Documentation

Domain – Telugu Books

Team:

HARSHAVARDHAN THATIPAMULA

NANDITHA MERUGU

BINGIMALLA YASASWINI

DATA COLLECTION

- We were given **Telugu Books** Domain. And our first task is to choose the reliable sources from where we can get the correct information/data.
- As far as in our team, we spent more time on collection of sources as it's been very difficult to find the sources for this domain.
- The below are the sites which we choose to scrap the data from:

https://www.avkf.org/BookLink/view_subjects.php?cat_id=34
https://www.chirukaanuka.com/collections/all-telugu-books
https://archive.org/details/books?&and[]=languageSorter%3A%22Telugu%22
https://kinige.com/

DATA SCRAPING

- From the above-mentioned websites, we scrapped the data from them using **Beautiful Soup** Python Library.
- The scrapped data is stored in Excel sheet and arranged according to the considered attributes.
- We tried finding data for the below attributes.

  `Language, పుస్తకం పేరు, బుక్ నేమ్, రచయిత, పబ్లిషర్, PLACE_OF_PUBLISH, PAGINATION, ITEM_RUPEES, YEAR_OF_PUBLISH, BINDING, ISBN`

- We faced trouble in scraping for few websites, the we used instant scrapping tools and also Octoparse to scrap the data from dynamic sites.
- We spent about 3 weeks working on data scraping part.
- The Link for the scrapped data list is below:

  https://1drv.ms/x/s!AogaObq4O9uxrExqyOe7bgi4iiUE

TEMPLATE CREATION

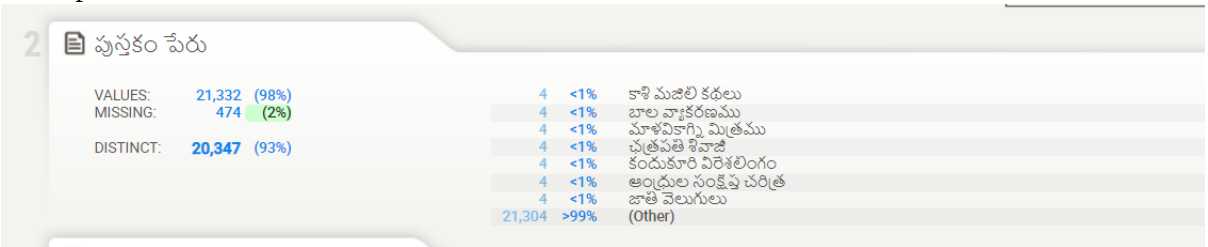
- By using the above-mentioned attributes, we created sentences.
- Example for the above-mentioned sentence:

  {{author}} వ్రాసిన {{BookName}} పుస్తకానికి {{publisher}} ప్రచురణకర్త.
- This is firstly done on paper got verified and then proceeded with translation and jinja.
- We worked on this section for 1 week. We also wrote sample article in tewiki. https://tewiki.iiit.ac.in/index.php?title=%E0%B0%86%E0%B0%B8%E0%B0%B0%E0%B0%BE

- The below is the link which directs to the sample sentences for each attribute. https://docs.google.com/document/d/1j-3Y5etNfzJdv6P_w6jPMvlVxzKjNzhaEC1GOmF5k1U/edit?usp=sharing

SWEETVIZ REPORT

- In the Data Scraping step, we ended up with the final excel file. Then using this excel file we generate the sweetviz report and look through the uniqueness data values for each attribute.
- The main use of this is to analyse the data.
- Example:



In the above attribute, there are 21,232 rows of data and from it we have 20,357 book names to be unique and 474 are of empty block.

- This is how we analyse the data for each attribute and make changes required in it and we done the same.
- We generated this in a day or so and worked on the changes to be made in the sheet for a week.

TRANSLATION/TRANSLITERATION

- We used Google translator for translation.
- We directly used the command below in each cell to translate the data from English to Telugu.

$$= GOOGLETRANSLATE(text, \text{"en"}, \text{"te"})$$

- After entire data got translated, then we teammates divided the entire work to check manually the entire data and correct it.
- The below is the link for final Excel sheet with translated data. https://github.com/nandithamerugu/books/blob/main/Final_excel_books.csv

JINJA TEMPLATE

- In the template creation, we wrote the sample structure of the article. Here, using jinja we need to structure the article.
- After writing the jinja template, we need to write the Python code for rendering the articles. After we run the render code, then it will display the articles in user interface.
- We worked for a week on jinja template.
- After the jinja template creation, we also reverified and corrected the grammatical mistakes with the help of mentors.

XML GENERATION

- This is the final step for the entire process. We need to generation out from the jinja file.
- We first generated 50 XML files i.e., 50 articles.
- We need to copy this XML content and paste it in the tewiki sandbox, then we can verify the preview of the article.
- In our domain, on of the article is displayed in tewiki as below:

Conclusion:

- The entire flow of the process is as follows:
  1. Data collection from reliable sources.
  2. Data scraping from the above collected sources.
  3. For the considered attributes, we need to fill the data using the scrapped data.

4. We need to write the sample article and get verified.
5. Then we need to generate sweetviz report and make relevant changes.
6. We need to translate the data.
7. Then we need to manually check the words after translation.
8. Jinja Template writing.
9. Generating the articles using render code.
10. If no changes, then we end up with xml file generation.
11. We need to copy this xml data into sandbox.

Thank You

We Really thank our mentors Shrestha mam and Vamshi Krishna sir for supporting us in every stage. And we also thank Kashyap sir for supporting us.