Research seminar HW2

Accomplished by Petr Yakovlev BASB 212, 24th March 2022

Key task

To analyze the textual dataset (more than 200 reviews from Amazon Movies), present top-10 most frequent words and visualize some interesting statistics

Dataset description

Positive reviews

101 text files each with the positive review on some movie

Negative reviews

101 text files each with the negative review on some movie

PySpark Initialization

To perform word count, we do the following:

- Use findspark() and build SparkSession()
- Initialize SparkContext()
- Import multiple files through textFile()

Ввод [1]:
```
# we download the findspark library and
import findspark
findspark.init()
```

Ввод [2]:
```
# we create Spark context in order to wo
from pyspark.sql import SparkSession
from pyspark.conf import SparkConf
```

Ввод [3]:
```
# Spark set-up
spark=SparkSession.builder\
    .master("local[*]")\
    .appName("WordCount")\
    .getOrCreate()
```

Ввод [4]:
```
# initialize Spark context
sc=spark.sparkContext
```

Ввод [5]:
```
# import files (101 text files with posi
file_positives = sc.textFile(r"C:\Users\
file_negatives = sc.textFile(r"C:\Users\

rdd_positives = file_positives.collect()
rdd_negatives = file_negatives.collect()

rdd_positives
rdd_negatives
```

Next, we clean the dataset

Ввод [6]:
```
# perform function to remove punctuation and transform all words
def lower_clean_str(x):
    punc='!"#$%&\'()*+,./:;<=>?@[\\]^_`{|}~-'
    lowercased_str = x.lower()
    lowercased_str = lowercased_str.replace('<br />', '')
    for ch in punc:
        lowercased_str = lowercased_str.replace(ch, '')
    return lowercased_str
```

Ввод [7]:
```
# we transform our instances
file_positives = file_positives.map(lower_clean_str)
file_negatives = file_negatives.map(lower_clean_str)
```

Ввод [8]:
```
# we use split function to separate the words
file_positives = file_positives.flatMap(lambda y: y.split(" "))
file_negatives = file_negatives.flatMap(lambda y: y.split(" "))
```

Ввод [9]:
```
# exclude whitespaces from the instances
file_positives = file_positives.filter(lambda x: x!='')
file_negatives = file_negatives.filter(lambda x: x!='')
```

Ввод [10]:
```
# to count the frequency of words we first need to apply transfro
positives_count = file_positives.map(lambda word:(word,1))
negatives_count = file_negatives.map(lambda word:(word,1))
```

- remove punctuation
- transform all words to lowercase
- separate the words
- exclude whitespaces

Word count

We use reduceByKey() to count the most frequent words

As you can see, the most frequent words are approximately the same for negative and positive dataset except for words ‘bad’ for negative dataset and ‘great’ for positive dataset

```
Ввод [14]: # we exclude stopwords values
positives_count = positives_count.filter(lambda x: x[1] not in stopwords).sortByKey(False)
negatives_count = negatives_count.filter(lambda x: x[1] not in stopwords).sortByKey(False)

Ввод [15]: #display the result for positive datasets
positives_count.sortByKey(False).take(10)

Out[15]: [(139, 'movie'),
 (130, 'film'),
 (113, 'one'),
 (63, 'see'),
 (58, 'good'),
 (54, 'even'),
 (49, 'great'),
 (46, 'time'),
 (44, 'like'),
 (44, 'show')]

Ввод [16]: #display the result for negative datasets
negatives_count.sortByKey(False).take(10)

Out[16]: [(177, 'film'),
 (136, 'movie'),
 (117, 'one'),
 (77, 'like'),
 (65, 'even'),
 (64, 'horror'),
 (55, 'good'),
 (51, 'bad'),
 (51, 'really'),
 (50, 'get')]
```

Ввод [17]:
```
from pyspark.sql.types import *
schema = StructType([StructField("frequency", IntegerType(), True),
                     StructField("word", StringType(), True)])
```

Ввод [18]:
```
counts_df_positives = spark.createDataFrame(positives_count, schema)
counts_df_negatives = spark.createDataFrame(negatives_count, schema)
```

Ввод [19]:
```
counts_df_positives.printSchema()
counts_df_negatives.printSchema()
```

```
root
 |-- frequency: integer (nullable = true)
 |-- word: string (nullable = true)

root
 |-- frequency: integer (nullable = true)
 |-- word: string (nullable = true)
```

Ввод [20]:
```
df_positives = counts_df_positives.toPandas()
df_negatives = counts_df_negatives.toPandas()
```

Ввод [21]:
```
df_positives
```

Out[21]:

	frequency	word
0	139	movie
1	130	film
2	113	one
3	63	see
4	58	good
...	...	...
3873	1	yells
3874	1	yoyo
3875	1	yuppie

We present some visualisations for the dataset next

We convert pyspark object to pandas dataframe and work with it

Word clouds

Negatives

```
Ввод [168]: plt.imshow(wc_negatives, interpolation='bilinear')
            plt.axis("off")
            plt.show()
```


```
Ввод [169]: plt.imshow(wc_positives, interpolation='bilinear')
            plt.axis("off")
            plt.show()
```

Positives

Ввод [125]:

```
from textblob import TextBlob

def preprocess(ReviewText):
    ReviewText = ReviewText.str.replace("(<br/>)", "")
    ReviewText = ReviewText.str.replace('(<a).*(>).*(</a>)', '')
    ReviewText = ReviewText.str.replace('(&amp)', '')
    ReviewText = ReviewText.str.replace('(&gt)', '')
    ReviewText = ReviewText.str.replace('(&lt)', '')
    ReviewText = ReviewText.str.replace('(\xa0)', ' ')
    return ReviewText
files_df['Review'] = preprocess(files_df['Review'])

files_df['polarity'] = files_df['Review'].map(lambda text: TextBlob(text).sentiment.polarity)
files_df['review_len'] = files_df['Review'].astype(str).apply(len)
files_df['word_count'] = files_df['Review'].apply(lambda x: len(str(x).split()))
```

We also can calculate some interesting statistics with aggregated datasets with more than 200 reviews

We can show the sentiment polarity of the review with one library in python and show the reviews with negative and positive sentiment

5 random reviews with negative sentiment polarity

5 random reviews with the highest negative sentiment polarity:

Seriously, I can't imagine how anyone could find a single flattering thing to say about this movie, much less find it in themse lves to write the glowing compliments contained in this comment section. How many methamphetamines was Bogdonovitch on during t he filming of this movie? Was he giving a bonus to the actor that spat his lines out with the most speed and least inflection o r thought? The dialogue is bad, the plot atrocious, even for a "screwball" comedy, and claims that the movie is an homage to cl assic film comedy is about the most inane thing I've ever heard. The cinematography is below the quality and innovation of that exhibited by the worst made-for-TV movies, the acting is awful (although I get the feeling that the fault for that lies squarel y in the lap of the director), and speaking of which, did I mention the direction is so haphazard and inscrutable that it defie s the definition of the word? The whole thing is a terribly unfunny (even in the much-beleaguered world of so-bad-it's-funny cl unkers), soul-sucking, waste of two hours of your life that you'll never get back. Be afraid, be very afraid...

An obscure horror show filmed in the Everglades. Two couples stay overnight in a cabin after being made a little uneasy by the unfriendliness of the locals. Who, or what, are the Blood Stalkers? After awhile they find out. Watch for the character of the village idiot who clucks like a chicken, he certainly is weird.

This film, which I rented under the title "Black Voodoo" should be avoided. I was expecting a blaxploitation/horror flick; but what I got was a very dull, standard "ghost extracts vegence". In this case the ghost was that of a religious cult leader who t ried to refuse treatment, but who's plea was ignored and he died in an operation. The result: his spirit posesses Nurse Sherry and forces her to commit acts of murder. The only voodoo connection was to one of the three black characters, in this case a bl inded ex-football player, who's mom practiced voodoo. The film is very slow and very dull. There is a very standard ending that provides on excitement, followed by a horrificly stupid ending (warning: SPOILER)

In which a woman actually manages to defend herself against murder charges by saying she was possessed. This movie is slow, and bad in a non-funny, just stupefyi ng way. Avoid it at all costs.

We brought this film as a joke for a friend, and could of been our worst joke to play. The film is barely watchable, and the ac ting is dire. The worst child actor ever used and Hasslehoff giving a substandard performance. The plot is disgraceful and at p oints we was so bored we was wondering what the hell was going on. It tries to be gruesome in places but is just laughable.

Just terrible

I found this movie really hard to sit through, my attention kept wandering off the tv. As far as romantic movies go..this one i s the worst I've seen. Don't bother with it.

5 random reviews with positive sentiment polarity

5 random reviews with the highest positive sentiment polarity:

The great Vincent Price has done many fantastic Horror films, some of which range among the greatest genre gems of all-time. Price's greatest achievements were doubtlessly his films in the 60s, with films such as Roger Corman's brilliant Poe-cycle (still the greatest Horror cycle of all-time), Michael Reeves' "Witchfinder General" (1968) or Ubaldo Ragona's "The Last Man on Earth" (1964) marking the ultimate highlights of this brilliant man's career. The films that made the man famous and thereby made him the immortal Horror icon he is, however date back to the 50s, with "House of Wax" (1953) marking his rise to stardom. "The Mad Magician" of 1954 follows a plot that is very similar to that of its successful predecessor. This is not to say, however, that this film isn't an original, delightfully macabre and absolutely wonderful gem itself. As the lines above may suggest, Vincent Price is my favorite actor, and, while I personally would not allow myself to miss anything the man has been in, none of my fellow fans of the man may miss this little gem.

Price stars as Don Galico (aka. Galico the Great), an underrated master magician and inventor of magic devices, whose boss, a sleazy businessman, stole his wife (Eva Gabor) from him. When the boss takes away one of Galico's ingenious inventions and gives it to his rival, The Great Rinaldi (John Emery), Galico snaps, and a murderous spree of revenge begins...

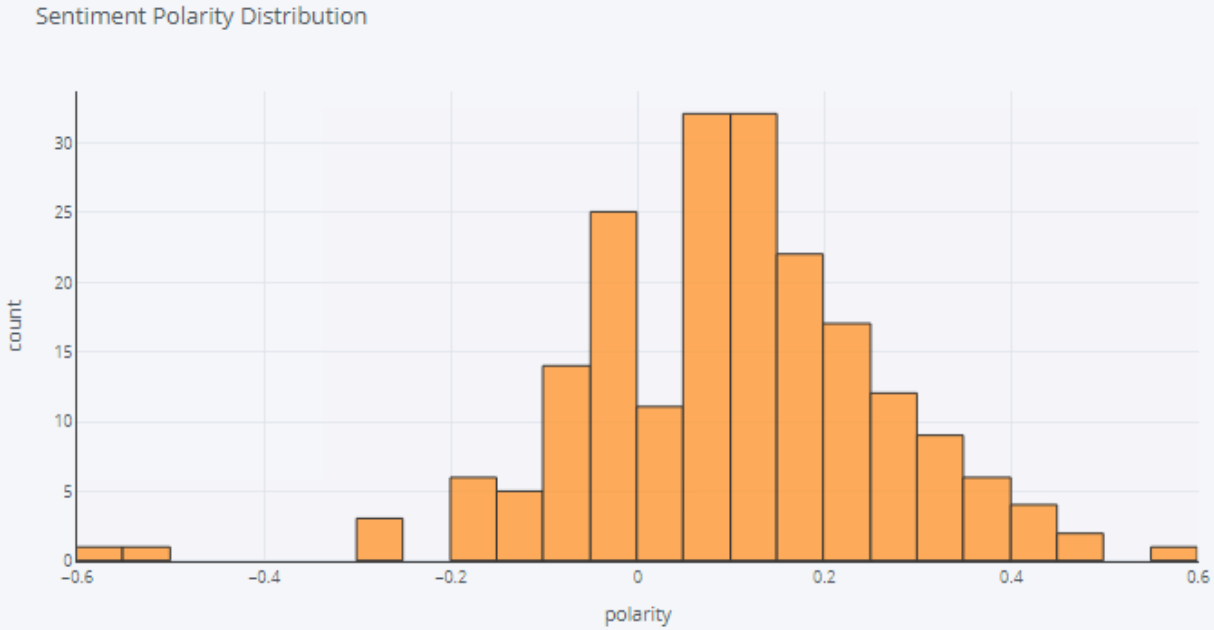
Don't we love Vincent Price when he's out for revenge? Some of his most famous and greatest films such as "The Abominable Dr. Phibes" (1971) or "Theater of Blood" (1973) were about absurd and delightfully macabre revenge murders, and this earlier film in his Horror career is another proof that no one takes revenge as Vincent Price does. This film provides a wonderfully eccentric leading role for Price, who, as always, delivers a brilliant performance, and guarantees 70 minutes of outrageously entertaining and macabre fun for every Horror fan. Another must-see for my fellow Price fans.

Liked Stanley & Iris very much. Acting was very good. Story had a unique and interesting arrangement. The absence of violence and porno sex was refreshing. Characters were very convincing and felt like you could understand their feelings. Very enjoyable movie.

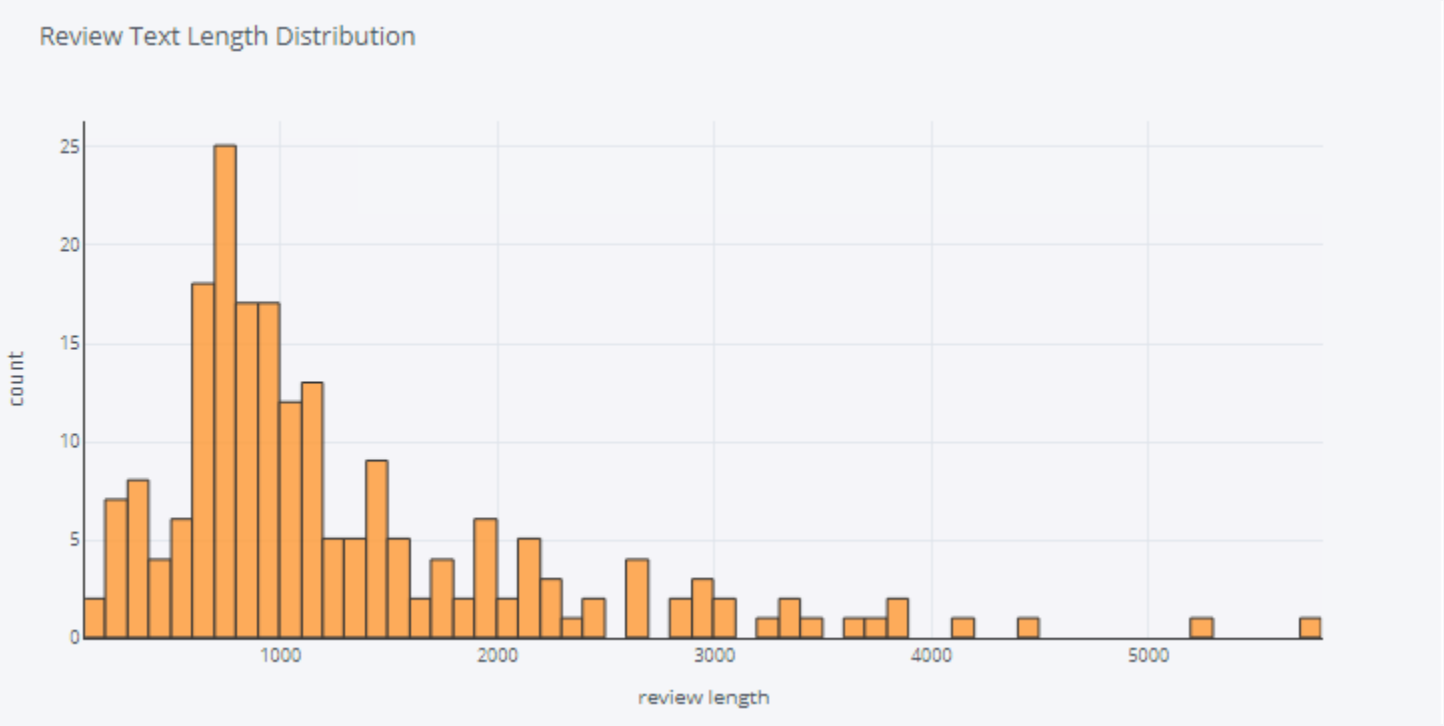
My children just happened to stop at this movie the other night and as things started to play out it really piqued my interest. I had to head out for bowling league so I had them record it for me on the dvr so I could watch the rest later. Well I just got done watching it and the front of my shirt must be soaked after crying buckets. It was an excellent movie even though I could almost feel the pain and anguish these girls were experiencing. And I never in a million years would have guessed the reason why Alissia had gone from this beautiful girl to an anti-social goth. This was probably WHY my shirt was soaked because I've experienced that same pain that Alissia was feeling. I too would not have sought out this movie, but I'm sure glad I saw it. Very moving, very touching. Great for those who love a good drama or tear-jerker.

My first child was born the year this program came out, and I played the record album for the boys every Christmas thereafter. When the CD came out, I bought about ten copies and still give them to friends and relatives as they start families...it invariably becomes their favorite Christmas album. I recently found several DVD's (made on DVD-R from video tapes, probably) for sale on eBay. The one I bought was an excellent copy, and it was so great to see the show again after more than 25 years. There are some songs on the show that were not on the album, and some of the songs on the album were studio versions of the same songs on

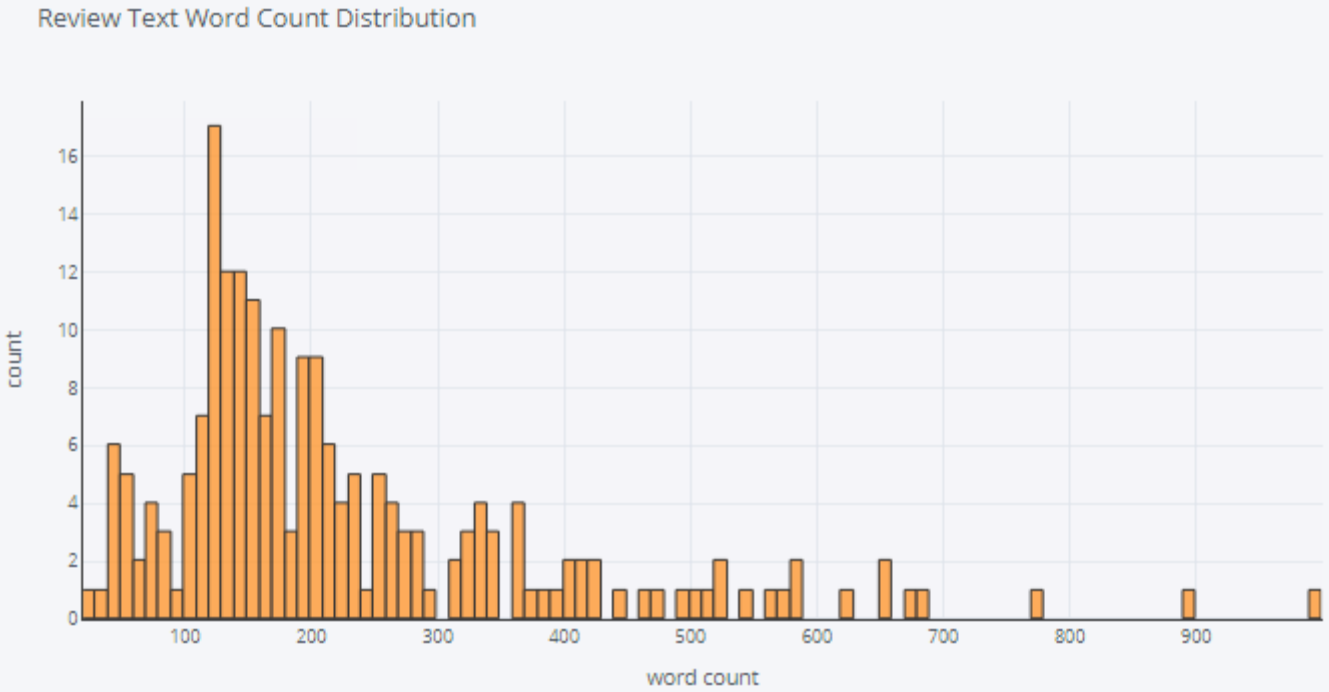
More statistics



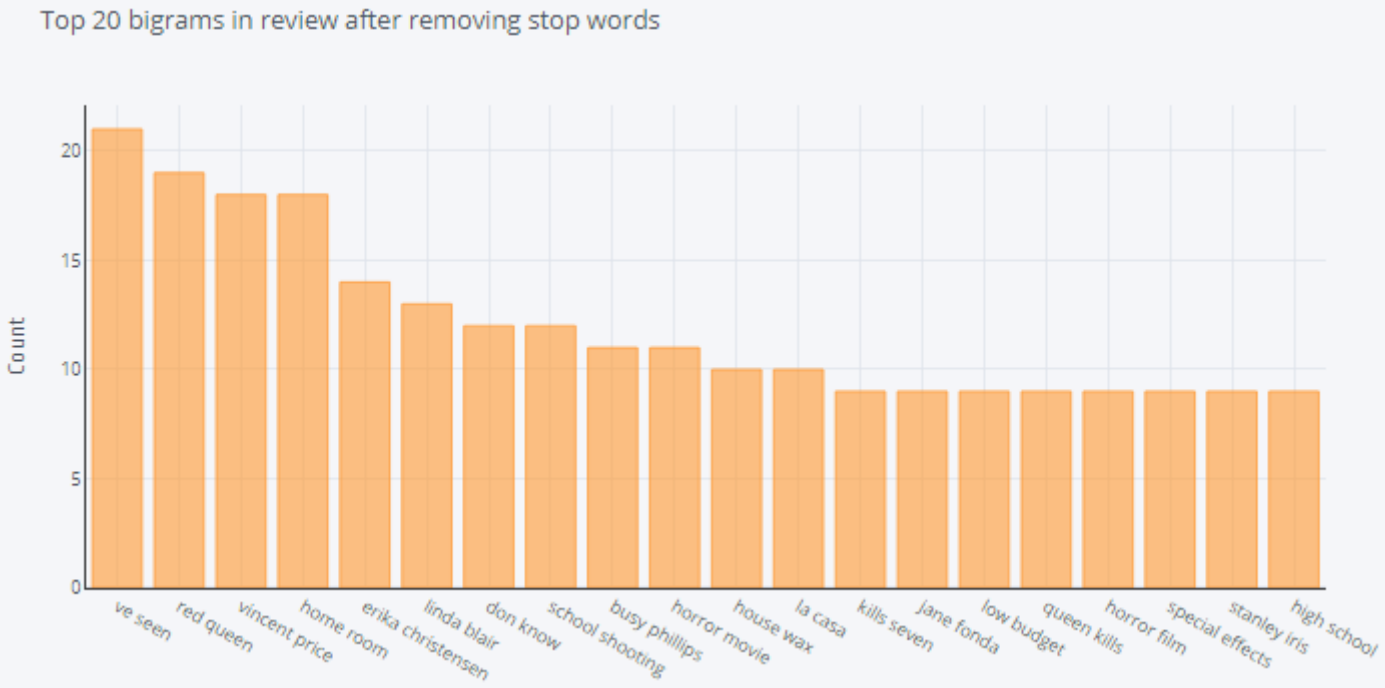
More statistics



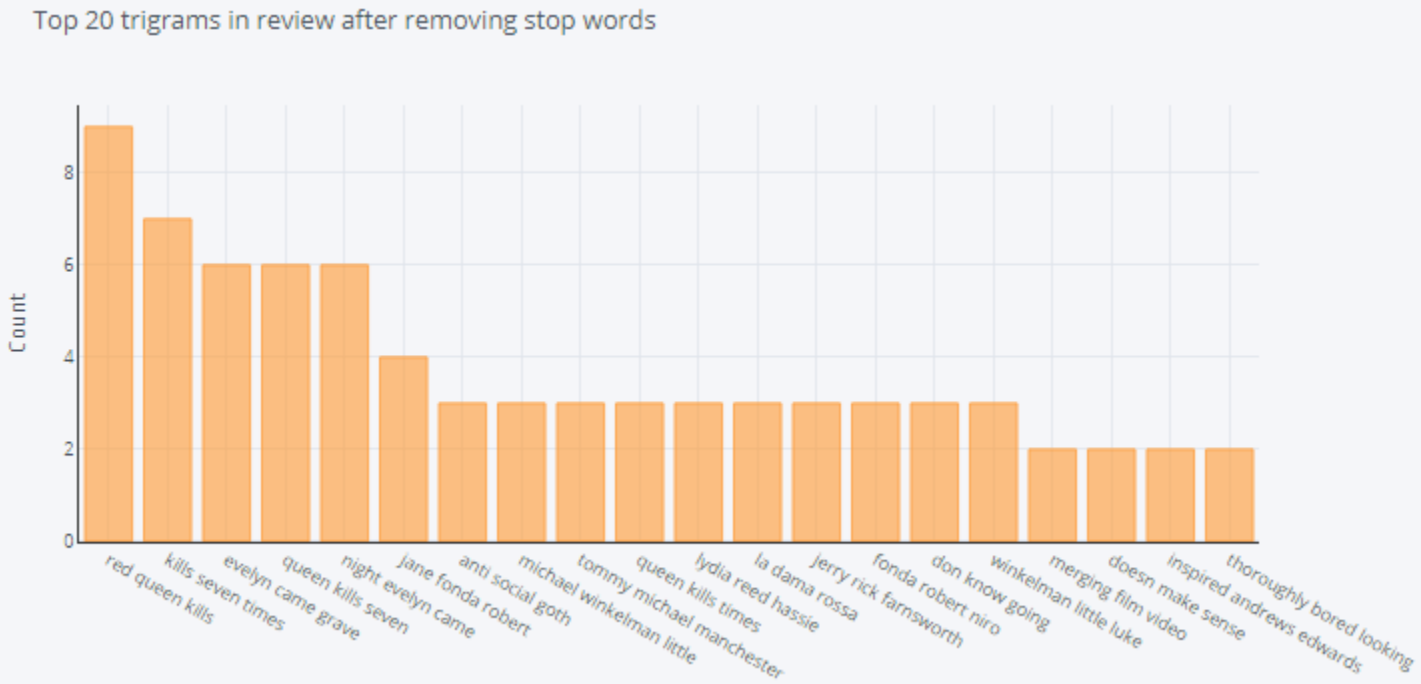
More statistics



More statistics



More statistics



Thanks!