Sentiment Analysis & NLP in Digital Marketing

Ever wondered how to tap into the thoughts and feelings of your target audience? ***Sentiment analysis*** powered by natural language processing ***(NLP) and Python*** is here to save the day! Let's dive into the world of sentiment analysis and explore some awesome tools and techniques that'll help you better understand your customers.

Sentiment Analysis & NLP: The Dynamic Duo

Sentiment analysis is a technique used to determine the emotions, opinions, and attitudes expressed in text data. With the help of NLP, a subfield of AI, and Python, a versatile programming language, we can now process and analyze vast amounts of textual data from social media, reviews, and more to gauge customer sentiment with remarkable accuracy.

Why Sentiment Analysis Matters in Digital Marketing

In the digital age, understanding customer sentiment is crucial for effective marketing. Here's why:

- Brand reputation management: Monitor customer feedback and address any issues promptly to maintain a positive brand image.
- Campaign performance evaluation: Assess the public's reaction to your marketing campaigns and make data-driven adjustments for better results.
- Competitor analysis: Gain insights into your competitors' strengths and weaknesses by analyzing their customer sentiment.

Getting Started with Sentiment Analysis in Python

Python offers various libraries for sentiment analysis and NLP, making it easy for digital marketers to dive into text data analysis. Here are some popular Python libraries for sentiment analysis:

- ***TextBlob***: A simple, beginner-friendly library that provides sentiment analysis functionality with just a few lines of code.
- ***VADER***: A specialized library for sentiment analysis on social media data, which is particularly useful for digital marketers.
- ***NLTK***: A powerful library for NLP tasks, including sentiment analysis, with extensive documentation and resources.

Description

This Dataset is an updated version of the Amazon review dataset released in 2014. As in the previous version, 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). In addition, this version provides the following features:

- More reviews:
  - The total number of reviews is 233.1 million (142.8 million in 2014).
- Newer reviews:
  - Current data includes reviews in the range May 1996 - Oct 2018.
- Metadata:
  - We have added transaction metadata for each review shown on the review page. Such information includes:
    - Product information, e.g. color (white or black), size (large or small), package type (hardcover or electronics), etc.
    - Product images that are taken after the user received the product.
  - Added more detailed metadata of the product landing page. Such detailed information includes:
    - Bullet-point descriptions under product title.
    - Technical details table (attribute-value pairs).
    - Similar products table.
- More categories:
  - Includes 5 new product categories.
- **You can also download the review data from our previous datasets.**
- Amazon review (2014)
- Amazon review (2013)

Citation

Please cite the following paper if you use the data in any way:

Justifying recommendations using distantly-labeled reviews and fined-grained aspects
Jianmo Ni, Jiacheng Li, Julian McAuley
Empirical Methods in Natural Language Processing (EMNLP), 2019
pdf

News

05/2021 We updated **high resolution image urls** to the metadata!

08/2020 We have updated the metadata and now it includes much less HTML/CSS code. Feel free to download the updated data!

Note

We provide a colab notebook that helps you parse and clean the data. For example:

- Load the metadata (e.g. as JSON or DataFrame)
- Check if title has HTML contents and filter them
- We provide a collab that helps you find target products and obtain their reviews!
- **We appreciate any help or feedback to improve the quality of our dataset! Feel free to reach us at jin018@ucsd.edu if you meet any following questions:**
- Unparsed HTML contents
- Duplicate items which have the same reviews

Amazon Fashion	reviews (883,636 reviews)	metadata (186,637 products)
All Beauty	reviews (371,345 reviews)	metadata (32,992 products)
Appliances	reviews (602,777 reviews)	metadata (30,459 products)
Arts, Crafts and Sewing	reviews (2,875,917 reviews)	metadata (303,426 products)
Automotive	reviews (7,990,166 reviews)	metadata (932,019 products)
Books	reviews (51,311,621 reviews)	metadata (2,935,525 products)
CDs and Vinyl	reviews (4,543,369 reviews)	metadata (544,442 products)
Cell Phones and Accessories	reviews (10,063,255 reviews)	metadata (590,269 products)
Clothing Shoes and Jewelry	reviews (32,292,099 reviews)	metadata (2,685,059 products)
Digital Music	reviews (1,584,082 reviews)	metadata (465,392 products)
Electronics	reviews (20,994,353 reviews)	metadata (786,868 products)
Gift Cards	reviews (147,194 reviews)	metadata (1,548 products)
Grocery and Gourmet Food	reviews (5,074,160 reviews)	metadata (287,209 products)
Home and Kitchen	reviews (21,928,568 reviews)	metadata (1,301,225 products)
Industrial and Scientific	reviews (1,758,333 reviews)	metadata (167,524 products)
Kindle Store	reviews (5,722,988 reviews)	metadata (493,859 products)
Luxury Beauty	reviews (574,628 reviews)	metadata (12,308 products)

Magazine Subscriptions	reviews (89,689 reviews)	metadata (3,493 products)
Movies and TV	reviews (8,765,568 reviews)	metadata (203,970 products)
Musical Instruments	reviews (1,512,530 reviews)	metadata (120,400 products)
Office Products	reviews (5,581,313 reviews)	metadata (315,644 products)
Patio, Lawn and Garden	reviews (5,236,058 reviews)	metadata (279,697 products)
Pet Supplies	reviews (6,542,483 reviews)	metadata (206,141 products)
Prime Pantry	reviews (471,614 reviews)	metadata (10,815 products)
Software	reviews (459,436 reviews)	metadata (26,815 products)
Sports and Outdoors	reviews (12,980,837 reviews)	metadata (962,876 products)
Tools and Home Improvement	reviews (9,015,203 reviews)	metadata (571,982 products)
Toys and Games	reviews (8,201,231 reviews)	metadata (634,414 products)
Video Games	reviews (2,565,349 reviews)	metadata (84,893 products)

Directory

Files

Complete review data

Please only download these (large!) files if you really need them. We recommend using the smaller datasets (i.e. k-core and CSV files) as shown in the next section.

raw review data (34gb) - all 233.1 million reviews

ratings only (6.7gb) - same as above, in csv form without reviews or metadata

5-core (14.3gb) - subset of the data in which all users and items have at least 5 reviews (75.26 million reviews)

Per-category data - the review and product metadata for each category.

To download the complete review data and the per-category files, the following links will direct you to enter a form. Please contact me if you can't get access to the form.

"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.

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 (item,user,rating,timestamp) tuples. Thus they are suitable for use with mymedialite (or similar) packages.

You can directly download the following smaller per-category datasets.

Amazon Fashion	5-core (3,176 reviews)	ratings only (883,636 ratings)
All Beauty	5-core (5,269 reviews)	ratings only (371,345 ratings)
Appliances	5-core (2,277 reviews)	ratings only (602,777 ratings)
Arts, Crafts and Sewing	5-core (494,485 reviews)	ratings only (2,875,917 ratings)
Automotive	5-core (1,711,519 reviews)	ratings only (7,990,166 ratings)
Books	5-core (27,164,983 reviews)	ratings only (51,311,621 ratings)
CDs and Vinyl	5-core (1,443,755 reviews)	ratings only (4,543,369 ratings)

Cell Phones and Accessories	5-core (1,128,437 reviews)	ratings only (10,063,255 ratings)
Clothing, Shoes and Jewelry	5-core (11,285,464 reviews)	ratings only (32,292,099 ratings)
Digital Music	5-core (169,781 reviews)	ratings only (1,584,082 ratings)
Electronics	5-core (6,739,590 reviews)	ratings only (20,994,353 ratings)
Gift Cards	5-core (2,972 reviews)	ratings only (147,194 ratings)
Grocery and Gourmet Food	5-core (1,143,860 reviews)	ratings only (5,074,160 ratings)
Home and Kitchen	5-core (6,898,955 reviews)	ratings only (21,928,568 ratings)
Industrial and Scientific	5-core (77,071 reviews)	ratings only (1,758,333 ratings)
Kindle Store	5-core (2,222,983 reviews)	ratings only (5,722,988 ratings)
Luxury Beauty	5-core (34,278 reviews)	ratings only (574,628 ratings)
Magazine Subscriptions	5-core (2,375 reviews)	ratings only (89,689 ratings)
Movies and TV	5-core (3,410,019 reviews)	ratings only (8,765,568 ratings)
Musical Instruments	5-core (231,392 reviews)	ratings only (1,512,530 ratings)
Office Products	5-core (800,357 reviews)	ratings only (5,581,313 ratings)
Patio, Lawn and Garden	5-core (798,415 reviews)	ratings only (5,236,058 ratings)
Pet Supplies	5-core (2,098,325 reviews)	ratings only (6,542,483 ratings)

Prime Pantry	5-core (137,788 reviews)	ratings only (471,614 ratings)
Software	5-core (12,805 reviews)	ratings only (459,436 ratings)
Sports and Outdoors	5-core (2,839,940 reviews)	ratings only (12,980,837 ratings)
Tools and Home Improvement	5-core (2,070,831 reviews)	ratings only (9,015,203 ratings)
Toys and Games	5-core (1,828,971 reviews)	ratings only (8,201,231 ratings)
Video Games	5-core (497,577 reviews)	ratings only (2,565,349 ratings)

Data format

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

Sample review:


```
{ "image": ["https://images-na.ssl-images-
amazon.com/images/I/71eG75FTJJL._SY88.jpg"], "overall": 5.0, "vote": "2",
"verified": True, "reviewTime": "01 1, 2018", "reviewerID": "AUI6WTTT0QZYS",
"asin": "5120053084", "style": { "Size:": "Large", "Color:": "Charcoal" },
"reviewerName": "Abbey", "reviewText": "I now have 4 of the 5 available colors of
this shirt... ", "summary": "Comfy, flattering, discreet--highly recommended!",
"unixReviewTime": 1514764800 }

{ "reviewerID": "A2SUAM1J3GNN3B", "asin": "0000013714", "reviewerName": "J.
McDonald", "vote": 5, "style": { "Format:": "Hardcover" }, "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
- `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

Metadata

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

metadata (24gb) - metadata for 15.5 million products

Sample metadata:


```
{ "asin": "0000031852", "title": "Girls Ballet Tutu Zebra Hot Pink", "feature":
["Botiquecutie Trademark exclusive Brand", "Hot Pink Layered Zebra Print Tutu",
"Fits girls up to a size 4T", "Hand wash / Line Dry", "Includes a Botiquecutie TM
Exclusive hair flower bow"], "description": "This tutu is great for dress up play
for your little ballerina. Botiquecute Trade Mark exclusive brand. Hot Pink Zebra
print tutu.", "price": 3.17, "imageURL": "http://ecx.images-
amazon.com/images/I/51fAmVkTbyL._SY300_.jpg", "imageURLHighRes":
"http://ecx.images-amazon.com/images/I/51fAmVkTbyL.jpg", "also_buy":
["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"], "salesRank": {"Toys & Games": 211836}, "brand": "Coxlures",
"categories": [["Sports & Outdoors", "Other Sports", "Dance"]] }
```


where

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 json.loads(l)
```

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 json.loads(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')
```

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
```