Methodology:

Will be using the websites Snapdeal and Flipkart for retrieving the product data. The website’s host all kinds of products and all variance of products. Products from the first page of both websites are taken and complied. Products will be of similar kind. Python programming language will be used for the project.

The first step is to get a product name from the user this will be done using “Tkinter” library. This will create a GUI for taking an input. This input will be considered as a keyword.

The second step is to form URLs using this keyword. Each e-commerce URL has a specific part where the keyword has to be entered. Using string concatenation the URL for a specific product is created.

URL=”"https://www.flipkart.com/search?q="+key+"&otracker=search&otracker1=search&marketplace=FLIPKART&as-show=on&as=off"”

Next we need to find out which HTML tags contain the relevant information for the project. We need to store these tags. As these tags will be used by web crawlers.

The next step is to set up a web crawler using Beautiful soup library. The crawler looks for the HTML tags which contain information related to the keyword representing the product. The crawler searches for the entire URL for matches of product information. This information is stored in a pandas dataframe.The entire process is repeated for the other e-commerce websites.

Once this information is gathered Ml techniques are applied to give out conclusions to the user. Information such as highest rating product, highest rating with lowest price and so on. This information will also be represented in graphical form such as graphs, pie charts using bokeh.

Flipkart and snapdeal was used because it provides consistent and well-structured webpages to crawl, while amazon webpages are mostly loaded using JavaScript which makes it difficult to crawl.

Algorithm:

Step1:

Using GUI, input is taken from the user. This input will be the name of the product. It will act as the keyword. The keyword is processed by turning all the letters into smaller case and filling up spaces if the keyword is a phrase.

Step2:

Once the keyword is prepared it is feed to the web crawlers. There are two web crawlers one for flipkart and another for snapdeal.Both the crawlers work at the same time and gather information related to the keyword in the webpages. Entries such as Name of the product, price of the product and the rating of the product is retrieved.

Step3:

The retrieved information is in raw format, it is processed using pandas dataframe.By removing unwanted entries and other fields. This data frame is stored in a CSV file.

Step4:

The CSV file is used to analyze the data gathered for a specific product. By using ML techniques the best product can be identified. Also the product with highest rating and lowest price, product with lowest price, product that is highest rated can be retrieved.

Step 5:

Finally the results can be obtained by slicing and dicing the csv file and also by using pandas data frame. The csv file can also be feed to ML algorithms for further processing.

For web crawling an algorithm called beautiful soup is used. Beautiful soup accesses the HTML of the webpage and extracts useful information/data from it. This technique is called web scraping.

**Steps involved in web scraping:**

1. Send a HTTP request to the URL of the webpage you want to access. The server responds to the request by returning the HTML content of the webpage. For this task, we will use a third-party HTTP library for python requests.
2. Once we have accessed the HTML content, we are left with the task of parsing the data. Since most of the HTML data is nested, we cannot extract data simply through string processing. One needs a parser which can create a nested/tree structure of the HTML data.
3. Now, all we need to do is navigating and searching the parse tree that we created, i.e. tree traversal. For this task, we will be using another third-party python library, [Beautiful Soup](http://www.crummy.com/software/BeautifulSoup/). It is a Python library for pulling data out of HTML and XML files.

Objectives:

1) To scrap major e-commerce websites, retrieving the best possible deal for a product by using web crawlers.

2) To use beautiful soup to find the highest rated and lowest price product in

E-commerce website.

3) To list the products from the 1st page of e-commerce websites. To validate the product retrieval by checking it’s if it’s the indeed the lowest priced and highest rated product.

The first part of the 1st objective is completed, that is retrieving the products from the website using web crawlers. Choosing the best product from the list of products generated from the crawler is yet to be done. It is the last milestone and will be done next. If the 7th milestone is complete then the entire 1st objective will be achieved.

During the 1st phase of the project all the functionalities of beautiful soup was understood. This allowed me to use beautiful soup for crawling webpages. Using this knowledge I was able to extract products from webpages. The first part of the 2nd objective was hence achieved while the 2nd part of 2nd objective is yet to be achieved.

The 3rd objective states to list the products from the 1st page of e-commerce websites that is crawled. This is already achieved while the 2nd part where the comparing of the obtained result and actual result is to be done. This can be achieved after completing the 2nd objective.3rd objective is dependent on the 2nd objective.3rd objective’s second part requires manual effort to find the highest rated and lowest priced product.

TimeLine:

Phase 1:

Milestone 1: To understand the objectives and design the architecture of the project. (4 days)

Milestone2: To research the structure of the e-commerce pages and plan on how to build a web crawler for the particular website. (1 week)

Milestone3: To understand how Beautiful soup works and the functionalities it provides. (1 week)

Phase 2:

Milestone 4: To form a keyword from the user input by performing string operations and preparing the keyword for web crawler format.(1 week)

Milestone 5: To build web crawlers for e-commerce websites.(2 weeks)

Milestone 6: To consolidate all the retrieved data from the websites.(1 week)

Milestone 7: To analyze the retrieved data (2 weeks)

Currently Phase 1 is complete. In phase 2 Milestone 4, 5 and 6 is complete. The web crawlers are currently retrieving the data from the website. The data retrieved are from the keyword provided from the user. The user input is vague hence it is processed and feed into the the web crawler as mentioned in milestone 4.Hence all the milestones till 7th milestone is complete. The final milestone of analyzing the data is yet to be completed.



