

with running instruction, details about the files in the folder, all text normalization details, shelve timing, packages used, test queries examples, and any other thoughts or concerns you like to share.

Running Instruction

You should run this project with following command

```
python boolean_index.py  
python boolean_query.py
```

Files in the folder

boolean_index.py

The python file to build inverted index and provide search function for the website.

boolean_query.py

The python file to generate and start the backend python flask framework of this website for user to generate the website and interact with our search engine.

corpus.json corpus11.json

A full film data original file and a test file.

corpus.db index_data.db stopWords.db

Database file for corpus, index_data and stopWords

Function in the program

loadJson

Load json file from original data

indexAllData

Function for build posting list

tokenize_stemming

Tokenize and stem input

storeStopWords

Down load stopwords from nltk and store it in our database

class SearchEngine

search

Input search query, return list of movie id.

preprocess

Check and delete the stopwords in the query. Check whether there is any unknown term in the query. return result of realquery, unknown term and stopwords.

findMovieId

Input a query (list of query words), output a list of movie id by intersecting posting list of all query words.

intersect

input two list, return the list of intersect movie id between two list.

get_movie_data

return the data of a movie by given id.

get_movie_snippet

return the title, text of a movie by given id.

shelve timing

Using time.clock() function to record the shelve building time in the main function.

```
if __name__ == '__main__':  
  
    start_time = time.clock()  
    print('Build Start!')  
    data = loadJson(data_path)  
    indexAllData(data)  
    storeStopWords(stopWords_path)  
    end_time = time.clock()  
    print('Build End!')  
    print('Build Time Use ' + str(end_time - start_time) + ' seconds')
```

Output after running

```
Build Start!  
Build End!  
Build Time Use 22.910499 seconds
```

Package Used

boolean_index.py

```
import shelve
import nltk
import json
from nltk.tokenize import RegexpTokenizer
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
import time
```

Here we use nltk.tokenize.RegexpTokenizer to tokenize all the words in the corpus(ignore all special mark '{',' ','!','?').

Use PorterStemmer as stemmer to stem words.

Download nltk.corpus.stopwords as our stopwords list

```
stemmer = PorterStemmer()
tokenizer = RegexpTokenizer(r'\w+')
```

```
def storeStopWords(stopWords_path):
    stopW_db = shelve.open(stopWords_path, writeback=False)
    stopW_db['stopWords'] = set(stopwords.words('english'))
    stopW_db.close()
```

boolean_query.py

```
from flask import *
```

Test queries examples

note: Here is some test query example to run in a full corpus database since I have already tested the correctness on my small test corpus.

2017 Film Search

Query:

Search Results:

Total hits: 467

Ignoring stopwords(s): []

Ignoring unknown word(s):

- 3 idiotas**
 - 3 Idiotas is a 2017 Mexican coming of age comedy–drama, which is a remake of the 2009 Indian film "3 Idiots". 3 Idiots is directed by Carlos Alfonso Dosal, Christian Vazquez, German Valdez and Martha Higareda. The film about two nerdy friends and their arch rival from college who a fun adventure determined to find a college roommate who disappeared without a trace on graduation day. * Alfonso Dosal as Pancho * Chris Isidoro * Sebastin Zurita as Emiliano * Paulina Dvila as Diana * Enrique Singer as Don Diego * Rodrigo Murray as Escalona 3 Idiotas was relea
- 8 Thottakkal**
 - 8 Thottakkal (The film starts with an orphaned boy, Sathya, getting framed by the employer for the murder of his wife and is sent to juvenile. I warden taught him and stays honest in his profession. Unlike the other policemen at his station, he refuses to be involved in bribes and is mad Thinking that Sathya is trying to become a hero by making him look bad, the inspector plans to shame Sathya. The inspector then arranges for loaded with 8 bullets. The inspector gives him one day to find his lost gun or face disciplinary action. With the help of a reporter friend, Meera, the gun attempts a robbery and in the process he accidentally shoots and kills a little girl. Meera, in danger of losing her job tells her superiors be arrested, and is tracked down by Pandian, the new handler of the case, Pandian says that if Sathya was a real policeman he should never s girlfriend a necklace for her birthday and takes all the fresh notes available, despite the boss's opposition. In the jewelry shop, he is exposed a his share. After a confrontation with the boss he tries to steal all the money and the leader shoots and kills him. The remaining two dispose the police attempt to arrest the boss with a ploy the boss shoots and kills the second robber and made sure there is no one left to connect him wi terrorist slip away as a constable. It turns out that Gunasekaran accepted a bribe from the terrorist and let him go, but this has already ruined Murthy sees the man who sold him the gun and leaves. Murthy kills him with the same gun when he follows Murthy to a restroom. However we Provident Fund. Sathya, now looking for clues in the murder, finds Murthy's application sanctioned and goes to his house to give it to him. Hov the family, by connecting the murder of the guy who sold the gun and Murthy's son's statement he figures out that Murthy is the robber and th him, after which he throws the money on the street. When Murthy returns home he finds out that the police have surrounded him. Pandian take He later tells the investigation board that Murthv shot Pandian dead and Sathva had killed Murthv in self–defense. Sathva is given his job back

As we can see, there are 467 hits in the search result.

Add some stopwords in it.

2017 Film Search

Query:

Search Results:

Total hits: 467

Ignoring stopword(s): ['as', 'you']

Ignoring unknown word(s):

1. [3 idiotas](#)

- 3 Idiotas is a 2017 Mexican coming of age comedy–drama, which is a remake of the 2009 Indian film "3 Idiots". 3 Idiotas is directed by Alfonso Dosal, Christian Vazquez, German Valdez and Martha Higareda. The film about two nerdy friends and their arch rival from college. It is a fun adventure determined to find a college roommate who disappeared without a trace on graduation day. * Alfonso Dosal as Pariso * Sebastin Zurita as Emiliano * Paulina Dvila as Diana * Enrique Singer as Don Diego * Rodrigo Murray as Escalona 3 Idiotas

2. [8 Thottakkal](#)

- 8 Thottakkal (The film starts with an orphaned boy, Sathya, getting framed by the employer for the murder of his wife and is sent to a prison. The warden taught him and stays honest in his profession. Unlike the other policemen at his station, he refuses to be involved in bribes. Thinking that Sathya is trying to become a hero by making him look bad, the inspector plans to shame Sathya. The inspector then arrests him and loads him with 8 bullets. The inspector gives him one day to find his lost gun or face disciplinary action. With the help of a reporter friend, the gun attempts a robbery and in the process he accidentally shoots and kills a little girl. Meera, in danger of losing her job tells her boss to be arrested, and is tracked down by Pandian, the new handler of the case, Pandian says that if Sathya was a real policeman he should have a girlfriend a necklace for her birthday and takes all the fresh notes available, despite the boss's opposition. In the jewelry shop, he is arrested. After a confrontation with the boss he tries to steal all the money and the leader shoots and kills him. The remaining two police attempt to arrest the boss with a ploy the boss shoots and kills the second robber and made sure there is no one left to confront the terrorist slip away as a constable. It turns out that Gunasekaran accepted a bribe from the terrorist and let him go, but this has already been exposed. Murthy sees the man who sold him the gun and leaves. Murthy kills him with the same gun when he follows Murthy to a restroom. He then goes to the Provident Fund. Sathya, now looking for clues in the murder, finds Murthy's application sanctioned and goes to his house to give it to him. The family, by connecting the murder of the guy who sold the gun and Murthy's son's statement he figures out that Murthy is the real killer.

we can see it can ignore the stop words 'as' and 'you'.

testing on a longer query:

2017 Film Search

Query:

Search Results:

Total hits: 19

Ignoring stopword(s): ['a']

Ignoring unknown word(s):

1. [Aby](#)
 - Aby is a 2017 Malayalam Drama film directed by prolific advertisement filmmaker Srikant Murali in his feature directorial debut. It is written by Santhosh Echikkanam and produced by Suv Sreenivasan plays the lead and titular character of Aby. Aju Varghese, Suraj Venjaramoodu play supporting characters. Aby is a boy with a dream that he wants to fly and make an aeroplane. The movie depicts how the hero struggles to make his dream come true and the obstacles he faces on his way to success and achieve glory. Aby is a boy who dreams to fly. The movie depicts elevated positions in an attempt to fly; often injuring himself in the process. He barely spoke a word, leading his folk to believe that he is mentally disabled. He went to a special-needs school to gain sympathy of potential customers to make them sign his policies. Eventually, his mother finds out of his doings and his absence from school because of this. A big fight follows in the morning. A few years later, Aby is an adolescent and works at a mechanical workshop; now a savant in electronics. His father takes the salary to feed his alcoholism. His neighbor and girl friend, Kunjootan. She requests Aby to help her make a science project for a competition, which Aby agrees to. The project ends with it getting big recognition and Anumol getting rewarded with a car. Aby's father arrives and makes a scene demanding the money as Aby made the project. A feud erupts between the neighbors. His father's alcoholism grows more and more volatile as he ends up selling his auto-rickshaw and salvage. Aby becomes frustrated and flees the town. Aby reaches Bangalore where he meets some workers who sell salvage and scrap for money. He helps fix their plane and brings nuisance to them. Soon, Aby becomes friends with them and joins their work. A few days later, he finds a discarded toy airplane, from where he meets G. K. Menon, an alcoholic airplane mechanic. Aby, with his wish to fly, discovering his technical intelligence in the process. Menon shares his basic knowledge with him, fuelling Aby's fascination. Soon, Menon receives a job to manufacture a large commercial aircraft. Aby, in conversations, learning more and more about aircraft in his free time. When the aircraft was built, Aby gets carried away and attempts to ride the aircraft and crashes it, earning Menon's disapproval. He goes away, where he flees back to his hometown. It is now 7 years since his return, and Anumol is helping the village expand. He has gathered materials, and begins work on a glider for himself. He helps Aby build a glider, which gradually grabs the attention of media and people from distant places. This earns her anger from her father who builds hatred towards Aby and his work. Confronted by onlookers, he is halted by law enforcement who stated that he did not have paperwork, making the flight illegal. Aby, with the help of Anumol and several of his townsfolk gathers the required documents, saving up, showing that he is turning into a better man and leaving his alcoholism. Menon even appears, being a certified airplane manufacturer, says that Aby has built the most efficient glider. Anumol's father attempts to sabotage the paperwork, where Anumol begs and pleads to let Aby fly and not take vengeance on his dreams. Anumol's father reconsiders his actions and decides to let the glider fly. The glider is now ready to go once again, with the correct paperwork filed in and now legal. However, another obstacle arises as the law enforcement gets a tip against Aby's disability paperwork. Anumol's father confesses to tipping against him, with Menon vouching for Aby's mental status. This promptly evolves into a huge quarrel involving the frustrated onlookers, which finally ends with the glider flying. The glider is now ready to go once again, with the correct paperwork filed in and now legal. However, another obstacle arises as the law enforcement gets a tip against Aby's disability paperwork. Anumol's father confesses to tipping against him, with Menon vouching for Aby's mental status. This promptly evolves into a huge quarrel involving the frustrated onlookers, which finally ends with the glider flying.

That seems great!

verify if intersecting the posting list from small to large

In order to verify we intersect the posting list from small to large in order to save time correctly, we add two print statements to print out the length of list we intersect in order in the following command.

```
def findMovieId(self, query):

    movieId = []

    if len(query) == 0 :

        return movieId

    posting_list = shelve.open(index_path, writeback=False)

    new_query = sorted(query, key = lambda word : len(posting_list[word]))

    movieId = posting_list[new_query[0]]

    if len(query) == 1 :
```

```

return movieId

print(len(movieId))

for word in new_query[1:]:
    movieId = self.intersect(movieId, posting_list[word])

return movieId

def intersect(self, idList1, idList2):
    print(len(idList2))

    i = 0

    j = 0

    res = []

    while (i
    if idList1[i] < idList2[j]:

        i = i + 1

        elif idList1[i] > idList2[j]:

            j = j + 1

            else:

                res.append(idList1[i])

                i = i + 1

                j = j + 1

    return res

```

Output


```
127.0.0.1 - - [27/Feb/2018 14:53:56] "GET / HTTP/1.1" 200 -  
214  
220  
467  
512  
127.0.0.1 - - [27/Feb/2018 14:54:07] "POST /results/1 HTTP/1.1" 200 -
```

That's correct!

Testing an unknown words

Input : I find a good friend asdfweq



2017 Film Search

Query:

Search Results:

Total hits: 0

Unknown words: ['asdfweq']

After testing, we verify our search engine works correctly! That's great!!

Further thoughts

I think later we can do some advanced search about this movie corpus. Combine text and title, location. And also we can make a positional index system later to enable positional query in the later assignment.

