IndicWiki Summer Internship – DIGITAL CAMERAS

Domain

The domain we worked on was **“Digital cameras”**, the aim of the project is to generate comprehensive articles for Telugu Wikipedia on 4000+ digital cameras, comprising all possible details of a particular camera.

Team

Member name	**Email**
Lakshmi Priya	lakshmipriya934@gmail.com
Yaswanth	yaswanthvakacharla@gmail.com

Data Collection

Sources/ Sites

We went through all possible websites and datasets where we can find information regarding digital cameras.

References:

Datasets:

https://www.kaggle.com/datasets/crawford/1000-cameras-dataset/code

Websites:

https://www.digicamdb.com/

https://pxlmag.com/db/

E-commerce websites (amazon, flipkart)

Among these data sources, we found that digicamdb is useful for us as it provides around 45 attributes of information for each of the available cameras.

The attributes are:

1. Camera_name
2. Brand
3. Model
4. Total_megapixels
5. Effective_megapixels
6. Sensor_type
7. Sensor_size
8. Sensor_resolution
9. Max_image_resolution
10. Crop_factor
11. Optical_zoom
12. Digital_zoom
13. ISO
14. Raw_support
15. Manual_focus
16. Normal_focus_range
17. Macro_focus_range
18. Focal_length

19. Aperture_priority
20. Max_aperture_priority
21. Metering
22. Exposure_compensation
23. Shutter_priority
24. Min_shutter_speed
25. Max_shutter_speed
26. Built_in_flash
27. External_flash
28. Viewfinder
29. White_balance_presets
30. Screen_size
31. Screen_resolution
32. Video_capture
33. Max_video_resolution
34. Storage_types
35. USB
36. HDMI
37. GPS
38. Wireless
39. Battery
40. Year
41. Also_known_as
42. Dimensions
43. Head_quarters

Tools used for Data collection

- **Selenium**
    - An automation testing-based Python library, primarily used for web scraping
    - **Issues:** No significant issues were encountered while working with this, except for the time delay issue - sometimes, web pages would not load elements on time and that would result in timeouts, exceptions, and end of execution of the program. To avoid this, an additional method was added to keep Selenium web driver's execution on hold, until some element would appear on the webpage.

- **BeautifulSoup**
    - This was used as an additive to Selenium to navigate through HTML elements and obtain text/information from them, hence, **no major issues** were observed.
    - Most of the data is scraped using this module.

Images

- Source: Wikimedia commons
- Since we collected data from digicamdb.com website instead of wikipedia, we had to scrape names of image files for cameras which exist in wikimedia commons. We used Special:MediaSearch tool to search for image and scrape using BeautifulSoup.
- **Issue#1:** Because the search gives relevant images, a lot of images didn't match with camera name. So, we had to delete them manually.
- **Issue#2:** A few search cases didn't give the correct image as the first image but the next image at some position. We might have missed a few images that are available in wikimedia commons as we scraped only the first image in every search.

Data Storing

- **Format #1:** Pickle, .pkl
- Why -
    - .pkl prevents automatic conversion of null values into ‘nan’ and other default data type conversions which usually happens with .csv or .xlsx files, and is hence, the format we’ve preferred to ensure data is kept intact and unadulterated.

- **Format #2:** Comma Separated Value, .csv
- Why -
    - csv is not a new/unreadable format to computers like .pkl is. Ease of access and editing of these files define why we’ve used .csv files for minor edits.

Data Cleaning

- The data we scraped is already consistent. So, we didn’t perform data cleaning.

Version Control

- For version control of files used in the project, we used the IIIT GitHub repository, which was especially useful when tasks associated with data cleaning and editing the jinja template had to be performed, when multiple people were working on various versions of the template/data which were sequentially updated.

Sample Article

- TeWiki Link

Sections

There can be only 2 sections possible in case of digital cameras. They are introduction and features. Here in this case, we have further divided features into some specific groups like screen related features, sensor related features.

- **Sections and their Description**
  - **Introduction(పరిచయం)**
    - Contains basic details about the camera such as camera_name, brand, model, launch_year, head_quarters etc.
  - **Sensor(సెన్సార్)**
    - Contains details about the camera sensor.
  - **Resolution(రిజల్యూషన్)**
    - Contains details about the resolution provided by the camera like screen resolution, max video resolution.
  - **Screen(స్క్రీన్)**
    - Contains details about the camera screen.
  - **Video(వీడియో)**
    - Contains details about video capturing capabilities of the the camera.
  - **Focus(ఫోకస్)**
    - Contains details about the camera focusing capabilities.
  - **Connectivity(కనెక్టివిటీ)**
    - Contains details about the camera connectivites like GPS, HDMI, USB etc.
  - **Physical_properties(భౌతిక లక్షణాలు)**
    - Contains details about the camera physical properties.

Jinja template creation

- GitHub Link

Edge cases

- **Making Sure all Edge Cases have been covered**
    - The first step was to prevent the rendering of sections/sentences if the attribute the sentence has been constructed based on has a null value for that record. So, in the initial stages of designing the template, all occurring null values in the database were identified and these were the very first edge cases added.
    - After this, at the end of every iteration of updating the sample article, and subsequently, the template, we tried rendering the article for tens of records, few with dense and few with sparse data. This would lead to the discovery of hidden edge cases which we would then include in the template and repeat the process.
- **Possible/ Common Edge Cases**
    - Null Values
    - Making grammatical corrections to sentences used in the template based on singular/plural value corresponding to the attribute forming it, if the attribute is numerical
- **Issues/Improvements Associated with Rendering**
    - To improve the readability of values corresponding to certain attributes, we had to render their English versions too in parentheses next to them.

      For instance: కెనాన్ ఈఒఎస్ ఆర్3(EOS R3) అనేది ఒక డిజిటల్ కెమెరా.
    - Undesired newline characters would occur in the absence of sentences owing to the values of the associated attributes being null, so these had to be meticulously checked and corrected.
    - Sometimes, despite the addition of spaces before punctuation marks like full stops and commas in the template, they wouldn't get added because of the whitespace elimination formatting used in the template. These also had to be carefully checked and added at all punctuation marks.

Categories, References

- **Identifying Categories**
    - Every camera brand itself is a category.
- **List of Categories**
    - Canon, Fujifilm, Nicon, Leica, Olympus, Panasonic, Pentax, Richo, Samsung, Sony,BenQ, Casio etc…

- **References - the criteria we defined to add them**
    - Every source from which data on a particular camera has been collected has been added as a reference.
    - To obtain this reference link, a generalized format of every link was looked up for, such that, if any of the primary keys (ISBN) is appended to the link, it would redirect to a page containing details of the camera, or search

      results of that camera on the site. For each camera, these generalized links with the primary key appended have been added as reference links.
    - The reference links appear at the end of the Wikipedia article, listing out all links enclosed in <ref></ref>, and these links have been appended at the end of sentences containing data obtained from that source, to add a citation. Hence, the reference tags listed at the end of the article are listed only if non-null data has been obtained from that reference link.

Infobox

- We have initially planned to make use of the existing Infobox:Camera template, but that does not contain all the attributes we wished to add to the infobox. We have written our own template with our desired attributes. But, in this case we were unable to render the data to infobox as an error named module error is coming.So, we decided to add the attributes which are common in our attributes and the attributes of the existing template.

  Those parameters are -

  - **| maker**, the brand name of the camera.
  - **| camera_name**, actual camera name.
  - **| data**, launch year of the camera.
  - **| sensor_type**, type of the sensor used.
  - **| sensor_size**, size of the sensor used.
  - **| res**, maximum resolution of the image the camera can produce.
  - **| shutter_speeds,** the minimum and maximum shutter speeds of the camera.
  - **| viewfinder**, the type of the viewfinder used in the camera.
  - **| metering**, the metering information of the camera.
  - **| dimensions**, the dimensional information of the camera.
  - **| battery**, the battery information of the camera.
  - **| weight**, the weight of the camera.
  - **| made_in**, where the headquarters of the camera located.

Translation/ Transliteration

Libraries

- DeepTranslit
  - Used for Transliteration
  - **Issue:** Did not have accurate transliterations of many independent English letters like “to”, “there”, “the”, etc. The workaround was manual correction.
- GoogleTrans
  - Was used for the initial translation of the data.
  - **Issue:** There is no issue in translation. Those that can be translated were translated and the remaining were left for transliteration.

Common Issues while translation, transliteration

- **Incorrect Transliteration caused by English data that wasn’t completely suitable for transliteration**
  - An example of this is “etc.” getting replaced by ఎæş
  - **Workaround:** Manual replacement of such transliterations with meaningful Words.

- **Incorrect Transliteration of abbreviated words**
  - An example of this is “EOS” getting transliterated to “ఏయస్”.
  - **Workaround:**
    - We need to manual transliterate each individual character to its corresponding telugu representation.

      For instance, “EOS” to “ఈ ఓ ఎస్

- **Incorrect Transliteration/Translation caused by non-English characters**
  - As mentioned above in a lot of other sections, transliteration and translation would heavily take a hit because of non-English characters used in various values.
  - **Workaround:** Manual identification and correction of such values.

We must carefully observe the translated values. Because “focal length” might get translated to “ఫోకల్ పొడవు”. Instead, it should be “ఫోకల్ దృష్టి”.We have to carefully observe such kind of translations and manually correct them.

XML Generation

- GitHub Link

After the template was finalized, the final step was to convert templates of articles in bulk into XML file(s) to be uploaded to tewiki, to publish the articles.

Procedural Details

Converting templates into XML files involved some key steps like the addition of **<mediawiki>** decorators at various places in the template collection, in abeyance with the syntax prescribed to upload XML files to tewiki, and adding tags of unique identifiers of every article such as **<sha1>** and **<id>** (on the basis of the range of ids we've been allowed to use), and other details such as **<timestamp>, <contributor>** (name and id of the user publishing these articles), and some other tags to complete the technical essentials of the same.

Of all the cameras in the database, the rendered template of each camera has been enclosed between the **<page> </page>** tags in the XML File, each "page" inclusive of the details such as user details, timestamp, id, etc. This is how the rendered templates of all cameras in the database have been enclosed and added to the same XML File. All the other tags in the file have been added to comply with the "mediawiki" syntax and enable its linked functionalities/pages when rendered to Tewiki.

Quality Review

As part of the quality review of the article, multiple iterations of reviewing were done to keep the quality and vocabulary of the article intact. Few instances of the quality review done are as follows -

- **Better Vocabulary**

  This was done to replace English versions of words (or) Telugu words with accurate Telugu versions of the same, such as “focal length” to “ఫోకల్ దృష్టి” etc.

- **Using Attributes only of Significant Value (in terms of content)**

  The attributes “Exposure Compensation”, “ISO” and “White Balance Presets” of the camera were eliminated, as they were either found to have added no value to the article about the camera or were not appropriate to be added as they were not informative, but promotional about the camera.

- **Improving Readability**

  To make values of some attributes comprehensible, such as camera name, storage types, battery, English versions of these values have been added in parentheses next to the Telugu versions of these values, allowing readers to view an accurate representation of the Telugu value in English.

- **Keeping the article concise**

  Shifting the emphasis from keeping the article populated with the maximum possible word count, to keeping the article concise was one of the objectives of the quality review.

  Similar corrections were done to keep the article detailed, yet brief - for instance, removing multiple occurrences of references for similar attributes.

Github Structure

Visit our GitHub repository here.

To navigate through the repository, please refer to the below guide.

- The Codes folder consists of the codes used for data scraping and XML Generation.
- The Template folder consists of the final Jinja Tempate.
- The Datasets folder consists of the data in excel format.
- The Sample_XML folder consists of sample XML file.
- The XML folder consists of the XML files for 1000 cameras each.