Final Report For IndicWiki Summer Internship Team Movies

Domain

- The domain we worked on is ‘Movies’. We tried to generate around 8900 articles on different movies from different languages on Telugu Wikipedia.

Team

- Chelpuri Abhijith
- Aravapalli Akhilesh
- Supraja Alleni
- Danda Jahnavi

Data Collection

- The main basis for data collection is IMDb rating with minimum number of votes(10000).

Sources/Sites

IMDb

- Format available - HTML
- Tools used - Parsehub, Selenium
- Attributes found - Title,release_year,rated,duration, genre,movie_rating,movie_director,synopsis,storyline,votes,gross,opening_weekend,cumulative, country, language, screenwriter, casting_by, film_editor, production_designer, composer,set_decoration,art_direction,producer, prod_company, filming_location, awards_nominated,awards_won, songs,sound_mix,colors,cast,starring, tagline, trivia, cinematographer

Wikidata

- Format available - HTML
- Tools used - Wptools (python library), selenium
- Attributes found - Wikipedia_links, wiki_ids, distribution_format, distributed_by, part_of_series, based_on, main_subject, narrative_location

Wikipedia

- Format Available - HTML
- Tools Used - Selenium
- Attributes found - Images

Tools used for Data Collection

ParseHub

- It is a GUI based data extraction tool, we observed that it takes very long time to extract data and also gives no option to switch between html pages so we shifted to another tool

WebScraper

- This is a similar tool to parseHub which does not have issues with time or switching between pages but it gives formatting errors when we download the extracted data → shifted to another tool

BeautifulSoup

- We were unable to understand at first how this works → shifted to another tool

Selenium

- Faced issues with webdriver it used, it sometimes does not load pages correctly resulting in misplacing of data in the csv file → manually corrected them as they were very little in number

Images

- Source - Wikipedia
- Issue → Workaround
    - We found out that images from IMDB are copyrighted. So, extracted the images from wikipedia for their respective movies. Hence, extracted images with the help of wikipedia links by passing the movie name as a parameter. But the problem is, not all the movie names in the dataset exactly match with the name of the movie mentioned in the wikipedia page.(Ex: Dangal (In our dataset extracted from IMDB website) does not match with the title named as Dangal(film)) in wikipedia, which has led to retrieving images only for about half of the films.
    - Since, in order to map the movie name properly, we found a way of getting the exact title of the film from wikidata. We extracted WikiIDs for all movies on the internet from WikiQuery SPARQL along with IMDb ID, and merged them both to create a database which has only the list of movies we want.As we already have imdb ids which are unique in our database which has been extracted from IMDB website.

Data Storing

- .csv - Can be edited using GUI and easier to manipulate.
- .pkl - This format is used for faster loading and better storage of versions.

Data Cleaning

Cases taken care of

- Removal of special characters → Used find and replace (VS Code GUI) to get rid of them.
- Some unneeded text placed in brackets → Used regex (in VS Code GUI) to find such text and removed that text
- Some attributes were placed in brackets → Used find and replace for the whole column to get rid of those brackets

- Data extracted for some attributes is intended to be in the form of lists or dictionaries, but .csv files do not support those datatypes → We needed to convert them into the appropriate datatypes using a python library called **ast** with the help of the method **literal_eval**.
- There were some null values, some cells were just empty for some attributes→ Used **fillna** method to fill all such cells with **NaN**
- Irregular and inconsistent data → manual removal or manual updation
- Normalization of the some attributes from data extracted(Ex: Color) - Observed the data and normalised.

Data Merging

- Primary key - IMDbID
- Process followed, Tool used - Merge methods from Pandas Dataframe
- FinalKB format - csv
- Final KB rows X columns - 8929(movies)*50(attributes)
- Final KB link - [github link]- https://github.com/indicwiki-iiit/Movies
- Issues → Workaround
    - There were multiple records for some movies because of merging multiple times → Manual correction

Version Control

Github

- We used the Github repo provided by IIIT, to update all the changes that are done in our project regularly(weekly).

DeepNote

- We used DeepNote for maintaining and editing code files (templates and render files) simultaneously.

Sample article

Link - {to be added after adding sample article to github}

Approach

- Large number of movie Wikipedia pages were studied, regarding the common sections to be included for a movie and finally decided on the sections that are to be included. Then the available attributes are divided among their respective sections.

Sections (Name → Description)

Infobox

- Movie name, director, writer, producer, actors starring in the movie, music(composer), cinematographer, editing (film editor of the movie), released (year of release), runtime (Duration of

the movie), budget of the movie, country, language, Gross, distributors and poster of the movie are the attributes included in the Infobox section

Intropara1

- In this paragraph, attributes listed are name of the movie, released year, Genre, Director, writers, producer, production company, starred actors, composer and film locations of the movie.

Intropara2

- And in the second paragraph, details regarding the budget, Rated (Film Certification), year of release, languages, countries, Rated, main subject of the film, synopsis, based on and distributors are the attributes which are mentioned for that particular movie.

Plot (కథ)

- Plot of the movie and narrative locations are provided in this section.

Cast and Crew (తారాగణం)

Cast

- Here cast of the film is given. Below are the character and their respective actor names are displayed in the form of a list:
    - Character1 గా actor1
    - Character2 గా actor2
    - Character3 గా actor3 …. ….
    - Character(N) గా actor(N)

Crew

- Here director, writers, producer, composer, film editor, cinematography, casting by (By whom the cast was selected), production designer, set decorator, and art director attribute's information is provided based on the availability of attributes for that movie.
    - దర్శకత్వం : {{ Director }}
    - కథా రచయిత / రచయితలు : {{ writers }}
    - నిర్మాత : {{ producer }}
    - సంగీతం : {{ composer }}
    - ఎడిటర్ : {{ film editor }}
    - ఛాయాగ్రహణం : {{ cinematography }}
    - క్యాస్టింగ్ : {{ casting by}}
    - నిర్మాణ రూపకల్పన : {{ production design }}
    - ఆర్ట్ డైరెక్టర్ : {{ set decorator}}
    - సెట్ డెకొరేషన్ : {{ art director }}

Songs (సంగీతం,పాటలు)

- In the songs section, music composer, sound mix and songs of that movie and the number of songs in the movie are also mentioned. Below is a table representation of song name, singer and duration of the song in the movie (When crossed a threshold length scrollable option is enabled):

Song (పాట)	Singer (గాయకుడు/గాయని)	Duration (సమయం)
May (మే)	Thomas Newman (తోమస్ న్యూమన్)	1:08
...	...	...

Technical Details (సాంకేతిక వివరాలు)

- Duration of the movie, sound mix, color (black and white/color), distribution format of the movie are mentioned in the technical details section.

Critical Response (ప్రతిస్పందన)

- Number of votes obtained for that particular movie in the imdb website and rating of the movie are present in this section.

Production Box Office (నిర్మాణం, బాక్స్ ఆఫీస్)

- production company, budget of the film were mentioned and also opening weekend, Gross , cumulative worldwide gross for the movie are given here.

Awards (అవార్డులు)

- A representation of Awards, details and Result (Winner/ Nominee) in the movie are mentioned in the form of a table in this section (Table becomes scrollable when the length is too much.)

పురస్కారము	అవార్డు వివరాలు	ఫలితము
ఫెలిక్స్ (Felix)	బెస్ట్ అడప్టెడ్ స్క్రీన్‌ప్లే ఫ్రాంక్ డారాబంట్	విన్నర్
అస్క్ స్క్రిప్టర్ అవార్డ్ (USC Scripter Award)	స్టెఫెన్ కింగ్ (ఆతోర్) ఫ్రాంక్ డారాబంట్ (స్క్రీయ్రుటర్)	విన్నర్
ఆస్కార్ (Oscar)	బెస్ట్ పిక్చర్ నికి మార్విన్	నామినేట్

Other Info (ఇతర విశేషాలు)

- Trivia and part of series for the movie are mentioned in this section.

References (మూలాలు)

- References of the movie are imdb , wikidata and Wikipedia, which are mentioned in this section.

Jinja template creation

Link - Movies/main_template.j2 at main · indicwiki-iiit/Movies (github.com)

Edge cases

- After rendering movie articles in the sandbox there were few cases missing for singular or plural sentences, gender specific sentences and few conditions were missing where more than one attribute is used in a sentence. From the observations made all other edge cases were covered are

- Possible/ Common edge cases
  - Singular, plural conditions
    - List of Attributes - Genre,Director,writers,composer,countries,languages,cast,film_locations,production_company,sound_mix,producer,cinematography,casting,production_design,set_decoration,art_design,narrative_location,distributed_by,distributed_format,part_of_series
  - Gender neutral conditions
    - List of Attributes - Director,writers,composer,producer,cinematography,cast,casting,production_design,set_decoration,art_design Multiple attributes conditions: If a sentence has 3 or 4 attributes, to check if all of the values are present we have used an if and else if conditions and if the attribute's value is missing another condition is added similarly to a sentence. For example: In the code below if the movie does not have both the attributes Votes and Rating then their respective sentences are not displayed. If either one or more of the attributes are present then the available attribute's respective sentences will be displayed.

```
{%- macro critics_response(Votes,Rating) -%}
{{ '\n==రేటింగ్స్ ==\n' }}
        {%- if Votes!="NaN" and Rating!="NaN" -%}
            {{Votes}} మంది విక్షకులు వేసిన ఓట్ల ఆధారంగా ఈ చిత్రానికి ఐ.ఎం.డీ.బి లో {{Rating}} రేటింగ్ లభించింది.{{" "}}
        {%- elif Votes=="NaN" or Rating!="NaN" -%}
            ఈ సినిమాకి ఐ.ఎం.డీ.బి లో లభించిన రేటింగ్ {{Rating}}.{{" "}}
        {%- endif -%}
{%- endmacro -%}
```

  - Table Scroll bar option for awards and songs: Whenever a movie has more than a particular amount of songs/awards in the table, then a scroll bar will be enabled.
  - Length of songs condition for Songs : If the movie has more than one song then it displays : ఈ చిత్రం లో మొత్తం {{ lensong }} పాటలు ఉన్నాయి. ఈ సినిమా పాటల జాబితా క్రింద ఇవ్వబడింది.And if there is only one song in the movie, it display : ఈ సినిమా పాటల జాబితా క్రింద ఇవ్వబడింది.

Issues while rendering edge cases:

- After translation and transliteration, "NaN" values were changed to "నాన్", this issue was observed while rendering the articles and therefore replaced the "నాన్" values to "NaN" again without disturbing the other telugu words(in VS Code GUI) which are having "నాన్" in them.
- Categories, References Methodology followed while listing categories: Have checked movie articles in both English Wikipedia and Telugu Wikipedia and then came up with patterns used in existing categories. So that a blue link is generated in the wikipedia sandbox if the categories exists.
  - List of categories:
    - Genre సినిమాలు
    - languages సినిమాలు
    - Release_Year languages సినిమాలు
    - Release_Year సినిమాలు

    - stars నటించిన సినిమాలు
    - Rated రేటింగ్ సినిమాలు
    - Director దర్శకత్వం వహించిన సినిమాలు
    - writers రచన అందించిన సినిమాలు
    - composer సంగీతం అందించిన చిత్రాలు
    - cinematography చిత్రీకరించిన సినిమాలు
    - countries సినిమాలు
    - production_company నిర్మాణం చేసిన సినిమాలు
    - colors సినిమాలు

- Cases considered while giving References: In the references section reliable sources are mentioned from where the attributes we planned on could be extracted. As most of the data is collected from the IMDB website, IMDB references for the Title page, introduction paragraph, Awards and Songs are given. As wikipedia images are used, Wikipedia references for images are also given. Attributes like production_design, set_decoration, art_design, distributed_by, distributed_format, part_of_series are extracted from wikidata using wptools and hence wikipedia references are given to these attributes.

Infobox

- Existing infobox templates from other articles are used and listed out the attributes in it.
- Issues while rendering Infobox:
    - At first, while rendering the articles, there was no fixed size to every image and hence, the image covered most of the page. After resizing the image it worked fine.
    - The attributes which have multiple values don't look organised in the infobox and hence converted those attributes into lists and fixed the issue.
    - After updating the sandbox in tewiki, it is not displaying the image used in the infobox, but was working fine in the Telugu wikipedia sandbox.

Transliteration

Used Deeptranslit(python library)

- Issues:
    - Transliteration for some words like The, To , Into,USA,UK etc.. Wrongly transliterated words:
        - The - తె → ది
        - To - టొ → టు
        - Into - ఇంటొ → ఇంటు
        - USA - ఉసా → యు.స్.ఎ
        - UK - ఉక్ → యు.కె
    - We manually found and replaced the mis-transliterated words using find and replace (VS Code GUI) Translation

Text blob(Python library) - Not used

- Issues:
    - Cannot be used for a bulk amount of text at a time.

    - Translation inappropriate for most of the data.

Translation

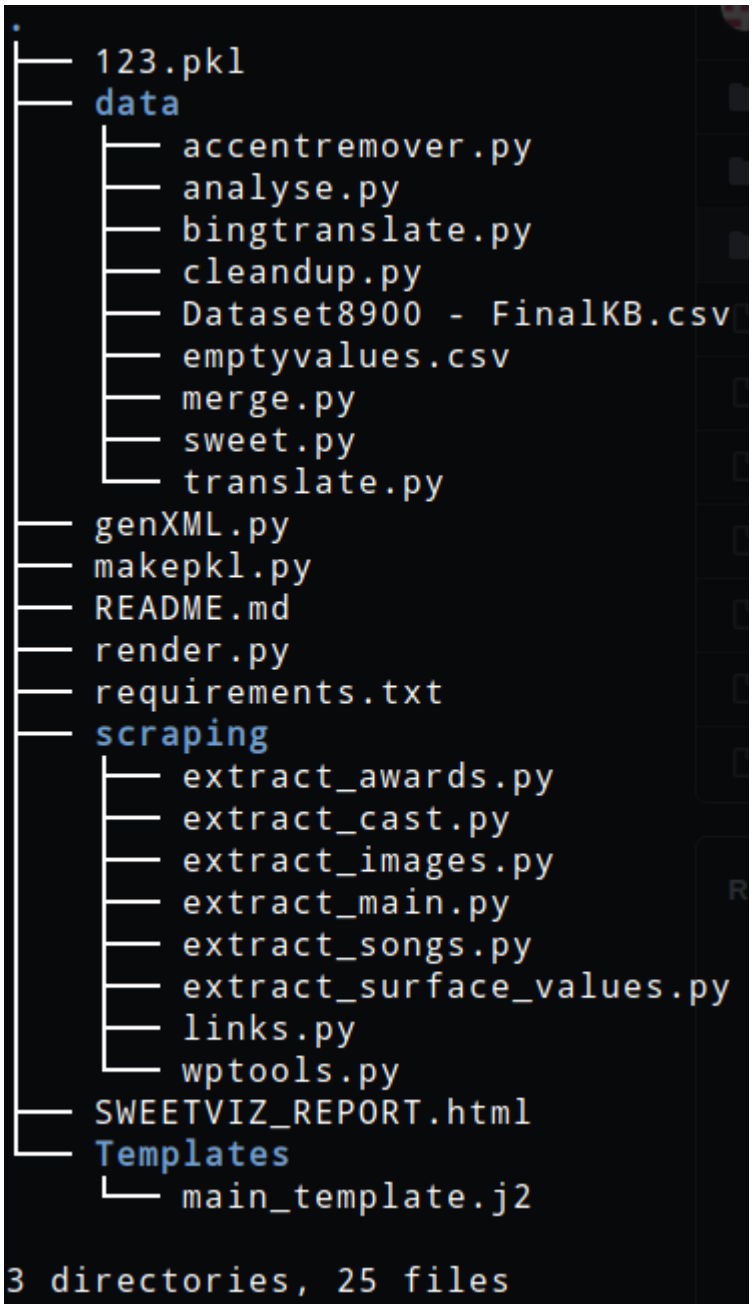
Bing Translate(Azure service): - Used

- We have used this library to translate all our required attributes as we found out that it is keeping a better idea of context while translating , but we have found some issues with this as mentioned below.
    - Issues:
        - Has translation character limit upto 2M characters → We borrowed some of our friends' IIIT accounts.
        - It is not a free resource for non-student-partner Microsoft accounts, so others need to mention their credit card details to login → We borrowed some of our friends' IIIT accounts.

Github

> Repository Link : https://github.com/indicwiki-iiit/Movies

Github Structure :



Details regarding structure

Templates

> Github folder Link: https://github.com/indicwiki-iiit/Movies/tree/main/Templates

- This folder contains the templates that are used for article generation
  - *main_template.j2* -- Contains the final Jinja2 Template for the article generation.

Data

> Github folder Link: https://github.com/indicwiki-iiit/Movies/tree/main/data

- *Dataset8900 - FinalKB.csv* -- This is the final dataset obtained after web scraping from the IMDB website using Selenium and wikidata using wptools.
- *Accentremover.py* -- Python Code to remove all the special characters with accents. ex. à,é
- *Analyse.py* -- Code used to analyse if the attribute holds any value for a particular record or not. So, for each attribute if there's a value. It gives 1 and if there is not any value present it shows 0 and this

is stored in a csv file.
- Bingtranslate.py_ -- Translates the required attributes to Telugu using Bing Translator API.
- *Cleanedup.py* -- Code to remove duplicate records which exist because of extracting data from Wikidata.
- *Emptyvalues.csv* -- Contains Boolean values sshowing availability a particular attribute from a particular movie.
- *Merge.py* -- This code is used for merging databases(csv files).
- *weet.py* -- The code present in this file is to generate a sweetviz report.
- *Translate.py* -- For transliterating attributes,this file is used.

Scraping

> Github folder Link: https://github.com/indicwiki-iiit/Movies/tree/main/scraping

- This folder contains files of codes written for all the extraction of attributes from different sources.
    - *extract_awards.py* -- This code is used for extracting data from the IMDB awards page for every movie
    - *extract_cast.py* -- This code is used for extracting crew details like set_decorator, art_director, film_editor, cinematography,production_design,composer,production_designer,producer from IMDB cast webpage.
    - *extract_images.py* -- This code is used for extracting images from wikipedia using wikipedia links extracted from wikidata.
    - *Extract_main.py* -- All the data like cast details, release date,director,writer, synopsis,trivia, storyline, tagline,film locations,production company, runtime,color,sound mix and all such attributes which could not be extracted from the surface level of imdb website, are extracted using this code
    - *extract_songs.py* -- This code is used for extracting songs from IMDB songs page.
    - *extract_surface_values.py* -- Attributes like genre,rating,rated,votes,gross,plot are extracted using IMDB webpages.
    - *links.py* -- This contains links for awards,songs,cast and crew, surface level attributes.
    - *Wptools.py* -- This code is used for extracting narrative location, distributed_by, distributed_format, part_of_series, main_subject and based_on attributes from wikidata using wptools.

123.pkl

> Github file Link: https://github.com/indicwiki-iiit/Movies/blob/main/123.pkl

- This is the final pickle file generated from the dataset.

Readme.md

> Github file Link: https://github.com/indicwiki-iiit/Movies#readme

- This file has links to the sample article created. It has English Wikipedia,Telugu Wikipedia and tewiki sandboxes. This also has a link to the documentation of what is done every week.

SWEETVIZ_REPORT.html:

> Github file Link: https://github.com/indicwiki-iiit/Movies/blob/main/SWEETVIZ_REPORT.html

- This is an exploratory data analysis report made to check how many unique values are present for each attribute and how many missing values the attribute has.

genXML.py

> Github file Link: https://github.com/indicwiki-iiit/Movies/blob/main/genXML.py

- This file contains the code for generating an XML file which has the data after rendering for an article.

makepkl.py:

> Github file Link: https://github.com/indicwiki-iiit/Movies/blob/main/makepkl.py

- This file contains code for reading the dataset and generating a pickle file

render.py

> Github file Link: https://github.com/indicwiki-iiit/Movies/blob/main/render.py

- This is the code used for rendering the movie articles using jinja2 template named main_template.j2 file in templates folder.

requirements.txt:

> Github file Link: https://github.com/indicwiki-iiit/Movies/blob/main/requirements.txt

- This contains all the packages and libraries that are necessary for building this project.