Task 1 Report

Ying Xue, Shuo Li, Zhen Gao

Objective

Crawl Twitter search results for movies as a corpus of information retrieval systems. Crawl at least 100K documents.

Resource

Twitter development API(tweepy), IMDB python package(Cinemagoer), Kaggle

Process

Due to twitter's restrictions on API usage, we can only use the search recent API. Only relevant tweets posted within seven days can be retrieved, and only 100 tweets can be crawled per query.

1. Generate query list

Using Cinemagoer, which is a python package provided by IMDB, we generated a query list of TOP 250 movies in IMDB

```
import imdb

ia = imdb.Cinemagoer()

top = ia.get_top250_movies()


with open('movie_list.txt', 'w', encoding='utf-8') as outfile:

    for movie in top:
        outfile.write(movie.get("title"))
        outfile.write("\n")
```

Then we used a dataset from a Kaggle contest https://www.kaggle.com/rounakbanik/the-movies-dataset/version/7?select=movies_metadata.csv, which contains 45000 movie titles. And we used the first 1200 movie titles.

2. Preprocessing

In order to query successfully, we preprocessed the list, removed the special characters, and turned all characters to lower case.

```
def read_document(name):
    with open(name, "r") as f:
        document = f.read()
    return document

def remove_punctuation(document):
    document = re.sub(r'[^\w\s]', '', document)
    return document.strip()

def case_folding(document):
    return document.lower()

movie_list = read_document("movie_list1.txt")
movie_list = remove_punctuation(movie_list)
movie_list = case_folding(movie_list)
movie_list = movie_list.split('\n')
print("query all " + str(len(movie_list)) + "movies")
```

3. Use Tweepy to crawl

We apply an Twitter development account and created an app to get the token to use related API. We can use this token to connect and pass the authentication.

```
client = tweepy.Client(bearer_token=bearer_token)
```

Then we use the search recent API and query all the movies in the list by their names.

```
response = client.search_recent_tweets(query = title, max_results = 100,
                                       tweet_fields=["author_id" ,"text","id","lang"])
```

We save three areas author_id which is the identifier of the author of the tweet, the text is the content of the tweet and id is the identifier of the tweet. "lang" is the identifier of the language of the tweets, since we only plan to develop a prototype, we didn't save tweets using languages other than English.

```
for tweet in tweets:
    data = {}
    data["author_id"] = tweet.author_id
    data["id"] = tweet.id
    data["text"] = tweet.text
    if(tweet.lang == "en"):
        result.append(data)
```

Results

We queried 237 movies in the IMDB Top 250 lists and 1090 movies from the Kaggle contest dataset(some of the movie titles can't be queried since they have invalid fields or there were no recent tweets about them). We collected more than 100 thousand tweets written in English.