Contents

[**Background and Motivation:** 2](#_Toc73570176)

[**Problem Statement:** 2](#_Toc73570177)

[Sources: 2](#_Toc73570178)

[An Example: 3](#_Toc73570179)

[Process Flow 3](#_Toc73570180)

[Data Acquisition: 3](#_Toc73570181)

[Web Scraping: 4](#_Toc73570182)

[Data Cleaning 4](#_Toc73570183)

[Data Standardization 4](#_Toc73570184)

[Data Mapping 4](#_Toc73570185)

[Data Storage 4](#_Toc73570186)

[Detailed Architecture Design: 5](#_Toc73570187)

[Name Matching 6](#_Toc73570188)

[API Development 6](#_Toc73570189)

[Deployment 7](#_Toc73570190)

[Technology Stack 7](#_Toc73570191)

[System Validation 7](#_Toc73570192)

[Team and Milestones: 7](#_Toc73570193)

[Milestone 1: Data Collection (4 Days) 7](#_Toc73570194)

[Stories 7](#_Toc73570195)

[Milestone 2: Data Mapping (5 Days) 7](#_Toc73570196)

[Stories 7](#_Toc73570197)

[Milestone 3: Data Indexing and Search (2 Days) 8](#_Toc73570198)

[Stories 8](#_Toc73570199)

[Milestone 4: UI and Presentation/Documentation (1 Day) 8](#_Toc73570200)

[Stories 8](#_Toc73570201)

[Final TODO: 8](#_Toc73570202)

# **Background and Motivation:**

One of the biggest changes in our lifestyle over the period of last decade is online shopping. COVID have further added a lot to this ever-growing industry of e-commerce. In our day-to-day life, for any need, first thing we do is to look at e-commerce websites.

Though this market is worth 84 billion US dollars, but still major market players in India are still handful of companies. So eventually, this is a huge competitive market, and over the period of last few years a lot of companies tried to get into this business but with no luck.

After a lot of research, our marketing team suggested Tech Team to come up with a product called **Best Gadget Finder**. This product will act as an aggregator and help user to take best decision with following set of features:

1. Compare prices of gadgets across various e-commerce platforms.
2. Average rating of product. Since some sites can have paid reviewers.
3. Availability of a product.
4. Discounts comparison across various products. Useful specially during festive seasons offers.
5. Reviews Summary aggregated from various platforms

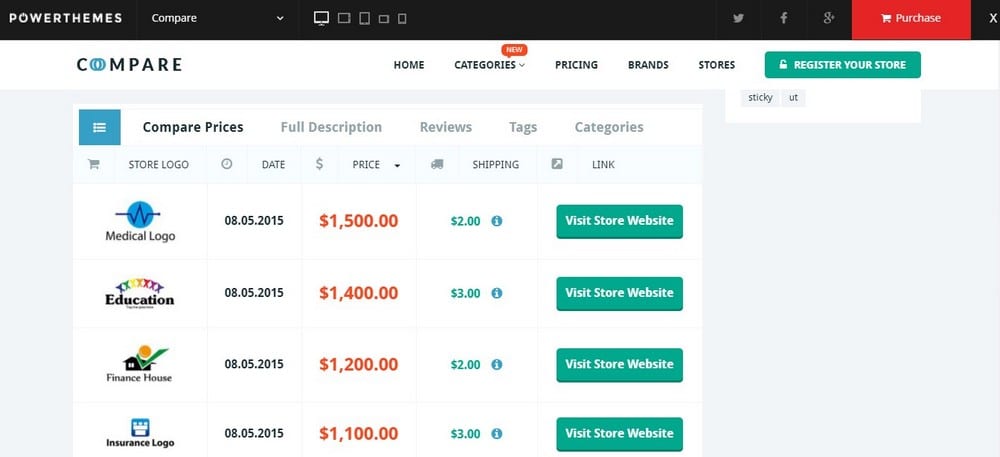
# **Problem Statement:**

**Best Gadget Finder**, a new startup to help users compare products at a single platform. It requires to pull data from various e-commerce websites and then compare product names across e-commerce websites. And then having a UI to search gadgets and see a detailed comparative UI.

## Sources:

* Amazon
* Flipkart
* Snapdeal

## An Example:



Process Flow

# Data Acquisition:

First step is to acquire the real time data from mentioned 3 websites. There can be many techniques to gather data such as open source platforms, APIs, RSS feeds or Web Scraping. We will be using Web Scraping for data collection process. Further we will have steps like Data cleaning, Standardization, Mapping, and Storage.

## Web Scraping:

Web scrapping engine’s responsibility is to scrape data from e-commerce websites and save extracted information in CSV format. A basic approach for web scrapping is as follows:

1. Search e-commerce website using product name. e.g. mobile phones.
2. Navigate result page for each result page.
3. From each page, click on each product link.
4. For each product page, extract following information using xpaths from beautiful soup.
   1. Product name.
   2. Product Description.
   3. Product Actual Price
   4. Product Discounted Price
   5. Product Rating
5. Save extracted information in a csv for each source.

## Data Cleaning

Source level data cleaning might be required as per the data collected. Some cleaning assumptions.

1. Text Casing.
2. Removing any unwanted characters such as end period, money symbol, rating keyword etc.
3. Converting prices into numerical. By converting string to number or removing commas etc.

## Data Standardization

Data from various sources need to be standardized. Such as:

1. Similar column names.
2. Adding removing extra columns

## Data Mapping

Finally, data needs to be merged and create a single CSV file with an extra column Source (Flipkart, Amazon etc.). Data from various sources need to be merged and put together.

## Data Storage

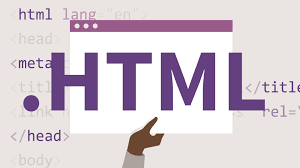
Final data will be stored in a SQL database with timestamp or CSV.

# Detailed Architecture Design:

**Text Matching**

**Scrapping Engine**

(Urllib, requests, beautifulsoup)



**Content Processing**





**API to search Gadgets**

**Output Generation and push to UI**

**Storing Standardized content to ES**

**Matching Names across sources**

**Extracting Relevant info**

**Web Scrapping**

**Sources to Scrape**

**User Interface**

****

**Flask API**

**Data Storage**

Elastic Search

Ngram Matching Distance

Edit Distance

Word2Vec

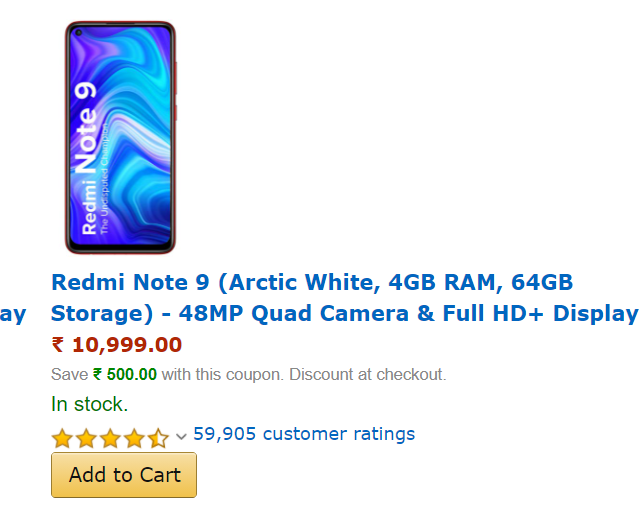
Cosine Similarity

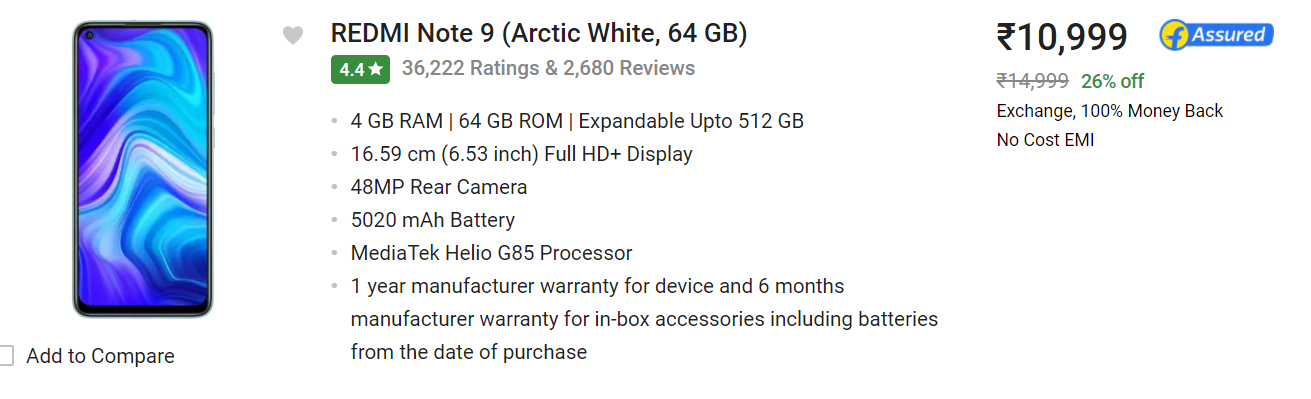
Information Standardization

Information Extraction

# Name Matching

As different sources can have different product names for a same product. Consider following products from amazon vs Flipkart.





These are 2 different products but have different names. For matching such names, we will use fuzzy matching algorithms such as cosine similarity, edit distance etc. Some study for fuzzy text matching algorithms.

* <https://towardsdatascience.com/surprisingly-effective-way-to-name-matching-in-python-1a67328e670e>
* <https://www.analyticsinsight.net/company-names-standardization-using-a-fuzzy-nlp-approach/>

# API Development

**Search Product: API to search product.**

* **Input**: A string (product name)
* **Output**: A Json with results from each source. Plus, Aggregated Rating

# Deployment

Application needs to be deployed using Flask APIs.

A basic web application to call UI and to show products comparisons for the end user.

# Technology Stack

* Python 3.7
* Flask
* NLTK
* Spacy
* Beautifulsoup
* Requests/urllib
* ElasticSearch

# System Validation

An unseen test data will be used to test the system

# Team and Milestones:

An unseen test data will be used to test the system

Milestone 1: Data Collection (4 Days)

Data collection using web scrapping, for 3 websites. Cleaning data and make it consumable for next milestone.

**Things to Learn:** Web Scraping, Data Collection

### Stories

|  |  |  |  |
| --- | --- | --- | --- |
| Story | Complexity | Team | Effort |
| Scrape Flipkart.com for phones | P1 | 1-2 | 2-3 days |
| Scrape snapdeal.com for phones | P1 | 1-2 | 2-3 days |
| Scrape amazon.com for phones | P1 | 1-2 | 2-3 days |
| Data Cleaning/Standardization for each source | P2 | 1-2/source | 1 day |

Milestone 2: Data Mapping (5 Days)

Mapping data using NLP algorithms and making a central database for all sources with products mapping.

**Things to Learn:** NLP, Spacy, NLTK, Machine Learning, TFiDF, sklearn and many more

### Stories

|  |  |  |  |
| --- | --- | --- | --- |
| Story | Complexity | Team | Effort |
| Names Preprocessing using NLP | P1 | Group | 1-2 days |
| Names Matching using fuzzy matching | P1 | 1-2 | 2-3 days |
| Names Matching using ML | P1 | 1-2 | 2-3 days |
| Merging Names Matching and creating a single dataset for all sources | P2 | Group | 1 day |

Milestone 3: Data Indexing and Search (2 Days)

Elastic Search setup and indexing data to ES. Build API to search products using Flask.

**Things to Learn:** Elastic Search, Flask, API development, Search

### Stories

|  |  |  |  |
| --- | --- | --- | --- |
| Story | Complexity | Team | Effort |
| Elastic Search Setup | P1 | 1-2 | 1 day |
| Indexing data into ES | P1 | 1-2 | 1 day |
| Flask API to Search Products and get results in json format | P1 | 1-2 | 1 day |
| Testing and bug fixes | P2 | Group | 1 day |

Milestone 4: UI and Presentation/Documentation (1 Day)

Final delivery of project, UI development and presentation.

**Things to Learn:** UI, HTML, JavaScript, Story Telling

### Stories

|  |  |  |  |
| --- | --- | --- | --- |
| Story | Complexity | Team | Effort |
| Basic UI to search and show results | P1 | 1-2 | 1 day |
| PPT and documentation | P2 | Group | 1 day |
| Code Push to Github | P2 | Group | 1 day |
| Project Presentation and Knowledge Sharing | P1 | Group | 1 day |

# Final TODO List:

1. Create a business story around the project. Why we need it. What are the current similar products available in the market.
2. Create a list of business insights. Usefulness of the project for business.
3. Future Scope of this project. How this project can be turned into a future product.
4. Current short comings of the project.