Mini Project Report On

Comparison Shopping Trend Analyzer.

Submission By:
Gaurav Gupta (RIT2012018)
Harshit Gupta (RIT2012048)
Kamal Soni (RIT2012076)
Nakshatra Maheshwari (RIT2012074)
Shashank Sharma (RIT2012075)
Shweta Choudhary (RIT2012025)

Under Guidance of
Dr. O.P. Vyas

Indian Institute of Information Technology, Allahabad

March 2015

CANDIDATE'S DECLARATION

We hereby declare that the project work entitled "**Comparison Shopping Trend Analyzer**" submitted to the IIIT Allahabad as 6th semester project, is a record of an original work done by us under the guidance of Dr. OP Vyas and has not been submitted to any other University or Institute or published earlier.

Place:
Date:

Gaurav Gupta (RIT2012018)
Harshit Gupta (RIT2012048)
Kamal Soni (RIT2012076)
Nakshatra Maheshwari (RIT2012074)
Shashank Sharma (RIT2012075)
Shweta Choudhary (RIT2012025)

CERTIFICATE FROM THE MENTOR

It is certified that this project report "**Comparison Shopping Trend Analyzer"** of mini project taken in 6th semester is the bona fide work of "Gaurav Gupta (RIT2012018), Harshit Gupta (RIT2012048), Kamal Soni (RIT2012076), Nakshatra Maheshwari (RIT2012074), Shashank Sharma (RIT2012075), Shweta Choudhary (RIT2012025)" who carried out the project work under my supervision.

Dr. O.P. Vyas

ACKNOWLEDGEMENTS

If words are considered as a symbol of approval and token of appreciation then let the words play the heralding role expressing my gratitude.

The satisfaction that accompanies that the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crown all efforts with success. We are grateful to our project mentor Mr. O P Vyas for the guidance, inspiration and constructive suggestions that helpful us in the preparation of this project. We also thank our colleagues who have helped in our mini-project.

Place:
Date:

Gaurav Gupta (RIT2012018)
Harshit Gupta (RIT2012048)
Kamal Soni (RIT2012076)
Nakshatra Maheshwari (RIT2012074)
Shashank Sharma (RIT2012075)
Shweta Choudhary (RIT2012025)

INDEX

1. INTRODUCTION

(a) Overview :

Price comparison have mushroomed over recent years and are now seen by many as a tool of consumer empowerment. They are slowly beginning to shift traditional asymmetries in information and power between a consumer and a supplier which focus primarily on information giving and advice. A new generation of services that build on the price comparison model is emerging. These include collective switching sites, group purchasing, mobile apps or more sophisticated consumption data analyzers. The growth of the price comparison tool market is undoubtedly fueled by consumer demand for third party services which can: save them time and money when navigating through the maze of deals and the complexity of products and services on the market take the hassle of switching, for example switching energy suppliers

There is probably no need to say that there is too much information on the Web nowadays. Search engines help us a little bit. What is better is to have something interesting recommended to us automatically without asking. Indeed, from as simple as a list of the most popular bookmarks on Delicious, to some more personalized recommendations we received on Amazon, we are usually offered recommendations on the Web.

(b) Objective :

To categorize products according to their specification in certain price range. (Price range is input by user).

* Only for Smart-Phones and Smart-TV's.

(c) Why this objective? :

Most people try to get best product under their affordable range irrespective of number of features.

2. Literature Survey

1) What is web- scraping?

   Data Scraping is a technique in which a computer program extracts information from websites.

2) What is Crawling of Web?
   Crawling of web refers to crawling of data from online sources via programs called bots/spiders.

3) Scrapy as a scraping program and spiders.
   Scrapy is a python library used to extract information from a given website by crawling the website.
   We write spiders or bots in python that run to crawl and gather the html source of a webpage from which scrapy extracts the required information.
4) Why we require data cleaning?

   The data gathered from scrapy may contain noise. For example: Expected Output:" Rs. 11000" but Actual Output: "there Rs. 11000". So we require data cleanising.

5) Why normalization of database?

   When we are gathering large amount of data, queries over it increases complexity. So, databases are to be normalized.

3. Methodology

Based on specification & ratings comparison of different models in that Price Range we will try to predict the best products.

(a) Extraction of Data :

To extract data from various websites we used scrapy library of python to crawl through the resulting HTML pages on client side and extract information based on pattern of their HTML tags.

But, scrapy can crawl only single page at a time, i.e. the page that is currently visible so, we included selenium web driver with scrapy to crawl through pages with Next Buttons, Ajax Scripts to load more products on that pages.

(b) Removing Noise :

Noise is some unwanted data. For example when collecting data via scrapy library, we may get some unwanted strings inside price.
Data Cleansing helps to resolve this problem.

(c) Create Adjacency Matrix of Product Ratings and use Matrix Factorization to get missing Values :

Matrix factorization can be used to discover latent features underlying the interactions between two different kinds of entities. (Of course, you can consider more than two kinds of entities and you will be dealing with tensor factorization, which would be more complicated.) And one obvious application is to predict ratings in collaborative filtering.

	D1	D2	D3	D4
U1	5	3	-	1
U2	4	-	-	1
U3	1	1	-	5
U4	1	-	-	4
U5	-	1	5	4

In a recommendation system such as Netflix or MovieLens, there is a group of users and a set of items (movies for the above two systems). Given that each users have rated some items in the system, we would like to predict how the users would rate the items that they have not yet rated, such that we can make recommendations to the users. In this case, all the information we have about the existing ratings can be represented in a matrix.

And the matrix obtained from the above process would look something like this:

	D1	**D2**	**D3**	**D4**
U1	4.97	2.98	2.18	0.98
U2	3.97	2.40	1.97	0.99
U3	1.02	0.93	5.32	4.93
U4	1.00	0.85	4.59	3.93
U5	1.36	1.07	4.89	4.12

(d) Select Best product as one with maximum ratings :

Now since we have ratings for all products sum it up and the product with maximum ratings can be considered the best.

(e) Regularization :

The above algorithm is a very basic algorithm for factorizing a matrix. There are a lot of methods to make things look more complicated. A common extension to this basic algorithm is to introduce regularization to avoid over fitting. This is done by adding a parameter and modify the squared error.

4. Hardware and Software Requirements

(a) Hardware Requirements:

None

(b) Software Requirements:

- Python Compiler
- PIP
- Scrappy Library,
- Selenium Webdriver (for multipages scroll and AJAX Websites)
- MYSQL (to store scrapped values),
- Notepad++ (Any IDE will be suitable), Dev C++ (gcc based IDE).
- Virtual Box and Ubuntu 14.04, LAMP-Server for web hosting.

5. Activity Time Chart

Time	Activity
January	Learning about Data Collection, python Language, php language.
February	Learning about Data Cleansing, Filling of missing ratings by Matrix Factorization, regular Expressions to remove noise.
March	Implementation of scrapy, removing noise from data using regulat expressions/Data Cleansing, matrix factorization and finally php based User Inerface to Query with MySQL database.

(a) Work done so far:

- **a.** Currently we are in success to extract data from websites
- **b.** Removing of the noise.
- **c.** PHP based UI for User Query.
- **d.** VB/.bat Scripts to crawl data at regular intervals.

(b) Work to be done after Mid-Semester:

a. Crawl data at regular intervals to gather large amount of Data.
b. Apply matrix factorization on ratings*.
c. Mine the above gathered data and analyze patterns to predict best product based on Features.
d. Prediction of increase/decrease of prices in Product based on previous pattern.

*** Ratings of product is extracted as overall rating per website as it was not possible to extract individual users' rating for every product and adjacency is created for websites vs Product.**

6. Screenshots of out work

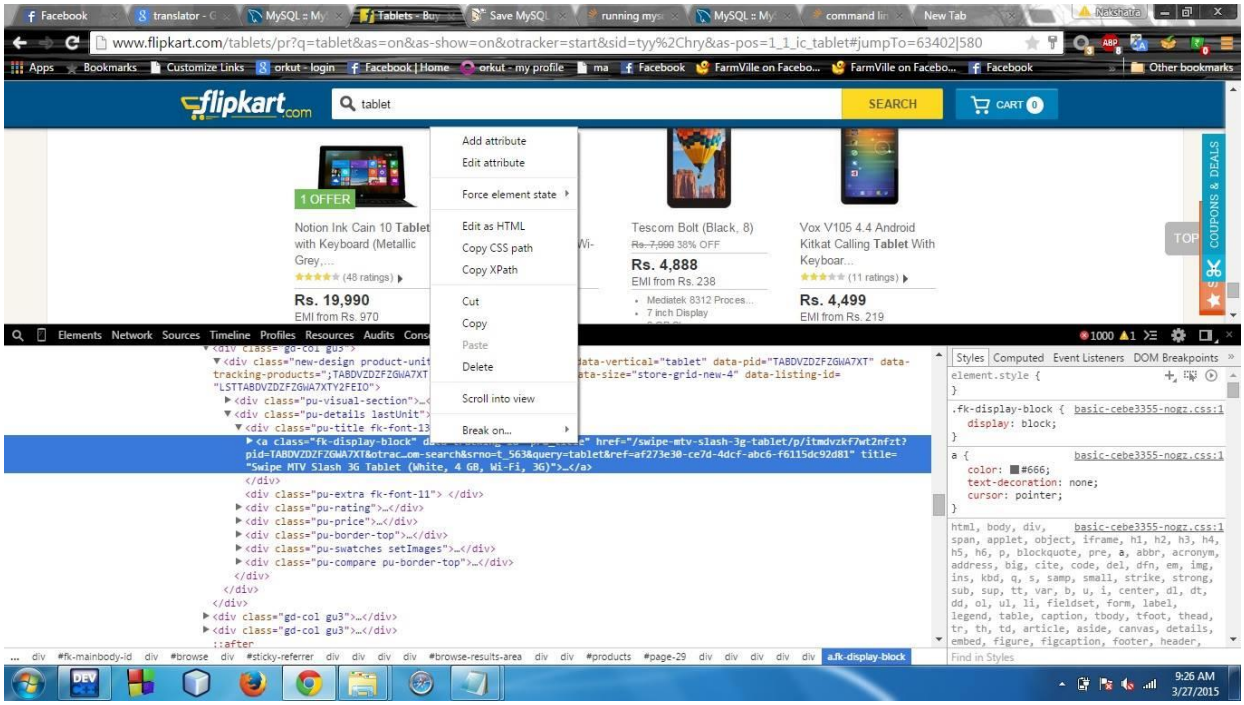

Fig: Getting X-path.

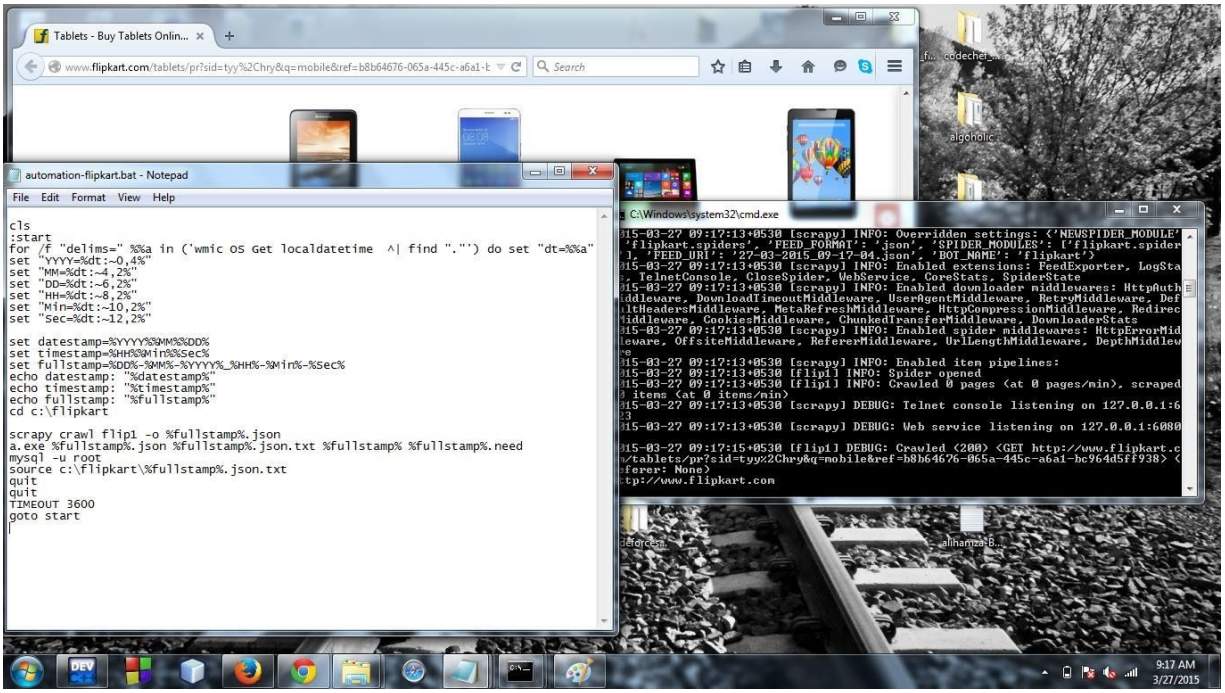

Fig : Data Extraction using Scrapy

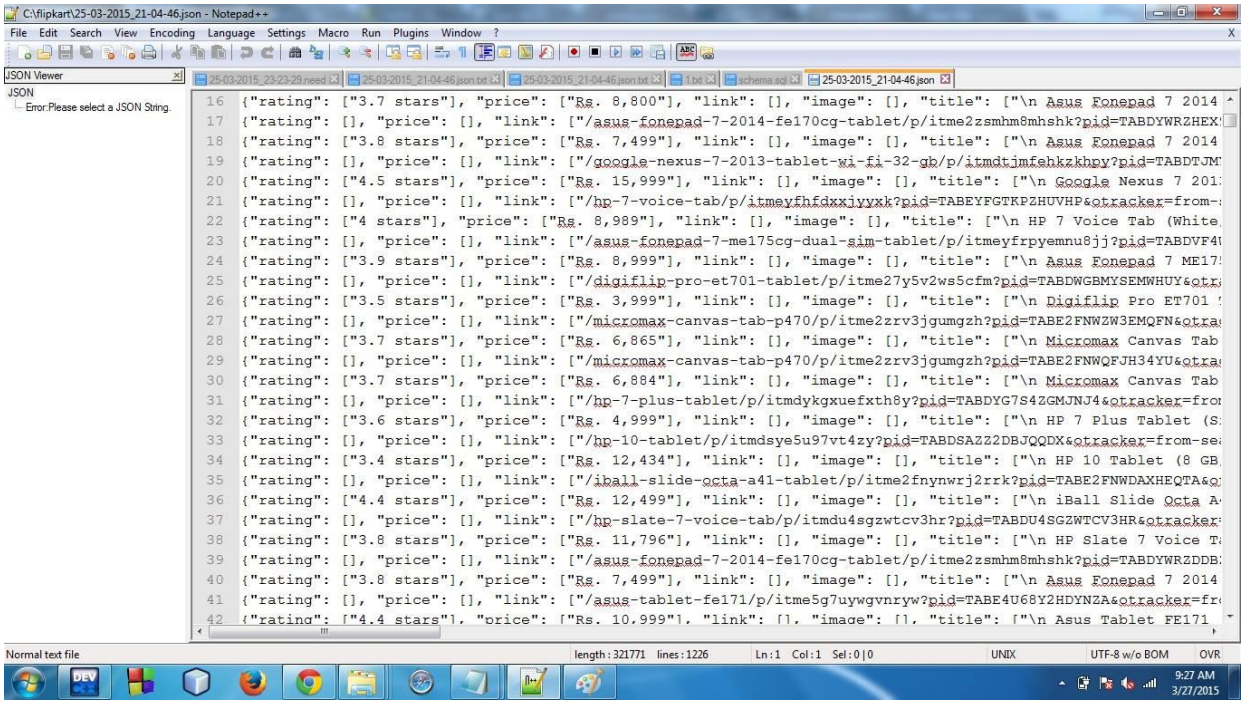

Fig: Extracted Raw Data

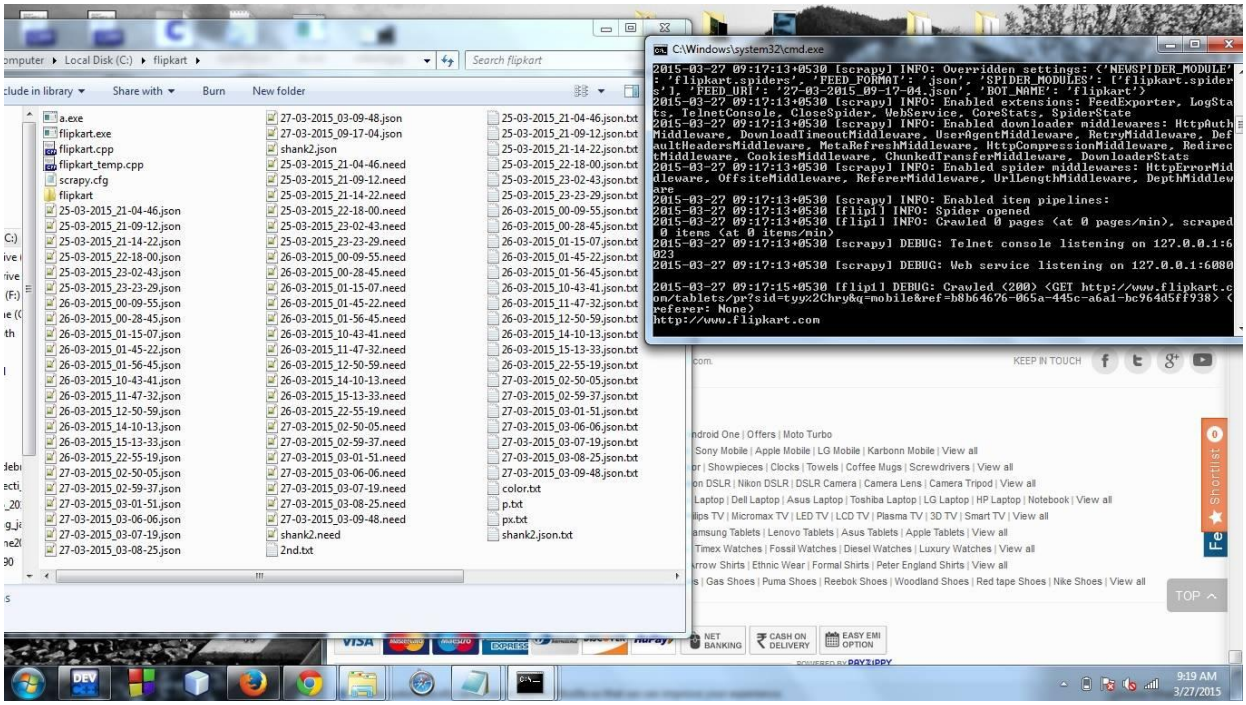

Fig: Data Extraction at Regular intervals

```
{"title": ["Spice Mi-520n (Black) with 3 Back Panels (Red, Yellow, White)"],
    "price": ["11,500.00"],"rating": [],"link"=[link],"image"=[image]}
INSERT INTO mobile VALUES("spice","mi 520n 3 back panels red yellow white",
    11500,NULL,"black","link","image-link");

{"title": ["OnePlus One (64GB, Sandstone Black)- Invite Only"],
    "price": ["from ","4,260.00"],"rating": ["3 out of 5 stars"],"link"=[link],"image"=[image]}
INSERT INTO mobile VALUES("oneplus","one",4260,3.0,"sandstone black","link","image-link");

{"title": ["Karbonn A11+(Black) smart-phone"],"price": ["Rs. 4,888.00"],
    "rating": ["4.6/5"],"link"=[link],"image"=[image]}
INSERT INTO mobile VALUES("karbonn","a11",4888,4.6,"black","link","image-link");
```

Fig: Data before and after Noise Removal.

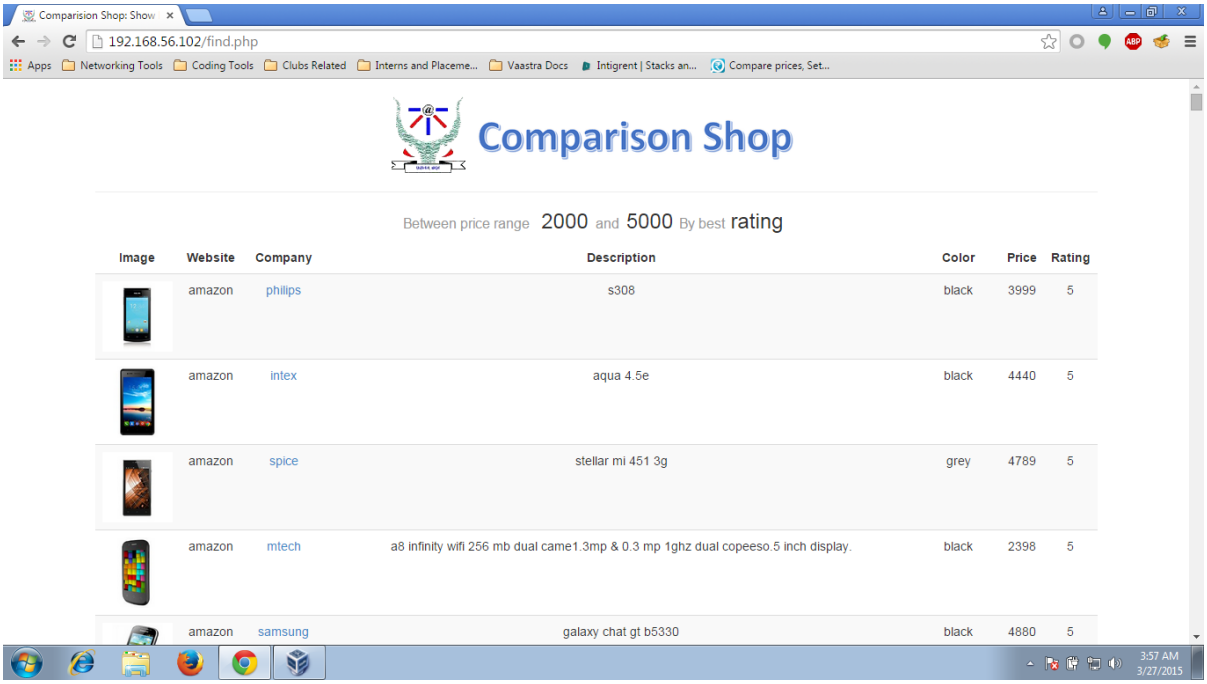

Fig: Resulting Output

7. Limitations

- Data may still have some noise.
- Problem with dynamic websites. (Websites keeps updating their format of data representation from time to time so the same scrapy script cannot be used again).
- Large databases are required to store data.
- High Speed Internet is required to scrap the data. (The VB/.bat Scripts run after certain threshold failing which no more data is extracted).

8. Refrences

- **Learning scrapy : http://doc.scrapy.org/en/latest/intro/tutorial.html**
- **Erhad Rahm* Hong Hai Do: "Data Cleaning: Problems and Current Approaches",Technical Paper, University of Leipzig, Germany.**
- **Yehuda Koren, Robert Bell and Chris Volinsky (2009): "Matrix Factorization Techniques for Recomender Systems",Yahoo Research, IEEE Computer Society.**
- **M. W. Berry, M. Browne, A. N. Langville, P. V. Pauca, and R. J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. Computational Statistics & Data Analysis, 52(1):155–173, September 2007**
- **Artus Krohn-Grimberghe, Lucas Drumond, Christoph Freudenthaler, Lars Schmidt-Thieme (2012): Multi-Relational Matrix Factorization using Bayesian Personalized Ranking for Social Network Data , Proceedings of the Fifth ACM International Conference on Web Search and Data Mining.**
- **Muller H., Freytag J.:"Problems, Methods, and Challenges in Comprehensive Data Cleansing", Humboldt-Universitat zu Berlin, Germany.**

9. Suggestions/Remarks (if any)