Documentation

Topic: Movies

Team: Abhijith | Akhilesh | Jahnavi | Supraja

**Sources**

1. Wikipedia
2. IMDb Website
3. Wikidata.

**Work-Flow:**

Sprint #1

|  |  |
| --- | --- |
| **Monday** | Listed Attributes and Referred sources and tools |
| **Tuesday** | Tried extracting data from Wikidata(SPARQL) and IMDb(parsehub) |
| **Wednesday** | Tried other tools like web scraper for IMDb |
| **Thursday** | using Selenium to extract data from Selenium and writing sample articles |
| **Friday** | used Selenium to extract remaining data |

Sprint #2

|  |  |
| --- | --- |
| **Monday** | Searched for more attributes & extracted data for a bigger set of movies. |
| **Tuesday** | Extracted awards, started writing code for songs’ extraction. |
| **Wednesday** | Extracted songs and gone through some libraries which can help us extract data from wikidata/wikipedia |
| **Thursday** | Merged all the data and created a Final\_DB and cleaned data completely |
| **Friday** |  |

**Methods used/will use:**

* Wikipedia - using ‘Wikipedia’/’Wikiextract’ library or ‘beautifulSoup’ library.
* Wikidata - using Wikidata Query (SPARQL).
* IMDb Website - using ParseHub(for 10 attributes) and Selenium.
* For cleaning - Sweetviz.

**Attributes:**

1. From IMDb -

movie\_name,movie\_release,movie\_rated(like R or PG-13 etc),movie\_duration ,movie\_genre,movie\_rating,movie\_ director,movie\_description ,movie\_votes,movie\_gross, country, language, screenwriter, cast\_member, film\_editor, production\_designer, composer, producer, prod\_company, filming\_location,, awards\_ach, awards\_nom, awards, songs.

1. From Wikidata -
   1. Distribution\_company,narrative\_loc,distribution\_format, distributed\_by, followed\_by, part\_of\_series, main\_subject. based\_on
2. From Wikipedia -

Critical Reception

**Attributes that are needed:**

1. Distribution\_company
2. Narrative Location
3. Distribution Format
4. Distribution\_by
5. Followed\_by
6. Part\_of\_series
7. Main\_subject
8. Based\_on
9. Critical Response

### **Total attributes list (as of sprint2):**

Records X attributes - 8930 X 40 (with one redundancy)

Imdb\_id, title, release\_year, duration, rating, rated, genre, director, synopsis, votes, gross, songs, poster, countries, languages, release\_date, film\_locations, budget, opening\_weekend, cumulative\_gross, production\_company, sound\_mix, colors, **(images)**, writers, cast, casting\_by, storyline, tagline, trivia, producer, composer, cinematographer, film\_editor, production-design, art\_direction, set\_decoration, nominee, winner, stars.