WEB BASED COLLEGE PLACEMENT ANALYSIS

Akhilesh Shinde[1], Aniket Shinde[2], Vivek Singh[3], Siddhesh Shivdikar[4], Sharvari Govilkar[5]

[1,2,3,4]Student, Dept. of Computer Engineering, Pillai College of Engineering, Maharashtra, India
[5]Head of Department, Dept. of Computer Engineering, Pillai College of Engineering, Maharashtra, India

---

Abstract - *According to current analysis the number of engineering pass outs from Mumbai and Pune universities are approximated to be 2 lakhs which raises problems regarding placement of the pass out students even after considering the students which apply for further higher education. Our project on web-based placement analysis provides base guidelines for placement officers in Mumbai and Pune universities. Our project provides the facility to different colleges to retrieve strategic information for better placement for their students. It reduces manual work and consumes less paperwork to reduce the time. Thus, this paper aims to scrap the data on the web as well as other primary sources and process and carry out the analysis for the year 2019.*

Key Words: **Web Scraping, Web Crawling, Custom search engine, NLP Natural Language Processing, Data pre-processing, Data visualization, Queries on data**.

1.INTRODUCTION

The project deals with extraction of the list of companies coming for placement of all the engineering colleges in Mumbai and Pune University and build up a company profile.

Companies choose a specific college where the placement will be held, there is a need to maintain all the paperwork, causing a large amount of space. If it is manually done, chances of missing, difficulty to handle the details of company data increases which in turn becomes tedious to analyze the placement data for future recruiting pool campus.

The purpose of the project "Web based placement analysis" the manual work makes the process slow and other problems such as inconsistency and ambiguity on operations. In order to avoid this, the web-based placement analysis system is proposed, where the recruiting company information in the colleges affiliated to Mumbai University and Pune University with regard to placement is analyzed efficiently. It intends to help in the analysis of the companies visiting various colleges.

Our objective considers the following proposing a system which is fully computerized, which removes all the drawbacks of the existing system. The proposed system is accessed throughout the colleges to retrieve recruiting company data in various colleges. Understanding the concepts of web scraping of various websites of colleges affiliated to Mumbai University and Pune University. Identification of evaluation metrics used for placement analysis of different companies building up a linking bridge between recruiting companies and various colleges by providing data that benefited each other.

The system is beneficial for the placement department of various colleges to analyze the company placement details and approach and accordingly plan placement drives for their colleges.

In the same manner, it helps the companies to analyze the statistics of placements of various colleges which in turn helps them for visiting the college in the future for recruitment.

The system scrapes the data from various websites of colleges and stores it in table or spreadsheet form. The Data cleaning process is performed in order to remove unwanted data. To analyze the data, various queries are fired to access the data and data visualization is done. In data visualization, analyzed data is represented in the form of bar plots, histograms, etc.

2. LITERATURE REVIEW

In the research publication Data Analysis by Web Scraping using Python by David Mathew Thomas, Sandeep Mathur, provides us with a detailed explanation of the scraping process along with data analysis strategies. The framework of the scraping process is one of the highlights which makes the process of web extraction easier to understand. Moreover all the crucial implementation strategies and methodologies are well explained with feasibility and applications.[4]

3. PROPOSED ARCHITECTURE

3.1 Implemented System.

The implemented systems "WEB BASED PLACEMENT ANALYSIS" consist of many steps regarding placement data which involves gathering data through web scraping from official college websites using python package libraries and then moving the pure placement data set to python Pandas data frames. The data frames are feasible and elegant to use the data set so that they can be easily cleaned as per our requirement and can present relevant data through graphical representation.

The challenges regarding scraped placement data from the official websites of the colleges is that the data available where not in the same format as some were html tables, pdfs and images respectively, those were handled using different packages of python.

The scraped data was brought into one proper format using data preprocessing and data cleaning techniques.

The proper processed data sets are converted to graphical representation using python libraries for extraction of strategic information to build company profile.

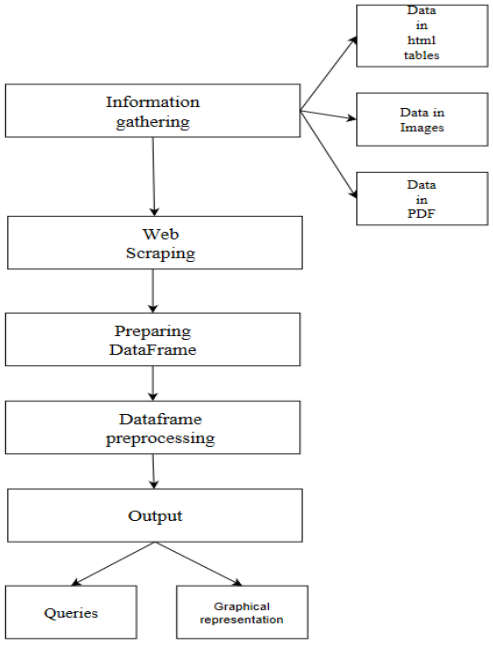

Fig -1: Workflow Diagram

3.2 TYPES OF INPUT

3.2.1 Tables in HTML

There are three types of tables in HTML which are table with span tag, table with div with heading tag, tables in pdf.

In order to extract data from table with span tag first
1. Import BeautifulSoup, requests and pandas packages in python.
2. Use the requests to get the target website URL.
3. Parse the source code.
4. The required information is stored in the table.
5. Find the table tag using the following syntax.
6. Navigate to the row of the table.
7. To extract all the data present in a row of specific column names use a for loop and save the text value. Replace '\n' and '\r' using the replace function if present.
8. Append the list to the data frame.
9. Convert to final CSV.

In order to extract data from table with div with heading tag
1. Import the packages BeautifulSoup, requests and pandas.
2. Use the requests to get the target website url.
3. Parse the source code.
4. The data to be extracted is stored in <head> tag of div.

3.2.2 Tables in PDF

In order to extract data from table in pdf
1. Import packages Camelot, pandas, pypdf2.
2. Insert the path of the pdf file.
3. Finding all the pdf files in the given path and storing them in a list.
4. It searches the above block of code using the OS package to search files that end with .pdf extension in the current directory.
5. And stores all the file names into the list which is used in automating the extraction process of tables.
6. Automation is done by looping through all the files extracted in a list from the directory.
7. Each file is then opened individually in pypdf2 and Camelot where pydf2 is used to get the number of pages in pdf.
8. Then the tables in all the pages are extracted using Camelot.
9. If the page has more than one table loop is used to fetch both the tables.
10. Then the extracted tables in pdf will be stored in csv and excel format in the directory where the program is stored.

3.3 Data Preprocessing and Visualization

Step 1:
After the phase information gathering the data which is present is preprocessed to remove irrelevant characters and symbols present in raw data. The preprocessing can be done by the following instructions.

```
df1=pd.read_csv("pce_companies.csv", index_col=[0])
```

Fig -2: Read CSV

```
df1.rename(columns={'Name of Company':'name of company'}, inplace=True)
```

Fig -3: Renaming Columns

```
temp= df1.append(df2, ignore_index=True)
```

```
In [87]:
temp
```

```
Out[87]:
```

	name of company	college name	region
0	zeus	Pillai College of Engineering	Navi Mumbai
1	Aditya Birla	Pillai College of Engineering	Navi Mumbai
2	capgemini	Pillai College of Engineering	Navi Mumbai
3	CMClimited	Pillai College of Engineering	Navi Mumbai
4	emco	Pillai College of Engineering	Navi Mumbai

Fig -4: Merging Data Frames

```
df4['name of company'].replace('', np.nan, inplace=True)
```

Fig -5: Replace Empty Space with NaN

```
df4.dropna(subset=['name of company'], inplace=True)
```

Fig -5: Remove NaN

Step 2:
After the phase of data preprocessing, the data visualization phase starts. This phase consists of firing queries on data sets to extract strategic information in form of graphical representation.

	name of company	college name	region	year
0	zeus	Pillai College of Engineering	Navi Mumbai	2018
1	Aditya Birla	Pillai College of Engineering	Navi Mumbai	2018
2	capgemini	Pillai College of Engineering	Navi Mumbai	2018
3	CMClimited	Pillai College of Engineering	Navi Mumbai	2018
4	emco	Pillai College of Engineering	Navi Mumbai	2018

Fig -5: Processed Data

Query 1: L&T Infotech visited colleges.

```
In [13]: df[df['name of company']=='L&T Infotech']['college name']

Out[13]: 314    Pillai HOC College of Engineering and Technology
         479    Pillai HOC College of Engineering and Technology
         826                 Ramrao Adik Institute of Technology
         Name: college name, dtype: object
```

Fig -5: Query 1 Result

Query 2: To count the number of occurrences of company in the dataset.

```
In [23]: df['name of company'].value_counts().head(5)

Out[23]: Name of Company    6
         Reliance Jio       5
         GEP                4
         Bitwise            4
         Zycus              4
         Name: name of company, dtype: int64
```

Fig -5: Query 2 Result

Query 3: The visualization of no. of companies visited in Mumbai, Thane, Pune and Navi Mumbai region.

```
: sns.barplot(x='region',y='No. of companies',data=df2)

: <matplotlib.axes._subplots.AxesSubplot at 0x249648f2e48>
```

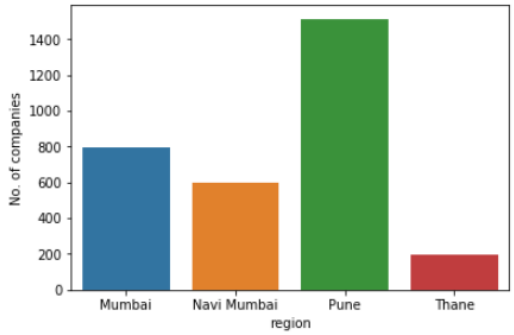

Fig -5: Query 3 Result

4. CONCLUSIONS

In this proposed system, the study of different Web Scraping techniques is presented. The preprocessing of data and data visualization is done. The main goal of this project was to explain how to use web scraping techniques to gather data from the web and represent it in a meaningful way. Queries were performed in order to fetch the data which was gathered to get possible outcomes. The comparative study of various techniques is presented in this report. This project can be used in colleges for their placement activities. With the help of this project, colleges

can rectify the problems because of which companies are not coming for placement and use for the betterment of the college placement. The overall outcome of this project is helping colleges and companies for placements activities by building company profiles.

From all the CSV data we have obtained, we generated 7 variations of the dataset which could be used for upcoming research topics related to placements offered by the universities. The Dataset can be further extended by adding current and next year recurring placement data.

We have published a dataset on Kaggle by name: College placement dataset. Link for dataset will be in present references.

ACKNOWLEDGEMENT

We would like to express a wealth of gratitude to Our guide Dr. **Sharvari Govilkar**. Her work ethic and dedication are unparalleled. We are grateful for her patience and guidance throughout our Project. Without her generous help, we could not finish our project. We are extremely indebted to her for accepting us as a student under her guidance.

Our warm appreciation is due to our **Dr. Sharvari Govilkar** Head of Department of Computers for providing all the necessary facilities to us. We feel highly indebted for giving permission to conduct the seminars related to our project so that we can clarify all our doubts.

We are very much indebted and very much grateful to our inspiration the principal of Pillai College of engineering I for his unflinching support, perseverant help, cherishing attitude, benign, gracious character, constant encouragement, valuable advice, ever willing and constant guidance.

We are immensely grateful to all of them for sharing their pearls of wisdom with us during this course of project-based learning.

REFERENCES

[1] Eleanor Clark, Kenji Araki, 2011, Text Normalization in Social Media: Progress,Problems and Applications for a Pre-Processing System of Casual English

[2] Thomas Gottron, Nedim Lipka2, 2011, A Comparison of Language Identification Approaches on Short, Query-Style Texts

[3] Siddhesh Shivdikar, "College Placement dataset", [Online],2021. Available: https://www.kaggle.com/siddheshshivdikar/college-placement/metadata

[4] David Mathews, Sandeep Mathur, "Data Analysis by web scraping using python", [online document], 2019. Available:https://www.researchgate.net/publication/335576922_Data_Analysis_by_Web_Scraping_using_Python

BIOGRAPHIES

"Siddhesh Shivdikar is Computer Engineering graduate from Pillai College of Engineering, Mumbai University. He has major interest in Full Stack Development. Machine Learning- AI and Data Analysis"

"Vivek Singh in 2021, is Associate Software Engineer at GEP Worldwide, Former Computer Engineering student at Pillai College of Engineering, Mumbai University. His area of interests are software development in WEB, Machine Learning and Data Science."

"Akhilesh Shinde, currently in year 2021 is Analyst at Capgemini, Computer Engineering pass out in 2020 from Pillai College of Engineering, Mumbai University. His area of interest are Data analysis and Machine Learning "

"Aniket Shinde, currently in the year 2021 is a System Engineer at TCS, Computer Engineering passout in 2020 from Pillai College of Engineering, Mumbai University. His area of interest Linux, DBMS and Machine Learning "

"Dr. Sharvari Govilkar is Head of Department of Computer Engineering in Pillai College of Engineering, New Panvel, Mumbai University.
More info:
https://www.pce.ac.in/faculty/faculty-directory/dr-sharvari-s-govilkar/ "