Movie Analysis Business Report

“The Dark Knight” and “Joker”

YU HUANG

2020-02-10

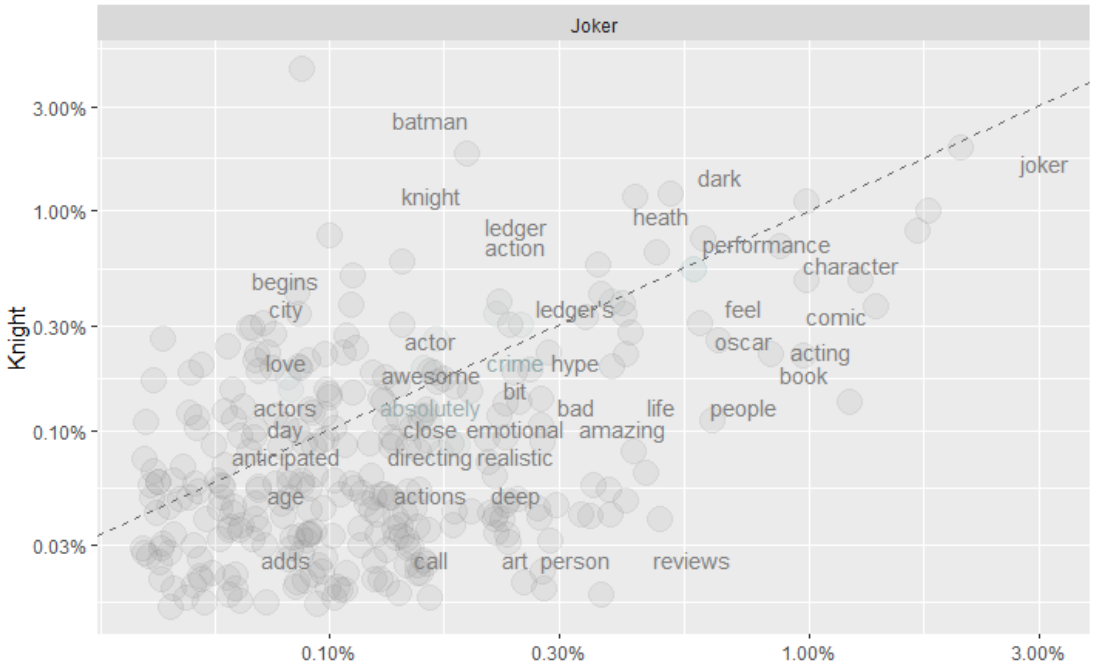
Background

“The Dark Knight” and the “Joker” may be so opposite as superhero and villain, but the two main characters have always been linked each other; however, they are so different. So, I want to analyze these two movies’ reviews, which character the reviewer prefer more and how they compare these two movies.

This report extracts the evaluation of reviewers from IMDB, aiming to analyze the reviews and tendency of reviewers towards these two movies. What is the main difference between two movies reviews?

Insights Support

Business insights of correlation: “The Dark Knight” and “Joker”.



In this correlation chart,

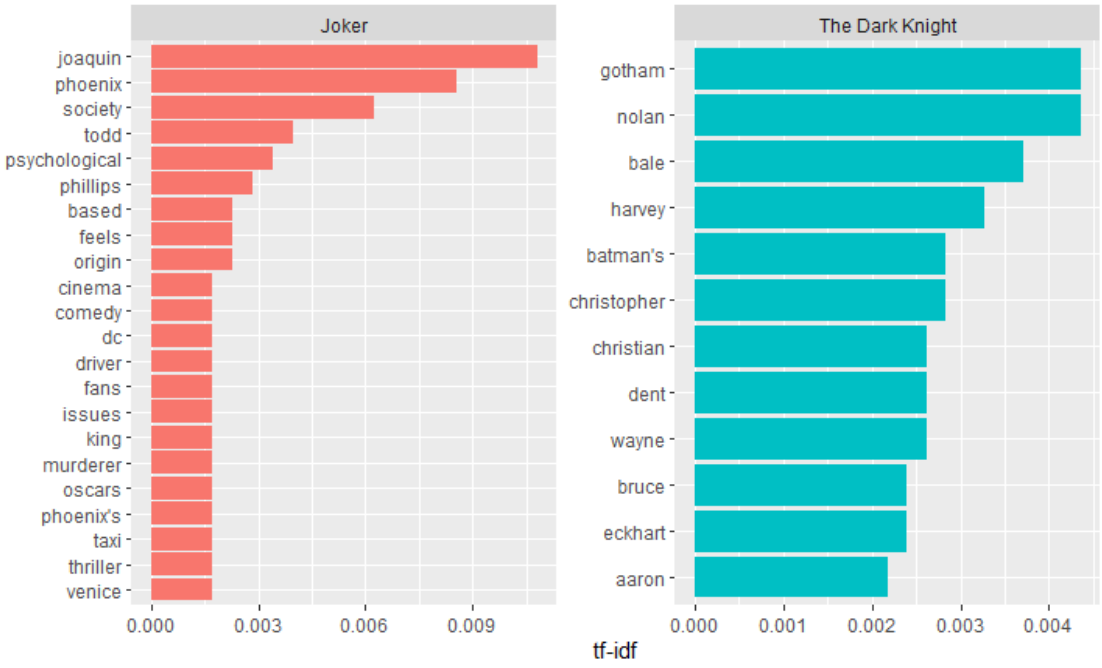
We can see the reviews of “The Dark Knight”, “Joker” correlation is 0.53.

Batman is the main character of “The Dark Knight”, so the proportion of some unique word are high in “The Dark Knight”, like “batman”,” knight”, “dark”, “city”, “superhero” are labels of Batman, so the reviewers who use these words are all referring “The Dark Knight” movie.

The movie “Joker” fully showed the drama, the realistic cruelty, and clowns became distorted and crazy, so “Drama”, “realistic”, “art”, “comic”, “crazy” are classic performance words of Joker in the movie “Joker”. The critics of these words are almost always evaluating the review of the Joker.

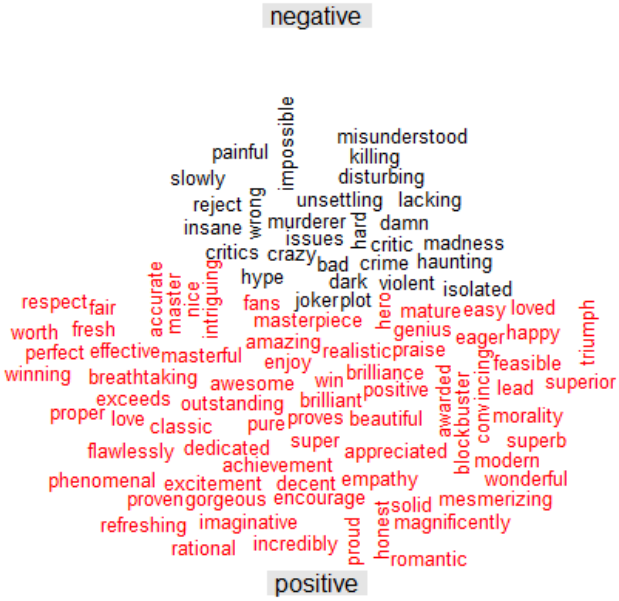
For both films, as Joker is batman's biggest enemy, so both of two movies reviews mentioned a lot of “joker” and “crime”. Even though the leading character in the first movie is a decent guy and the leading character in the second movie is a villain, most of reviewers like both of Batman and Joker from their common words “brilliant”, “awesome” in word cloud.

Business insights from TF-IDF: “The Dark Knight” and “Joker”.



From the TF-IDF visualization plot, we can see the Joker's reviewers more likely focus on society and emotional feels. That make sense because the joker's personality ranges from brutal to funny to diablo's disordered intelligence, and the joker's film is more socially reflective. Joker shape a pessimistic insight into human nature, which accurately points to the darkest and weak side of human nature. The movie "The Dark Knight" reviewers more likely talked about the batman himself.

Business insights from "bing" sentiment word cloud: "Joker".

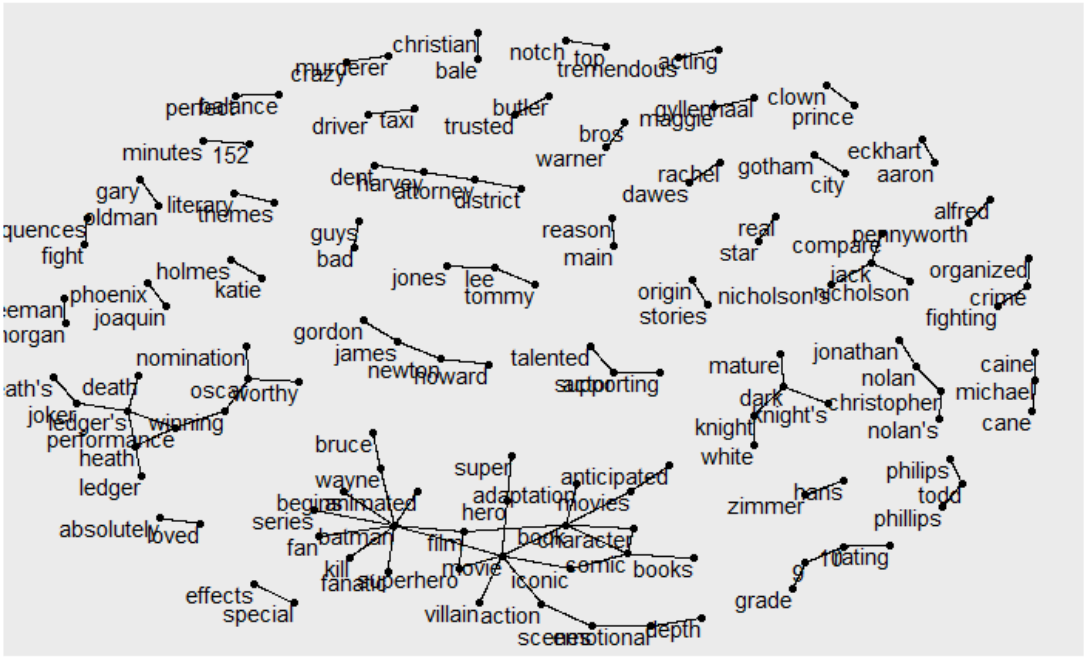


From this sentiment word cloud, we can see the movie is about a world without love, childhood trauma and compassion. Admittedly, the joker has received mostly positive reviews. we also see there are much more positive words than negative words, the fascination with "bad guys" like clowns is also part of the insights. From the positive words “romantic”, “amazing”, “genius” etc. we can see reviewers romanticize batman's Nemesis—Joker.

Business insights from “nrc” sentiment word cloud: “The Dark Knight”.

We can see the sentiment cloud from movie “The Dark Knight”, the reviewers prefer “trust “ sentiment from this movie. Probably the batman has a lot of money and justice enough.

Business insights from spider chart: combine the “Joker” and “The Dark Knight”.



From this spider chart, we can see some insights here. Some words connection like “ death-joker-ledger-oscar-worthy”. So we can understand that the review is about the Joker’s actor Ledger, whom dead after performance. But he still got the Oscar reward.

Recommendations

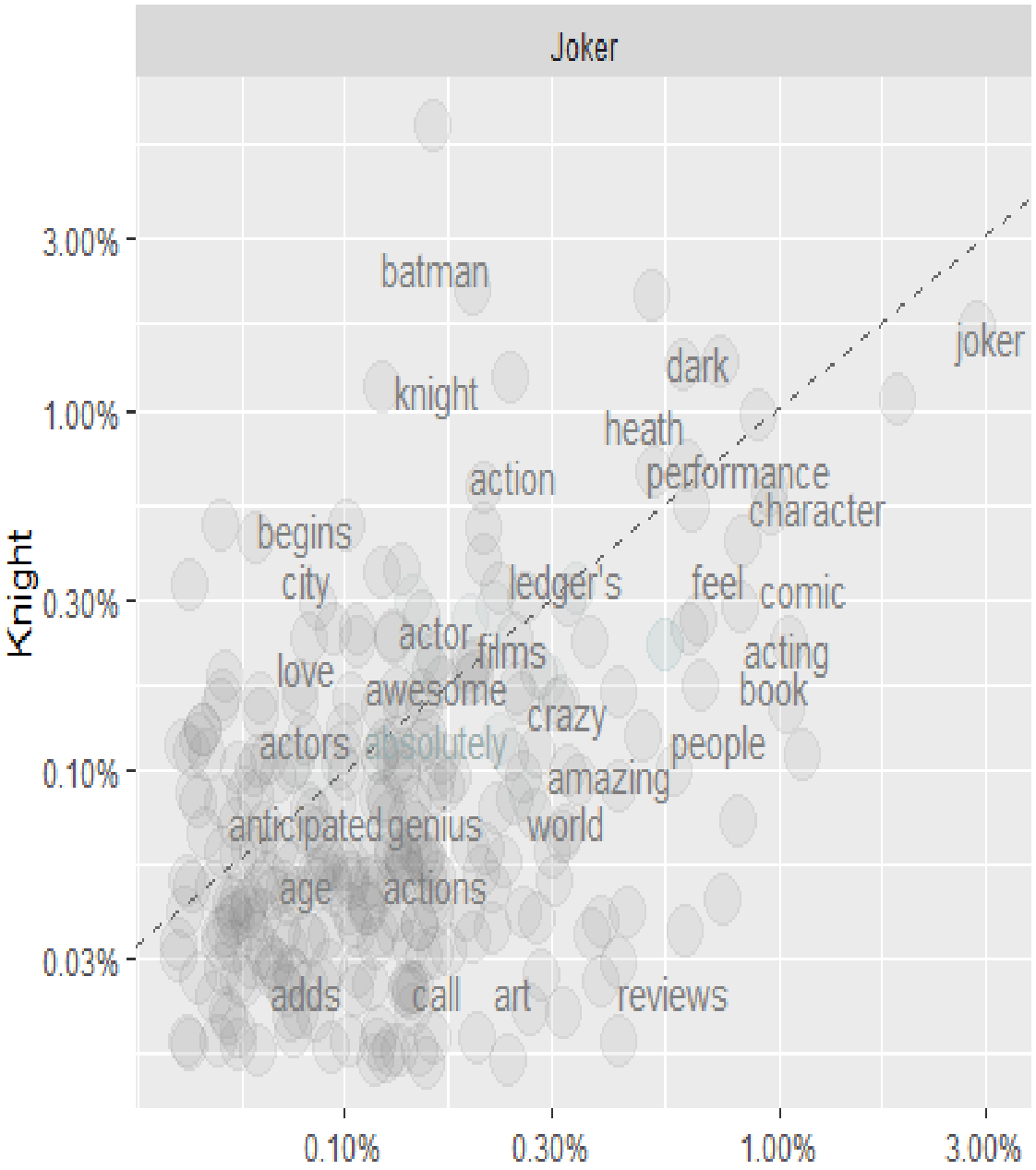
The core of a good movie can be full of hope, truth, goodness and beauty, or it can be enlightening and thought-provoking. The fight of the positive character batman and black tech ;Middle-sized character two-face justice to corruption; The joker's lunacy and calculation, the interplay between the good guys and the bad guys, make this a good popcorn movie even if it doesn't have a core, let alone a discussion of justice, darkness, and humanity. Unlike batman, the joker seeks absolute chaos, absolute darkness, which is the exact opposite of the order and light that batman seeks. Both of two movies are deserved high review score.

Appendix

#Correlation of "The Dark Knight" and "Joker" :

```
library(magrittr)

library(rvest)

library(dplyr)

library(tidyr)

library(tidyverse)

library(tidytext)

library(stringr)
library(ggplot2)
library(scales)

library(reshape2)

library(wordcloud)



Joker <- xml2::read_html("https://www.imdb.com/title/tt7286456/reviews?
ref_=tt_urv")
Joker_review <- Joker %>%
  html_nodes('.text') %>%
  html_text()
#View(Joker_review)

The_Dark_Knight <-xml2::read_html("https://www.imdb.com/title/tt0468569
/reviews?ref_=tt_urv")
The_Dark_Knight_review <- The_Dark_Knight %>%
  html_nodes('.text') %>%
  html_text()
#View(The_Dark_Knight_review)

df_Joker <- tibble(id=1:25, text=Joker_review)
#View(df_Joker)

df_The_Dark_Knight <- tibble(id=1:25, text= The_Dark_Knight_review)
#View(df_The_Dark_Knight)

cust_stop <- data_frame(
  word=c("movie","film"),
```

```
  lexicon=rep("custom",each=2)
)

Joker_review_frequencies <- df_Joker %>%
  unnest_tokens(word, text)%>%
  anti_join(stop_words) %>%
  anti_join(cust_stop)

The_Dark_Knight_review_frequencies <- df_The_Dark_Knight %>%
  unnest_tokens(word, text)%>%
  anti_join(stop_words) %>%
  anti_join(cust_stop)

frequency <- bind_rows(mutate(Joker_review_frequencies, movie="Joker"),
                       mutate(The_Dark_Knight_review_frequencies, movie="
Knight")
                       )  %>%  #closing bind_rows
  mutate(word=str_extract(word, "[a-z']+")) %>%
  count(movie, word) %>%
  group_by(movie) %>%
  mutate(proportion = n/sum(n)) %>%
  select(-n) %>%
  spread(movie, proportion) %>%
  gather(movie, proportion, `Joker`)

cor.test(data=frequency[frequency$movie == 'Joker',],
         ~proportion + `Knight`)

##  Pearson's product-moment correlation
## data:  proportion and Knight
## t = 11.104, df = 306, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4511788 0.6110741
## sample estimates:
##       cor
## 0.5359147

ggplot(frequency, aes(x=proportion, y=`Knight`,
                      color = abs(`Knight`- proportion))) +
  geom_abline(color="grey40", lty=2) +
  geom_jitter(alpha=.1, size=5.5, width=0.3, height=0.3) +
  geom_text(aes(label=word), check_overlap = TRUE, vjust=1.5) +
  scale_x_log10(labels = percent_format())+
  scale_y_log10(labels= percent_format())+
  scale_color_gradient(limits = c(0,0.0001), low = "darkslategray4", hi
gh = "gray75")+
  facet_wrap(~movie, ncol=2) +
  theme(legend.position = "none") +
  labs(y="Knight", x=NULL)
```

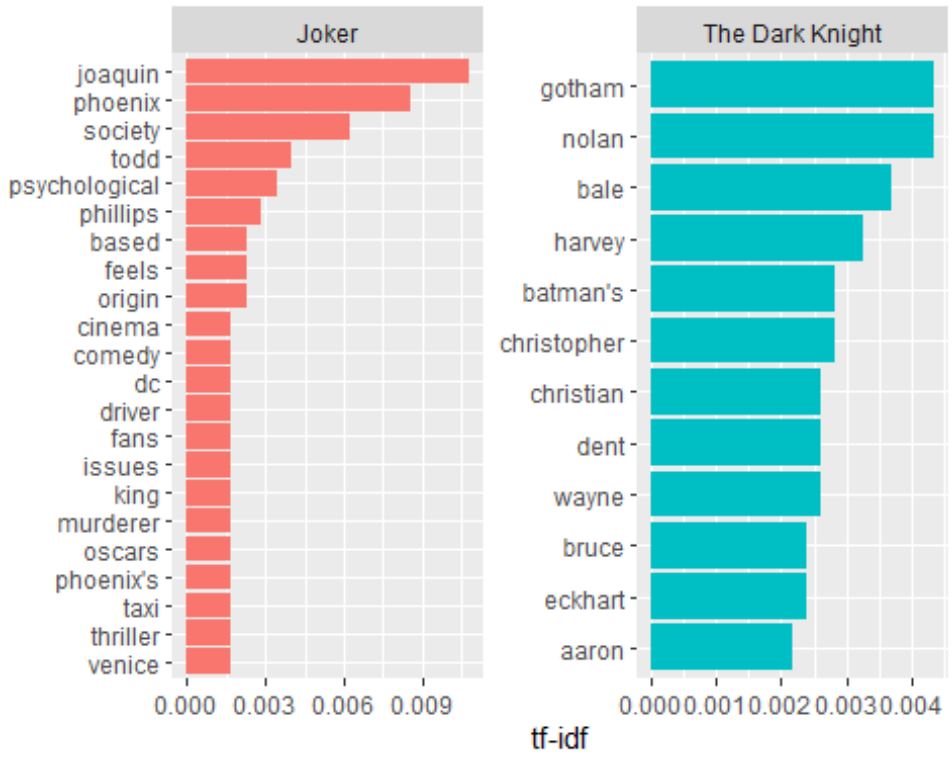


#Movie "The Dark Knight" and "Joker" TF-IDF analysis:

```
Joker_review_frequencies <- df_Joker %>%
  unnest_tokens(word, text)%>%
  anti_join(stop_words) %>%
```

```
  anti_join(cust_stop) %>%
  count(word, sort=TRUE)

The_Dark_Knight_review_frequencies <- df_The_Dark_Knight %>%
  unnest_tokens(word, text)%>%
  anti_join(stop_words) %>%
  anti_join(cust_stop) %>%
  count(word, sort=TRUE)

df <- bind_rows(
  mutate(Joker_review_frequencies,movie="Joker"),
  mutate(The_Dark_Knight_review_frequencies,movie="The Dark Knight")
)

movies_words <- df %>%
  bind_tf_idf(word, movie, n) #book is location info

movies_words # we get all the zeors because we are looking at stop word
s ... too common

## # A tibble: 2,260 x 6
##    word            n movie      tf   idf  tf_idf
##    <chr>       <int> <chr>   <dbl> <dbl>   <dbl>
##  1 joker          38 Joker 0.0312  0     0
##  2 joaquin        19 Joker 0.0156  0.693 0.0108
##  3 character      15 Joker 0.0123  0     0
##  4 phoenix        15 Joker 0.0123  0.693 0.00854
##  5 comic          14 Joker 0.0115  0     0
##  6 acting         13 Joker 0.0107  0     0
##  7 book           12 Joker 0.00985 0     0
##  8 society        11 Joker 0.00903 0.693 0.00626
##  9 time           11 Joker 0.00903 0     0
## 10 performance    10 Joker 0.00821 0     0
## # ... with 2,250 more rows

movies_words %>%
  arrange(desc(tf_idf))

## # A tibble: 2,260 x 6
##    word              n movie               tf   idf  tf_idf
##    <chr>         <int> <chr>            <dbl> <dbl>   <dbl>
##  1 joaquin          19 Joker           0.0156  0.693 0.0108
##  2 phoenix          15 Joker           0.0123  0.693 0.00854
##  3 society          11 Joker           0.00903 0.693 0.00626
##  4 gotham           20 The Dark Knight 0.00629 0.693 0.00436
##  5 nolan            20 The Dark Knight 0.00629 0.693 0.00436
##  6 todd              7 Joker           0.00575 0.693 0.00398
##  7 bale             17 The Dark Knight 0.00534 0.693 0.00370
##  8 psychological     6 Joker           0.00493 0.693 0.00341
##  9 harvey           15 The Dark Knight 0.00471 0.693 0.00327
```

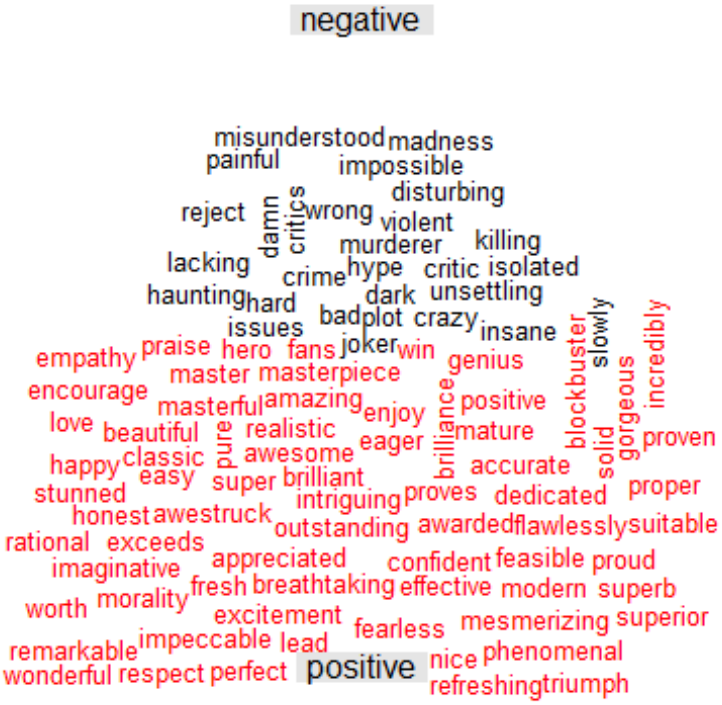
```
## 10 phillips          5 Joker           0.00411 0.693 0.00285
## # ... with 2,250 more rows


##############
# Looking at the graphical apprach:
movies_words %>%
  arrange(desc(tf_idf)) %>%
  mutate(word=factor(word, levels=rev(unique(word)))) %>%
  group_by(movie) %>%
  top_n(12) %>%
  ungroup %>%
  ggplot(aes(word, tf_idf, fill=movie))+
  geom_col(show.legend=FALSE)+
  labs(x=NULL, y="tf-idf")+
  facet_wrap(~movie, ncol=2, scales="free")+
  coord_flip()

## Selecting by tf_idf
```



#Sentiment cloud of "The Dark Knight" and "Joker" :

```
afinn <- get_sentiments("afinn")
nrc <- get_sentiments("nrc")
bing <- get_sentiments("bing")
```

```
Joker_review_frequencies <- df_Joker %>%
  unnest_tokens(word, text)%>%
  anti_join(stop_words) %>%
  anti_join(cust_stop) %>%
  inner_join(get_sentiments("bing")) %>%  #pizza flavor word cloud
  count(word, sentiment, sort=TRUE) %>%
  acast(word ~sentiment, value.var="n", fill=0) %>%
  comparison.cloud(colors = c("black", "red"),
                   max.words=100,
                   scale = c(0.8,0.8),
                   fixed.asp=TRUE,    #True 将长宽比例固定
                   title.size=1
  )
```




```
The_Dark_Knight_review_frequencies <- df_The_Dark_Knight %>%
  unnest_tokens(word, text)%>%
  anti_join(stop_words) %>%
  anti_join(cust_stop)%>%
  inner_join(get_sentiments("bing")) %>%  #pizza flavor word cloud
  count(word, sentiment, sort=TRUE) %>%
  acast(word ~sentiment, value.var="n", fill=0) %>%
  comparison.cloud(colors = c("black", "red"),
                   max.words=100,
                   scale = c(0.6,0.6),
                   fixed.asp=TRUE,
```

```
                    title.size=1
  )

The_Dark_Knight_review_frequencies <- df_The_Dark_Knight %>%
  unnest_tokens(word, text)%>%
  anti_join(stop_words) %>%
  anti_join(cust_stop)%>%
  inner_join(get_sentiments("nrc")) %>%  #pizza flavor word cloud
  count(word, sentiment, sort=TRUE) %>%
  acast(word ~sentiment, value.var="n", fill=0) %>%
  comparison.cloud(colors = c("black", "red","pink","yellow","orange","
grey","blue","green"),
                   max.words=100,
                   scale = c(0.8,0.8),
                   fixed.asp=TRUE,    #True 将长宽比例固定
                   title.size=1
  )
```

Spider plot of “The Dark Knight” and “Joker”

```
bigrams_separated_Joker <- df_Joker %>%
  unnest_tokens(bigram, text, token = "ngrams", n=2)%>%
  separate(bigram, c("word1", "word2"), sep = " ")

bigrams_separated_Knight <- df_The_Dark_Knight %>%
  unnest_tokens(bigram, text, token = "ngrams", n=2)%>%
  separate(bigram, c("word1", "word2"), sep = " ")
```

```r
bigram_counts_Joker <- bigrams_separated_Joker %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word) %>%
  count(word1, word2, sort = TRUE)

bigram_counts_Knight <- bigrams_separated_Knight %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word) %>%
  count(word1, word2, sort = TRUE)
negation_tokens <- c("no","never","without","not")

negated_words_Joker <- bigrams_separated_Joker %>%
  filter(word1 %in% negation_tokens) %>%
  inner_join(get_sentiments("afinn"), by=c(word2="word")) %>%
  count(word1, word2, value, sort=TRUE) %>%
  ungroup()

negated_words_Knight <- bigrams_separated_Knight %>%
  filter(word1 %in% negation_tokens) %>%
  inner_join(get_sentiments("afinn"), by=c(word2="word")) %>%
  count(word1, word2, value, sort=TRUE) %>%
  ungroup()

library(igraph)

library(ggraph)

## Warning: package 'ggraph' was built under R version 3.6.2

bigram_graph <- bind_rows(bigrams_separated_Knight,
                          bigrams_separated_Joker)%>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word) %>%
  count(word1, word2, sort = TRUE) %>%
  filter(n>1) %>%
  graph_from_data_frame ()

bigram_graph

## IGRAPH 8ce55d4 DN-- 134 100 --
## + attr: name (v/c), n (e/n)
## + edges from 8ce55d4 (vertex names):
##  [1] dark       ->knight      heath      ->ledger      comic      ->
book
##  [4] batman     ->begins      christopher->nolan       harvey     ->
dent
##  [7] bruce      ->wayne       christian  ->bale        joaquin    ->
phoenix
## [10] aaron      ->eckhart     gotham     ->city        batman     ->
```

```
movie
## [13] heath      ->ledger's    ledger's    ->performance 10          ->
10
## [16] ledger's   ->joker       michael     ->caine       morgan      ->
freeman
## [19] attorney   ->harvey      book        ->movie       district    ->
attorney
## [22] gary       ->oldman      jack        ->nicholson   superhero   ->
movie
## + ... omitted several edges
```

```
ggraph(bigram_graph, layout = "fr") +
  geom_edge_link()+
  geom_node_point()+
  geom_node_text(aes(label=name), vjust =1, hjust=1)
```