IndicWiki Summer Internship

TED SPEAKERS

[Domain](#_Domain) [Team](#_Team)

Data Collection [Sources/ Sites](#_Sources/_Sites)

[Tools used for Data collection](#_Tools_used_for)

Data Cleaning

Sample Article

Jinja template creation

Translation/ Transliteration

XML Generation

Final data

# Domain

The domain we worked on was **“TED SPEAKERS”**, the aim of the project being generating comprehensive articles for Telugu Wikipedia on 1800+ speakers, comprising all possible details on a speaker.

# Team

|  |  |
| --- | --- |
| **Member** | **Email** |
| Bhargavi Inukurthi | inukurthibhargavi238@gmail.com |
| Sai Ram Gollapalli | Sairamgollapalli2001@gmail.com |

Data Collection

## Sources/ Sites

We searched for data in various websites, and we found a website which is solely dedicated in providing TED speaker details to its viewers. Apart from this website, we scraped data from Wikimedia as well including Kaggle which contained a exclusive dataset related to TED TALKS which is merged with other two datasets to create a attribute heavy dataset for generating articles.

* TED.com
  + Link: <https://www.ted.com/>
  + Format of data available – Text
  + Tools used – Beautiful Soup (Python Libraries)
  + Attributes found
* The attributes for a TED speaker were 5 – 7 which are very limited for generating a article but they are included in the Dataset while merging
* Wikimedia
  + Link: <https://www.wikipedia.com/>
  + Format of data available - Web Pages (HTML)
  + Tools used - Beautiful Soup (Python Libraries)
  + Attributes found
    - We have found around 110 attributes from the Wikipedia, but for a speaker on an average there are 7-10 attributes. So we removed at the remaining attributes during cleaning.
* Kaggle
  + Link: https://www.kaggle.com/
  + Format of data available - Database
  + Tools used – Excel, SAS
  + Attributes found
    - We have found around 20 attributes from the Kaggle.

## Tools used for Data collection

### Beautiful Soup

* + This was used to navigate through HTML elements and obtain text/information from them, hence, **no major issues** were observed.
* **Python pandas**
  + A famous python library which is used to efficiently handle datasets present in csv or xlsx formats.
* **SAS**
* This is a software tool used for Data organization.

Data Cleaning

* Data cleaning has been done on the data collected and removed the attributes which are not necessary.
* Remove the rows with a smaller number of attributes.
* Cleaned the duplicate rows and corrected them.
* After cleaning, we created structured data.

Translation and Transliteration

* After cleaning the data, translation and transliteration is done.
* To translate the data, google translator is used. We translated each column by uploading the file to the Google’s translator website - <https://translate.google.co.in/?sl=auto&tl=te&op=translate>
* Apart from it, Bing Translator and Yandex are also used

Sample Article

* Generated sample articles for TED Speaker
  + Sample article for a TED speaker with more attributes article – BILL GATES

Jinja Template Generation

Ted\_Template.j2

XML generation

Ted\_speakers.xml

Final Data:

Ted\_Speakers.xlsx