

# Project Report

AKSHAY D. KULKARNI (A20448255) – [akulkarni10@hawk.iit.edu](mailto:akulkarni10@hawk.iit.edu)

VARUN G. SHANBHAG (A20452310) – [vshanbhag1@hawk.iit.edu](mailto:vshanbhag1@hawk.iit.edu)

---

# REPORT

## Overview

To analyze a vast dataset of product selling websites (Amazon) and accurately display the Top 10 Products based on customer reviews, Show trends based on reviews and products associated with it and, Recommended products among those by analyzing customer patterns and history.

## Requirement & Implementation

- As we are a digital marketing company, we aim to display top products for a particular website, in our case, Amazon, to increase product selling websites profit. (<http://jmcauley.ucsd.edu/data/amazon/>)
- We have performed analysis on a large, actively maintained, and updated dataset from UCSD in following fashion (Sample json snippet below)

```
Digital_Music_5.json
1 {"reviewerID": "A3EBHHCZ06V2A4", "asin": "5555991584", "reviewerName": "Amaranth \"music fan\"", "helpful": [3, 3], "reviewText": "It's hard to believe \"Memory of Trees\" came out 11 years ago;it has held up well over the passage of time.It's Enya's last great album before the New Age/pop of \"Amarantine\" and \"Day without rain.\" Back in 1995,Enya still had her creative spark,her own voice.I agree with the reviewer who said that this is her saddest album;it is melancholy,bittersweet,from the opening title song.\"Memory of Trees\" is elegaic&majestic.;\"Pax Deorum\" sounds like it is from a Requiem Mass,it is a dark threnody.Unlike the reviewer who said that this has a \"disconcerting\" blend of spirituality&sensuality;;I don't find it disconcerting at all.\"Anywhere is\" is a hopeful song,looking to possibilities.\"Hope has a place\" is about love,but it is up to the listener to decide if it is romantic,platonic,etc.I've always had a soft spot for this song.\"On my way home\" is a triumphant ending about return.This is truly a masterpiece of New Age music,a must for any Enya fan!", "overall": 5.0, "summary": "Enya's last great album", "unixReviewTime": 1158019200, "reviewTime": "09 12, 2006"}
2 {"reviewerID": "AZFWAXJG90JXV", "asin": "5555991584", "reviewerName": "bethtexas", "helpful": [0, 0], "reviewText": "A classically-styled and introverted album, Memory of Trees is a masterpiece of subtlety. Many of the songs have an endearing shyness to them - soft piano and a lovely, quiet voice. But within every introvert is an inferno, and Enya lets that fire explode on a couple of songs that absolutely burst with an expected raw power.If you've never heard Enya before, you might want to start with one of her more popularized works, like Watermark, just to play it safe. But if you're already a fan, then your collection is not complete without this beautiful work of musical art.", "overall": 5.0, "summary": "Enya at her most elegant", "unixReviewTime": 991526400, "reviewTime": "06 3, 2001"}
3 {"reviewerID": "A38IRLOX2T4DFF", "asin": "5555991584", "reviewerName": "bob turnley", "helpful": [2, 2], "reviewText": "I never thought Enya would reach the sublime heights of Evacuee or Marble Halls from 'Shepherd Moons.' 'The Celts, Watermark and Day...' were all pleasant and admirable throughout, but are less ambitious both lyrically and musically. But Hope Has a Place from 'Memory...' reaches those heights and beyond. It is Enya at her most inspirational and comforting. I'm actually glad that this song didn't get overexposed the way Only Time did. It makes it that much more special to all who own this album.", "overall": 5.0, "summary": "The best so far", "unixReviewTime": 1058140800, "reviewTime": "07 14, 2003"}
4 {"reviewerID": "A22IK3I6U76GX0", "asin": "5555991584", "reviewerName": "Calle", "helpful": [1, 1], "reviewText": "This is the third
```

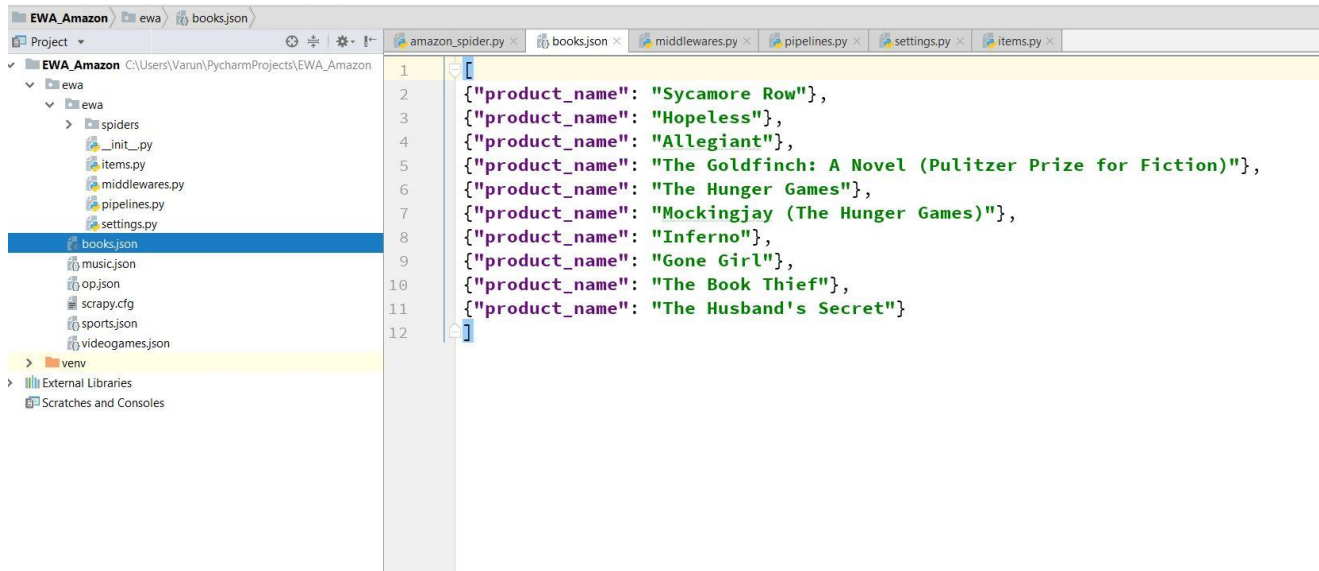
- This dataset had more than 7million records for each category.
  - With our python script, we analyze and pick the top 10 products for the following categories: Books, Sports, Video Games, Music based on the number of reviews, and ratings. ([https://github.com/Varun-Shanbhag-007/CSP\\_EWA/blob/master/DataExtraction.ipynb](https://github.com/Varun-Shanbhag-007/CSP_EWA/blob/master/DataExtraction.ipynb))
-

```
print ("}")
```

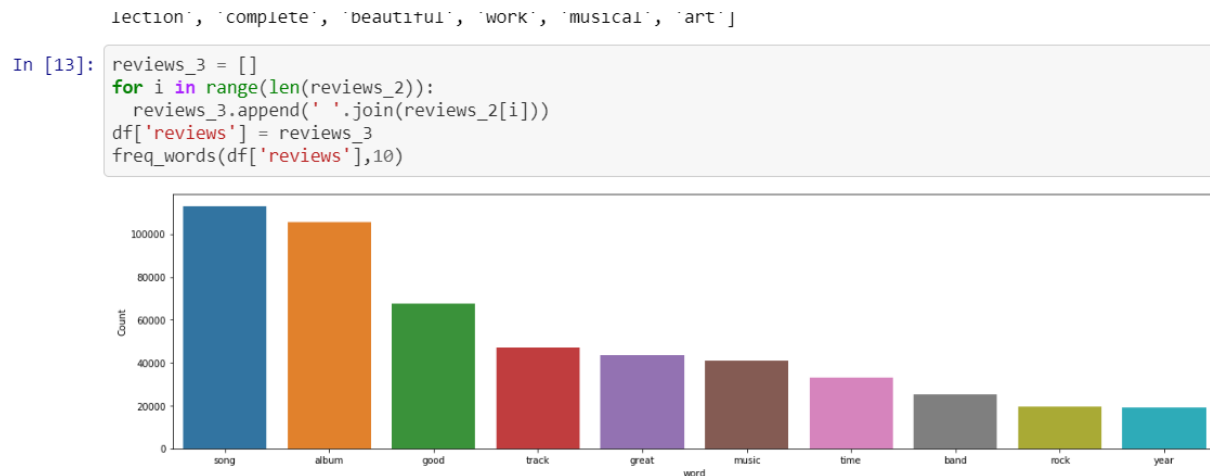
```
{ "music" : [
{"product": "http://www.amazon.com/dp/B0007NFL18", "fiveStar" : 127 , "oneStar" : 145 , "img_url" : "http://images.amazon.com/images/P/B0007NFL18.01._PE30_PI_SCLZZZZZZZ_.jpg" },
{"product": "http://www.amazon.com/dp/B000084T18", "fiveStar" : 196 , "oneStar" : 75 , "img_url" : "http://images.amazon.com/images/P/B000084T18.01._PE30_PI_SCLZZZZZZZ_.jpg" },
{"product": "http://www.amazon.com/dp/B00006690F", "fiveStar" : 165 , "oneStar" : 39 , "img_url" : "http://images.amazon.com/images/P/B00006690F.01._PE30_PI_SCLZZZZZZZ_.jpg" },
{"product": "http://www.amazon.com/dp/B00004T9UF", "fiveStar" : 182 , "oneStar" : 20 , "img_url" : "http://images.amazon.com/images/P/B00004T9UF.01._PE30_PI_SCLZZZZZZZ_.jpg" },
{"product": "http://www.amazon.com/dp/B00005YW4H", "fiveStar" : 159 , "oneStar" : 31 , "img_url" : "http://images.amazon.com/images/P/B00005YW4H.01._PE30_PI_SCLZZZZZZZ_.jpg" },
{"product": "http://www.amazon.com/dp/B0006ZQ9BS", "fiveStar" : 154 , "oneStar" : 31 , "img_url" : "http://images.amazon.com/images/P/B0006ZQ9BS.01._PE30_PI_SCLZZZZZZZ_.jpg" },
{"product": "http://www.amazon.com/dp/B0000AGWFX", "fiveStar" : 154 , "oneStar" : 27 , "img_url" : "http://images.amazon.com/images/P/B0000AGWFX.01._PE30_PI_SCLZZZZZZZ_.jpg" },
{"product": "http://www.amazon.com/dp/B00005054Q", "fiveStar" : 156 , "oneStar" : 13 , "img_url" : "http://images.amazon.com/images/P/B00005054Q.01._PE30_PI_SCLZZZZZZZ_.jpg" },
{"product": "http://www.amazon.com/dp/B00000163G", "fiveStar" : 161 , "oneStar" : 5 , "img_url" : "http://images.amazon.com/images/P/B00000163G.01._PE30_PI_SCLZZZZZZZ_.jpg" },
{"product": "http://www.amazon.com/dp/B00065XJ52", "fiveStar" : 96 , "oneStar" : 63 , "img_url" : "http://images.amazon.com/images/P/B00065XJ52.01._PE30_PI_SCLZZZZZZZ_.jpg" }
]}
```

- We do scrapping of website using Scrapy (Python) for extracting title of product based on processed data above.

```
(base) C:\Users\Varun\PycharmProjects\EWA_Amazon\ewa>scrapy crawl amazon_s -o books.json
2019-11-16 15:34:33 [scrapy.utils.log] INFO: Scrapy 1.8.0 started (bot: ewa)
2019-11-16 15:34:33 [scrapy.utils.log] INFO: Versions: lxml 4.2.5.0, libxml2 2.9.8, cssselect 1.1.0, parsel 1.5.2, w3lib 1.21.0, Twisted 19.10.0, Python 3.7.1 (default, Dec 10 2018, 22:54:23) [MSC v.1915 64 bit (AMD64)], pyOpenSSL 18.0.0 (OpenSSL 1.1.1a 20 Nov 2018), cryptography 2.4.2, Platform Windows-10-10.0.18362-SP0
2019-11-16 15:34:33 [scrapy.crawler] INFO: Overridden settings: {'BOT_NAME': 'ewa', 'FEED_FORMAT': 'json', 'FEED_URI': 'books.json', 'NEWSPIDER_MODULE': 'ewa.spiders', 'ROBOTSTXT_OBEY': True, 'SPIDER_MODULES': ['ewa.spiders'], 'USER_AGENT': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/61.0.3163.100 Safari/537.36'}
2019-11-16 15:34:34 [scrapy.extensions.telnet] INFO: Telnet Password: f66bdf556bd503b6
2019-11-16 15:34:34 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.corestats.CoreStats',
 'scrapy.extensions.telnet.TelnetConsole',
 'scrapy.extensions.feedexport.FeedExporter',
 'scrapy.extensions.logstats.LogStats']
2019-11-16 15:34:35 [scrapy.middleware] INFO: Enabled downloader middlewares:
['scrapy.downloadermiddlewares.robotstxt.RobotsTxtMiddleware',
 'scrapy.downloadermiddlewares.httpauth.HttpAuthMiddleware',
 'scrapy.downloadermiddlewares.downloadtimeout.DownloadTimeoutMiddleware',
 'scrapy.downloadermiddlewares.defaultheaders.DefaultHeadersMiddleware',
 'scrapy.downloadermiddlewares.useragent.UserAgentMiddleware']
```



- We also do trend analysis and show trends, as shown below.( [https://github.com/Varun-Shanbhag-007/CSP\\_EWA/blob/master/Trends.ipynb](https://github.com/Varun-Shanbhag-007/CSP_EWA/blob/master/Trends.ipynb))



- We save the user pattern and history of his visits and recommend the products accordingly using surprise API.

---

```
# Print the recommended items for each user
for uid, user_ratings in top_n.items():
    print(uid, [iid for (iid, _) in user_ratings])

out = open(pr_file_path+'output.csv', 'w', newline='')
output=csv.writer(out)

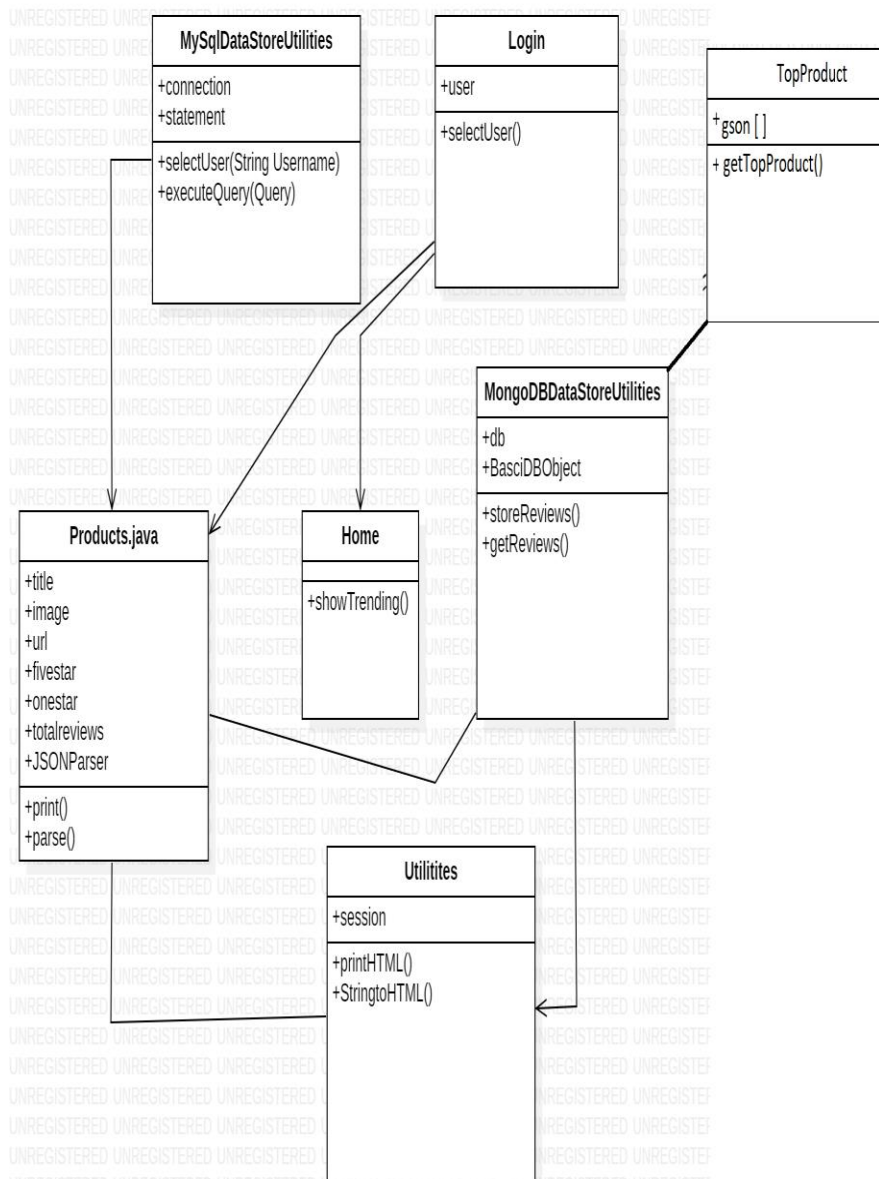
for uid, user_ratings in top_n.items():
    output.writerow([uid, [iid for (iid, _) in user_ratings]])

out.close()
```

```
one ['Maglula UpLULA Magazine Speed Loader', 'Rothco 550lb. Type III Nylon Paracord']
x ['PlayStation 4 500GB Console', 'The Last of US']
```

---

## Class Diagram





## Project Flow

