IRE Major Project Mid Document

Team 19:

1. Pratyanshu Pandey (2019101025)
2. Pavan Baswani (2021701035)
3. Prince Singh Tomar (2019101021)

INDEX

Problem Statement:

Collecting domain specific data given a seed set of structured and unstructured sources

Given a domain, the task is to build a structured, clean and high quality dataset for the domain. This will involve searching for structured and unstructured sources for the domain, scraping the web pages to get the raw data, extracting tabular information from the raw data, processing and formatting this tabular information to give a structured, clean and high quality json output.

Data Structure:

The data we aim to collect is tabular in nature with **140,280** rows, each row containing data for a company.
Each company has **32 major attributes** that are represented by columns. Some of these attributes are subdivided into sub-attributes each with its own column. In this manner the total number of **columns is 58**.
A **cell** in this table generally holds a single value but it can also be **multivalued, textual or empty.**

The following table lists all the column names, sub columns if they exist, sources from which data for these columns are extracted and the current status of our work for that column.

S.No.	Column Name	Sub Columns	Sources
1	CIN/LLPIN/IDS	-	zaubacorp.com mca.gov.in seed data
2	Company Name		zaubacorp.com mca.gov.in seed data
3	Date of Registration		zaubacorp.com mca.gov.in seed data
4	Registration Number		zaubacorp.com mca.gov.in
5	Age of Company		zaubacorp.com
6	Company Status		zaubacorp.com mca.gov.in seed data
7	Company Class		zaubacorp.com mca.gov.in seed data
8	Company Category		zaubacorp.com mca.gov.in seed data
9	Company Subcategory		zaubacorp.com mca.gov.in seed data
10	Listing Status		zaubacorp.com mca.gov.in
11	Number of Members		zaubacorp.com

			mca.gov.in
12	Previous Names		zaubacorp.com
13	Previous CINs		zaubacorp.com
14	Number of Employees		zaubacorp.com
15	Authorized Capital		zaubacorp.com mca.gov.in seed data
16	Paid Up Capital		zaubacorp.com mca.gov.in seed data
17	Registered State		zaubacorp.com mca.gov.in seed data
18	RoC		zaubacorp.com mca.gov.in seed data
19	Principal Business Activity		zaubacorp.com mca.gov.in
20	Address		zaubacorp.com mca.gov.in seed data
21	Email		zaubacorp.com mca.gov.in
22	Website		zaubacorp.in
23	Date of Last Annual General Meeting		zaubacorp.in
24	Date of Latest Balance Sheet		zaubacorp.in
25	Important People	DIN	zaubacorp.com mca.gov.in
		Name	
		Appointment Date	
		Cessation Date	
		Designation	
26	Prosecution Details	Defaulting Entities	zaubacorp.com
		Court Name	

		Prosecution Section	
		Date of Order	
		Status	
27	Charges/Borrowing Details	Charge ID	zaubacorp.com mca.gov.in
		Creation Date	
		Modification Date	
		Closure Date	
		Assets under Charge	
		Amount	
		Charge Holder	
28	Trademarks	Name/ID	zaubacorp.com
		Class	
		Application Date	
		Status	
		Goods and Services Description	
		Applicant Address	
29	Trading Details	BSE	nseindia.com bseindia.com
		NSE	
		Revenue	
		Market Capitalization	
30	Establishment Details	Name	zaubacorp.com
		City	
		PinCode	
		Address	
		Coordinates/Link	
31	Area Served		wikipedia.org
32	Subsidiaries List		wikipedia.org

Data Sources and Methodologies:

General Approach:

1. For every source that we decided to work with, the general approach is to look at the source code of the web page and what data can be accessed in what manner.
2. Some data is freely available while others are behind a cache or a login mechanism. Some data is in tabular form while others are in text inside html tags.
3. We used selenium and BeautifulSoup to overcome these walls to get the data.
4. We write the code making sure that it can handle exceptions well and generates an output log for manual analysis of errors.
5. The log file is then checked manually after execution to find any errors. These errors have been few in number and have been handled manually.

zaubacorp.com :

Creating exhaustive list of Companies

1. From the list of all companies registered in Telangana we extracted the company name, CIN, status and link to the full page for the company.
2. We performed uniqueness and integrity checks for the list of companies and removed the duplicates.
3. This list now had **140,280** which is close to double of the seed data available to us.
4. We made sure that all the companies present in the seed data were present in the new list as well. This list is now the exhaustive list of all companies.

Scrape Company data from zaubacorp.com

1. For accessing the full company page on zaubacorp we used "*selenium*". It helped in the automation of the process and made it easier. We first logged into the site by sending username and password to the required fields of the website.
2. The problem arose when we had to solve the captcha for that we accessed the region where the mathematical equations were presented and solved it using the eval function in python.
3. After logging in we used the driver.get() function to access the required webpage. Since the data on zauba corp is organised in tabular fashion, which helped in automating the process.
4. We extracted the required tables and did some processing on it. We cleaned the data to remove unnecessary noise, For example, to remove advertisements.

5. After all cleaning the data was stored in a dictionary and later the dictionary was exported in json format.

Scrap Trademarks Data

1. From the list of companies we generated the url to get the trademarks registered by the companies.
2. Since login was not necessary to access this part of the website, we simply used requests.get() to get the complete web page.
3. The data was not in a table but rather within tags so we used the BeautifulSoup library to extract it.
4. Integrity and uniqueness checks were performed on the data.
5. This data was stored in a dictionary and exported in json format.

mca.gov.in :

1. The website mca.gov.in requires the login with cin and captcha present in the image to access the company data in tabular form.
2. Initially, the screenshot of the whole page is saved and cropped the image where the captcha is located using the python package **PIL**.
3. With the help of the "*pytesseract*" package, the text present in the cropped image is extracted.
4. Using the selenium web driver, the fields are located with their ids and sent the keys (cin and captcha text) to submit the form.
5. Once the form is submitted, the form will be evaluated with the login details and redirected to the company details page. The exceptions (invalid cin or captcha) are handled to run the code without any interrupt/crash.
6. Finally, all the company details are downloaded in excel format with the selenium driver with button click.

wikipedia.org :

1. Extracting data from the wikipedia pages of the Companies will be done in the 2nd half.
2. For extracting the data we planned on using "*Wikipedia's API*", which itself is a great API for scraping wikipedia.
3. If that fails we will use selenium to access the desired wikipedia webpage and extract the data from the page.
4. We will use the data for getting datas like Area served, Subsidiaries etc.

Google search :

1. A lot of smaller companies have their own websites but details about the website are only easily available for bigger companies in Wikipedia or zaubacorm.com.

2. A simple google search with the company name as a query however reveals the website and other basic information compiled from various sources.
3. We will perform a google search for each company and extract data from the blocks created by google as well as look at the top 10 links and determine if any of them is a link to the official company website (by extracting the keywords/metadata of the websites present in top 10).

nseindia.com/bseindia.com:

1. The list of all companies on NSE or BSE is available in excel format on their website.
2. The excel sheet contains the company name as the primary key.
3. We have downloaded excel and during the data processing phase, we will extract relevant information for the relevant company from the excel sheets using the match of company names.

Findings and Data Extracted:

We have extracted the following data and it is available in the data folder of the below mentioned github repository.

1. Zaubacorp.com
    a. Creating Exhaustive List of Companies: Complete
    b. Scrap Company Data: In Progress - Code is deployed and the scraping is 30% complete
    c. Scrap Trademarks Data: Complete
2. Mca.gov.in - In Progress - Code is ready and deployed and the amount of excel data extracted is 40%
3. wikipedia.org - analysis and coding started
4. Google search - analysis and coding started
5. nseindia.com/bseindia.com - Raw data downloaded

Code and Data Link:

The code is present in the src folder and the extracted data is present in the data folder.

Github Link: https://github.com/IRE-Project/Data-Collector

Difference and Progress as compared to Scope Document:

Difference:

1. Instead of collecting all sorts of raw data and then analyzing the columns we switched these steps and analyzed the sources to get a list of columns. We then started with collection of data.
2. Instead of keeping data collection and processing as 2 completely independent tasks we have introduced some amount of pre-processing during the data collection phase itself. These include uniqueness check, integrity check etc.
3. Instead of just settling with tabular data we are also extracting some data that is present within html tags and not tables.
4. For the domain that we are working with we no longer feel the necessity of using named entity recognition to extract data from unstructured sources, so we will not be using this methodology.

Progress:

Given the changes in the methodologies and the amount of data and the domain we are working with, we feel we have made good progress and are in line to finish the project within the deadline.

END