Group LLZ(21)

Rui Li

Qiulin Luo

Weikun Zhuang

**Final Project Report**

**Task and Its Importance**

In life, we often encounter situations where we can only remember part of a film, such as an actor who participated in the film, a scene or a classic line, but can't remember the name of the film. In this case, we usually want to get the name of the film by simply describing some relevant information about the film and putting it in the search engine. In this project, our group will be implementing a system to query film names through relevant information of the film, such as scene descriptions, film content keywords, etc.

The queries for this search engine system can be continuous descriptive sentences or keywords separated by spaces. The continuous sentences can be more convenient for retrieving movie lines and movie scenes, while the keywords can be queried for movie genres, etc. The system will return a series of ordered files with related movie names according to the query content, and show the plot summary.

The part of the system that we’ll focus on implementing will be retrieval model optimization. Specifically, we are interested in exploring how to effectively extract keywords from a query that define some of the important elements of a movie that the user is looking for. This extracted information is then retrieved at a deeper and more accurate level using an improved retrieval model. Since the amount of details contained in queries can vary based on the completeness of the user's recollections, we will explore how to process the information to better facilitate retrieving search results. After implementation, we will also make sure to carefully construct criteria for judging relevance of the search results in order to evaluate our retrieval model.

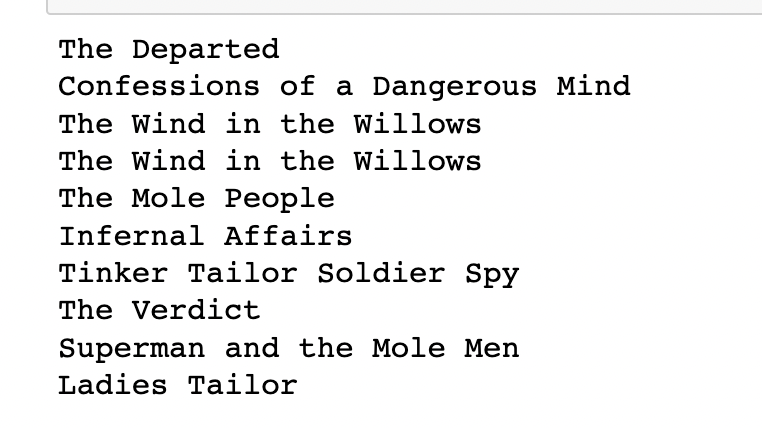
When users only input some key information, this system can return the information about the movies matching this information, including the movie name, and a story summary provided by Wikipedia. And we sort the returned movie according to the correlation between the movie and the input information, Therefore, our model can quickly help users find the movies and related information they want.

The collection of documents that we will draw results from comes from Kaggle and the dataset is titled [“Wikipedia Movie Plot”](https://www.kaggle.com/datasets/jrobischon/wikipedia-movie-plots) . It contains information about movies that are from multiple countries and span more than a century. The dataset provides ample information about plot, genre, cast, director, etc. for us to process and index. However, with that said, this dataset doesn’t include movies for recent years and due to the sheer amount of movies around the world, this is more of an exploration of movie information retrieval.

**Query & Search Result**

The query input is related information of the film, which can be a certain plot of the film, such as weightlessness fighting in the corridor; it can be a plot summary, such as entering other people’s dreams to steal secrets and reshape their dreams; it can also contain the names of the leading actors or film types, such as action, adventure, suspense, thriller, science fiction; it can be related to a tagline that has appeared in the film, or the name of the director, screenwriter or producer; the film distribution company, etc. It can also be a combination of two or more of the above, such as film type and the director ( action, adventure, science fiction and director is Christopher Nolan).

The search result would include the movie title, such as “Inception”; and the storyline provided by Wikipedia for users to check if this is indeed related to the search intended in the first place. For a more detailed example before we go into details of result organization, user can input “mole in police” and the system will return something similar to the following:



For the results above, the top result is the most relevant document since in the movie “The Departed”, one of the main characters is an undercover spy inside the Massachusetts State Police.

The sample queries we created for evaluation will take into account searching for a specific movie (one relevant result) and some other searches that have more than one relevant answer. Some queries are exclusively about the plot but some will add information about cast, genre, cast, etc. However, Annotated results are not exhaustive. Only scanning some of the results in the dataset.

**Result Organization:**

The search results were ranked by relevance in decreasing order, The retrieved list of documents are structured snippets that tidily display the movie title and plot (containing relevant keywords/summaries/synopsis/storyline).

The query results will show users the top 10 documents most relevant to the input content according to the calculated correlation each time (The number of documents displayed is also adjustable). Depending on the query, The movie documents in the result list are sorted according to their relevance to input content. In other words, depending on the clarity of the user’s input information, the results that the user wants may not appear at the top of the ranked list or even when the input information is less or too vague, the results the user wants may not appear in the result list.

**Evaluation Metrics**

* + System should process the queries and retrieve a list of ten documents in a user-friendly way.
  + Documents are organized and contain movie title, and plot descriptions that mention keywords or exact match of the query.
  + Documents are ranked by relevance in decreasing order.
  + Retrieved documents in the ranked list containing documents that match the intent of the query and if there could be more than 1 answer, about half of the documents retrieved should match this intent of the query.
  + Relevant documents should be within the first 5 positions of the retrieved list.
  + System should work for queries for keywords, sentences and the combination of keywords and sentences.
  + The response time for a given query should be a few seconds

**Implementation**

* + Found and collected dataset for movie information from Kaggle
  + Constructed and sourced possible movie info queries
  + Preprocessed the original dataset before further implementation
  + Executed tokenization for each record of the preprocessed data
  + Conducted stemming for each record of the dataset
  + Calculated the term frequency, tf-idf and BM25 values
  + Parameter tuning and optimization of the BM25 model to establish a weighting mechanism.
  + Pre-processing of queries, extraction of relevant information and establishment of scoring mechanisms.
  + Get query results, perform model testing and evaluation.

**Performance**

We used binary relevance judgment when evaluating documents - non-relevant documents are marked as 0 and relevant documents are marked as 1. From the annotations, we computed the MAP@10 values to get the accuracy of the results returned by the query. Since the movie dataset is too large and we weren’t able to find a suitable query-document dataset , we only annotated the binary relevance of the first ten results returned and there could be more relevant documents in the collection (except for the queries that only have one relevant answer). The final calculated value of MAP@10 is 0.828

**Appendix**

To access the dataset and code file, please visit the following link:

<https://drive.google.com/drive/folders/15tbbgZ3ebyIMsx7EvYyof05EWQVL7fVQ?usp=sharing>

**Evaluation Set**

**Query 1**

Input: “Dream within a dream nolan sci-fi”

Description: The user is searching for the name of the 2010 movie “Inception” directed by Christopher Nolan.

Narrative: In this case the relevant document should only be “Inception” since it’s implied here that Nolan refers to the director Chrisptopher Nolan.

Candidate results:

Title | Relevance judgment (0 non relevant, 1 relevant)

Inception 1

A Tree Grows in Brooklyn 0

Walking and Talking 0

Hop Harrigan 0

Dead Poets Society 0

Gangs of Chicago 0

Orca 0

Picture Snatcher 0

The Man Without a Country 0

Interstellar 1

**Query 2**

Input: “psycho killer”

Description: Identify the movies that have a psycho killer as the main antagonist.

Narrative: Relevant documents should involve a psycho killer in their plot. They can be in any genre and they don’t have to be in the psychological/drama/crime categories. Non-relevant documents will not have mentions of killers.

Candidate results:

Title | Relevance judgment (0 non relevant, 1 relevant)

Sawaari 1

Utkarsha 1

Cold Feet 1

Monster 1

Behind the Mask: The Rise of Leslie Vernon 1

Angel 1

Deadball 1

A Fantastic Fear of Everything 1

Cherry Falls 1

The Punisher 1

**Query 3**

Input: “Abusive drummer tempo”

Description: This query is searching for the 2014 movie “Whiplash”. One of the main character is an abusive drummer that terrorize the lead character. One of the taglines is “Not quite my tempo”.

Narrative: The relevant document that the system should retrieve is the movie “Whiplash” because the “tempo” here refers to the famous tagline. Other documents (movies) that simply have drummers of erratic behavior in their plot will not be counted as relevant.

Candidate results:

Title | Relevance judgment (0 non relevant, 1 relevant)

Whiplash 1

Swapaanam 0

That Thing You Do! 0

James Black and the New Treasure and Your Lie in April 0

The Runaways 0

Sankofa 0

Get Him to the Greek 0

Still Crazy 0

Walk Hard: The Dewey Cox Story 0

The Wild Thornberrys Movie 0

**Query 4**

Input: “mole inside boston police ”

Description: This query is searching for the 2006 movie “The Departed”, the plot of which involves a spy in the MSP and the clash of “good and evil”

Narrative: The relevant document should be the movie “departed” since it’s specific about the region of the police department. Other movies should not be considered as relevant document.

Candidate results:

Title | Relevance judgment (0 non relevant, 1 relevant)

The Departed 1

Confessions of a Dangerous Mind 0

The Wind in the Willows 0

The Mole People 0

Infernal Affairs 0

Tinker Tailor Soldier Spy 0

The Verdict 0

Superman and the Mole Men 0

Ladies Tailor 0

The Brink’s Job 0

**Query 5**

Input: “explore another earth for human sci-fi”

Description: User is searching up science fiction movies that involve the narrative that the human race is facing critical situations on earth and they need to find a new home.

Narrative: The relevant document should include references to finding a new earth, no matter what the premises are. Vague mention of space traveling will not be considered as relevant.

Candidate results:

Title | Relevance judgment (0 non relevant, 1 relevant)

Godzilla vs. Mothra 0

Journey to the Center of Time 0

Explorers 1

Alien vs. Predator 1

Rocketship X-M 1

The Shape of Things to Come 1

My Favorite Martian 0

Godzilla vs. Mechagodzilla 0

Avatar 1

The Last Winter 0

**Query 6**

Input: “CIA agents shootings”

Description: Users may be searching for crime thrillers about CIA agents and shooting cases.

Narrative: Films that do not feature CIA agents and do not have a shooter element will not be marked as relevant.

Candidate results:

Title | Relevance judgment (0 non relevant, 1 relevant)

Operation Avalanche 1

Spy 1

Scorpio 1

Bad Company 1

2000 AD 1

Real Men 1

The Ipcress File 1

Mission: Impossible 1

The Delta Factor 1

In the Line of Fire 1

**Query 7**

Input: “nuclear war post-apocalyptic”

Description: Users may want to look for movies about world nuclear war, because nuclear war is about post-apocalyptic elements, so it is also possible to search for the rest of post-apocalyptic related movies.

Narrative: Movies with any depiction of nuclear war or other post-apocalyptic related elements are considered relevant, and movies that are not are very well identified.

Candidate results:

Title | Relevance judgment (0 non relevant, 1 relevant)

One Night Stand 0

WarGames 1

The Space Children 1

The Sum of All Fears 1

Thirteen Days 1

When the Wind Blows 1

Last War! The Last War 1

Steel Rain 1

Red Planet Mars 0

Dance to Death 1

**Query 8**

Input: “Things that happen on vacation”

Description: Users may want to search for movies that have a plot that unfolds on vacation, which can be either light-hearted or crisis-ridden.

Narrative: The movie plot must contain a vacation-related storyline.

Candidate results:

Title | Relevance judgment (0 non relevant, 1 relevant)

The Vanishing 1

Gunturodu 0

Second Honeymoon 1

Naku Penta Naku Taka 0

By Your Leave 1

Suddenly It’s Magic 1

Hailesa 1

Happy Husbands 0

Happy Husbands 0

Rekka 0