The Lion King Movie Twitter Analysis

CMPE 256 Summer 2019 Individual Project
Xiaoting Jin 013842192

Data Collection

Data Preprocessing

Daily Frequency of Tweets

Word Cloud

Hashtags Frequency

Sentiment Analysis of Tweets

Semantic Meaning

Conclusion

The Lion King is a 2019 American computer-animated film and was released on July 19, 2019, in the United States, I still haven't got the chance to enter a theater to enjoy this movie, so I decided to collect tweets from Twitter.com about this film and present the findings of my analysis. The movie gets an IMDb score of 7.1/10 rated by 75,983 people.

Data Collection

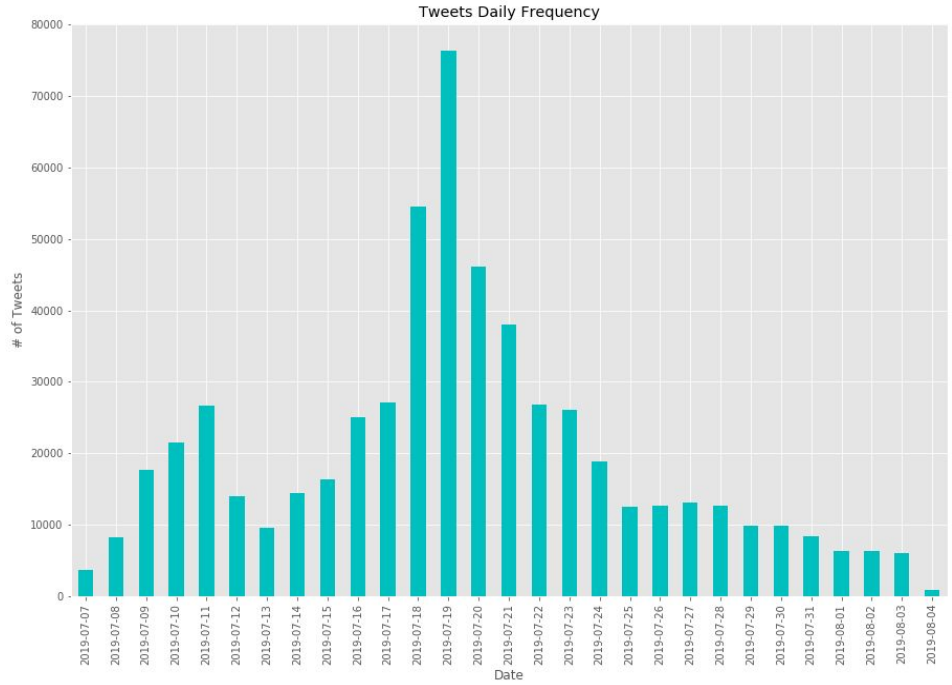
To collect data, I use a tool called twint to scrape Tweets which include the keyword, the lion king, from Twitter and the total number of records collected after removing duplicated records is 569,169, the date range is from July 07, 2019 to Aug 04, 2019.

Data Preprocessing

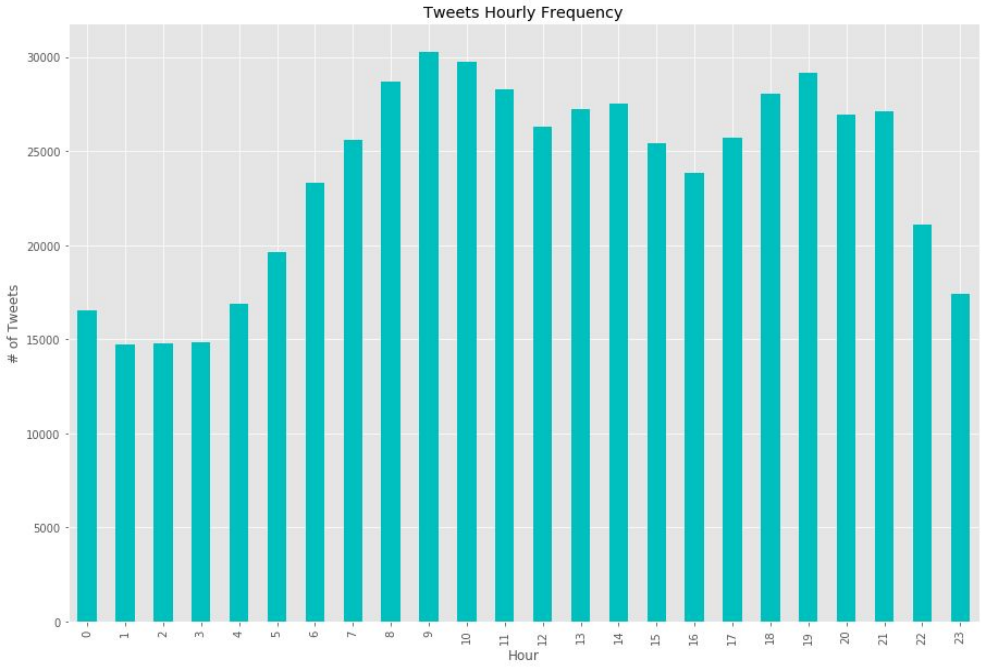
In the preprocessing stage, I read the .json file and created a pandas DataFrame to process the data. For text processing, special characters, most common words and stop words are removed to get more meaningful texts.

Daily Frequency of Tweets

The bar chart below shows all tweets from the July 7th to Aug. 4th. The number of tweets started to climb up from July 1 and reached its peak on July 19, which is exactly the movie's release date. The frequency dipped after July 19 and became flatter after one week.

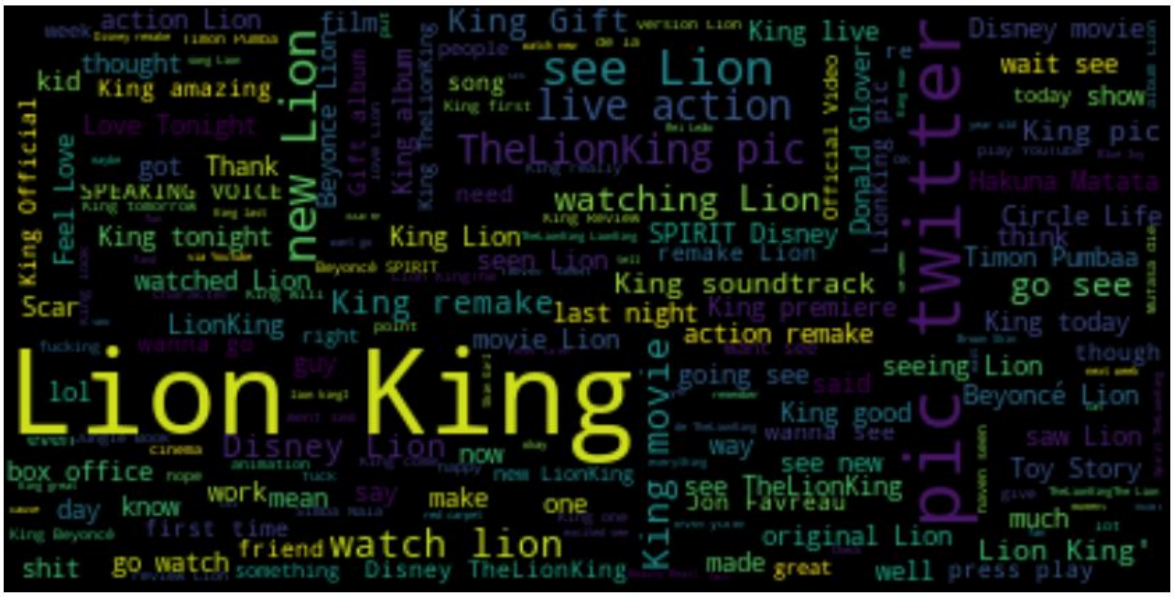


The busiest time of a day is in the late morning between 9 am and 10 am and in the late evening between 18 pm and 19 pm. The silent time of the day is between 0 am and 4 am.

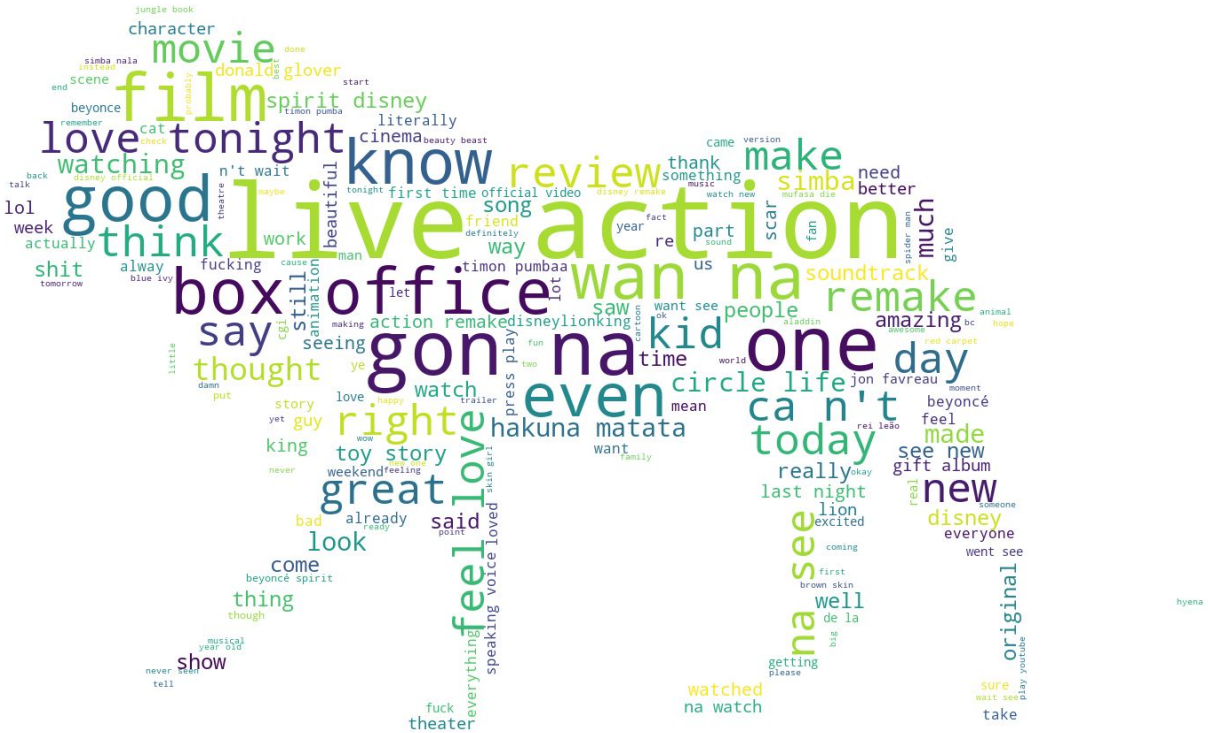


Word Cloud

The entire text includes a few most frequent words such as “lion”, “king”, “movie”, “twitter”, which makes the graph less informative, so I removed these common words and stop words.

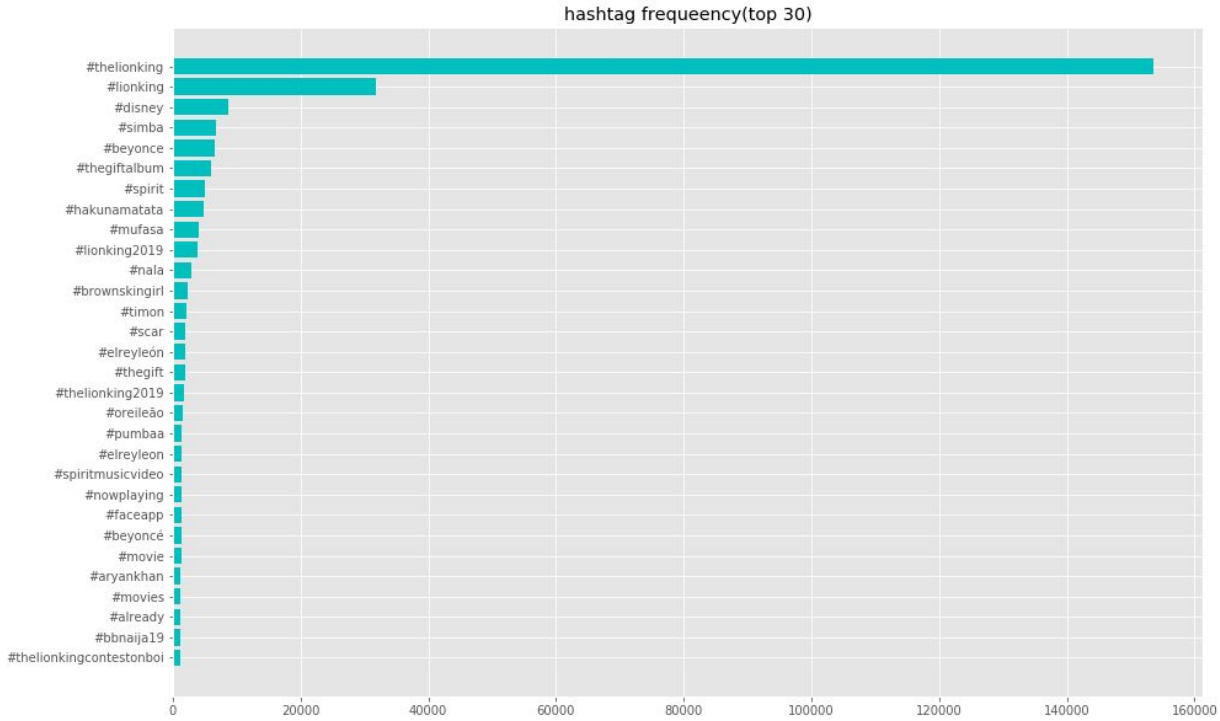


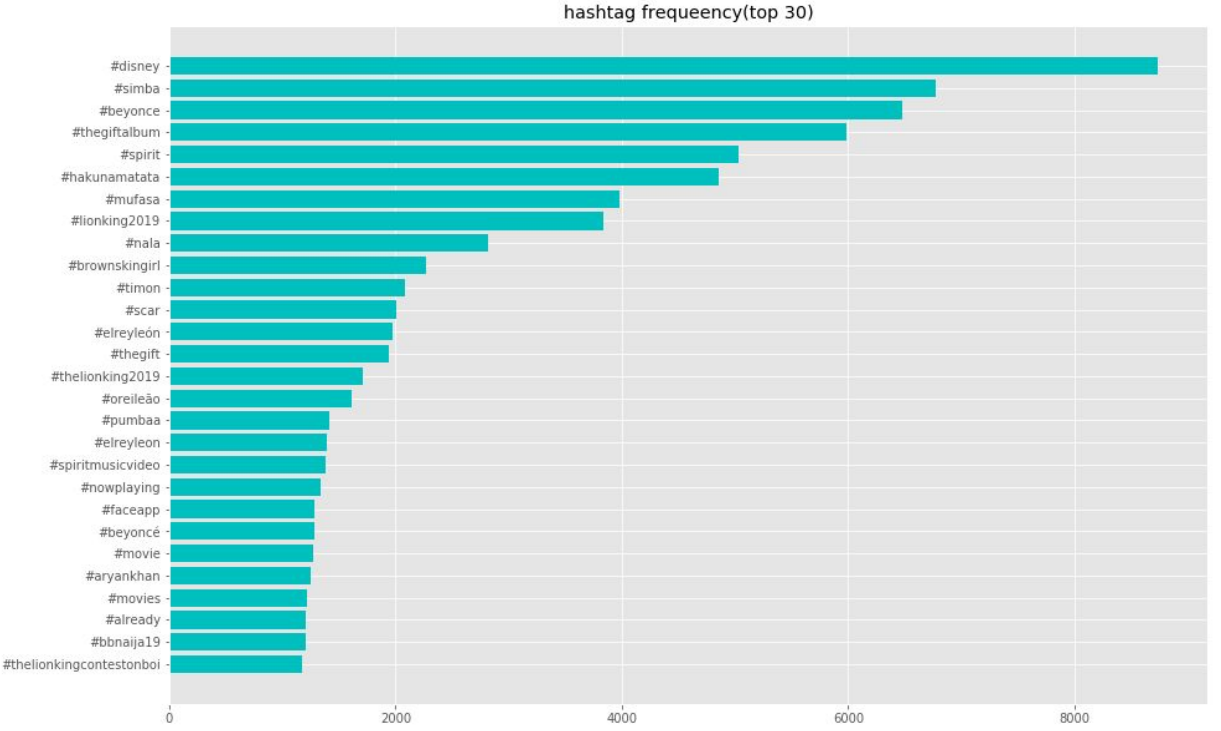
After tokenization and removing common words, the word cloud graph looks like below.



Hashtags Frequency

Predictably, the hashtag #thelionking and #lionking appear most, so I removed both of them and plotted another graph. The remaining top 10 hashtags are #disney, #simba, #beyonce, #thegiftalbum, #spirit, #hakunamatata, #mufasa, # lionking2019, #nala and #brownskingirl.





Sentiment Analysis of Tweets

I performed sentiment analysis using Google's Cloud NLP API, the score of a document's sentiment indicates the overall emotion of a document and the score is between -1.0 and +1.0 where 1.0 means more positive and -1.0 means more negative. However, the API has a quota

limit on request per day and per minute per user, so I only got 169 tweets' sentiment score, and the average score is 0.1562.

```
print(df[df.id == 1157992220995117056].tweet)
print(sentiment_scores[1157992220995117056])
```

```
1    No spoilers pls. Thanks and God bless! — watch...
Name: tweet, dtype: object
0.4000000059604645
```

```
print(df[df.id == 1157988870526578689].tweet)
print(sentiment_scores[1157988870526578689])
```

```
37    The hype around the new Lion King makes me sad...
Name: tweet, dtype: object
-0.699999988079071
```

```
print(df[df.id == 1157984313473814528].tweet)
print(sentiment_scores[1157984313473814528])
```

```
95    During interval 🥰😘🤗 movie resumes 😻 #TheLionKi...
Name: tweet, dtype: object
0.800000011920929
```

Semantic Meaning

I use Word2Vec to build up word embedding and find out the words that are most similar to "thelionking" based on a cosine metric similarity measure. The nearest neighbors of "thelionking" are "disney studios", "hakuna matata", "live action", "jonfavs", "circle of life", "timon and pumbaa", "jon_favreau", "the jungle book", "donald glover", and "james earl jones".

```
[3]: model.wv.most_similar(['thelionking'],topn=20)

[3]: [('lionking', 0.638075053691864),
      ('disneylionking', 0.5510846376419067),
      ('lionking2019', 0.35245639085769653),
      ('disneystudios', 0.35045847296714781),
      ('hakunamatata', 0.3497157394886017),
      ('liveaction', 0.33293676376342773),
      ('theliongking', 0.32415902614593506),
      ('jonfavs', 0.3092525899410248),
      ('circleoflife', 0.30613574385643005),
      ('timonandpumbaa', 0.30438920855522156),
      ('thelionking2019', 0.2745281457901001),
      ('…', 0.27296748757362366),
      ('thelionking_19', 0.2725541591644287),
      ('jon_favreau', 0.26087382435798645),
      ('thejunglebook', 0.2597157657146454),
      ('donaldglover', 0.2580404281616211),
      ('jonfavreau', 0.25580212473869324),
      ('jamesearljones', 0.25269845128059387),
      ('thelionking…', 0.25025674700737),
      ('disneyanimation', 0.24993756413459778)]
```

The adjectives regarding this movie are “great”, “awesome”, “amazing”, “beautiful”, “fantastic”, “wonderful”, “definitely”, which statistically demonstrates how people felt about this movie and it seems like it is an enjoyable movie!

```
: adjectives = [x for x in my_list if x[0].find('lion') < 0]

: adjectives

: [('great', 0.4150570034980774),
   ("'s", 0.38416287302970886),
   ('awesome', 0.3832293748855591),
   ('movie', 0.37753015756607056),
   ('amazing', 0.36635076999664307),
   ('film', 0.35671499371528625),
   ('circleoflife', 0.3545438051223755),
   ('course', 0.3540392518043518),
   ('&', 0.34989655017852783),
   ('beautiful', 0.34782469272613525),
   ('one', 0.34089893102645874),
   ('also', 0.33774811029434204),
   ('fantastic', 0.32908082008361816),
   ('experience', 0.32907021045684814),
   ('however', 0.3194894790649414),
   ('explore', 0.31895163655281067),
   ('king', 0.31839412450790405),
   ('touch', 0.3102189898490906),
   ('hakunamatata', 0.3050859570503235),
   ('disneystudios', 0.30401352047920227),
   ('especially', 0.30273616313934326),
   ('😍🐂🦒😊', 0.30225884914398193),
   ('wonderful', 0.3014325797557831),
   ('alongside', 0.29471302032470703),
   ('definitely', 0.2945542633533478)]
```

Conclusion

After this individual project, I realized the text preprocessing is a severely important topic in NLP, poor preprocessing would result in poor result, and I didn’t perform very well in this part. Hope this project could steer me towards realizing the importance of appropriate text preprocessing and gain more experience in NLP.