# CS 179g Final report

Team 8:

Mahdi Aouchiche
Yongfeng Liang
Hongan Zhang
Yanjun Zhu

Our dataset:

Link for the Shared Folder: Metacritic Movies Data

## Table of content

## Part 1 - Data collection

### Requirements

Collecting around 1GB of data from one of the followings:

- Crawl the web using Scrapy from website: *metacritic.com*

TA comments to address:

- How many reviews did you get in total?

We got 575057 reviews in total

- What's the min/max/average # reviews per movie?

We are not sure how to answer this question. According to metacritic.com, there are always a couple critic reviews for each movie, and some movies may not have any user reviews. Some popular movies can have hundreds of reviews.

- You may also include the movie description, release date, countries, languages. I think these are important as well.

We have included movie descriptions (summary) and release dates. We do not include countries and languages because they are not available for all movies.

### Design

We want to get movie data from the website *metacritic.com*. We used a seed url to crawl all the links in the website to gather information about all the movies listed in the website (about 14488 movies). We also crawled another dataset for the review: user and critics reviews (about 575057 reviews), which we want to analyze to find out the user sentiment using a model.

### Implementation

Using scrapy to crawl data about movies such as title, ratings, release date, distributer, reviews,actors, meta scores, user scores, positive scores, negative scores and so on.

We implement two scrapy crawlers to crawl the detailed information of each movie, and the critic and user reviews of all the movies. The data format are two JSON files: *movies.json* and *reviews.json*

- Movie spider: crawls through a list of movies (100 items) per page.We follow the movie detail link for each movie to get its metadata including (title, ratings, release date, actors, etc…)

- Review spider: traverse through the list of movies on the site, we follow through its critic and user review pages to get all the rating score, distribution of ratings, and comment of each movie.

## Evaluation

Overall we are satisfied with the data we got for each movie so far. However, since the data is compact and in text format. We Are able to get roughly ~ 1GB of data from the entire collection of movies in metacritic.com.

Properties of data:

1. Movies data:
   Title, id, rating, scores (user and meta), actors, genre, distributor, runtime, etc.
2. Reviews data: (both critic and user reviews)
   Author of user reviews, rating score, given review, etc.

Performance:

- Runtime: The crawler to crawl the movies data along with the critic reviews took 36657.27 seconds. Since we have set a delay of approximately 1-5 seconds in between requests to avoid getting blocked from the site, a longer runtime will be expected.
- Successes/misses: There are a total of 14482 movies on the site, we are able to crawl 14471 items, the overall success rate is 99.92% with the missing rate of 0.08%.
- 'retry/count': 947
- 'retry/max_reached': 7
- 'retry/reason_count/429 Unknown Status': 89
- 'retry/reason_count/504 Gateway Time-out': 858

Challenges:

- It gets difficult to avoid rejected requests for data and the websites block the crawling bots.
- All of our data is text, so it is not easy to gather 2GB of data.
- Using a delay to send requests makes the crawling process very long to be able to gather data.
- Websites tend to have many different layouts and designs which makes it difficult to crawl and get the same type of data we want, this adds a huge overhead in the cleaning process.

## Screenshots

*Figure: Movie data*


```
▼ root: [] 14488 items
  ▼ 0:
    ▼ info:
        id: 6
        title: "Three Colors: Red"
        distributor: "Miramax"
        date: "November 23, 1994"
      ▼ starring: [] 2 items
          0: "Irène Jacob"
          1: "Jean-Louis Trintignant"
        summary: "Krzysztof Kieslowski closes his Three Colors trilogy in grand fashion, with an incandescent meditation on fate and ch
          model in Geneva whose life dramatically intersects with that of a bitter retired judge, played by Jean-Louis Trintignant. Mea
          alousy and betrayal unfolds. Red is an intimate look at forged connections and a splendid final statement from a remarkable f
      ▼ director: [] 1 item
          0: "Krzysztof Kieslowski"
      ▼ genres: [] 3 items
          0: "Drama"
          1: "Mystery"
          2: "Romance"
        rating: "R"
        runtime: "99 min"
        awards: [] 0 items
      ▶ stream_link: [] 2 items
        image: "https://static.metacritic.com/images/products/movies/7/cb3524a14b6375d42934270defe3aaca-250h.jpg"
    ▼ meta_score:
        score: "100"
      ▼ distribution:
          positive: "9"
          mixed: "0"
          negative: "0"
        review_link: "https://www.metacritic.com/movie/three-colors-red/critic-reviews"
    ▼ user_score:
        score: "8.6"
      ▼ distribution:
          positive: "152"
          mixed: "4"
          negative: "9"
        review_link: "https://www.metacritic.com/movie/three-colors-red/user-reviews"
```


*Figure: Critics review data*

```
▼ 570096:
    review_type: "Critic Review"
    release_date: "December 6, 2019"
    movie_name: "Portrait of a Lady on Fire"
    meta_score: "95"
    review_score: "90"
    review: "Visually ravishing ... [A] piercingly intelligent treatise on art, agency and queer love in the 18th century."
▼ 570097:
    review_type: "Critic Review"
    release_date: "December 6, 2019"
    movie_name: "Portrait of a Lady on Fire"
    meta_score: "95"
    review_score: "89"
    review: "What the film excels at, however, is the anticipatory desire. It builds slowly, concluding with a stunning sequence that is all breathless remembrance and self-s
      oth wordless and impalpable. The film will seem the height of romantic desire to some, but will be a slow burn for others."
▼ 570098:
    review_type: "Critic Review"
    release_date: "December 6, 2019"
    movie_name: "Portrait of a Lady on Fire"
    meta_score: "95"
    review_score: "88"
    review: "A film in which everything feels stunningly fresh, raw and new."
▼ 570099:
    review_type: "Critic Review"
    release_date: "December 6, 2019"
    movie_name: "Portrait of a Lady on Fire"
    meta_score: "95"
    review_score: "86"
    review: "Even more powerful than Sciamma's portrayal of a feminine portrait of solidarity and desire is the statement that art is not exclusive to those who make it."
```


*Figure: User review data*


```
▼ root: [] 575057 items
  ▼ 0:
      review_type: "User Review"
      release_date: "January 12, 1940"
      movie_name: "The Shop Around the Corner"
      user_review_score: "6.9"
      reviewer_name: "Unchartedhero"
      review_score: "10"
      review: "One of the best stylish German director Ernst Lubstich Movies and I actually recommend to everyone"
  ▼ 1:
      review_type: "User Review"
      release_date: "January 12, 1940"
      movie_name: "The Shop Around the Corner"
      user_review_score: "6.9"
      reviewer_name: "HaskettFamily"
      review_score: "9"
      review: "This has to be one of the great movies of all time. The acting is wonderful and the premise, which, because it is an old movie, is very original, if fun and entert
        more modern movie, "You've Got Mail" is a remaking of this movie. That one wasn't nearly as good. :-)"
  ▼ 2:
      review_type: "User Review"
      release_date: "July 2, 2021"
      movie_name: "Summer of Soul (...Or, When the Revolution Could Not Be Televised)"
      user_review_score: "6.9"
      reviewer_name: "RatedRex"
      review_score: "10"
      review: "Easily one of the best music documentaries of all time. It's better than "WattStax," which came out in 1973. And it's more culturally relevant than "Woodstock." It
        n time when America was coming to the realization that change was inevitable."
```


*Figure: Crawling Result*

```
2021-10-28 08:14:18 [scrapy.core.engine] INFO: Closing spider (finished)
2021-10-28 08:14:18 [scrapy.extensions.feedexport] INFO: Stored json feed (14471 items) in: movies_with_reviews2.json
2021-10-28 08:14:18 [scrapy.statscollectors] INFO: Dumping Scrapy stats:
{'downloader/request_bytes': 9543974,
 'downloader/request_count': 30364,
 'downloader/request_method_count/GET': 30364,
 'downloader/response_bytes': 784016858,
 'downloader/response_count': 30364,
 'downloader/response_status_count/200': 29092,
 'downloader/response_status_count/301': 318,
 'downloader/response_status_count/429': 92,
 'downloader/response_status_count/504': 862,
 'dupefilter/filtered': 14260,
 'elapsed_time_seconds': 36657.265427,
 'feedexport/success_count/FileFeedStorage': 1,
 'finish_reason': 'finished',
 'finish_time': datetime.datetime(2021, 10, 28, 8, 14, 18, 367742),
 'httpcompression/response_bytes': 5034076733,
 'httpcompression/response_count': 29092,
 'httperror/response_ignored_count': 7,
 'httperror/response_ignored_status_count/429': 3,
 'httperror/response_ignored_status_count/504': 4,
 'item_scraped_count': 14471,
 'log_count/DEBUG': 44836,
 'log_count/ERROR': 7,
 'log_count/INFO': 628,
 'memusage/max': 166469632,
 'memusage/startup': 56541184,
 'request_depth_max': 146,
 'response_received_count': 29099,
 'retry/count': 947,
 'retry/max_reached': 7,
 'retry/reason_count/429 Unknown Status': 89,
 'retry/reason_count/504 Gateway Time-out': 858,
 'robotstxt/request_count': 1,
 'robotstxt/response_count': 1,
 'robotstxt/response_status_count/200': 1,
 'scheduler/dequeued': 30363,
 'scheduler/dequeued/memory': 30363,
 'scheduler/enqueued': 30363,
 'scheduler/enqueued/memory': 30363,
 'start_time': datetime.datetime(2021, 10, 27, 22, 3, 21, 102315)}
2021-10-28 08:14:18 [scrapy.core.engine] INFO: Spider closed (finished)
```

## **Contribution**

Movies data - Yanjun Zhu, Yongfeng Liang, Hongan Zhang, Mahdi Aouchiche
Critic reviews - Yanjun Zhu, Hongan Zhang, Yongfeng Liang,Mahdi Aouchiche
User reviews - Yongfeng Liang, Yanjun Zhu, Hongan Zhang,Mahdi Aouchiche
Jupyter Lab setup - Yanjun Zhu, Yongfeng Liang, Hongan Zhang, Mahdi Aouchiche
Brainstorm design - Yanjun Zhu, Yongfeng Liang, Hongan Zhang, Mahdi Aouchiche
Preparing reports - Yanjun Zhu, Yongfeng Liang, Hongan Zhang, Mahdi Aouchiche
Checking data results - Yanjun Zhu, Yongfeng Liang, Hongan Zhang, Mahdi Aouchiche

**Part 2 - Spark Data Processing and store in MySQL Report**

**Requirements**

Collect, clean, and analyze data and store it in SQL databases.

**Design**

**Movies Rating Analysis:**

By obtaining data from web crawlers, these original data (movies.json and reviews.json) are further cleaned to retain the desired analysis data. For example, MetaSore, UserScore, PostiveScore, NegativeScore and MixedScore are obtained after cleaning the scoring data of movies(Clean_New_Data.json).The data after cleaning is analyzed. First, Weight values of PostiveScore, NegativeScore and MixedScore are processed by implementing Spark. Next, normalization of Metascore and Userscore is performed by using pyspark. Finally, all standardized data are input into the model to calculate the movie recommendation score. By referring to the recommendation score, we can get the popularity of each film or recommend it to the audience.

**Content-based filtering:**

There are 14,488 movies in the dataset, one of the attributes we found common are the genres. Since there are 27 unique genres in the dataset. We use tf-idf to quantify the importance of a genre and construct a feature vector for each movie. Next, we use cosine similarity to compute a movie with every other movie in the database to get a list of top 10 similar movies.

**Review Sentiment Analysis:**

There are a total of 575,057 approximate reviews which include critic and user reviews. We want to analyze the sentiment of each text by calculating its polarity score. The polarity score is a float within the range [-1.0, 1.0].

**Implementation**

**This portion is for the movies.json file cleaning and analysis**

There was not much to clean in this file but we used Spark to store the data in a dataframe and then normalize the columns that contain the scores (the user score is on the scale of 0 to 10, and the critics score is on a scale of 0 to 100). We then store the cleaned data into a JSON file. Then, we use pyspark dataframe to further process and analyze the data with operations such as join, remove duplicate, and drop.Some data needs to be normalized. When we got all the data which we needed, we stored the data in the database by using SQL tables finally. Since there are a large number of reviews, we use spark to load the data into a dataframe and for each individual review we calculate the polarity score to further determine if the review has a positive, neutral or negative sentiment.We generate the bar chart to explore the frequency of each genre in the data (see figure below) to discover distributions of different genres for the movie recommendations.

```
from pyspark.sql.functions import col
df = (sparkdf
       .withColumn('ID',col('ID').cast('int'))
       .withColumn("Metascore", col("Metascore").cast("float"))
       .withColumn("Userscore", col("Userscore").cast("float"))
       .withColumn("UserScorePositive", col("UserScorePositive").cast("float"))
       .withColumn("UserScoreMixed", col("UserScoreMixed").cast("float"))
       .withColumn("UserScoreNegative", col("UserScoreNegative").cast("float"))
       .dropna()
       )
```

*Figure: Using pyspark to cast data type*

```
dfPosi=Posi.drop('UserScorePositive','UserScoreMixed','UserScoreNegative','Sum','date')
dfPosi.show()
```

```
+---+--------------------+--------------------+---------+---------+------------------+
| ID|               title|              genres|Metascore|Userscore|          Positive|
+---+--------------------+--------------------+---------+---------+------------------+
|  6|   Three Colors: Red|[Drama, Mystery, ...|    100.0|      8.6|0.9212121212121213|
|  4|          Casablanca|[Drama, Romance, ...|    100.0|      8.9|0.9448818897637795|
|  3|         Rear Window| [Mystery, Thriller]|    100.0|      8.8|0.9358974358974359|
| 12|          Intolerance|    [Drama, History]|     99.0|      7.4|0.7222222222222222|
| 10|         City Lights|[Drama, Comedy, R...|     99.0|      8.9|0.9622641509433962|
|  9| Singin' in the Rain|[Comedy, Romance,...|     99.0|      8.7|0.9037433155080213|
|  8|           Notorious|[Drama, Thriller,...|    100.0|      8.0|0.8648648648648649|
| 18|      Some Like It Hot|   [Comedy, Romance]|     98.0|      8.4|0.9032258064516129|
| 16|The Treasure of t...|[Adventure, Drama...|     98.0|      8.5|0.9207920792079208|
|  1|         Citizen Kane|    [Drama, Mystery]|    100.0|      8.4| 0.872093023255814|
|100|           Lady Bird|     [Drama, Comedy]|     94.0|      7.6|0.7915407854984894|
| 99|         We Were Here|       [Documentary]|     94.0|      6.9|0.6428571428571429|
| 98|       The Gunfighter|           [Western]|     94.0|      7.4|0.8888888888888888|
| 97|      Apocalypse Now|[Action, Drama, War]|     94.0|      8.8| 0.917037037037037|
```

*Figure: Using pyspark to drop columns which is unnecessary  for processing data*

## Math model to calculate recommendation scores

The weight value of user score and Meta score is calculated based on the weighted average, and the recommendation score is calculated based on the weighted sum of user score, Meta score and average of sum of Mixed, Positive and Negative. Because the results of the two scores are considered comprehensively, the recommendation score is more reasonable.

```
Result = normaldf.withColumn(
    "RecommedScore",
     col("NormalizedMetaScore")* 0.50 + col("Userscore")*0.50 + ((col('Mixed')*0.1 + col('Positive')* 0.1+
                                                              col('Negative')*0.1)/3)
    )
```

Recommendation scores = (MetaScore * 0.5 weighted) + (UserScore * 0.5 weighted) +{(Mixed) * 0.1 weighted + Positive * 0.1 weighted + Negative * 0.1 weighted )/3}

**This portion is for the reviews.json file cleaning and analysis**

In “clean_the_reviews.py” file we used spark to clean the reviews.json file, we extracted review_clean as tokens without punctuation and removed stop words from the raw reviews. The first table is the output file which we used to populate the mysql database. We also used to compare the sentiment score with the overall critic and user scores. The second and the 3rd tables represent the critic reviews and user reviews respectively. The row count, max, min, and average values are calculated using spark functions. The results are displayed at the bottom of the tables. The sentiment score calculated does not match each review score but overall we are happy with the result using the raw reviews to calculate the sentiment score, which is very similar to using the clean reviews column.

Show cleaned reviews with sentiments

| release_date | movie_name | review_type | reviewer_name | review_score | review | review_clean | sentiment_score |
|---|---|---|---|---|---|---|---|
| January 12, 1940 | The Shop Around t... | User Review | Unchartedhero | 10 | One of the best s... | one best stylish ... | 0.375 |
| January 12, 1940 | The Shop Around t... | User Review | HaskettFamily | 9 | This has to be on... | one great movies ... | 0.463 |
| July 2, 2021 | Summer of Soul (.... | User Review | RatedRex | 10 | Easily one of the... | easily one best m... | 0.472 |
| July 2, 2021 | Summer of Soul (.... | User Review | Faxthtax | 5 | If this gets a DV... | gets dvd release ... | 0.038 |
| July 2, 2021 | Summer of Soul (.... | User Review | churchman58 | 10 | One word....WOW! ... | one wordwow time ... | 0.213 |

only showing top 5 rows

Show cleaned critic reviews with sentiments

| release_date | movie_name | review_type | reviewer_name | review_score | review | review_clean | sentiment_score |
|---|---|---|---|---|---|---|---|
| January 17, 1940 | Gone with the Wind | Critic Review | null | 100 | A towering landma... | towering landmark... | 0.567 |
| January 17, 1940 | Gone with the Wind | Critic Review | null | 100 | Well, even if it ... | well even essenti... | -0.077 |
| January 17, 1940 | Gone with the Wind | Critic Review | null | 100 | One of the truly ... | one truly great f... | 0.378 |
| January 17, 1940 | Gone with the Wind | Critic Review | null | 100 | Some elements see... | elements seem gro... | -0.177 |
| January 17, 1940 | Gone with the Wind | Critic Review | null | 100 | To see Gone With ... | see gone wind big... | 0.0 |

only showing top 5 rows

Show cleaned user reviews with sentiments

| release_date | movie_name | review_type | reviewer_name | review_score | review | review_clean | sentiment_score |
|---|---|---|---|---|---|---|---|
| January 12, 1940 | The Shop Around t... | User Review | Unchartedhero | 10 | One of the best s... | one best stylish ... | 0.375 |
| January 12, 1940 | The Shop Around t... | User Review | HaskettFamily | 9 | This has to be on... | one great movies ... | 0.463 |
| July 2, 2021 | Summer of Soul (.... | User Review | RatedRex | 10 | Easily one of the... | easily one best m... | 0.472 |
| July 2, 2021 | Summer of Soul (.... | User Review | Faxthtax | 5 | If this gets a DV... | gets dvd release ... | 0.038 |
| July 2, 2021 | Summer of Soul (.... | User Review | churchman58 | 10 | One word....WOW! ... | one wordwow time ... | 0.213 |

only showing top 5 rows

This results in this part are processed using 2 spark threads:

```
Total crawled reviews   = 575057
Total cleaned reviews   = 575057
Total critic reviews    = 258238
Total user reviews      = 316819


Min sentiment score     = -1.0
Max sentiment score     = 1.0
Average sentiment score = 0.14253025363970745


Min critics score       = 0.0
Max critics score       = 100.0
Average critics score   = 62.49895832526584


Min  user score         = 0.0
Max  user score         = 10.0
Average  user score     = 6.589424876664594


SPARK using 2 threads:


        Spark data cleaning time        = 42.917057037353516 seconds
        Spark calculation time          = 334.89447498321533 seconds
        Total time for spark processing = 377.81153202056885 seconds
```

The next result is the same file run again using only 1 thread:

```
Total crawled reviews   = 575057
Total cleaned reviews   = 575057
Total critic reviews    = 258238
Total user reviews      = 316819


Min sentiment score     = -1.0
Max sentiment score     = 1.0
Average sentiment score = 0.14253025363970745


Min critics score       = 0.0
Max critics score       = 100.0
Average critics score   = 62.49895832526584


Min  user score         = 0.0
Max  user score         = 10.0
Average  user score     = 6.589424876664594


SPARK using 1 thread:


        Spark data cleaning time        = 44.57609152793884 seconds
        Spark calculation time          = 341.2775456905365 seconds
        Total time for spark processing = 385.85363721847534 seconds
```

In our case the data is not big enough to illustrate a large gap in the run time between spark using 2 threads or 1 thread, which is usually the case. For both our runs it took approximately 6 min and 45 seconds to clean and process the data.

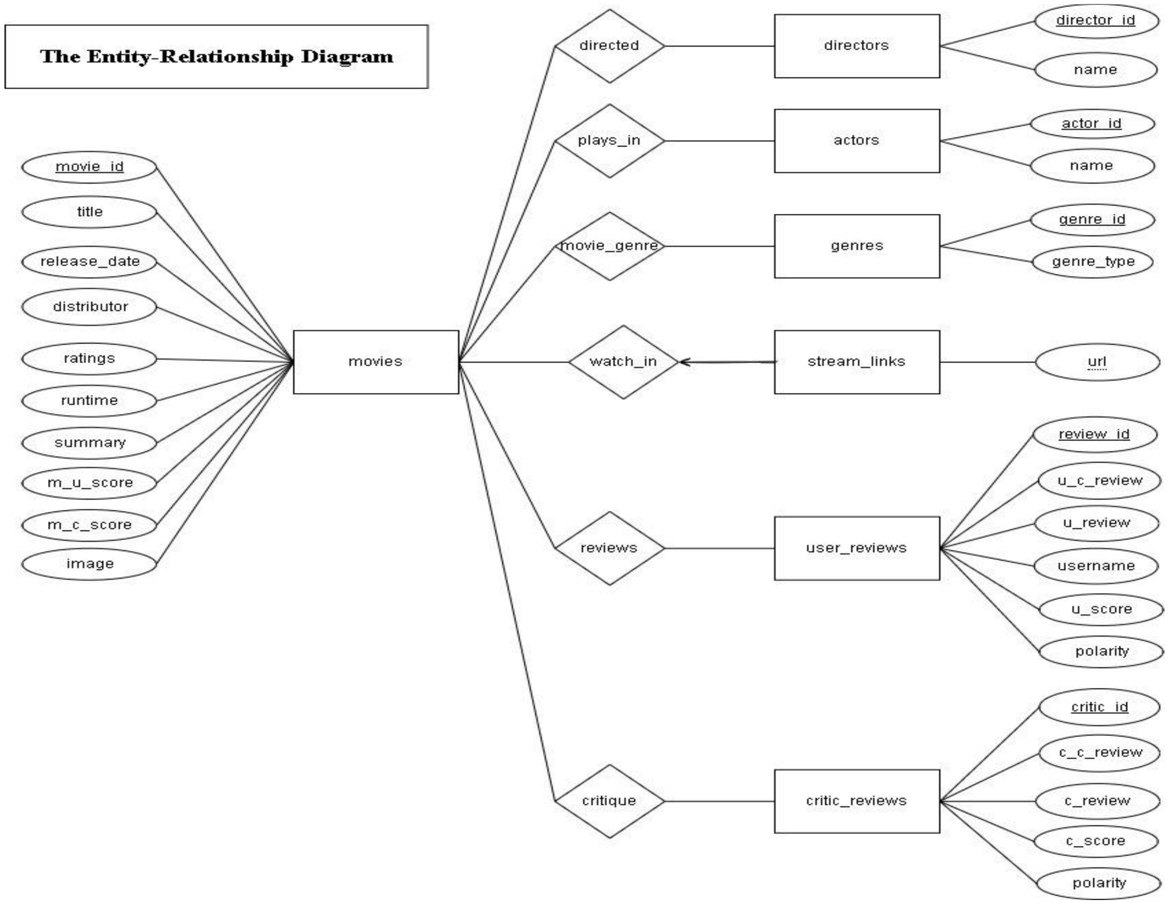
## Evaluation

By calculating the recommendation score output by the model, we can see that it is generally consistent with user score and Meta score, but not very accurate with each review. The recommendation score calculated by the model is more reasonable, mainly because the weight of input value is taken into comprehensive consideration.The weight value of user score and Meta score is calculated based on the weighted average, and the recommendation score is calculated based on the weighted sum of user score and Meta score. Because the results of the two scores are considered comprehensively, the recommendation score is more reasonable.

## Screenshots

### SQL relational diagram:

We designed the database to be able to handle all the crawled data and the different attributes for each movie. Each movie has a different amount of its attributes and so we had to make many tables to accommodate the queries for the user interface. The database is created and populated by the "Create_SQL_Database.py" which deletes the existing tables when the file is run and creates them fresh again and then populated. The cs179g mysql database contains 12 tables in total as shown in the diagram below which has 1 weak entity "stream links".




The MySQL database creation took approximately 22 min to download, some attributes such as the number of actors, directors, movies and reviews are displayed below.

```
Total time to create MySQL DB   = 1288.4311833381653 s
Number of actors         = 33654
Number of directors      = 7680
Number of genres         = 27
Number of user reviews   = 316731
Number of critic reviews = 257758
Number of total reviews  = 600523
Number of movies         = 14488
Number of movies w/o release date = 89
(your_venv) ubuntu@ip-172-31-20-178:~/users/maouc001$
```

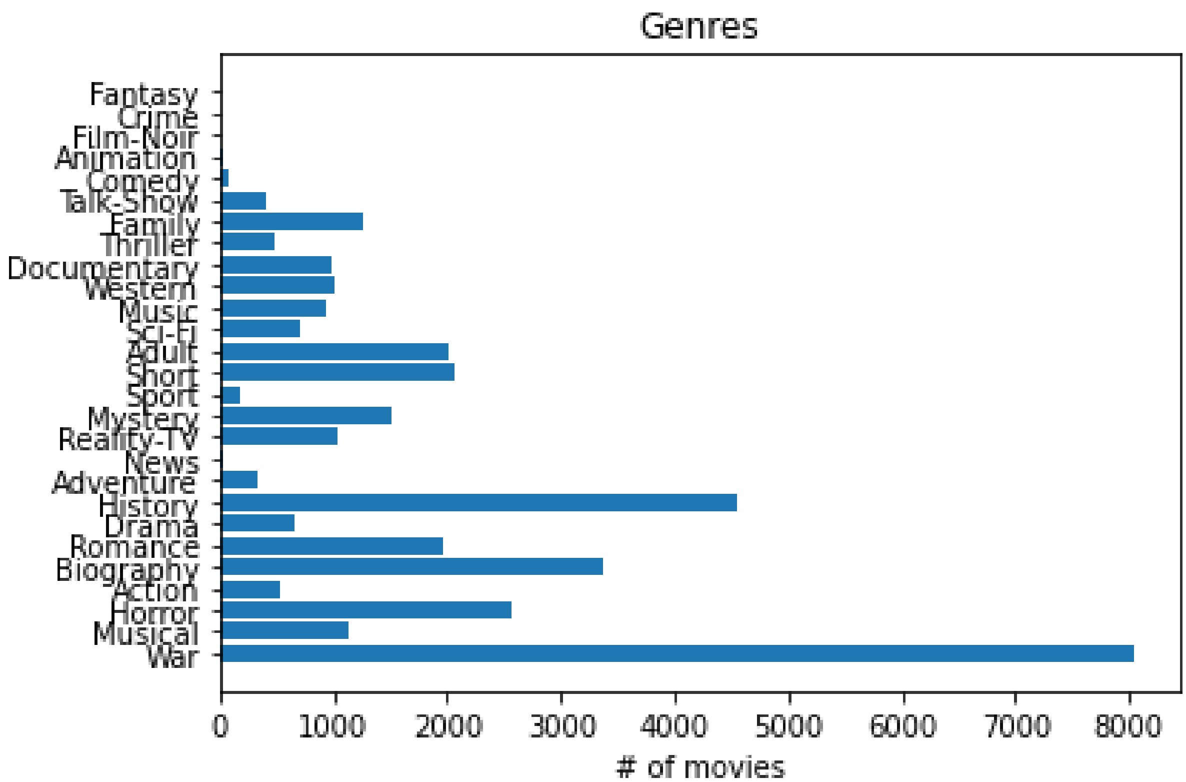



*Figure: The frequency of Each Genre in the data*

```
[274] from pyspark.sql import SparkSession
      spark=SparkSession.builder.appName('Sc').getOrCreate()
```

```
sparkdf =spark.createDataFrame(df)
sparkdf.printSchema()
sparkdf.show()
```

```
root
 |-- ID: long (nullable = true)
 |-- title: string (nullable = true)
 |-- genres: array (nullable = true)
 |    |-- element: string (containsNull = true)
 |-- date: string (nullable = true)
 |-- Metascore: string (nullable = true)
 |-- Userscore: string (nullable = true)
 |-- UserScorePositive: string (nullable = true)
 |-- UserScoreMixed: string (nullable = true)
 |-- UserScoreNegative: string (nullable = true)

+---+-----------------+--------------------+-----------------+---------+---------+-----------------+--------------+-----------------+
| ID|            title|              genres|             date|Metascore|Userscore|UserScorePositive|UserScoreMixed|UserScoreNegative|
+---+-----------------+--------------------+-----------------+---------+---------+-----------------+--------------+-----------------+
|  6| Three Colors: Red|[Drama, Mystery, ...|November 23, 1994|      100|      8.6|              152|             4|                9|
|  5|          Boyhood|             [Drama]|    July 11, 2014|      100|      7.5|            1,869|           274|              354|
```

*Figure: Using pyspark to implement data* e + Text

RAM
Disk

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 3 | Rear Window | [Mystery, Thriller] | September 1, 1954 | 100 | 8.8 | 365 | 15 | 10 |
| 2 | The Godfather | [Drama, Thriller,... | March 11, 1972 | 100 | 9.2 | 3,293 | 93 | 115 |
| 12 | Intolerance | [Drama, History] | September 5, 1916 | 99 | 7.4 | 13 | 1 | 4 |
| 11 | Moonlight | [Drama] | October 21, 2016 | 99 | 7.1 | 1,244 | 175 | 317 |
| 10 | City Lights | [Drama, Comedy, R... | March 7, 1931 | 99 | 8.9 | 102 | 1 | 3 |
| 9 | Singin' in the Rain | [Comedy, Romance,... | April 11, 1952 | 99 | 8.7 | 169 | 13 | 5 |
| 8 | Notorious | [Drama, Thriller,... | September 6, 1946 | 100 | 8.0 | 64 | 7 | 3 |
| 18 | Some Like It Hot | [Comedy, Romance] | March 29, 1959 | 98 | 8.4 | 140 | 11 | 4 |
| 17 | Pan's Labyrinth | [Drama, Mystery, ... | December 29, 2006 | 98 | 8.7 | 2,414 | 188 | 150 |
| 16 | The Treasure of t... | [Adventure, Drama... | January 24, 1948 | 98 | 8.5 | 93 | 6 | 2 |
| 1 | Citizen Kane | [Drama, Mystery] | September 4, 1941 | 100 | 8.4 | 600 | 52 | 36 |
| 100 | Lady Bird | [Drama, Comedy] | November 3, 2017 | 94 | 7.6 | 786 | 117 | 90 |
| 99 | We Were Here | [Documentary] | September 9, 2011 | 94 | 6.9 | 27 | 6 | 9 |
| 98 | The Gunfighter | [Western] | June 23, 1950 | 94 | 7.4 | 8 | 1 | 0 |
| 97 | Apocalypse Now | [Action, Drama, War] | August 15, 1979 | 94 | 8.8 | 619 | 35 | 21 |
| 96 | The Apartment | [Drama, Comedy, R... | June 15, 1960 | 94 | 8.8 | 40 | 0 | 1 |
| 95 | Meet Me in St. Louis | [Drama, Comedy, R... | January 1, 1945 | 94 | 6.0 | 2 | 2 | 1 |

only showing top 20 rows

```
from pyspark.sql.functions import col
df = (sparkdf
          .withColumn('ID',col('ID').cast('int'))
          .withColumn("Metascore", col("Metascore").cast("float"))
```

✓ 1s completed at 2:42 PM

*Figure: Display Normalized Mata scores and User Scores*

**Part 3 - Web Page Frontend**

**Requirements**

Build Web interface using a web programming framework (Django) to explore the preprocessed or analyzed data. For other front-end related web design techniques, we used HTML as well.

**Design**

The web interface will have a Home page. It has a Home button, a search button, a reset button, 2 drop down menus and a text input box. The goal of this home page is to ask the user to select a category of movie search between movie name, actor, distributor, genre, and director that they want to search from. They can choose to order the movies by the release date (default display) or descending order by user or critic score. Then, enter the keyword(s) in the input box and hit the search button. This will redirect the user to a search result page.

In the result page, There is a home button on top that redirects back to the home page. Below it is a text that tells how many results are returned. All the relevant movies are displayed in rows. It shows the information of each movie including the image, title, metascore, user score, summary, genre, distributor, etc. Additionally, there is a button under each movie's title called "Show more info". Once the user clicks on any of those buttons it will redirect to the movie's detailed page.

In the movie's detailed page, more information is shown in addition to the result page. The user can see the stream links, critic reviews, user reviews and the recommended similar movies in this page.

**Implementation**

First, we follow the TA's steps to set up Django, start the web interface project, and modify some of the files.

Next, we create a home page using the template provided by the TA. Then, we modify the HTML code to add a background image, a few buttons, and inputs.

Finally in the views.py, we create two functions: search and more that correspond to two web pages. The home page calls the search function and the search function calls the more function. The structure of these two functions are similar to lab8's which composes MySQL queries and input these queries to the MySQL database. We hardcode most of the HTML codes in strings as well as to organize and display the results returned from the database. We then return the complete HTML code to Django to form a webpage.

## Evaluation

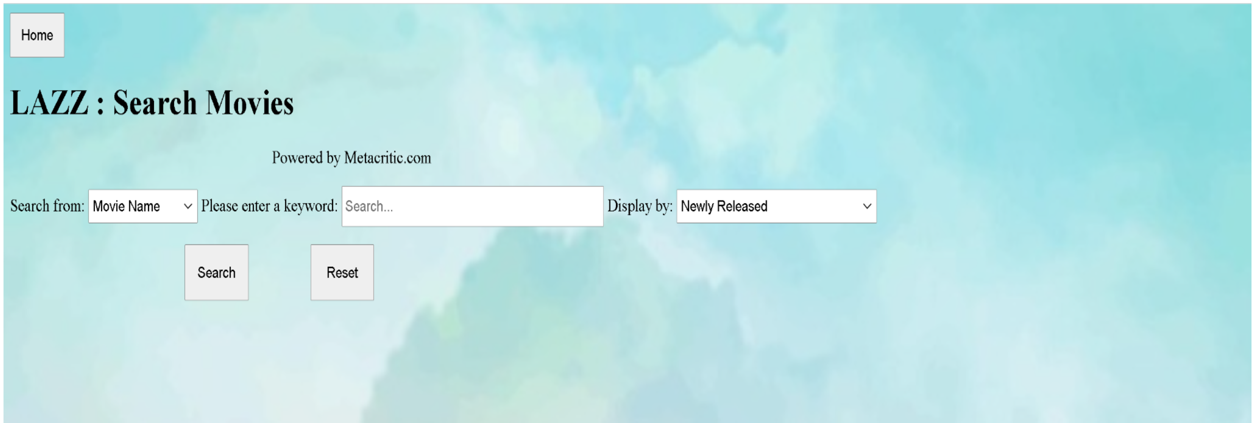
The data are directly retrieved from the database tables to display to the front end which can reduce the time to load the webpage. The overall time to road the search result page… seconds

The search using some non alphabetic and non numeric characters still creates the issue of SQL injection is still an issue, we tried to overcome it by checking the user input in the search using a regex to eliminate some characters and display an error message to the user. We also tried to take advantage of the Django csrf Middleware parameter and the {% csrf_token %} after a user input to catch some of the sql injection types.

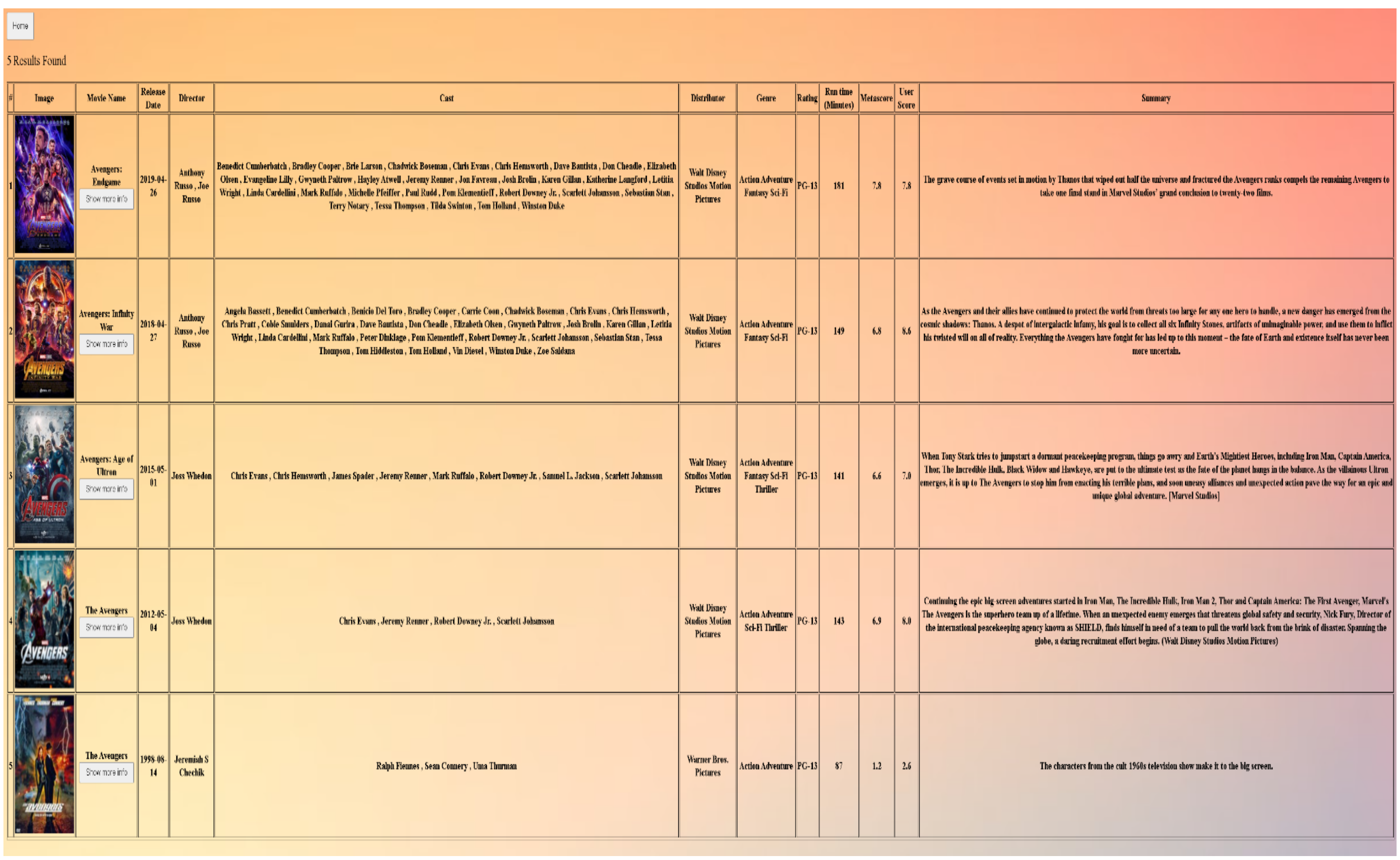
The search by each category works fine and returns correct results. Ideally we wanted to display the available movie genre available when searching by the genre, but we did not implement this option, so the user has to type all or part of the movie genre to get results which is the same as all other searches.

## Screenshots

Home page with various of searching options:




Result of the search displayed by release date in descending order: (if none selected it is used by default)



Home

5 Results Found

| # | Image | Movie Name | Release Date | Director | Cast | Distributor | Genre | Rating | Run time (Minutes) | Metascore | User Score | Summary |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | Avengers: Endgame Show more info | 2019-04-26 | Anthony Russo , Joe Russo | Benedict Cumberbatch , Bradley Cooper , Brie Larson , Chadwick Boseman , Chris Evans , Chris Hemsworth , Dave Bautista , Don Cheadle , Elizabeth Olsen , Evangeline Lilly , Gwyneth Paltrow , Hayley Atwell , Jeremy Renner , Jon Favreau , Josh Brolin , Karen Gillan , Katherine Langford , Letitia Wright , Linda Cardellini , Mark Ruffalo , Michelle Pfeiffer , Paul Rudd , Pom Klementieff , Robert Downey Jr. , Scarlett Johansson , Sebastian Stan , Terry Notary , Tessa Thompson , Tilda Swinton , Tom Holland , Winston Duke | Walt Disney Studios Motion Pictures | Action Adventure Fantasy Sci-Fi | PG-13 | 181 | 7.8 | 7.8 | The grave course of events set in motion by Thanos that wiped out half the universe and fractured the Avengers ranks compels the remaining Avengers to take one final stand in Marvel Studios' grand conclusion to twenty-two films. |
| 2 | | Avengers: Infinity War Show more info | 2018-04-27 | Anthony Russo , Joe Russo | Angela Bassett , Benedict Cumberbatch , Benicio Del Toro , Bradley Cooper , Carrie Coon , Chadwick Boseman , Chris Evans , Chris Hemsworth , Chris Pratt , Cobie Smulders , Danai Gurira , Dave Bautista , Don Cheadle , Elizabeth Olsen , Gwyneth Paltrow , Josh Brolin , Karen Gillan , Letitia Wright , Linda Cardellini , Mark Ruffalo , Peter Dinklage , Pom Klementieff , Robert Downey Jr. , Scarlett Johansson , Sebastian Stan , Tessa Thompson , Tom Hiddleston , Tom Holland , Vin Diesel , Winston Duke , Zoe Saldana | Walt Disney Studios Motion Pictures | Action Adventure Fantasy Sci-Fi | PG-13 | 149 | 6.8 | 8.6 | As the Avengers and their allies have continued to protect the world from threats too large for any one hero to handle, a new danger has emerged from the cosmic shadows: Thanos. A despot of intergalactic infamy, his goal is to collect all six Infinity Stones, artifacts of unimaginable power, and use them to inflict his twisted will on all of reality. Everything the Avengers have fought for has led up to this moment – the fate of Earth and existence itself has never been more uncertain. |
| 3 | | Avengers: Age of Ultron Show more info | 2015-05-01 | Joss Whedon | Chris Evans , Chris Hemsworth , James Spader , Jeremy Renner , Mark Ruffalo , Robert Downey Jr. , Samuel L. Jackson , Scarlett Johansson | Walt Disney Studios Motion Pictures | Action Adventure Fantasy Sci-Fi Thriller | PG-13 | 141 | 6.6 | 7.0 | When Tony Stark tries to jumpstart a dormant peacekeeping program, things go awry and Earth's Mightiest Heroes, including Iron Man, Captain America, Thor, The Incredible Hulk, Black Widow and Hawkeye, are put to the ultimate test as the fate of the planet hangs in the balance. As the villainous Ultron emerges, it is up to The Avengers to stop him from enacting his terrible plans, and soon uneasy alliances and unexpected action pave the way for an epic and unique global adventure. [Marvel Studios] |
| 4 | | The Avengers Show more info | 2012-05-04 | Joss Whedon | Chris Evans , Jeremy Renner , Robert Downey Jr. , Scarlett Johansson | Walt Disney Studios Motion Pictures | Action Adventure Sci-Fi Thriller | PG-13 | 143 | 6.9 | 8.0 | Continuing the epic big-screen adventures started in Iron Man, The Incredible Hulk, Iron Man 2, Thor and Captain America: The First Avenger, Marvel's The Avengers is the superhero team up of a lifetime. When an unexpected enemy emerges that threatens global safety and security, Nick Fury, Director of the international peacekeeping agency known as SHIELD, finds himself in need of a team to pull the world back from the brink of disaster. Spanning the globe, a daring recruitment effort begins. (Walt Disney Studios Motion Pictures) |
| 5 | | The Avengers Show more info | 1998-08-14 | Jeremiah S Chechik | Ralph Fiennes , Sean Connery , Uma Thurman | Warner Bros. Pictures | Action Adventure | PG-13 | 87 | 1.2 | 2.6 | The characters from the cult 1960s television show make it to the big screen. |

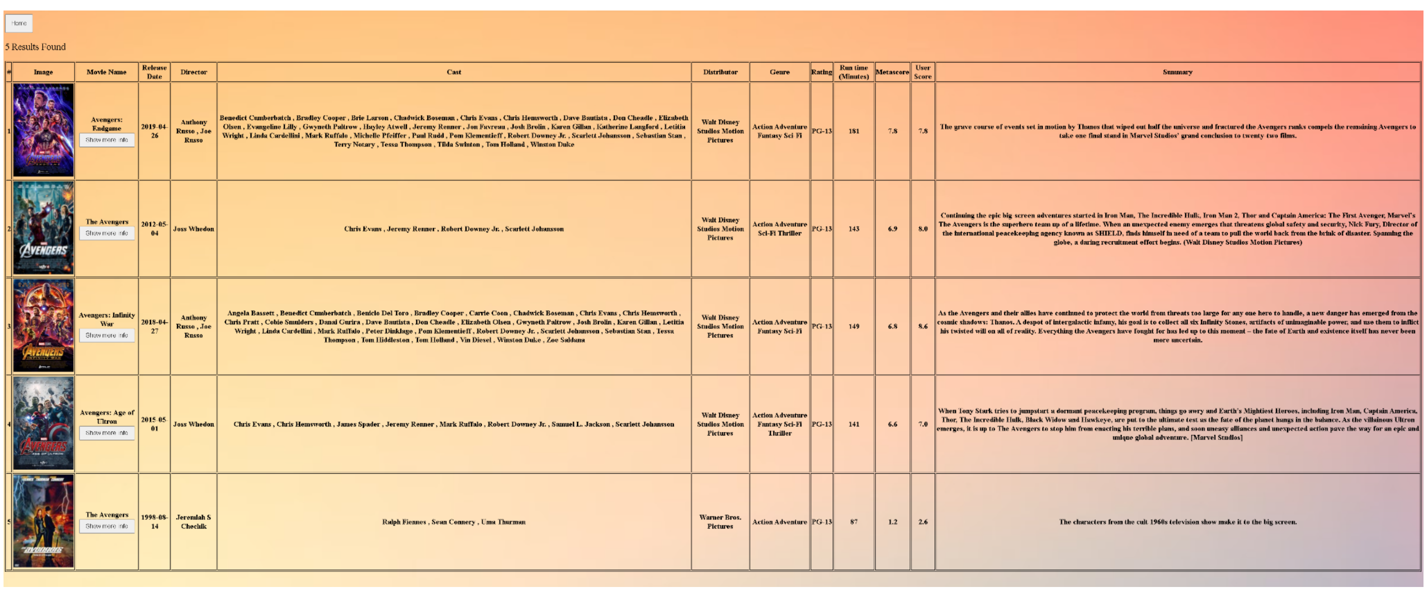
Result of the search displayed by user score in descending order:

Home

5 Results Found

| # | Image | Movie Name | Release Date | Director | Cast | Distributor | Genre | Rating | Run time (Minutes) | Metascore | User Score | Summary |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | Avengers: Infinity War Show more info | 2018-04-27 | Anthony Russo , Joe Russo | Angela Bassett , Benedict Cumberbatch , Benicio Del Toro , Bradley Cooper , Carrie Coon , Chadwick Boseman , Chris Evans , Chris Hemsworth , Chris Pratt , Cobie Smulders , Danai Gurira , Dave Bautista , Don Cheadle , Elizabeth Olsen , Gwyneth Paltrow , Josh Brolin , Karen Gillan , Letitia Wright , Linda Cardellini , Mark Ruffalo , Peter Dinklage , Pom Klementieff , Robert Downey Jr. , Scarlett Johansson , Sebastian Stan , Tessa Thompson , Tom Hiddleston , Tom Holland , Vin Diesel , Winston Duke , Zoe Saldana | Walt Disney Studios Motion Pictures | Action Adventure Fantasy Sci-Fi | PG-13 | 149 | 6.8 | 8.6 | As the Avengers and their allies have continued to protect the world from threats too large for any one hero to handle, a new danger has emerged from the cosmic shadows: Thanos. A despot of intergalactic infamy, his goal is to collect all six Infinity Stones, artifacts of unimaginable power, and use them to inflict his twisted will on all of reality. Everything the Avengers have fought for has led up to this moment – the fate of Earth and existence itself has never been more uncertain. |
| 2 | | The Avengers Show more info | 2012-05-04 | Joss Whedon | Chris Evans , Jeremy Renner , Robert Downey Jr. , Scarlett Johansson | Walt Disney Studios Motion Pictures | Action Adventure Sci-Fi Thriller | PG-13 | 143 | 6.9 | 8.0 | Continuing the epic big-screen adventures started in Iron Man, The Incredible Hulk, Iron Man 2, Thor and Captain America: The First Avenger, Marvel's The Avengers is the superhero team up of a lifetime. When an unexpected enemy emerges that threatens global safety and security, Nick Fury, Director of the international peacekeeping agency known as SHIELD, finds himself in need of a team to pull the world back from the brink of disaster. Spanning the globe, a daring recruitment effort begins. (Walt Disney Studios Motion Pictures) |
| 3 | | Avengers: Endgame Show more info | 2019-04-26 | Anthony Russo , Joe Russo | Benedict Cumberbatch , Bradley Cooper , Brie Larson , Chadwick Boseman , Chris Evans , Chris Hemsworth , Dave Bautista , Don Cheadle , Elizabeth Olsen , Evangeline Lilly , Gwyneth Paltrow , Hayley Atwell , Jeremy Renner , Jon Favreau , Josh Brolin , Karen Gillan , Katherine Langford , Letitia Wright , Linda Cardellini , Mark Ruffalo , Michelle Pfeiffer , Paul Rudd , Pom Klementieff , Robert Downey Jr. , Scarlett Johansson , Sebastian Stan , Terry Notary , Tessa Thompson , Tilda Swinton , Tom Holland , Winston Duke | Walt Disney Studios Motion Pictures | Action Adventure Fantasy Sci-Fi | PG-13 | 181 | 7.8 | 7.8 | The grave course of events set in motion by Thanos that wiped out half the universe and fractured the Avengers ranks compels the remaining Avengers to take one final stand in Marvel Studios' grand conclusion to twenty-two films. |
| 4 | | Avengers: Age of Ultron Show more info | 2015-05-01 | Joss Whedon | Chris Evans , Chris Hemsworth , James Spader , Jeremy Renner , Mark Ruffalo , Robert Downey Jr. , Samuel L. Jackson , Scarlett Johansson | Walt Disney Studios Motion Pictures | Action Adventure Fantasy Sci-Fi Thriller | PG-13 | 141 | 6.6 | 7.0 | When Tony Stark tries to jumpstart a dormant peacekeeping program, things go awry and Earth's Mightiest Heroes, including Iron Man, Captain America, Thor, The Incredible Hulk, Black Widow and Hawkeye, are put to the ultimate test as the fate of the planet hangs in the balance. As the villainous Ultron emerges, it is up to The Avengers to stop him from enacting his terrible plans, and soon uneasy alliances and unexpected action pave the way for an epic and unique global adventure. [Marvel Studios] |
| 5 | | The Avengers Show more info | 1998-08-14 | Jeremiah S Chechik | Ralph Fiennes , Sean Connery , Uma Thurman | Warner Bros. Pictures | Action Adventure | PG-13 | 87 | 1.2 | 2.6 | The characters from the cult 1960s television show make it to the big screen. |

Result of the search displayed by metacritics score in descending order:



Home

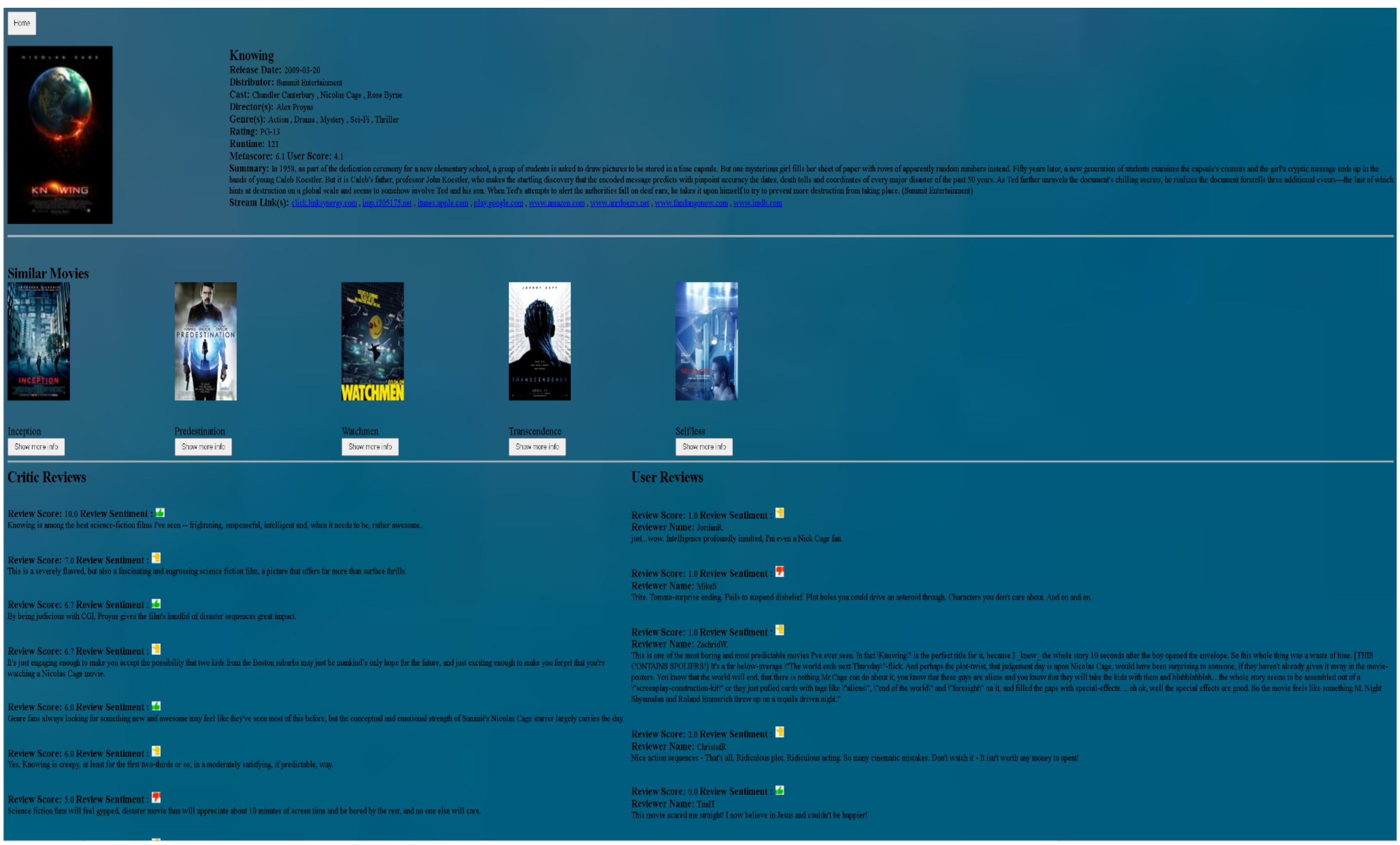
5 Results Found

| # | Image | Movie Name | Release Date | Director | Cast | Distributor | Genre | Rating | Run time (Minutes) | Metascore | User Score | Summary |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | Avengers: Endgame Show more info | 2019-04-26 | Anthony Russo , Joe Russo | Benedict Cumberbatch , Bradley Cooper , Brie Larson , Chadwick Boseman , Chris Evans , Chris Hemsworth , Dave Bautista , Don Cheadle , Elizabeth Olsen , Evangeline Lilly , Gwyneth Paltrow , Hayley Atwell , Jeremy Renner , Jon Favreau , Josh Brolin , Karen Gillan , Katherine Langford , Letitia Wright , Linda Cardellini , Mark Ruffalo , Michelle Pfeiffer , Paul Rudd , Pom Klementieff , Robert Downey Jr. , Scarlett Johansson , Sebastian Stan , Terry Notary , Tessa Thompson , Tilda Swinton , Tom Holland , Winston Duke | Walt Disney Studios Motion Pictures | Action Adventure Fantasy Sci-Fi | PG-13 | 181 | 7.8 | 7.8 | The grave course of events set in motion by Thanos that wiped out half the universe and fractured the Avengers ranks compels the remaining Avengers to take one final stand in Marvel Studios' grand conclusion to twenty-two films. |
| 2 | | The Avengers Show more info | 2012-05-04 | Joss Whedon | Chris Evans , Jeremy Renner , Robert Downey Jr. , Scarlett Johansson | Walt Disney Studios Motion Pictures | Action Adventure Sci-Fi Thriller | PG-13 | 143 | 6.9 | 8.0 | Continuing the epic big-screen adventures started in Iron Man, The Incredible Hulk, Iron Man 2, Thor and Captain America: The First Avenger, Marvel's The Avengers is the superhero team up of a lifetime. When an unexpected enemy emerges that threatens global safety and security, Nick Fury, Director of the international peacekeeping agency known as SHIELD, finds himself in need of a team to pull the world back from the brink of disaster. Spanning the globe, a daring recruitment effort begins. (Walt Disney Studios Motion Pictures) |
| 3 | | Avengers: Infinity War Show more info | 2018-04-27 | Anthony Russo , Joe Russo | Angela Bassett , Benedict Cumberbatch , Benicio Del Toro , Bradley Cooper , Carrie Coon , Chadwick Boseman , Chris Evans , Chris Hemsworth , Chris Pratt , Cobie Smulders , Danai Gurira , Dave Bautista , Don Cheadle , Elizabeth Olsen , Gwyneth Paltrow , Josh Brolin , Karen Gillan , Letitia Wright , Linda Cardellini , Mark Ruffalo , Peter Dinklage , Pom Klementieff , Robert Downey Jr. , Scarlett Johansson , Sebastian Stan , Tessa Thompson , Tom Hiddleston , Tom Holland , Vin Diesel , Winston Duke , Zoe Saldana | Walt Disney Studios Motion Pictures | Action Adventure Fantasy Sci-Fi | PG-13 | 149 | 6.8 | 8.6 | As the Avengers and their allies have continued to protect the world from threats too large for any one hero to handle, a new danger has emerged from the cosmic shadows: Thanos. A despot of intergalactic infamy, his goal is to collect all six Infinity Stones, artifacts of unimaginable power, and use them to inflict his twisted will on all of reality. Everything the Avengers have fought for has led up to this moment – the fate of Earth and existence itself has never been more uncertain. |
| 4 | | Avengers: Age of Ultron Show more info | 2015-05-01 | Joss Whedon | Chris Evans , Chris Hemsworth , James Spader , Jeremy Renner , Mark Ruffalo , Robert Downey Jr. , Samuel L. Jackson , Scarlett Johansson | Walt Disney Studios Motion Pictures | Action Adventure Fantasy Sci-Fi Thriller | PG-13 | 141 | 6.6 | 7.0 | When Tony Stark tries to jumpstart a dormant peacekeeping program, things go awry and Earth's Mightiest Heroes, including Iron Man, Captain America, Thor, The Incredible Hulk, Black Widow and Hawkeye, are put to the ultimate test as the fate of the planet hangs in the balance. As the villainous Ultron emerges, it is up to The Avengers to stop him from enacting his terrible plans, and soon uneasy alliances and unexpected action pave the way for an epic and unique global adventure. [Marvel Studios] |
| 5 | | The Avengers Show more info | 1998-08-14 | Jeremiah S Chechik | Ralph Fiennes , Sean Connery , Uma Thurman | Warner Bros. Pictures | Action Adventure | PG-13 | 87 | 1.2 | 2.6 | The characters from the cult 1960s television show make it to the big screen. |

Selecting a movie from the search results using the ‘show more info’ button will display the info page below which contains the complete info for the selected movie. It also displays up to 5 recommended movies based on the genres of the selected movie. After the recommended movies, the info page displays the reviews given by the critics (on the left) and the users (on the right). Each review displays its score if available, otherwise N/A is displayed. Next to the review score, we display the sentiment analysis of the reviews in 3 categories. Positive reviews are represented by a thumbs up, sentiment polarity is in range: [0.3, 1.0]. Neutral reviews are represented by a neutral thumb, sentiment polarity is in range: [-0.3, +0.3] . Negative reviews are represented by a thumb down, sentiment polarity is in range: [-1.0, - 0.3] .

The available stream links to watch the movie are provided as the website where to watch the movie, clicking on the preferred website will pop a new tab where the user is directed to watch the movie.

The recommended movies also have a show more button which will display the movie info and stream links on this page if the user chooses a recommended movie. And in turn it will populate other movie recommendations again and the selected movie critic and user reviews information with scores and sentiments.




Challenges:

_ The time is limited to only display a number of movies and a number of reviews, So we implemented the option to show all the movies returned by the search and all the movie reviews, the user can scroll down to see everything that is requested.

**Contribution**

Store analyzed data in MySQL: Mahdi Aouchiche, Yongfeng Liang, Hongan Zhang, Yanjun Zhu

Web Interface Design: Mahdi Aouchiche, Yongfeng Liang, Hongan Zhang, Yanjun Zhu

Web Interface Implementation: Mahdi Aouchiche, Yongfeng Liang, Hongan Zhang, Yanjun Zhu

To run the project:

| | |
|---|---|
| Go to the directory: | cd final_submit/part3/mysite |
| Run the commends: | python3 manage.py migrate |
| | python3 manage.py runserver 0.0.0.0:8080 |
| Then open a web browser: | http://cs179g-fall-2021-08.cs.ucr.edu:8080/ |