

***Feature Extraction and Analysis of E-Commerce portal (Amazon product data) using Machine Learning***

# 

# B9DA105 Applied Research Process for Data Analytics



**Name: Akash Vaykar**

**Student ID: 10510107**

**Prof: Dr. Shahram Azizi**

**Index**

[Introduction: 1](#_Toc15915573)

[Theoretical Background and the review of world literature: 1](#_Toc15915574)

[Dataset Information: 2](#_Toc15915575)

[Conventions and Methodology: 4](#_Toc15915576)

[Methodological directions: 5](#_Toc15915577)

[Tools and Programming Libraries: 5](#_Toc15915578)

[Application Implementation to real world datasets: 6](#_Toc15915579)

[Flow Chart design: 6](#_Toc15915580)

[Conclusions and Future work: 7](#_Toc15915581)

[References: 8](#_Toc15915582)

Introduction:

Internet data is savvy. Everything is in data in E-commerce websites about product information, price details, technical specifications, product reviews etc. and Shopping online is became very popular nowadays. Ever wondered, when everything is on web with too much information, internet users have more often come across information that are not relevant for their use. Understanding and Exploring product that includes prices and reviews information before purchasing is too difficult when there are many sellers on same platform. It is due lot of information and systems recommendations that encourages new sellers to propose their products on websites. With its unique play role of business, it is obvious that customer fail to get quality and right price of product. Which leads to customer Unsatisfaction with numerous reasons.

This research aims to propose new insight view of product though comparatively using Sentimental analysis. With respect to reviews, sellers service, product prices, return policies and etc.. Identifying key aspects and recommending better system that will impact existing platform and service to customer.

Theoretical Background and the review of world literature:

The digital technological innovation has been recognized as one essential aspect of the E-commerce businesses comprehensive solution to improve the experience, productivity, recommendations, security and service-oriented outcomes of online shopping. Examples that include but not limited to, bad experience with product quality, price comparisons with respect to other sellers, Inadequate information of product details, product returns or cancellation terms and etc. all impacts on customer satisfaction. This project aims to provide better initiative sense of understanding all factors and improving customer satisfaction with respect to product prices, comparisons, reviews etc.

Many researchers have developed techniques carefully that identifies behaviours of customers and encourages online activities due to its smart, convenient and reliable in nature. Customers of online shopping seeks to get better product for better price along with many opinions. Investigating key aspects that benefits customer and existing platform is a challenging task of our project.

Most of internet data is unstructured and sometimes hard to interpret data that is scraped. Lots of regex and scraping techniques will have to be used to crawl, fetch and structure it. Recent or past researchers introduced different methodologies (like Jsoup, java library) to scrap web data and process information for Structural data.

Initial step crawling technique is used to scrap data of product from amazon.com. That will face many issues due to complex in nature of website. Breaking the logic and fetching the data is also a key part of our project. After many pre-preparations, extractions, framing data for right analysis will take 50% of project work. Eventually, the scraped data will be used for Comparisons, Predictions and drawing inferences using Statistical and Machine Learning algorithms.

We are presenting a new way to crawl, extract information and analysis of E-commerce (Amazon) product big data with numerous features. Based on website structure and securities, crawl bot detects data with limited content and that used subjectively for analytical purposes. Several python libraries like *urllib.request, urllib.parse, urllib.error, Beautifulsoup, ssl certificates and scrapy and etc. helps to extract data with different formats. Expected file formats would be json, csv, xml, html or text.* Using Machine Learning methodology, prices inferences and Opinion mining (Sentimental Analysis) will be performed to key insights of product.

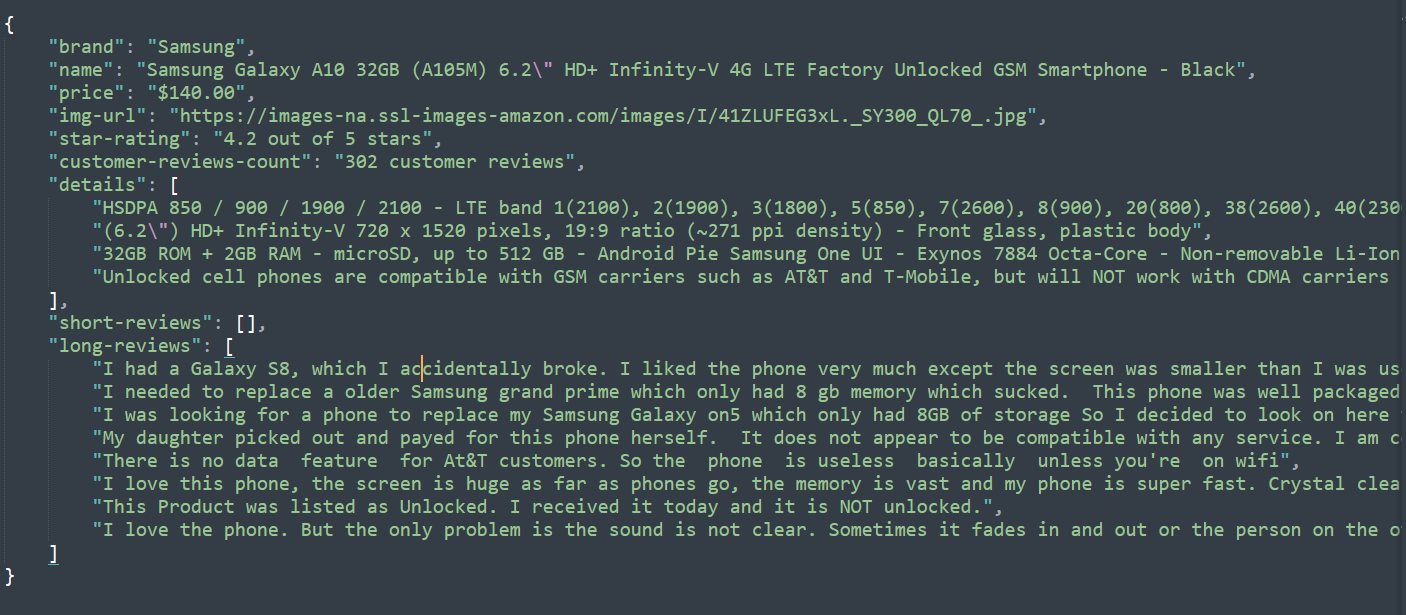
Lastly all results will be shown comparatively of chosen products that recognizes and attract users to place orders and new learn to seller at same time.

Dataset Information:

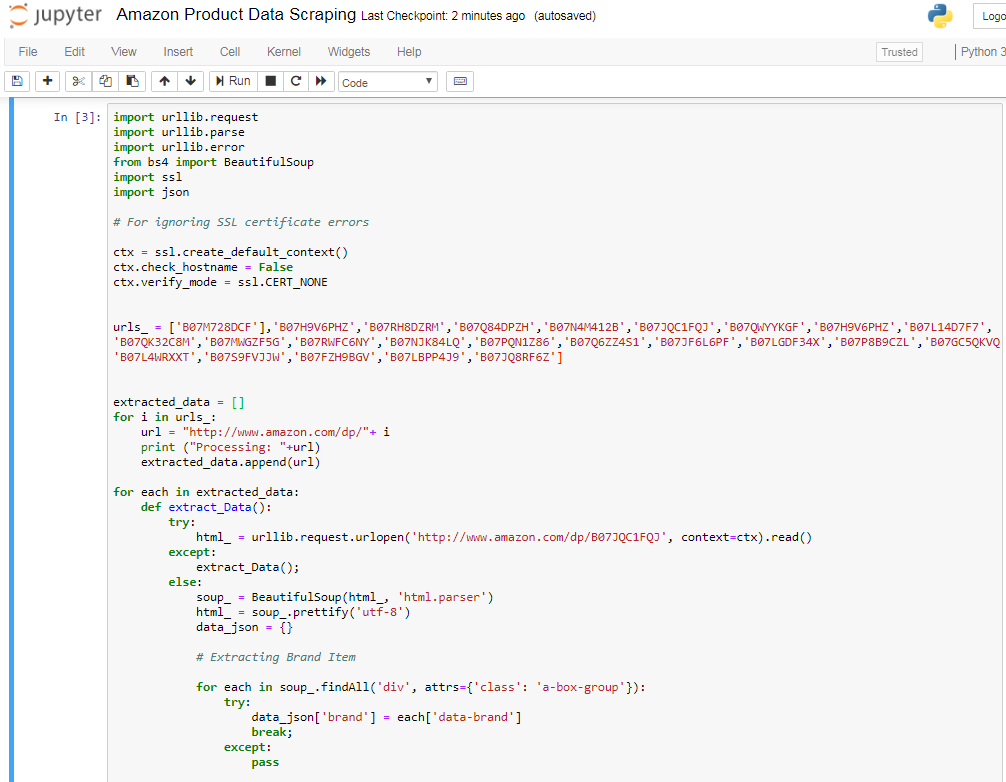
A multinational & well-established technology company *Amazon* focuses on e-commerce, cloud computing, digital streaming and artificial intelligence. Our research focused on more about e-commerce products in particular in [www.*amazon.com*](http://www.amazon.com)*.*

Product details like prices, technical information, ratings, reviews and etc. are scraped and thus helps to make analyzations and draw effective conclusions and inferences using Statistics and Machine Learning algorithms.

Sample scraped data in Json format:



Sample code:



It is a challenging task to scrape the data from an E-commerce site due to its complex structure and design with its critical logic play to avoid hacks and cracks of confidential information by anonymous users for safety and security reasons. After all many cycles of code enhancements, achieved to fetch sample required information for this project execution. This will help us to investigate product data and draw effective realizations of better experience to customer and existing platform.

Conventions and Methodology:

Recent researchers influenced different critical identifications of product view systems and compared with other e-commerce sites of same product with its effective changes in all scenarios. Our project aims to identify smartphones data, its prices, reviews and ratings. Thus data will be comparatively investigated and draw some inferences.

Key variables of E-commerce product:

Title: Product title

Description: includes product main information

Price ($): Product price, original price

Product reviews: Scraped and placed in .json format file. later it will be read through coding

Ratings: Beautifulsoup and urllib.request

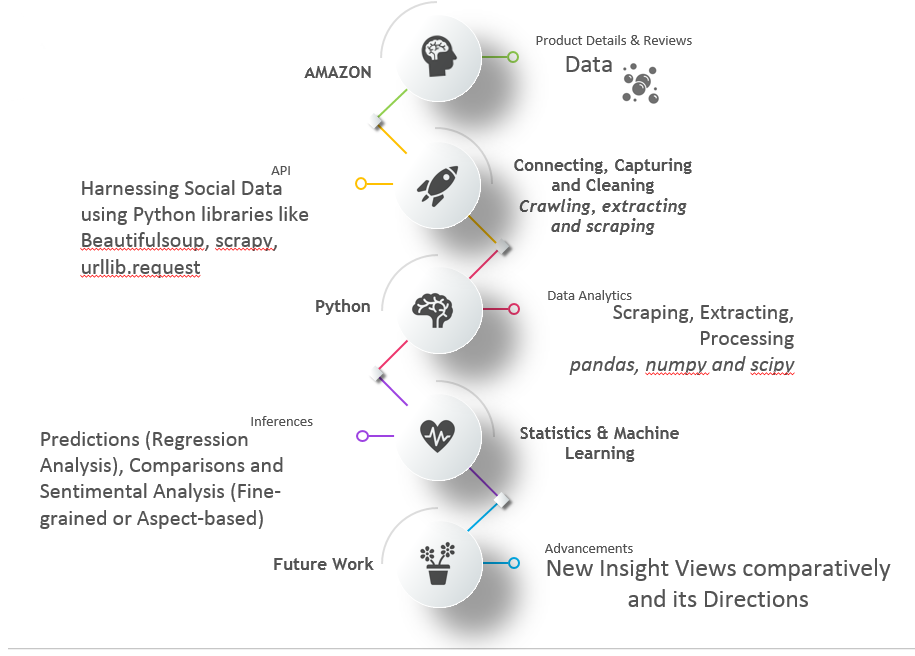
Other variables may include: Number of ratings, transformed variables in dataset.

*Crawling technique:* Scraping a well-established, high security data from amazon.com is not that easy. Many active users failed and exhausted while extracting required information. Libraries *urllib.request* and *Beautifulsoup* are used effectively and played logical task to extract data from site. Due to website high complex in nature, code may vary certain situations and face critical errors. Played smart role in terms of extracting information though it is raising HTTP error. Thus, data is will be used for pre-processing, cleaning and apply necessary transformations.

*Data Pre-processing:* Extracted data will be first taken into cleaning process and design perspective structure for apply statistics and draw some inferences comparatively. Here data science libraries such *pandas, numpy and scipy* are effectively used to structure the data. To get better intuitive sense of data, visualizations plays a key role. *matplotlib.pyplot, seaborn and Tableau* tools make advanced graphs help the customer to take decision of product purchase. This stage will implement thorough understanding of key aspects of product, its intentions to price and ratings.

*Statistics and Modeling*: Although Sentimental analysis (a Machine Learning algorithm) is used by many researchers for prediction performances of product review data, same algorithm will be used for critical identifications of reviews comparatively and suggest better system in this stage which was not performed before by other researchers.

Methodological directions:



Tools and Programming Libraries:

Python Web Scraping : Beautiful Soup, Urllib. request, Scrapy (optional).

Processing and Cleaning : Data Science libraries include *pandas, numpy, scipy*

Data Source : Amazon.com (USA)

Sample Link [*https://*](NULL)[*www.amazon.com/s?k=smartphones&ref=nb\_sb\_noss\_2*](https://www.amazon.com/s?k=smartphones&ref=nb_sb_noss_2)

Features : Feature Selection methodology

Inferences and Predictions : Statistics, Sentimental Analysis (Fine-grained and Aspect-based Sentimental Analysis) using Machine Learning, Regression analysis for inferences about product prices and other parameters.

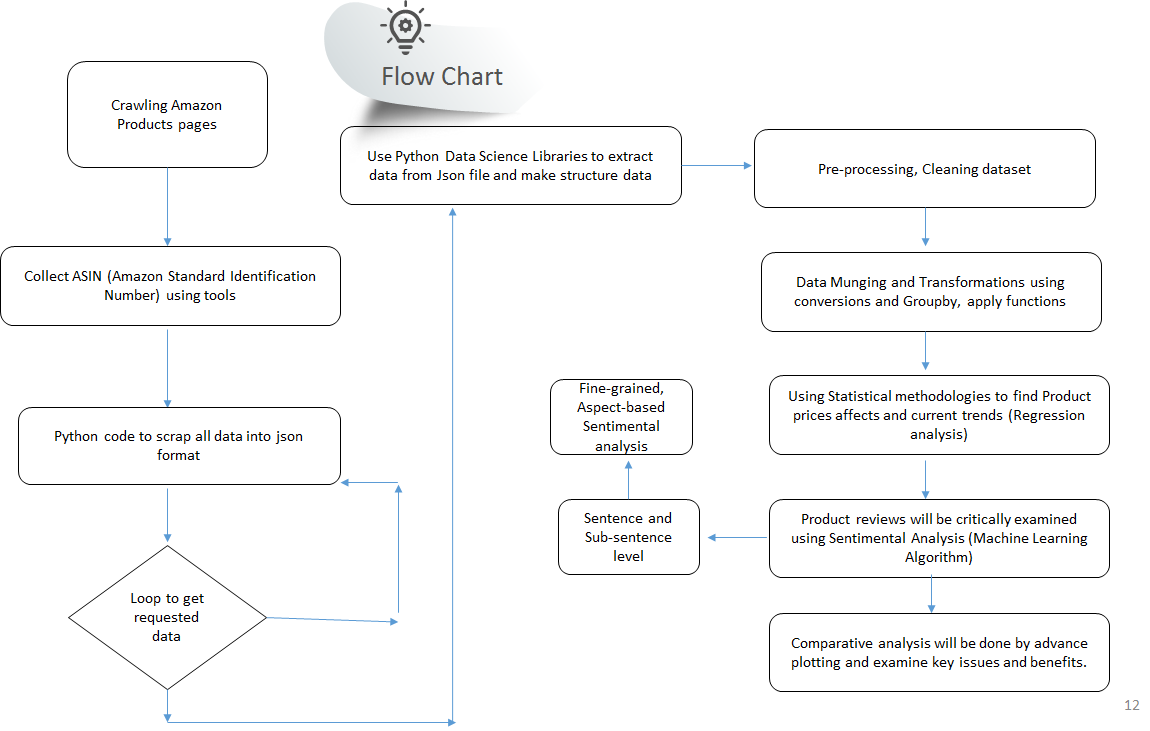
Visualization tools : matplotlib. pyplot, seaborn and Tableau

Application Implementation to real world datasets:

With this research, findings shows clear identification of product details comparatively with good quality, prices, and reviews and with overall rating of a seller recommendations to customers.

New insight views can be introduced by keeping in mind both platform and customer benefit. So that System influences more purchases with good confidence level.

Flow Chart design:



Conclusions and Future work:

Our research investigation influences existing system either directly or indirectly by performing other consistence research on product data. Future works may also include following points.

a). Users intentions and experiences can be identified and propose new recommender systems

b). Continuous enhancements lead to new representation on E-commerce site for better experience

c). Balancing customer, sellers (with guidelines) and platform and enriching businesses.

References:

1. D. Rahul, S. Mrudav, P. Priyanka and P. Bhagirath."A Novel Approach to Web Scraping Technology,"*International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 5, Issue. 5, 2015.
2. A. Pratiksha, S.R. Tandan, T. Priyanka and M. Rohit."Web Information Retrieval Using Python and BeautifulSoup,"*International Journal for Research in Applied Science & Engineering Technology (IJRASET)*. vol. 4, Issue VI, 2016.
3. Sung-min Kim and Young-guk Ha, "Automated discovery of small business domain knowledge using web crawling and data mining," 2016 International Conference on Big Data and Smart Computing (BigComp), Hong Kong, 2016, pp. 481-484.
4. Jie Wang, Shuo Yang, Yuezhi Wang and Cheng Han, "The crawling and analysis of agricultural products big data based on Jsoup," 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Zhangjiajie, 2015, pp. 1197-1202.
5. Z. Singla, S. Randhawa and S. Jain, "Statistical and sentiment analysis of consumer product reviews," 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, 2017, pp. 1-6.
6. Xiaosong Jiang and Yijun Huang, "Research on Data Pre-process and Feature Extraction Based on Wavelet Packet Analysis," 2006 6th World Congress on Intelligent Control and Automation, Dalian, 2006, pp. 5850-5853.
7. GAFFAR KHAN, Abdul. “Electronic Commerce: A Study on Benefits and Challenges in an Emerging Economy.”*Global Journal of Management And Business Research*, Vol 16, No 1-B (2016).