Honors Project CLD

Padakanti Srijith

2019114002

Machine translated Telugu Wikipedia for plants

GitHub link for my project: https://github.com/srijith9862/honors-project-CLD

Data for my project (.csv file): Here

Aim:

The aim of this Honors project is to increase the number of Telugu Wikipedia pages in any domain. The domain I chose is plants domain. The total number of plant Telugu Wikipedia pages I created are 7416.

Motivation:

My motivation to increase the number of Telugu Wikipedia pages is to increase the Telugu database for future use.

Related work:

1. Data Scraping from plants related websites.
2. Machine translating the plants database.
3. Prepared a database for plants using 40 attributes and 7416 types.
4. Post editing and formatting the machine translated data.
5. Final plants database. Here

Methodology:

1. First collect the data to make a dataset.
2. Use machine translation to translate the database into Telugu database.
3. Correct the Telugu database after machine translation.
4. The .csv file for plants dataset. Here
5. Use the jinja2 template to write a template for the Wikipedia page.
6. Use macro for writing the template.
7. Use python to create a .xml file for uploading to the sandbox.

Conclusion:

Created XML files for 7416 Wikipedia pages based on different plants.

Future Development:

1. Will perfect the existing Telugu data using DeepTrans or CoreNLP.
2. Will change it into more formal data by using phrases instead of sentences while translation.
3. Add more Data to the existing plant data from various websites.
4. Refine the data into more meaningful data.
5. Do this for various domains.