IndicWiki Summer Internship

Indian Companies



Domain Team

Data Collection Sources/ Sites

Tools used for Data collection

Data Cleaning

Sample Article

Jinja template creation

Translation/ Transliteration

XML Generation

Final data

# Domain

The domain we worked on was **“Indian Companies”**, the aim of the project being generating comprehensive articles for Telugu Wikipedia on 60000+ Indian Companies, comprising all possible details on a particular company, such as its industry, revenue and a lot more, which we have been successfully able to do.

# Team

|  |  |
| --- | --- |
| **Member** | **Email** |
| Leela Vardhan Siramdasu | vardhanpslv@gmail.com |
| Lavanya Gajulapalli | lavanyagajulapalli123@gmail.com |
| Bhargav Sai Chakravarti Tammineedi | bhargavtammineedi249@gmail.com |

Data Collection

## Sources/ Sites

Owing to the absence of a unified knowledge base on companies, data collected from several sources had to be conglomerated, to have a significantly large database, ranging over 60000 companies.

* Zauba Corp
  + Link: <https://www.zaubacorp.com/>
  + Format of data available - Web Pages (HTML)
  + Tools used - Selenium (Python Libraries)
  + Attributes found
    - CIN (Primary Key of a company)
    - Name, Status, Registration Number, RoC, Category
    - Data of Incorporation, Age of Company, Activity
    - Authorized capital, Paid up capital, Directors
* Wikipedia
  + Link: <https://www.wikipedia.com/>
  + Format of data available - Web Pages (HTML)
  + Tools used - Selenium, Beautiful Soup (Python Libraries)
  + Attributes found
    - Industry, Sector, Headquarters, Founded Year
    - Services, Revenue, Net income, Products

## Tools used for Data collection

### Selenium

* + An automation testing-based Python library, primarily used for web scraping
  + **Issues:** No significant issues were encountered while working with this, except for the time delay issue - sometimes, web pages would not load elements on time and that would result in timeouts, exceptions, and end of execution of the program. To avoid this, an additional method was added to keep Selenium web driver’s execution on hold, until some element would appear on the webpage.

### Beautiful Soup

* + This was used as an additive to Selenium to navigate through HTML elements and obtain text/information from them, hence, **no major issues** were observed.

Data Cleaning

* Data cleaning has been done on the data collected and removed the attributes which are not necessary.
* Remove the rows with less number of attributes.
* Cleaned the duplicate rows and corrected them.
* After cleaning, we created structured data.

Translation and Transliteration

* After cleaning the data , translation and transliteration is done.
* To translate the data google translator present in deep translit is used.
* Look at [this](https://pypi.org/project/deeptranslit/) document on how to use deep translit.

Sample Article

* Generated sample articles for some companies
  + Bse Nse companies sample articles : [sample1](https://tewiki.iiit.ac.in/index.php?title=%E0%B0%85%E0%B0%A6%E0%B0%BE%E0%B0%A8%E0%B1%80_%E0%B0%8E%E0%B0%82%E0%B0%9F%E0%B0%B0%E0%B1%8D%E2%80%8C%E0%B0%AA%E0%B1%8D%E0%B0%B0%E0%B1%88%E0%B0%9C%E0%B1%86%E0%B0%B8%E0%B1%8D_%E0%B0%B2%E0%B0%BF%E0%B0%AE%E0%B0%BF%E0%B0%9F%E0%B1%86%E0%B0%A1%E0%B1%8D),[sample2](https://tewiki.iiit.ac.in/index.php?title=%E0%B0%87%E0%B0%A8%E0%B1%8D%E0%B0%AB%E0%B1%8B%E0%B0%B8%E0%B0%BF%E0%B0%B8%E0%B1%8D_%E0%B0%B2%E0%B0%BF%E0%B0%AE%E0%B0%BF%E0%B0%9F%E0%B1%86%E0%B0%A1%E0%B1%8D),[sample3](https://tewiki.iiit.ac.in/index.php?title=%E0%B0%AD%E0%B0%BE%E0%B0%B0%E0%B0%A4%E0%B0%BF_%E0%B0%8E%E0%B0%AF%E0%B0%BF%E0%B0%B0%E0%B1%8D%E2%80%8C%E0%B0%9F%E0%B1%86%E0%B0%B2%E0%B1%8D_%E0%B0%B2%E0%B0%BF%E0%B0%AE%E0%B0%BF%E0%B0%9F%E0%B1%86%E0%B0%A1%E0%B1%8D),[sample4](https://tewiki.iiit.ac.in/index.php?title=%E0%B0%B9%E0%B0%BF%E0%B0%82%E0%B0%A6%E0%B1%81%E0%B0%B8%E0%B1%8D%E0%B0%A5%E0%B0%BE%E0%B0%A8%E0%B1%8D_%E0%B0%AF%E0%B1%82%E0%B0%A8%E0%B0%BF%E0%B0%B2%E0%B1%80%E0%B0%B5%E0%B0%B0%E0%B1%8D_%E0%B0%B2%E0%B0%BF%E0%B0%AE%E0%B0%BF%E0%B0%9F%E0%B1%86%E0%B0%A1%E0%B1%8D)
  + Psu companies sample articles : [sample\_article](https://tewiki.iiit.ac.in/index.php?title=%E0%B0%B5%E0%B0%BE%E0%B0%A1%E0%B1%81%E0%B0%95%E0%B0%B0%E0%B0%BF:Vardhan/%E0%B0%AA%E0%B1%8D%E0%B0%B0%E0%B0%AF%E0%B1%8B%E0%B0%97%E0%B0%B6%E0%B0%BE%E0%B0%B2/%E0%B0%AD%E0%B0%BE%E0%B0%B0%E0%B0%A4%E0%B1%8D_%E0%B0%B9%E0%B1%86%E0%B0%B5%E0%B1%80_%E0%B0%8E%E0%B0%B2%E0%B0%95%E0%B1%8D%E0%B0%9F%E0%B1%8D%E0%B0%B0%E0%B0%BF%E0%B0%95%E0%B0%B2%E0%B1%8D%E0%B0%B8%E0%B1%8D_%E0%B0%B2%E0%B0%BF%E0%B0%AE%E0%B0%BF%E0%B0%9F%E0%B1%86%E0%B0%A1%E0%B1%8D)
  + Notable wiki Companies sample articles : [sample article](https://tewiki.iiit.ac.in/index.php?title=%E0%B0%B5%E0%B0%BE%E0%B0%A1%E0%B1%81%E0%B0%95%E0%B0%B0%E0%B0%BF:Bhargav/%E0%B0%AA%E0%B1%8D%E0%B0%B0%E0%B0%AF%E0%B1%8B%E0%B0%97%E0%B0%B6%E0%B0%BE%E0%B0%B2)
  + Remaining 57k Companies sample articles : [sample1](https://tewiki.iiit.ac.in/index.php?title=%E0%B0%B5%E0%B0%BE%E0%B0%A1%E0%B1%81%E0%B0%95%E0%B0%B0%E0%B0%BF:Lavanya/%E0%B0%AA%E0%B1%8D%E0%B0%B0%E0%B0%AF%E0%B1%8B%E0%B0%97%E0%B0%B6%E0%B0%BE%E0%B0%B2/%E0%B0%9C%E0%B1%8D%E0%B0%AF%E0%B1%81%E0%B0%B5%E0%B0%BE%E0%B0%B0%E0%B1%8D%E0%B0%97%E0%B0%BF_%E0%B0%AB%E0%B0%BE%E0%B0%B0%E0%B1%8D%E0%B0%AE%E0%B0%B0%E0%B1%8D_%E0%B0%AA%E0%B1%8D%E0%B0%B0%E0%B1%8A%E0%B0%A1%E0%B1%8D%E0%B0%AF%E0%B1%82%E0%B0%B8%E0%B0%B0%E0%B1%8D_%E0%B0%95%E0%B0%82%E0%B0%AA%E0%B1%86%E0%B0%A8%E0%B1%80_%E0%B0%B2%E0%B0%BF%E0%B0%AE%E0%B0%BF%E0%B0%9F%E0%B1%86%E0%B0%A1%E0%B1%8D)

Jinja Template Generation

* Created Jinja template for the collected data using the attributes
* For four different companies four jinja templates were created
  + Jinja template for nse bse : [nse bse jinja template](https://github.com/indicwiki-iiit/Indian_Companies/blob/main/indian%20companies%20bse%20nse/templates/company.j2)
  + Jinja template for Psu : [PSU jinja template](https://github.com/indicwiki-iiit/Indian_Companies/blob/main/templates/PSU_companies.j2)
  + Jinja template for Notable wiki companies : [wiki companies jinja template](https://github.com/Bhargav-git-hub/Indicwiki-internship/blob/main/Wiki%20companies/Templates/introduction.j2)
  + Jinja template for 57k companies : [other indian companies jinja template](https://github.com/indicwiki-iiit/Indian_Companies/blob/main/templates/remaining%20Indian%20companies.j2)

XML generation

* Generated XML for all the companies using jinja template.
* Code for generating XMLfile : [Link for code](https://github.com/indicwiki-iiit/Indian_Companies/blob/main/indian%20companies%20bse%20nse/XMLgen.ipynb)
* Xml file for 100 companies : [XML file](https://github.com/indicwiki-iiit/Indian_Companies/blob/main/indian%20companies%20bse%20nse/xml%20files/sample100_company.xml)
* Page id of BSE NSE companies present in wikipedia : 716000 - 716450
* Page id of BSE NSE companies not present in wikipedia : 716500 -720092
* Page id of Psu : 721000-721356
* Page id of Notable companies : 722000-722104
* Page id of 57k : 722105 - 779610

Final Data:

* NSE BSE Companies : [wiki bse nse](https://github.com/indicwiki-iiit/Indian_Companies/blob/main/indian%20companies%20bse%20nse/data/final_wiki_71.xlsx), [Ramaining nse bse](https://github.com/indicwiki-iiit/Indian_Companies/blob/main/indian%20companies%20bse%20nse/data/nonwiki_8.xlsx)
* PSU companies : [PSU data (eng)](https://github.com/indicwiki-iiit/Indian_Companies/blob/main/data/PSU_eng.csv) , [PSU data (tel-eng)](https://github.com/indicwiki-iiit/Indian_Companies/blob/main/data/PSU_tel_eng.csv)
* Notable wiki companies : [Wiki data](https://github.com/Bhargav-git-hub/Indicwiki-internship/blob/main/Wiki%20companies/Data/masterwiki.xlsx)
* 57k Companies: [other companies (eng)](https://github.com/indicwiki-iiit/Indian_Companies/blob/main/data/Indian_firms_eng.csv), [other companies (tel-eng)](https://github.com/indicwiki-iiit/Indian_Companies/blob/main/data/indian_firms_tel_eng.csv)