IndicWiki Summer Internship - [FOOTBALL_PLAYERS]

Domain

We are a team of 5 members and we have worked on football players' domain. We tried to generate articles on 4376 football players, including both male and female players.

Team :

1. VALLEPU SAITEJA - vallepusaiteja@gmail.com
2. NALLURI PAVAN SAI - nalluripavansai@gmail.com
3. RIKKA PRANEETH - 18311A12G2@sreenidhi.edu.in
4. V.ABHINAYA -abhinayavankadari@gmail.com
5. VEMULA SREE HARSHA - harshavemula1922@gmail.com

Data Collection

We have collected data from transfermarkt website , kaggle and fifa index (2021)

Sources/ Sites

- TRANSFERMARKT
    - Format of data available - unstructured,tables
    - Tools used -selenium, scrapy, beautifulsoup
    - Attributes found -awards, national career, debuts, birth place, career stats, age and coach ,player name.

- KAGGLE(sofifa)
    - Format of data available - structured (dataset)
    - Attributes found-stats_by_competition(competition, appearances, goals, assists), yellow_cards, second_yellow_cards, red_cards,minutes_played), ball_control, dribbling, marking, slide_tackle, stand_tackle, aggression, reactions,interceptions, vision, composure, acceleration, stamina, strength, balance, agility, sprint_speed, jumping,Achievements,national_team,national_team_position,national_rating,national_jersey_number,international_reputation,highest_market_value, wage_euro,club_team, club_rating, club_position, club_jersey, club_joined_date, contract_end_year overall_rating, potential,full_name, dob, birth_place, nationality,positions,dob, height, weight, positions, current_team,jersy_ number.
- FIFA INDEX (2021 female players)
    - Format of data available - unstructured
    - Tools used -selenium, scrapy, beautifulsoup
    - Attributes found -.name,nation,height,weight,preferred foot,dob,age,positions,kit number,ball control,dribbling, marking,stand tackle,slide tackle,aggression,reactions,interception,vision,composure,crossing,shortpass, longpass,acceleration,stamina,strength,balance,sprint speed,agility,jumping,heading,shot power,finishing,long shots,curve,FK_ACC,penalties,volleys,GK_positioning,GK_diving,GK_handling,GK_kicking,Gk_reflexes,images.

Tools used for Data collection

- Selenium

We have used selenium for extracting data from transfermarkt

- issue →workaround.

We got a timeout error while scraping a large amount of data. We have solved it by using the time.sleep function.

It is difficult to scrape large amount of data which is time taking process so we partitioned data in different mini segments which is easily scraped by using selenium

- Beautiful soup
  - Issue → workaround

Beautiful soup is a very good tool for web scraping but it is not working for some websites showing 404 page not found like transfermarkt so instead we used selenium for scraping for this website.

- Scrapy
  - Issue → workaround

This is also another tool which is very useful for scraping. When we want to use scrapy, Urls should be known in advance so for that we have collected all the urls in advance

Images

We have used beautiful soup for extracting images

- We have extracted images from wiki commons using player name as a query search in order to scrape images.
- While scraping we didn't get images for some players so we used an alternate name of a player as a query search inorder to scrape those images.

Data Storing

- .csv- Because it is structured data which helps us to retrieve data easily whenever we need in order to make a template.
- .pkl - This format is used for faster loading and better storage of versions and also Pickle will only write out any single object once, making it effective to store recursive structures ,and can serialize pretty much any python object.

Data Cleaning

Cases taken care of

- Removing unwanted attributes

We had Identified unwanted attributes so we removed them by checking whether it is useful in paraphrasing or not.

- Removing latin words

  We had a problem with latin words while we were translating it. So in order to clean those latin words we used a unicodedata module for replacing them.

- Removing special symbols
  While we are working with data cleaning we come to know that some special symbols are found in our dataset. In our dataset for cleaning which can be simply removed by find and replace in the dataset.

Data Merging

- Primary key - player ID,player name.
- Process followed-initially we had three datasets by using player name as a primary key we merge dataset 2 and 3 and result is merged with dataset 1 by using player ID as a primary key .
- Tool used - .initially we used pandas(merge method) in python for merging
- FinalKB format -comma separated values(csv file)
- Final KB rows X columns - male(4008 X 64),female(368 x 45)
- Final KB link - [IIIT github link is appreciated]
- Don't have primary key→ merged together with two primary keys

Version control

- Version control used between the team [if multiple, list all of them]

Sample article

- Link - [IIIT github link is appreciated]

Sections

- Sections are made based on the english wikipedia articles
- infobox→ it contains biodata of the player like player name, birth place, club name,jersey number, height, national team etc.
- Introduction → this section contains a brief introduction about the players by containing details like full name,dob,nationality,clubname etc.
- Personal life → this section is all about player personal life by displaying details like birth place,date of birth.
- Early life →early life is a section whose main theme is to describe debuts of the player in each competition and coach names
- Club career → this section contains all about player present club details.
- International career →in this details like international team and international debuts ,and highest market value and wage euros etc. are going to be described in this section
- Playing style → this section contains details like preferred foot, player positions, skill ratings and traits.
- Career ratings → this section mainly contains all the ratings and red yellow cards in his career.

- Awards → this section describes all the awards received by the player along with their count.
- References → it contains all the references from where we have collected the data
- Representational format if any [tables, lists, etc] and Why

Jinja template creation

- Link - [IIIT github link is appreciated]

Edge cases

- After rendering we have observed that all the conditions are checked in the template and we didn't find any mistakes in the article ,we have rendered for the players who are having nan values also and we have observed that all the edge cases were covered
- There are some mistakes while transliterating abbreviations and other names which can be done by manually
- There are no edge cases while rendering

Categories, References

- Listing categories : Based on the attributes we have in the data set and referred some of the articles in wikipedia which are similar to our data set and listed them
- categories should be relevant to the article and it should be reliable content.
- Categories :
    - Country name
    - Fifa(ఫిఫా)
    - Tewiki football players(తెవికీ ఫుట్‌బాల్ క్రీడాకారులు)
    - Football players(ఫుట్‌బాల్ క్రీడాకారులు)
    - players(క్రీడాకారులు)
    - games(ఆటలు)
- We need to make sure that reference should contain details about that para or sentence ant it should be genuine

Infobox

- We have used existing infobox template for our article

Translation/ Transliteration

Libraries

- deep translit - transliteration - Used
    - We have used deep translit for transliteration purpose and it didn't transliterated abbreviations and country names correctly so for that we have used google_new_trans and bing for transliteration,

- Google_new_trans - transliteration - Used
    - We have used it for It for transliteration purposes.It didn't translate abbreviations and also it return some names as it is in english.thats why we used bing translator
- Bing - transliteration - Used
    - We mainly used it for transliterating abbreviations and also for some names.we didn't face any much issues by using bing translator.

XML Generation

- Executing the 'render.py' file would generate wikitext for every cricket player whose details are present in the finalized dataset. This wikitext is dumped into a single xml file with the help of file 'genXML.py'.
- The single XML file 'footballplayersxml.xml' contains wikitext of every football player whose details were collected. This xml file is imported to wikipedia to generate the articles for different football players.
- While creating the XML, predefined entities(< > & ' ") were also taken care of by replacing them with appropriate strings(< > & ' ") respectively. These entities are replaced with the strings mentioned above in all of the Wikitext.

Quality Review

- We rendered different football players articles (about 15) and ensured that all the edge cases are handled properly.
- We sent those rendered articles to our friends for review and took suggestions to improve the quality of the articles
- We regularly had review meetings with our mentor (Sai Teja) and enriched the structure and quality of the article, by making corrections based on suggestions.
- As suggested by language experts, we made appropriate corrections in the template.
- Entire translated/transliterated dataset was sent for review to ensure that there are no mistakes in the transliterated/translated data.
- We rendered many articles and checked for uniformity (spaces between words, paragraph breaks etc.) in the template, and made necessary corrections based on observations.

Common Issues while translation, transliteration

- Found some issues while transliterating abbreviations and country names →for country names we used bing and google translator, for abbreviations we did them manually.
- Transliterating special characters and roman numbers → special characters are done manually by find and replace, same for the roman numbers.

Further Enhancement

➔ There is a scope to add players personal details (family details)
➔ There is a chance to add more details about player club career like club debuts
➔ There is a scope to add awards for a player who doesn't have
➔ There is a chance to add more details to about his early life
➔ There is a scope to add images for players