "time": "3 years ago",
   "Rating": 5,
   "text": "Absolutely love this place.",
   "pics": [
     {"id": "AF1QipOlejvRhkB1g-v52UczxYMD7uebcZIhKC9uGud",
      "url": ["https://lh5.googleusercontent.com/p/"]},
   ],
   "link": "https://www.google.com/maps/reviews/"}
]
```

Download link

See our [data folder](#) containing all related files. The file *image_review_all.json* contains the full dataset, while *filter_all_t.json* is a subset with filtered review sentences that have higher correlation with images. Code is available in our [Github repository](#).

Citation

Please cite the following if you use the data:

Steam Video Game and Bundle Data

Description

These datasets contain reviews from the Steam video game platform, and information about which games were bundled together.

Basic statistics

Reviews: 7,793,069
Users: 2,567,538
Items: 15,474
Bundles: 615

Metadata

- reviews
- purchases, plays, recommends ("likes")
- product bundles
- pricing information

Example (bundle)

```
{
  "bundle_final_price": '$29.66',
  "bundle_url": "http://store.steampowered.com/bundle/1482/?utm_source=SteamDB...",
  "bundle_price": '$32.96',
  "bundle_name": "Two Tribes Complete Pack!",
  "bundle_id": '1482',
  "items": [{"genre": 'Casual, Indie', 'item_id': '38700',
  'discounted_price': '$4.99', 'item_url':
  'http://store.steampowered.com/app/38700', 'item_name': 'Toki Tori'},
  {'genre': 'Adventure, Casual, Indie', 'item_id': '201420',
  'discounted_price': '$14.99', 'item_url':
  'http://store.steampowered.com/app/201420', 'item_name': 'Toki Tori
  2+'}, {"genre": 'Strategy, Indie, Casual', 'item_id': '38720',
  'discounted_price': '$4.99', 'item_url':
  'http://store.steampowered.com/app/38720', 'item_name': 'RUSH'},
  {'genre': 'Action, Indie', 'item_id': '38740', 'discounted_price':
  '$7.99', 'item_url': 'http://store.steampowered.com/app/38740',
  'item_name': 'EDGE'}],
  "bundle_discount": '10%'
}
```

Download links

[Version 1: Review Data \(6.7mb\)](#)

[Version 1: User and Item Data \(71mb\)](#)

[Version 2: Review Data \(1.3gb\)](#)

[Version 2: Item metadata \(2.7mb\)](#)

[Bundle Data \(92kb\)](#)

Citation

Please cite the following if you use the data:

Self-attentive sequential recommendation

Wang-Cheng Kang, Julian McAuley

ICDM, 2018

[pdf](#)

Item recommendation on monotonic behavior chains

Mengting Wan, Julian McAuley

RecSys, 2018

[pdf](#)

Goodreads Book Reviews

These datasets contain reviews from the Goodreads book review website, and a variety of attributes describing the items. Critically, these datasets have multiple levels of user interaction, ranging from adding to a "shelf", rating, and reading.

Basic statistics

Items:	1,561,465
Users:	808,749
Interactions:	225,394,930

Metadata

- reviews
- add-to-shelf, read, review actions
- book attributes: title, isbn
- graph of similar books

Example (interaction data)

```
{  
  "user_id": "8842281e1d1347389f2ab93d60773d4d",  
  "book_id": "130580",  
  "review_id": "330f9c153c8d3347eb914c06b89c94da",  
  "isRead": true,  
  "rating": 4,  
  "date_added": "Mon Aug 01 13:41:57 -0700 2011",  
  "date_updated": "Mon Aug 01 13:42:41 -0700 2011",  
  "read_at": "Fri Jan 01 00:00:00 -0800 1988",  
  "started_at": ""  
}
```

Download links

See our [project page](#) for download links.

Citation

Please cite the following if you use the data:

Item recommendation on monotonic behavior chains
Mengting Wan, Julian McAuley
RecSys, 2018
[pdf](#)

Goodreads Spoilers

These datasets contain reviews from the Goodreads book review website, along with annotated "spoiler" information from each review.

Basic statistics

Books:	25,475
Users:	18,892
Reviews:	1,378,033

Metadata

- reviews
- ratings
- spoilers
- see also metadata from the complete [Goodreads](#) dataset

Example (spoiler data)

Sentences are annotated as "1" if the sentence contains a spoiler, "0" otherwise.

```
{  
    'user_id': '01ecla320ffd6b2dd47833f2c8e4fb',  
    'timestamp': '2013-12-28',  
    'review_sentences': [[0, 'First, be aware that this book is not for  
    the faint of heart.'],  
    [0, 'Human trafficking, drugs, kidnapping, abuse in all forms -  
    this story contains all of this and more.'],  
    ...,  
    [0, '(ARC provided by the author in return for an honest  
    review.)']],  
    'rating': 5,  
    'hasSpoiler': False,  
    'book_id': '18398089',  
    'review_id': '4b3ffeaaf14310ac6854f140188e191cd'  
}
```

Download links

See our [project page](#) for download links.

Citation

Please cite the following if you use the data:

Fine-grained spoiler detection from large-scale review corpora

Mengting Wan, Rishabh Misra, Ndapa Nakashole, Julian McAuley

ACL, 2019

[pdf](#)

Pinterest Fashion Compatibility

This dataset contains images (scenes) containing fashion products, which are labeled with bounding boxes and links to the corresponding products.

Basic statistics

Scenes:	47,739
Products:	38,111
Scene-Product Pairs:	93,274

Metadata

- product IDs
- bounding boxes

Example (fashion.json)

```
{  
    "product": "0027e30879ce3d87f82f699f148bff7e",  
    "scene": "cdab9160072dd1800038227960ff6467",  
    "bbox": [  
        0.434097,  
        0.859363,  
        0.560254,  
        1.0  
    ]  
}
```

Download links

See our [project page](#) for download links, and for instructions as to how the product images can be collected from Pinterest.

Citation

Please cite the following if you use the data:

Complete the Look: Scene-based complementary product recommendation

Wang-Cheng Kang, Eric Kim, Jure Leskovec, Charles Rosenberg, Julian McAuley

Clothing Fit Data

Description

These datasets contain measurements of clothing fit from *ModCloth* and *RentTheRunway*.

Basic statistics

	Modcloth	Renttherunway
Number of users:	47,958	105,508
Number of items:	1,378	5,850
Number of transactions:	82,790	192,544

Metadata

- ratings and reviews
- fit feedback (small/fit/large etc.)
- user/item measurements
- category information

Example (RentTheRunway)

```
{  
    "fit": "fit",  
    "user_id": "420272",  
    "bust_size": "34d",  
    "item_id": "2260466",  
    "weight": "137lbs",  
    "rating": "10",  
    "rented_for": "vacation",  
    "review_text": "An adorable romper! Belt and zipper were a little hard to navigate in a full day of wear/bathroom use, but that's to be expected. Wish it had pockets, but other than that-- absolutely perfect! I got a million compliments.",  
    "body_type": "hourglass",  
    "review_summary": "So many compliments!",  
    "category": "romper",  
    "height": "5' 8\"",  
    "size": 14,  
    "age": "28",  
    "review_date": "April 20, 2016"  
}
```

Download links

[Modcloth](#) (8.5mb)

[Renttherunway](#) (31mb)

Citation

Please cite the following if you use the data:

Decomposing fit semantics for product size recommendation in metric spaces

Rishabh Misra, Mengting Wan, Julian McAuley

RecSys, 2018

[pdf](#)

Product Exchange/Bartering Data

Description

These datasets contain peer-to-peer trades from various recommendation platforms.

Basic statistics

Tradesy Ratebeer Gameswap

Number of users: 128,152 2,215 9,888
Number of transactions: 68,543 125,665 3,470

Metadata

- peer-to-peer trades
- "have" and "want" lists
- image data (tradesy)

Example (tradesy)

```
{  
  'lists':  
  {  
    'bought': ['466', '459', '457', '449'],  
    'selling': [],  
    'want': [],  
    'sold': ['104', '103', '102']  
  },  
  'uid': '2'  
}
```

Download links

[Tradesy](#) (3.8mb)

See the [project page](#) for ratebeer, gameswap (and other) datasets

Citation

Please cite the following if you use the data:

Bartering books to beers: A recommender system for exchange platforms
Jérémie Rappaz, Maria-Luiza Vladarean, Julian McAuley, Michele Catasta
WSDM, 2017
[pdf](#)

VBPR: Visual bayesian personalized ranking from implicit feedback
Ruining He, Julian McAuley
AAAI, 2016
[pdf](#)

Behance Community Art Data

Description

Likes and image data from the community art website Behance. This is a small, anonymized, version of a larger proprietary dataset.

Basic statistics

Users:	63,497
Items:	178,788
Appreciates ("likes"):	1,000,000

Metadata

- appreciates (likes)
- timestamps
- extracted image features

Example ("appreciate" data)

Each entry is a user, item, timestamp triple:

```
276633 01588231 1307583271  
1238354 01529213 1307583273  
165550 00485000 1307583337  
2173258 00776972 1307583340  
165550 00158226 1307583406  
1238354 01540285 1307583495  
2459267 01578261 1307583509
```

```
165550 00264669 1307583518  
165550 00171501 1307583536
```

Code to read image features

```
import struct
def readImageFeatures(path):
    f = open(path, 'rb')
    while True:
        itemId = f.read(8)
        if itemId == '': break
        feature = struct.unpack('f'*4096, f.read(4*4096))
        yield itemId, feature
```

Download links

See our [data folder](#) containing all Behance files. The folder also contains additional documentation.

Citation

Please cite the following if you use the data:

Vista: A visually, socially, and temporally-aware model for artistic recommendation
Ruining He, Chen Fang, Zhaowen Wang, Julian McAuley
RecSys, 2016
[pdf](#)

Social Recommendation Data

Description

These datasets include ratings as well as social (or trust) relationships between users. Data are from [LibraryThing](#) (a book review website) and [epinions](#) (general consumer reviews).

Basic statistics

	Librarything	Epinions
Number of users:	73,882	116,260
Number of items:	337,561	41,269
Number of ratings/feedback:	979,053	181,394
Number of social relations:	120,536	181,304

Metadata

- reviews
- price paid (epinions)
- helpfulness votes (librarything)
- flags (librarything)

Example (LibraryThing reviews)

```
{
    'work': '3067',
    'flags': [],
    'unixtime': 1160265600,
    'stars': 4.5,
    'nhelpful': 0,
    'time': 'Oct 8, 2006',
    'comment': 'great storytelling in this novel about a couple crossed by a time travelling disorder',
    'user': 'justine'
}
```

Example (LibraryThing social network)

```
Rodo anehan
Rodo sevilemar
Rodo dingsi
Rodo slash
```

```

RelaxedReader AnnRig
RelaxedReader bookbroke
RelaxedReader Bumpersmom
RelaxedReader DivaColumbus
RelaxedReader AnnRig
RelaxedReader bookbroke
RelaxedReader BookWorm2729
RelaxedReader Bumpersmom

```

Download links

[LibraryThing](#) (594mb)

[epinions](#) (66mb)

Citation

Please cite the following if you use the data:

SPMC: Socially-aware personalized Markov chains for sparse sequential recommendation

Chenwei Cai, Ruining He, Julian McAuley

IJCAI, 2017

[pdf](#)

Improving latent factor models via personalized feature projection for one-class recommendation

Tong Zhao, Julian McAuley, Irwin King

Conference on Information and Knowledge Management (CIKM), 2015

[pdf](#)

Other Non-Recommender-Systems Datasets

Description

Below are various datasets collected by my lab that are not related to recommender systems specifically. Formats of these datasets vary, so their respective project pages should be consulted for further details.

DogWhistle: Cant Understanding Data

DogWhistle is a Chinese dataset collected from the historical records for an online game. It provides hidden words and the cant for them, with human answers. The dataset is suitable for semantic similarity evaluation for large language models.

Basic statistics

	train	dev	test
Games:	9,817	1,161	1,143
Rounds:	76,740	9,593	9,592
Word Combinations:	18,832	2,243	2,220
Unique words:	1,878	1,809	1,820
Cant:	230,220	28,779	28,776

Metadata

- cant and the hidden words
- cant history
- human answers

Example (insider subtask)

```

0 高铁,周末,无情,条纹 冷漠,休息,斑马 冷漠 2
1 高铁,周末,无情,条纹 冷漠,休息,斑马 休息 1
2 高铁,周末,无情,条纹 冷漠,休息,斑马 斑马 3

```

Download links

Please refer to our [leaderboard page](#) for download instructions.

Citation

Please cite the following if you use the data:

Blow the Dog Whistle: A Chinese Dataset for Cantonese Understanding with Common Sense and World Knowledge

Canwen Xu, Wangchunshu Zhou, Tao Ge, Ke Xu, Julian McAuley, Furu Wei
NAACL, 2021

[pdf](#)

Video Game Data

Description

Step charts from the video game *Dance Dance Revolution*, and audio files from the NES platform.

Basic statistics

Num songs (DDR): 223 (7 hours)
Num charts (DDR): 1,102
Num games (NES): 397
Num songs (NES): 5,278 (46 hours)
Num notes (NES): 2,325,636

Download links

See the project pages for [Dance Dance Convolution](#) and [NES MDB](#) for further details and links to the data

Citation

Please cite the following if you use the data:

Dance Dance Convolution

Chris Donahue, Zachary Lipton, Julian McAuley
ICML, 2017

[pdf](#)

The NES Music Database: A symbolic music dataset with expressive performance attributes

Chris Donahue, Henry Mao, Julian McAuley
International Society for Music Information Retrieval Conference (ISMIR), 2018

[pdf](#)

Multi-aspect Reviews

Description

These datasets include reviews with multiple rated dimensions. The most comprehensive of these are beer review datasets from Ratebeer and Beeradvocate, which include sensory aspects such as taste, look, feel, and smell.

Basic statistics

	Ratebeer	BeerAdvocate
Number of users:	40,213	33,387
Number of items:	110,419	66,051
Number of ratings/reviews:	2,855,232	1,586,259
Timespan:	Apr 2000 - Nov 2011	Jan 1998 - Nov 2011

Metadata

- reviews
- aspect-specific ratings (taste, look, feel, smell, overall impression)
- product category
- ABV

Example (ratebeer)

```
beer/name: John Harvards Simcoe IPA
beer/beerId: 63836
beer/brewerId: 8481
beer/ABV: 5.4
beer/style: India Pale Ale (IPA)
review/appearance: 4/5
review/aroma: 6/10
review/palate: 3/5
review/taste: 6/10
review/overall: 13/20
review/time: 1157587200
review/profileName: hopdog
review/text: On tap at the Springfield, PA location. Poured a deep
and cloudy orange (almost a copper) color with a small sized off
white head. Aromas or oranges and all around citric. Tastes of
oranges, light caramel and a very light grapefruit finish. I too
would not believe the 80+ IBUs - I found this one to have a very
light bitterness with a medium sweetness to it. Light lacing left on
the glass.
```

Download links

[BeerAdvocate](#) (433mb)

[RateBeer](#) (388mb)

[Sentences with aspect labels \(annotator 1\)](#) (758kb)

[Sentences with aspect labels \(annotator 2\)](#) (759kb)

Citation

Please cite the following if you use the data:

Learning attitudes and attributes from multi-aspect reviews

Julian McAuley, Jure Leskovec, Dan Jurafsky

International Conference on Data Mining (ICDM), 2012

[pdf](#)

From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews

Julian McAuley, Jure Leskovec

WWW, 2013

[pdf](#)

Social Circles

Description

These datasets contain social connections and "circles" from Facebook, Twitter, and Google Plus.

Basic statistics

	Facebook	Twitter	Google Plus
Number of networks:	10	133	1,000
Number of nodes:	4,039	106,674	192,075
Number of circles:	193	479	5,541

Metadata

- social connections
- circles (sets of friends sharing a common property)
- user metadata

Example (Kaggle egonet data)

```
UserId: Friends
1: 4 6 12 2 208
2: 5 3 17 90 7
```

Download links

See SNAP [facebook](#), [twitter](#), and [Google Plus](#) data, as well as the [Kaggle competition](#) based on the same data.

Citation

Please cite the following if you use the data:

Learning to Discover Social Circles in Ego Networks
Julian McAuley, Jure Leskovec
Neural Information Processing Systems (NIPS), 2012
[pdf](#)

Reddit Submissions

Description

Submissions of reddit posts (and in particular resubmissions of the same content) along with metadata.

Basic statistics

Num of submissions (images):	132,308
Num of unique images:	16,736
Timespan	July 2008 - January 2013

Metadata

- timestamps
- upvotes/downvotes
- post title, subreddit, etc.

Example

```
#image_id,unixtime,rawtime,title,total_votes,reddit_id,...
```

number_of_downvotes,localtime,score,number_of_comments,username
1005,1335861624,2012-05-01T15:40:24.968266-07:00,I immediately regret
this decision,27,t296r,20,pics,7,1335886824,13,0,ninjaroflmaster
1005,1336470481,2012-05-08T16:48:01.418140-07:00,"Pushing your friend
into the water,Level: 99",18,tds4i,16,funny,2,1336495681,14,0,hme4
1005,1339566752,2012-06-13T12:52:32.371941-07:00,I told him. He
Didn't Listen,6,v0cma,4,funny,2,1339591952,2,0,HeyPatWhatsUp
1005,1342200476,2012-07-14T00:27:56.857805-07:00,Don't end up as this
guy.,16,wjivx,7,funny,9,1342225676,-2,2,catalyst24

Download links

[resubmissions data \(7.3mb\)](#)
[raw html of resubmissions \(1.8gb\)](#)
See also the [SNAP project page](#).

Citation

Please cite the following if you use the data:

Understanding the interplay between titles, content, and communities in social media
Himabindu Lakkaraju, Julian McAuley, Jure Leskovec
ICWSM, 2013
[pdf](#)

Questions and comments to [Julian McAuley](#)