**Internship Report**

Submitted in Partial Fulfillment of the requirements for the final year practicum of **MASTER OF SCIENCE IN INFORMATION TECHNOLOGY**

****

By

Chada Pratyusha Mayura

Roll.No : 2020501076

**Organization**

Indicwiki, IIIT Hyderabad



**ACKNOWLEDGEMENT**

I express my deep gratitude to ***Mr. Anandaswarup Vadapalli, Registrar - MSIT Program and AI & ML specialization faculty*** for permitting me to do the internship with the organization.

I am highly indebted to ***Dr. Vasudeva Varma - Project Head and Chief Investigator, Dr. Radhika Mamidi - Co-Project Investigator, and Mr. Praveen Garimella - Co-Project Investigator*** for allowing me to do the internship in the organization.

I also thank ***Mrs. PSL Prasanna - Project Manager and Mr. Krupal Kasyap - Project Consultant*** for guiding me throughout my internship period and helping me to learn and giving me the freedom to explore.

I am extremely grateful to the colleagues who helped me in learning the methodology and also technology which helped my successful completion of this internship.

**Chada Pratyusha Mayura**

**2020501076**

**TABLE OF CONTENTS**

Goal and Purpose

Basic Pipeline of Project

Domains Information

Data Collection

Sourcing

Scrapping

Data Storing

Data visualization

Template creation

Translation/Transliteration

XML Generation

Github

Description of Internship Experience

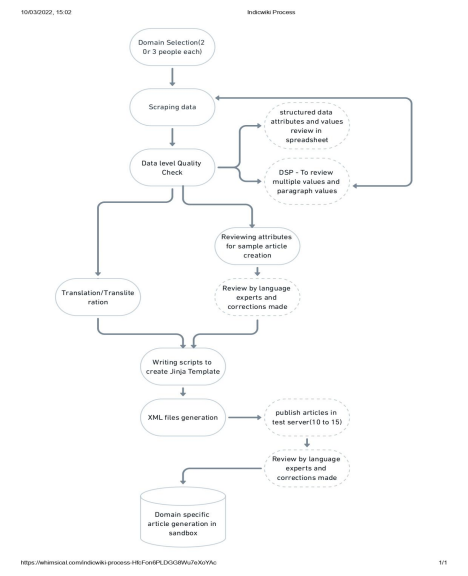
**Goal & Purpose:**

Wikipedia is the most widely used resource of the Encyclopedic Knowledge, Education, and E-Literacy platform on the Web. It is a free, online encyclopedia with over 7 million articles only in English. India is in second place with 117 million views per day behind the USA with 189 million views as of March 2022.

Most of the people in India who can read in only their native language or Hindi are not able to benefit from this fantastic tool for access to knowledge and learning. International Institute of Information Technology(IIIT), Hyderabad has started the Project IndicWiki to enhance the content in Indian Language Wikipedias starting with Telugu and Hindi language Wikipedia.

This project aims to generate approximately 1M wiki articles in Telugu language given a particular domain of interest. Every wiki article generated should meet a minimum word count of 500 words and also have references, infobox, categories, and images.

**Basic Pipeline of Project**

****

**Domain Information**

We have worked on 2 domains.

**RELIGIOUS PLACES**

The domain we worked on was *Religious Places* across the world, which involved the generation of 20K+ Wikipedia articles in Regional Language (in our case it was Telugu).

**Subdomains**

Our domain includes many subdomains that we had to work on individually. The subdomains are:

1. Mosques

2. Churches

3. Temples

4. Religious Buildings

5. Gurudwaras

6. Jain Temples

**Softwares :**

The outcome of this domain was to collect data and articulate the information about various antiviruses and softwares for wikipedia. There are numerous softwares which are currently in use, although there was no particular website which describes the information in a native language (i.e Telugu). Therefore we took up the task to generate the essay’s for about 11500 softwares and nearly 100 antiviruses. Furthermore, we’ve used python at each and every stage of the process(i.e scrapping,cleaning,merging) apart from that we’ve used few libraries to translate and transliterate the data (i.e anuvad,deep translit).

**Data collection**

**Sourcing**

The next step is to search for reliable sources where we can collect the data. These websites should include information from government websites. The data collection should be done on the basis of attributes. These attributes should be able to generate a 500-word wikipedia article. For the tourist destinations domain, we have used the google travel website and wikipedia as a source to collect the data. For the programming languages domain, we have used hopl, wikidata, and wikipedia as a source to collect the data.

**Scrapping**

After sourcing, we need to work on scrapping the data. For scrapping the data they are many python libraries like beautiful soup, selenium, scrapy, tabula, etc. For the tourist destinations and programming languages domain, we have used beautiful soup to scrape the data. For the programming language domain, we have used wikipedia, wikidata, and SPARQL to scrape the data. For

scrapping of images for both the domains we have used wikidata and the Wikimedia Commons website.

**Data Storing**

We have used excel and google sheets to store the domain data.

**Data Visualization**

There is a tool called sweet viz through which we have visualized the data. It is an open-source python library that generates beautiful, high-density visualizations to kickstart EDA(Exploratory data analysis) with just two lines of code. Output is a fully self-contained HTML application.

**Template Creation**

Before the creation of the template, we need to write a sample article for the collected data. For writing a sample article one should have a thorough understanding of the attributes. We need to write sentences in Telugu. For writing the Religious Places & Softwares sample articles we have referred to wikipedia articles and taken some randomized sentences from there.

**Note:** Github link for the Softwares is - [link](https://github.com/indicwiki-iiit/Softwares-and-Antiviruses)

The next step after the creation of the sample article is to create a template. We have used the Jinja2 template engine for creating templates. For creating a jinja template we need to create macros that are used for organizing the template into sections according to the context. In order to follow the randomization of sentences, we have written multiple sentences for each attribute. We have included the Infobox section in the jinja template. This infobox section contains facts and statistics that include the information of the basic summary of the article.

**Translation/Transliteration**

In order to translate and transliterate the data first, we need to figure out which column should be translated and which column should be transliterated. For example, if the word is the same in both languages i.e., English, or Telugu, then it is transliterated. Hyderabad → హైదరాబాద్ is a transliteration and School → పాఠశాల is translation since school has its own Telugu word పాఠశాల and not స్కూల్.

For the tourist destinations domain, we have used google translate and bing translator to translate the data. For the programming languages domain, we have google translator for translating the data.

**XML Generation**

The final and important step of this project is XML generation. It is the process of generating an XML file containing article text. The article text is generated from the jinja template filled with attribute values from the data frame/excel sheet. Some requirements for the script to generate XML are Wikipedia username, userid, and pageid. Username and userid can be found on the user's Wikipedia account and pageid can be any unique number. In the script, the values in the data frame are passed to the jinja template and it is rendered to substitute those values in the template to form a valid text.

**Github**

All the scripts used for scraping, cleaning, translation, transliteration, Jinja template, and XML generation can be accessed on GitHub.

Religious Places GitHub link - [link](https://github.com/indicwiki-iiit/Religious-Places)

Softwares GitHub link - [link](https://github.com/indicwiki-iiit/Softwares-and-Antiviruses)

**Description Of Internship Experience**

This internship has enhanced my data science skills a lot. I enjoyed scraping the data from various sources. I also had hands-on experience in scraping data using selenium. In the translation process, I have worked with python libraries like deep translit and anuvaad. I also learned how to write a jinja template.

In this internship, I have also worked on creating a User manual by mentioning a clear process diagram on the overall process that will be done in creating articles for one particular domain ([manual](https://docs.google.com/document/d/1v6VDjpEcXQngobHZqCC8TMLK_61W3wfZ0wNBDF5nHkY/edit)). I have also elaborated and documented all the technicalities that were needed by a person who starts from scratch. Overall this internship has given me good exposure on both frontend and backend.