Internship Report

Submitted in Partial Fulfillment of the requirements for the final year practicum of

MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

By
Vamsi Krishna Ramisetti
Roll.No : 2020501089

Organization
Indicwiki, IIIT Hyderabad

ACKNOWLEDGEMENT

I express my deep gratitude to ***Dr. Arun Kumar Parayatham, Registrar - MSIT Program and Data science 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 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.

Vamsi Krishna Ramisetti
2020501089

TABLE OF CONTENTS

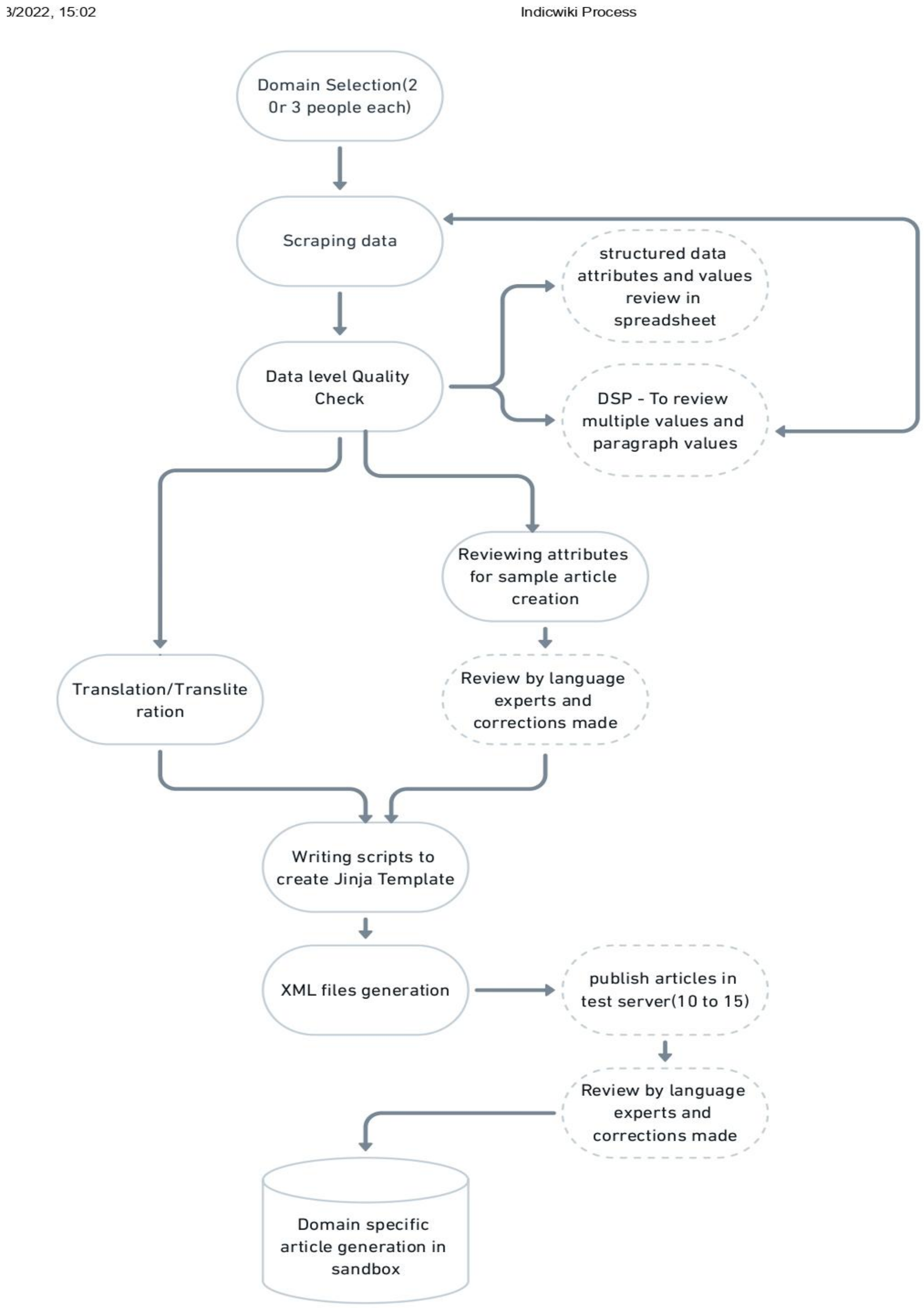
Goal & Purpose:

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


3/2022, 15:02
Indicwiki Process
Domain Selection(2 Or 3 people each)
Scraping data
structured data attributes and values review in spreadsheet
Data level Quality Check
DSP - To review multiple values and paragraph values
Reviewing attributes for sample article creation
Translation/Transliteration
Review by language experts and corrections made
Writing scripts to create Jinja Template
XML files generation
publish articles in test server(10 to 15)
Review by language experts and corrections made
Domain specific article generation in sandbox

Domain Information

We have worked on 2 domains. The first domain is Tourist destinations. In tourist destinations, we have collected data for about 20000 cities and we have collected the information on 17 attributes. The attributes of tourist destinations include name, city, latitude, longitude, city name, state name, country name, full address, continent name, destination type, visitors per year, destination rating, temperatures, best time to visit, how to reach, destination description, Images. We have collected the data from multiple sources.

The second domain we have worked on is Programming languages. In programming languages, we have collected data for about 8650 languages. The number of attributes for this domain is 14. The attributes of programming languages include name, author/developer, year, country, application_domain, category, description, current_version, type, company/organization, extension, paradigm, typing_discipline, influenced_by, platform, references.

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 sweetviz 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 tourist destinations and programming languages sample articles we have referred to wikipedia articles and took some randomized sentences from there.

Note: Github link for the tourist destinations sample article is - link

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.

Tourist destinations GitHub link - link
Programming languages GitHub link - link

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 scrapping 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 a Domain-Specific Platform(DSP). It is a platform that is used to display the domain data in a structured format. In DSP we have worked on frontend and backend. For the front end, we have used react js and developed an mcd using firebase. Firebase is a platform that allows users to authenticate and store their data. We have created a user module and admin module in firebase. The Admin module operations include storing domain data, authenticating users, and making users admins. The User can only log in and view the uploaded data. In the backend, we have python for creating user tables. For storing the data we have set up the Mongo DB database.

Overall this internship has given me good exposure on both frontend and backend.