Progress Report : Wikipedia Article Generation Domain: Indian Medicinal Plant

Suraj Kumar

July 2022

1 Introduction

All the steps outlined in the IndicWiki manual were followed to create the Hindi Wikipedia articles on Indian Medicinal Plants. As such, only some external details not covered in depth in the manual and some details that were specific to my domain have been outlined below.

2 Data Source & Collection Strategy

Data scrapping needs to be done only after consulting privacy policies and also understanding if the information can be used in Wikipedia articles (whether they are copyrighted or not). Web etiquette needs to be followed during scrapping, like using time.sleep

Total 3291 medicinal plants' data were scrapped from various websites. They are listed below:

- https://indiabiodiversity.org/group/medicinal_plants : To pull features like diagnostic desciption, global distributions & references
- flowersofindia.net : To pull names in different language, description & uses
- iucnredlist.org : To pull habitat, geography, uses, population.
- wikidata.org : To pull taxonomy information and get few IDs which is available on other websites.
- English Wikipedia : To pull introduction
- wikicommons - To pull plants' image names

We have used Selenium with python to do all the scraping activities and loaded a single dataframe. Each row of dataframe contains the different collected features of a single plant, where column represents each feature.

Some data preprocessing were done to clean the data and remove unwanted/junk characters. Also, few features were broken down into multiple features like names, references. Finally we used 52 features to generate the article content.

3 Translation & Transliteration

Google Translate library is being used for English to Hindi translation. For transliteration, indic_transliteration library is used. For technical words, along with hindi transliteration, we have kept their original English word as well for better readibility. For few words, translation & transliteration were generating special characters, which were cleaned and corrected manually.

4 Jinja Template and Sample Article Creation

English and Hindi templates were created in the form of word documents first, which were later modified based on the feedback of the annotator and Kashyap sir. The sentences from this word document were later used to create the Jinja template. For a sample article, I generated sentences and created an article on Hindi Wikipedia Sandbox for review. Based on feedback, correction were made in the Jinja template and XML Generation code as well.

XML generation code is written based on XML data of an existing article on Wikipedia. It is advisable to use existing Infobox template in our Jinja template which helps in rendering the XML without any issues.

We should keep a practice of using a unique category name for all the articles we are generating. It helps reviewer to identify all articles in one go by searching with just category name.

5 XML Dump and Importing Articles

With the help of Jinja template and XML generation code, XML dump of first 50 artciles has been created and uploaded to Wikipedia Sandbox for review. Once reviewer verifies that the all the sections (ex. Infobox, Intro, Reference etc.) of article gets loaded correctly, full XML dump of plants is loaded into Wikipedia.