

Project Report
A mini search system on Movie review
dataset
Information Retrieval

Submitted

By

Meda Rukmini (S20180010102),

UG3,

Indian Institute of Information Technology, Sri City.

18-12-2020.

The Reel

A mini search system on movie review data set

Abstract

Now a days, watching a movie is one of the first choices for a group of people to hangout. Choosing a movie depends on various factors like reviews, story, actors etc.,. In this project, I collected information from public sources, created a suitable index and efficient retrieval system from scratch. I have explored fundamental algorithms in indexing, retrieving and ranking in information retrieval system. Initially developed with a data set of 1000 documents, it scales pretty well till 10000 documents. A positional index is developed with ranked retrieval. The indexing takes reasonable time for 1000 – 10000 documents but a long time for 100,000 documents.

Data set collection

After exploring many sources like Kaggle, imdb etc., I have chosen two data sets from the following sources.

1. <http://ai.stanford.edu/~amaas/data/sentiment/>

This data set was used for sentiment analysis and was divided into train and test data sets for negative, positive labels and unsupervised data. I have combined all the text documents into a single data set and renamed them sequentially using a python script. This contains one lakh documents.

2. <https://md-datasets-cache-zipfiles-prod.s3.eu-west-1.amazonaws.com/38j8b6s2mx-1.zip/>

This data set contains 1000 reviews, 100 each from 10 different movies. I combined them sequentially into a single data set using a python script.

Parts of the search system

Indexing -> Querying -> Ranking

The project has three parts. They are:

1. Indexing
2. Querying
3. Ranking

Indexing

As an initial step for building the search system, I have implemented positional index from scratch in C++. I used standard template library in accomplishing it. The index has vocabulary, postings and position lists as main parts. It is implemented using a map. It is a map of a string and a pair of integer and binary search tree. Binary search tree has been implemented for postings. A grow able array called vector in standard template library is used for position lists.

Querying

Two types of querying is implemented. They are:

1. One word Queries
2. Free Text Queries

One word Queries

One word queries are pretty straight forward to implement. To avoid storing the index in memory, I have stored the lines at which each term is present to directly fetch the term's postings. I have used a map for it. Using the map, I fetch the line number and read postings from disk at that particular line. Since, number of queries are not very largely expected in this project, this is an efficient method for the system. After reading the line, it is parsed into its components and the respective documents are printed in the console.

Free Text Queries

Free Text Queries are implemented using implementation of one word queries. Each query term's postings are retrieved through one word queries function and the obtained postings of every term are intersected. The obtained list of documents is displayed in the console.

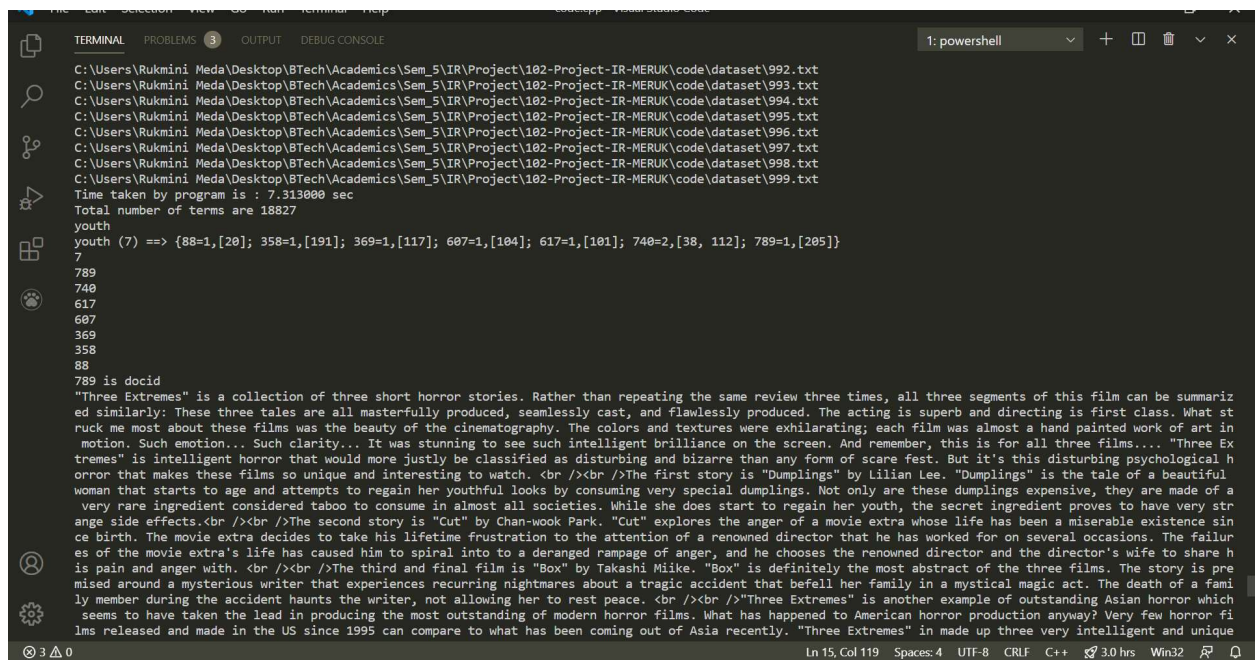
Ranking

I have implemented tf-idf weighting for terms. No libraries (except for STL) are used and completely implemented from scratch in C++. I have implemented cosine similarity for ranking the documents with respect to the query.

Conclusion

The basic algorithms for the index and retrieval system are implemented from scratch. Retrieval works efficiently whereas the indexing takes a long time for one lakh documents due to reasonably large size of each document. It takes nearly 10s for the algorithm to run for 1000 documents and more than an hour for one lakh documents.

Results



```
TERMINAL PROBLEMS 3 OUTPUT DEBUG CONSOLE 1: powershell
C:\Users\Rukmini Mada\Desktop\BTech\Academics\Sem_5\IR\Project\102-Project-IR-MERUK\code\dataset\992.txt
C:\Users\Rukmini Mada\Desktop\BTech\Academics\Sem_5\IR\Project\102-Project-IR-MERUK\code\dataset\993.txt
C:\Users\Rukmini Mada\Desktop\BTech\Academics\Sem_5\IR\Project\102-Project-IR-MERUK\code\dataset\994.txt
C:\Users\Rukmini Mada\Desktop\BTech\Academics\Sem_5\IR\Project\102-Project-IR-MERUK\code\dataset\995.txt
C:\Users\Rukmini Mada\Desktop\BTech\Academics\Sem_5\IR\Project\102-Project-IR-MERUK\code\dataset\996.txt
C:\Users\Rukmini Mada\Desktop\BTech\Academics\Sem_5\IR\Project\102-Project-IR-MERUK\code\dataset\997.txt
C:\Users\Rukmini Mada\Desktop\BTech\Academics\Sem_5\IR\Project\102-Project-IR-MERUK\code\dataset\998.txt
C:\Users\Rukmini Mada\Desktop\BTech\Academics\Sem_5\IR\Project\102-Project-IR-MERUK\code\dataset\999.txt
Time taken by program is : 7.313800 sec
Total number of terms are 18827
youth
youth (7) ==> {88=1,[20]; 358=1,[191]; 369=1,[117]; 607=1,[104]; 617=1,[101]; 740=2,[38, 112]; 789=1,[205]}
7
789
740
617
607
369
358
88
789 is docid
"Three Extremes" is a collection of three short horror stories. Rather than repeating the same review three times, all three segments of this film can be summariz
ed similarly: These three tales are all masterfully produced, seamlessly cast, and flawlessly produced. The acting is superb and directing is first class. What st
ruck me most about these films was the beauty of the cinematography. The colors and textures were exhilarating; each film was almost a hand painted work of art in
motion. Such emotion... Such clarity... It was stunning to see such intelligent brilliance on the screen. And remember, this is for all three films.... "Three Ex
tremes" is intelligent horror that would more justly be classified as disturbing and bizarre than any form of scare fest. But it's this disturbing psychological h
orror that makes these films so unique and interesting to watch. <br /><br />The first story is "Dumplings" by Lilian Lee. "Dumplings" is the tale of a beautiful
woman that starts to age and attempts to regain her youthful looks by consuming very special dumplings. Not only are these dumplings expensive, they are made of a
very rare ingredient considered taboo to consume in almost all societies. While she does start to regain her youth, the secret ingredient proves to have very str
ange side effects.<br /><br />The second story is "Cut" by Chan-wook Park. "Cut" explores the anger of a movie extra whose life has been a miserable existence sin
ce birth. The movie extra decides to take his lifetime frustration to the attention of a renowned director that he has worked for on several occasions. The failur
es of the movie extra's life has caused him to spiral into to a deranged rampage of anger, and he chooses the renowned director and the director's wife to share h
is pain and anger with. <br /><br />The third and final film is "Box" by Takashi Miike. "Box" is definitely the most abstract of the three films. The story is pre
mised around a mysterious writer that experiences recurring nightmares about a tragic accident that befell her family in a mystical magic act. The death of a fami
ly member during the accident haunts the writer, not allowing her to rest peace. <br /><br />"Three Extremes" is another example of outstanding Asian horror which
seems to have taken the lead in producing the most outstanding of modern horror films. What has happened to American horror production anyway? Very few horror fi
lms released and made in the US since 1995 can compare to what has been coming out of Asia recently. "Three Extremes" in made up three very intelligent and unique
```

Fig. Sample output of running the program, includes indexing process and a single one word query.

Index can also be viewed at the output.txt file.

References

1. IR lecture slides
2. Introduction to information retrieval by Christopher Manning
3. <http://www.ardendertat.com/2012/01/11/implementing-search-engines/>