데이터 수집

데이터 출처

https://amazon-reviews-2023.github.io/

데이터 설명

McAuley Lab에서 수집한 2023년도 대규모 아마존 리뷰 데이터셋

포함하는 정보:

1. **User Reviews** (*평점, 내용, 유용성 평가 등*)
2. **Item Metadata** (*제품 설명, 가격, 원본 이미지 등*)
3. **Links** (*유저-아이템 / 함께 산 아이템 그래프*)

카테고리

22~23년도 리뷰데이터

1번_ 소규모 : **All_Beauty**

2번_대규모 : **All_Beauty , Beauty_and_Personal_Care,Health_and_Personal_Care**

사이즈:

Year	#Review	#User	#Item	#R_Token	#M_Token	#Domain	Timespa
2023	571.54M	54.51M	48.19M	30.14B	30.78B	33	May'96 - Sep'23

카테고리:

Category	#User	#Item	#Rating	#R_Token	#M_Token
All_Beauty	632.0K	112.6K	701.5K	31.6M	74.1M
Amazon_Fashion	2.0M	825.9K	2.5M	94.9M	510.5M
Appliances	1.8M	94.3K	2.1M	92.8M	95.3M
Arts_Crafts_and_Sewing	4.6M	801.3K	9.0M	350.0M	695.4M
Automotive	8.0M	2.0M	20.0M	824.9M	1.7B
Baby_Products	3.4M	217.7K	6.0M	323.3M	218.6M
Beauty_and_Personal_Care	11.3M	1.0M	23.9M	1.1B	913.7M
Books	10.3M	4.4M	29.5M	2.9B	3.7B
CDs_and_Vinyl	1.8M	701.7K	4.8M	514.8M	287.5M
Cell_Phones_and_Accessories	11.6M	1.3M	20.8M	935.4M	1.3B
Clothing_Shoes_and_Jewelry	22.6M	7.2M	66.0M	2.6B	5.9B
Digital_Music	101.0K	70.5K	130.4K	11.4M	22.3M
Electronics	18.3M	1.6M	43.9M	2.7B	1.7B
Gift_Cards	132.7K	1.1K	152.4K	3.6M	630.0K
Grocery_and_Gourmet_Food	7.0M	603.2K	14.3M	579.5M	462.8M
Handmade_Products	586.6K	164.7K	664.2K	23.3M	125.8M
Health_and_Household	12.5M	797.4K	25.6M	1.2B	787.2M
Health_and_Personal_Care	461.7K	60.3K	494.1K	23.9M	40.3M
Home_and_Kitchen	23.2M	3.7M	67.4M	3.1B	3.8B
Industrial_and_Scientific	3.4M	427.5K	5.2M	235.2M	363.1M
Kindle_Store	5.6M	1.6M	25.6M	2.2B	1.7B

Category	#User	#Item	#Rating	#R_Token	#M_Token
Magazine_Subscriptions	60.1K	3.4K	71.5K	3.8M	1.3M
Movies_and_TV	6.5M	747.8K	17.3M	1.0B	415.5M
Musical_Instruments	1.8M	213.6K	3.0M	182.2M	200.1M
Office_Products	7.6M	710.4K	12.8M	574.7M	682.8M
Patio_Lawn_and_Garden	8.6M	851.7K	16.5M	781.3M	875.1M
Pet_Supplies	7.8M	492.7K	16.8M	905.9M	511.0M
Software	2.6M	89.2K	4.9M	179.4M	67.1M
Sports_and_Outdoors	10.3M	1.6M	19.6M	986.2M	1.3B
Subscription_Boxes	15.2K	641	16.2K	1.0M	447.0K
Tools_and_Home_Improvement	12.2M	1.5M	27.0M	1.3B	1.5B
Toys_and_Games	8.1M	890.7K	16.3M	707.9M	848.3M
Video_Games	2.8M	137.2K	4.6M	347.9M	137.3M
Unknown	23.1M	13.2M	63.8M	3.3B	232.8M

User Reviews의 데이터 필드

필드	타입	설명
rating	float	평점 1.0-5.0
title	str	리뷰 제목
text	str	리뷰 내용
images	list	유저가 올리는 제품의 실제 이미지 각 이미지 사이즈는 상이하고 (small, medium, large), small_image_url, medium_image_url, large_image_url로 표기
asin	str	제품 ID
parent_asin	str	제품의 상위 ID 다른 색, 스타일, 사이즈들의 제품들은 보통 같은 상위 ID
user_id	str	리뷰어 ID
timestamp	int	리뷰한 시간 (유닉스 시간)
verified_purchase	bool	구매 확인
helpful_vote	int	리뷰의 유용성 투표

Item Metadata의 데이터 필드

필드	타입	설명
main_category	str	제품의 메인 카테고리
title	str	제품 이름
average_rating	float	제품 페이지에 표기되는 평균 평점
rating_number	int	평점 수
features	list	글머리표 형식으로 된 제품 특징
description	list	제품 설명
price	float	크롤링 당시의 제품 가격 (US $)
images	list	제품 이미지들 각 이미지의 사이즈가 상이하고 이미지의 위치는 “variant”필드로 표기
videos	list	제목과 url이 포함되어있는 제품 비디오들
store	str	판매처 이름
categories	list	제품의 계층적 분류
details	dict	제품 상세 설명 (재질, 브랜드, 크기 등)
parent_asin	str	제품의 상위 ID
bought_together	list	웹사이트에서 추천하는 같이사면 좋은 제품들