**Data minning and analyze data about IEEE research papers**

¹Vishwajeet Shinde ²Anirudha Shende 3 Hrishikesh Vaze 4 Prajwal Singh

GUIDED BY : Prof.Varsha Powar

1. Abstract

This system extracts data from different IEEE paper’s websites with help of automation via selenium using python. This system uses xpath with selenium for web scraping to get data that contain different attributes of IEEE papers from its website. Then it converts output data to a CSV file on which preprocessing is done with help of pandas and NumPy.

It will help us to automate data mining and create excel files containing all types of attributes related to IEEE research papers. This describes how web scraping and natural language processing can be utilized to get correct data related to research papers from different websites. It demonstrates how web scraping can be useful for extrapolating large amounts of data from publicly available web pages to extract data from a wide array of sources and to further information in the field. We discuss how natural language processing can be used to reliably obtain salient information from textual data, and how it can complement qualitative analysis.

Visualization and Clustering can be applied to given data attributes like citation and journal name to get journal wise clusters and it will help to understand data.

1. Introduction

Research papers is best way to demonstrate any topic to all world and now many people are publishing their research papers in different websites such as IEEE etc. A research paper is an essay in which you explain what you have learned after exploring your subject in depth. In this you include information from resources such as books, articles, interviews, and websites. You also use your own ideas, knowledge, and opinions.

In different colleges teachers and students are publishing their papers on the internet but it's huge in numbers and so it takes a huge amount of time to keep a record of each and every paper. For that automation is a good way which helps us to save a lot of time and our work. So this system is to automate this work i.e to keep record of every research paper with a single link of that paper. With help of link /url given this system uses selenium with python for web scraping which gets different attributes related to papers like author, data,ISSN,DOI title, publication etc. this data helps us to keep record of every paper and gives information about that papers in short. So by using xpath and other selenium html parser we get that data and save it in a dataframe. After that this system create a csv file of that data and preprocess it with different preprocessing techniques like checking empty values, data transformation, visualization and correlation

analysis. It basically corrects the format of given data which might contain some wrong values and outliers or noises. To analyze data about research papers and find out current trends in topics and most used papers and their departments and many more can be done using NLP.

After getting data in correct format by using different visualization tools like Tableau public and orange get different visualizations to understand the pattern and data relatively. Then we can apply machine learning model to get inference of data.The Paper is organized in the following manner: the (I) section includes the introduction of the paper(II) second section includes the literature survey, (III) third section includes the methods used (IV) includes the work contributions and the last section (V) includes the references used.

1. Problem Statement

Data mining with web scraping and retrieving data in the correct format. Analyze data about research papers(IEEE). Making inference of given data using visualization and ml algorithm.

1. Objective

Extract data of ieee research papers from ieee websites to collect data by using web scraping, preprocessing data and making inferences of it. The goal of this project is to retrieve data about ieee papers from given links and analyze it.

Another goal is to avoid manual error and workload in manual data entry in excel sheets by using automations and machine learning.

After getting data in the correct format, visualize data with different tools and make inferences of it by using ml algorithms.

1. Dataset Description

Current dataset of the project consists of the values of attributes we are required to fill in the excel sheet. It consists of values about IEEE research papers like ISSN, DOI, Authors details, date of publications and link etc.

Preprocessed dataset has values in correct format which on visualization gives us better understanding of data about citations and journal type.

# Dataset values: Authors

# Title of papers

# Journal name

# Citations

# DOI

# Pageno

# Month of publications

# Year of publications

# ISSN

# URL

# Data Preprocessing

1] Dealing with errors in data.

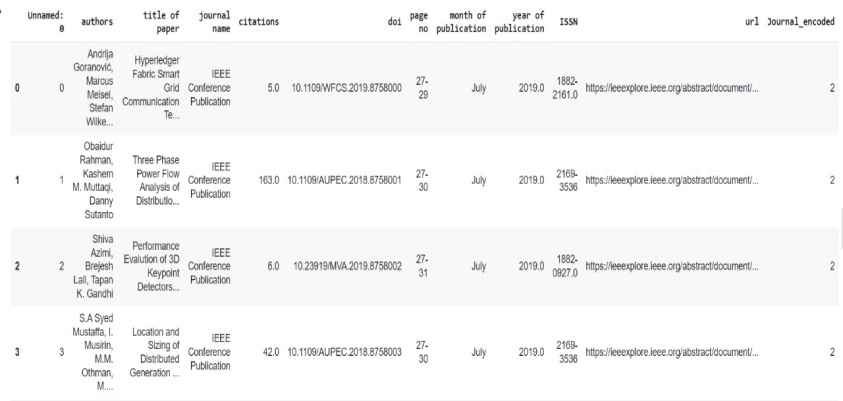
2] Dealing with missing values and noise.

3] For data visualization we used various bar graphs and pie charts.

# Related Works:

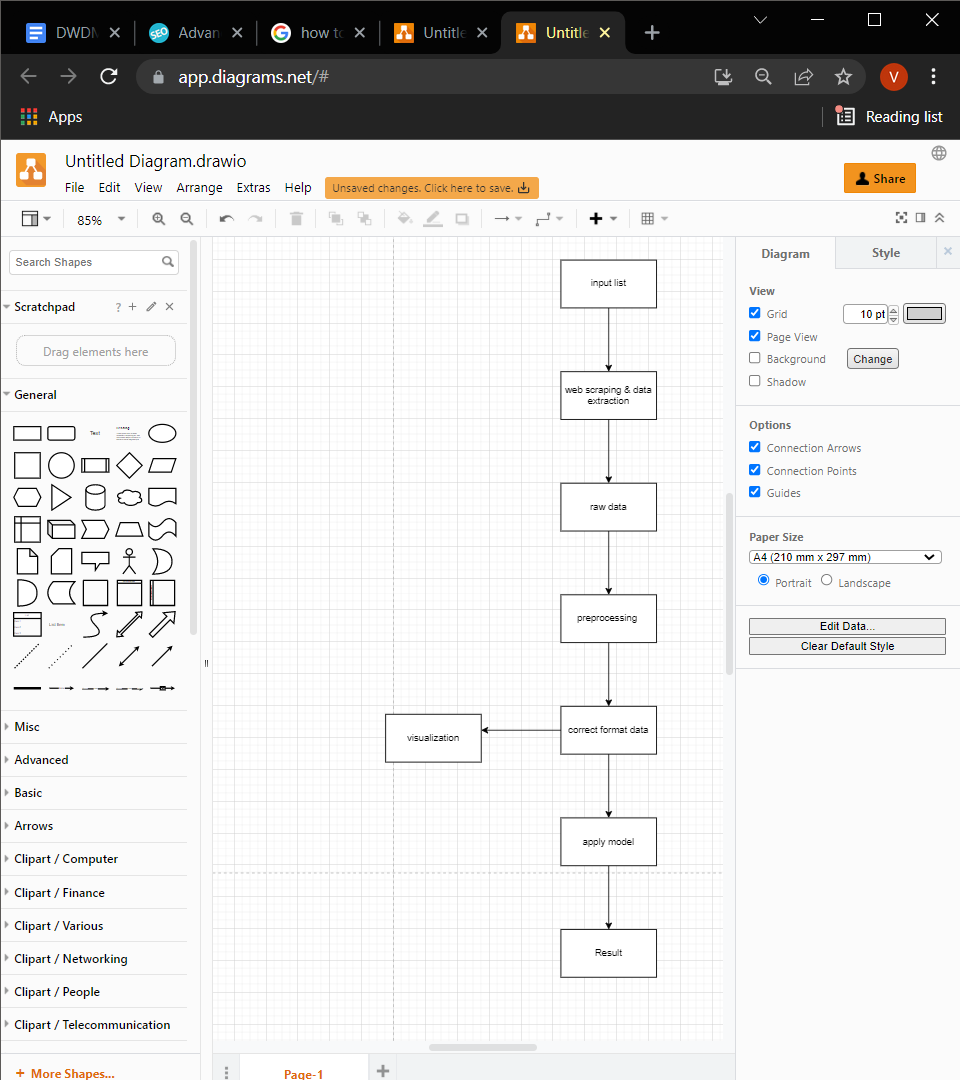
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Serial Number | Research Paper Info | Dataset | Technology/  Tools | Conclusion | Limitations/  Future Work |
| 1 | Web scrapping and storing data in a database, a case study of the used cars market | Cars Dataset | BeautifulSoup  Pandas  Numpy  Python | it deals with the process of web scraping data from different locations on the Internet for the purpose of collecting and analyzing data | It can be used in further analyzes. |
| 2 | WEB-BASED COLLEGE PLACEMENT ANALYSIS  Akhilesh Shinde, Aniket Shinde, Vivek Singh, Siddhesh Shivdikar, Sharvari Govilkar  Published 5 October 2021 | College Placement dataset | Python  Matplotlib  BeautifulSoup  Pandas  Numpy | This project can be used in college for their placement activities. colleges  can rectify the problems because of which companies are not coming for placement and use for the betterment of the college placement | 1. The Dataset can be further extended by adding current and next year recurring placement data 2. Working for more accuracy |
| 3 | Data Analysis by Web Scraping using Python  David Mathews, Sandeep Mathur | Retail | Beautiful Soup  Pandas  Matplotlib | Thus categorized retail dataset. | . Hidden web data need synthetic and semantic matching to fully achieve automatic integration. |
| 4 | Utilizing Web Scraping and Natural Language Processing to Better Inform Pedagogical Practice  Stephanie Lunn | Dataset from indeed.com | Beautiful Soup  LXML  WordCloud  Pandas  Matplotlib | It demonstrates techniques that could be utilized for numerous applications to further knowledge in CSE and analysis of it. | More detailed analysis |
| 5 | Comparison of E-commerce Products using web mining  By  Riya Shah | Dataset created for ecommerce price data | Django  Mongodb | It will help users in decision making while buying products online. This website will facilitate users to analyze prices that are present on different e-commerce shopping websites | Can be combined together to get automatic suggestions from the website itself for searched products.  Further use of AI/NLP |

##### Dataset and features



##### 

##### Architecture



##### fig1:System Architecture

##### methodologes

**Data extraction :**

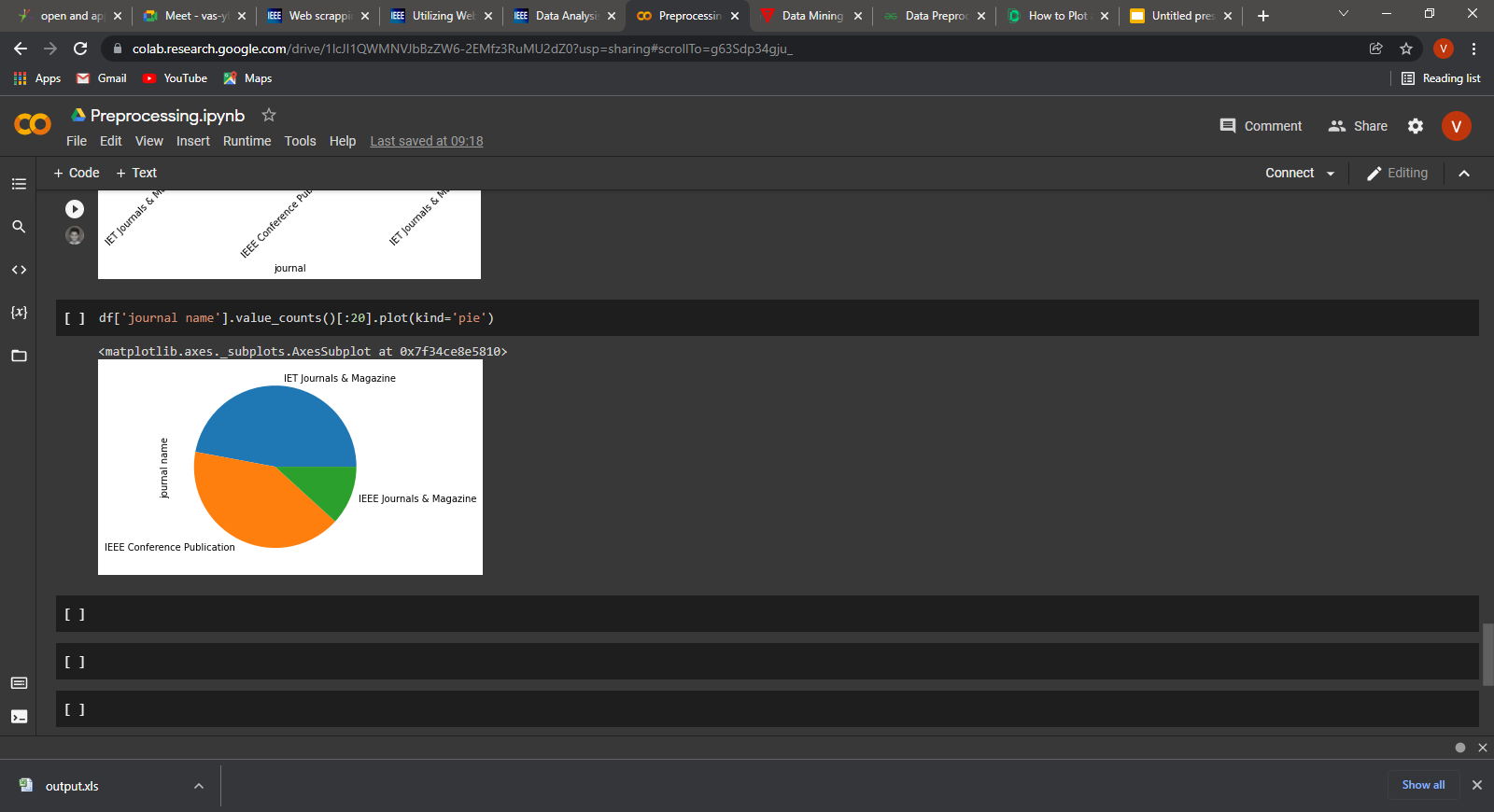
Web scraping is data extraction methodology which will extract data from websites using its HTML code. With help of selenium we can use automated web scraping for data extraction to get the data with its html tag.

For getting data in require format application of selenium which will get the data present on website by searching its html tag in its source code.

**Data preprocessing :**

Applying preprocessing techniques on raw data to convert it into standard data without errors and noise is called as data preprocessing.

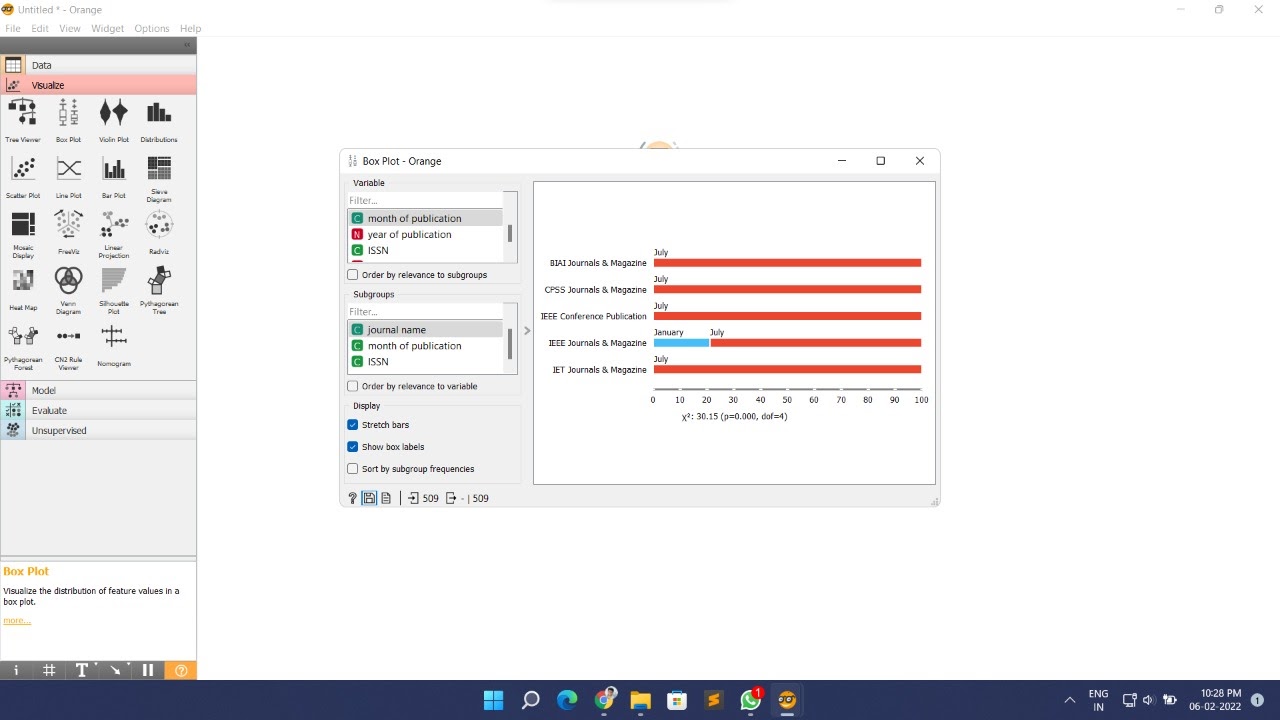
As we get data in csv format from web scraping but this data is not in correct format with respect to sementics and grammar, we tried different preprocessing techniques which will transform data in useful and correct format.

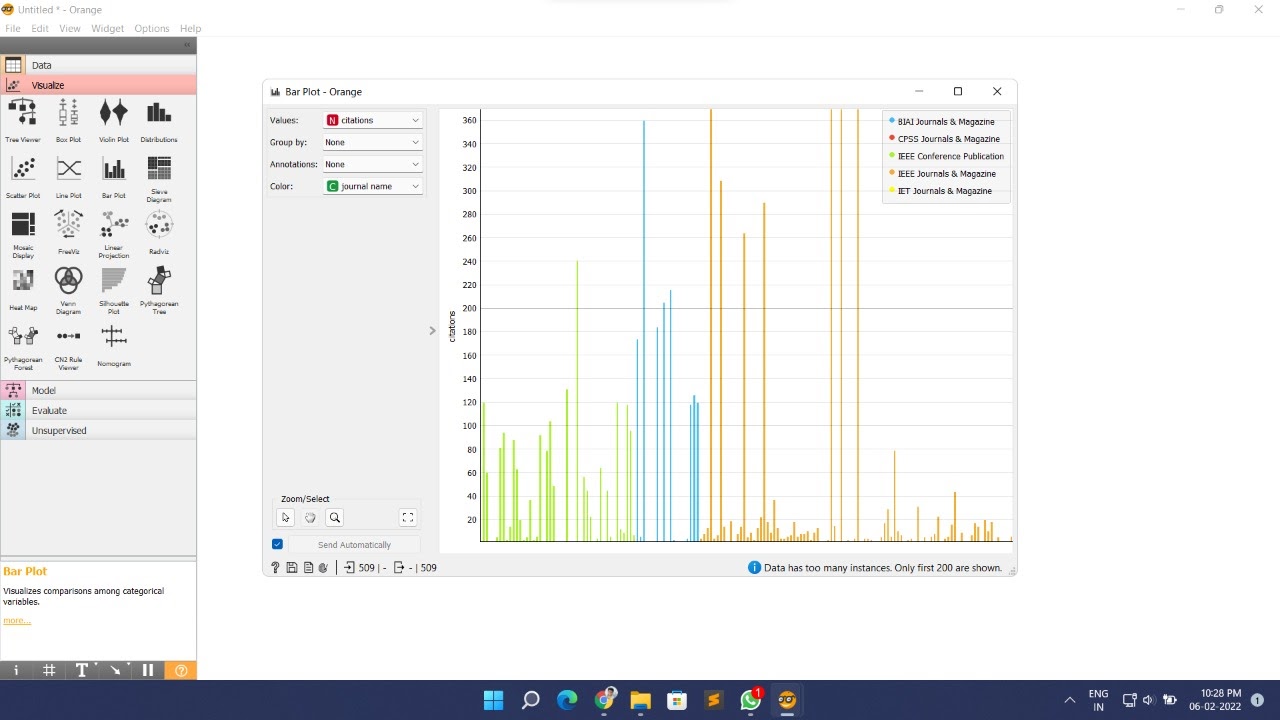


Visualization :

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

So in this dataset number of citations and journal type are attributes which can be visualized to better understand relations in them.

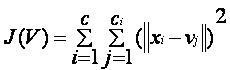




K means :

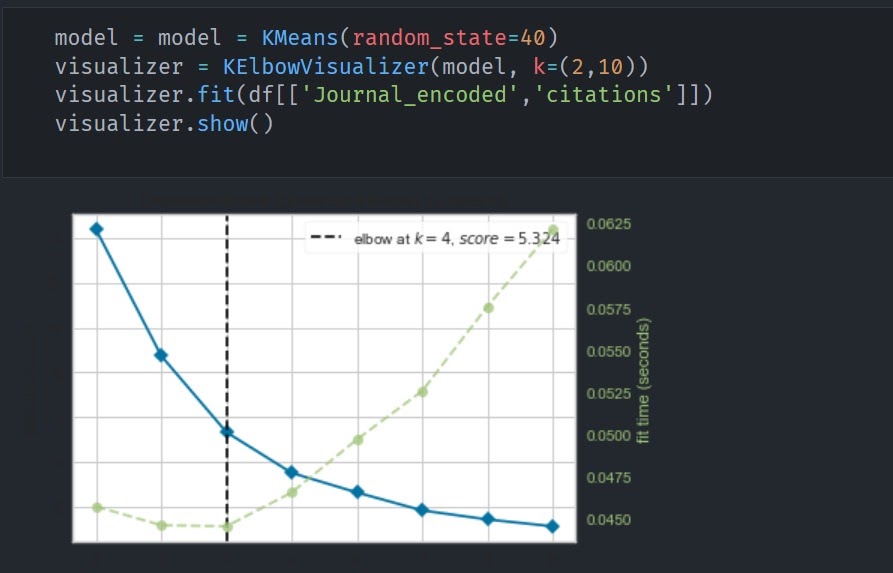
K-Means clustering intends to partition n objects into k clusters in which each object belongs to the cluster with the nearest mean. This method produces exactly k different clusters of greatest possible distinction.

As we applied k mean clustering we can get clusters which are categorized journal wise and citations are clustered accordingly. So it will give us more understanding of data



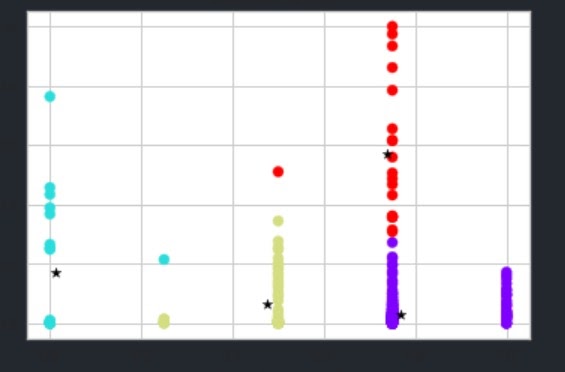
Elbow method:

Elbow method helps to get optimum numbers of clusters for given set of data. By iterating in range given for k values and find maximum score for k number of clusters .

****

1. RESULTS-

According to k means we get here clusters of journals with citations given below :

****

Silhouette score for given clusters is 0.73

Inertia for given clusters is 5.34

##### Conclusion

According to the requirements of the project this system automatically extracted data in raw format using web scraping (selenium) and preprocessed it using different preprocessing techniques. After that for better understanding of data we used visualization from different tools like orange etc.

After understanding and getting data in correct format we applied a machine learning model like Kmeans for clustering the given papers according to citations.

So thus automated data extraction and getting data in correct format for making inference of it is implemented.

1. Future scope

This system has done web scraping to extract data from IEEE websites and transform raw data to correct format by using preprocessing and for further understanding of data it uses an ML model like kmeans.

But for the future we can make a system that can extract data from not only websites but also from given information from that research paper which will require NLP text mining.

1. Work Distribution/Contribution:

Vishwajeet shinde: Research Papers, Literature Survey, web scraping, Model Building

Anirudha shende : web scraping, Data Preprocessing, Data Understanding, Model Building

Hrishikesh vaze: Research Papers, Literature Survey, web scraping, Model Building

Prajwal Singh: web scraping, Data Preprocessing, Data Understanding, Model Building

##### References

David Mathew Thomas Sandeep Mathur IEEE/International conference of Electronics, Communication and Aerospace Technology

Data Analysis by Web Scraping using Python <https://ieeexplore.ieee.org/document/8822022>

Stephanie Lunn; Jia Zhu; Monique Ross IEEE/Frontiers in Education (FIE) Conference

Utilizing Web Scraping and Natural Language Processing to Better Inform Pedagogical Practice

<https://ieeexplore.ieee.org/document/9274270>

Riya Shah, Karishma Pathan, Anand Masurkar, Shweta Rewatkar, Prof. (Ms.) P.N.Vengurlekar

International Journal of Scientific and Research Publications, Volume 6, Issue 5, May 2016 640

ISSN 2250-3153

Comparison of E-commerce Products using web mining

<http://www.ijsrp.org/research-paper-0516/ijsrp-p53102.pdf>