IndicWiki Summer Internship

Airports



[Domain](#_Domain) [Team](#_Team)

Data Collection [Sources/ Sites](#_Sources/_Sites)

[Tools used for Data collection](#_Tools_used_for)

Data Cleaning

Sample Article

Jinja template creation

Translation/ Transliteration

XML Generation

Final data

# Domain

The domain we worked on was **“Airports”**, the aim of the project being generating comprehensive articles for Telugu Wikipedia on 20000+ airports, comprising all possible details on an airport

# Team

|  |  |
| --- | --- |
| **Member** | **Email** |
| Srihitha Reddy T | srihitha.reddy23@gmail.com |
| P Nikhil Chandra | nikhilchandrap24@gmail.com |
| Maddi Varun | maddivarun9@gmail.com |

Data Collection

## Sources/ Sites

We searched for data in various websites, and we found a website which is solely dedicated in providing airport and flight details to its viewers – OurAirports. Apart from this website, we scraped data from Wikipedia info boxes as well.

* OurAirports
  + Link: <https://ourairports.com/help/data-dictionary.html>
  + Format of data available – csv files
  + Tools used - Python Libraries
  + Attributes found
    - ID, Ident (Primary Key of an airport)
    - Name, Country, Region, IATA, and ICAO codes,
    - Municipality, Type of airport, scheduled services, home link
    - Airport\_frequency, runway length and width, surface of the runway and others
* Wikipedia
  + Link: <https://www.wikipedia.com/>
  + Format of data available - Web Pages (HTML)
  + Tools used - Beautiful Soup (Python Libraries)
  + Attributes found
    - Sector of the airport (type), owner, operator
    - Serves, land area, passengers, aircraft movements

## Tools used for Data collection

### Beautiful Soup

* + This was used to navigate through HTML elements and obtain text/information from them, hence, **no major issues** were observed.
* **Python pandas**
  + A famous python library which is used to efficiently handle datasets present in csv or xlsx formats

Data Cleaning

* Data cleaning has been done on the data collected and removed the attributes which are not necessary.
* Remove the rows with a smaller 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 is used. We translated each column by uploading the file to the Google’s translator website - <https://translate.google.co.in/?sl=auto&tl=te&op=translate>
* Apart from it, Bing Translator and Yandex are also used

Sample Article

* Generated sample articles for some airports
  + Sample article for a wiki\_link 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:Srihitha/%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

Jinja Template Generation

* Created Jinja template for the collected data using the attributes
* The jinja template for our articles is here.

XML generation

We have written python file to generate xml from template and generate 1 sample xml and then 50 xml articles and finally all the xml articles.

Final Data:

Our final data set consists of both translated and English datasets

* English data set – Mastersheet.xlsx
* Final Dataset – Final\_translated.xlsx