2020/5/12 NLP Over Movie Data

NLP Over Movie Data

Code ▼

Jinhang Jiang

Load data and find cut off

Hide

```
tag<- read.csv("tag.csv")
quantile(tag$RELEVANCE, probs = seq(0,1,0.10))</pre>
```

```
    0%
    10%
    20%
    30%
    40%
    50%
    60%
    70%
    80%
    90%
    100%

    0.00025
    0.01200
    0.02000
    0.02925
    0.04075
    0.05650
    0.07925
    0.11375
    0.17100
    0.29050
    1.00000
```

Hide

```
tag.top<-tag[tag$RELEVANCE>=0.2905,]
text<- tag.top$TAG</pre>
```

Load packages

Hide

```
library("NLP")
library("tm")
library("SnowballC")
library("RColorBrewer")
library("wordcloud2")
```

Clean text

Hide

```
docs<-tm_map(docs, removeWords, c("good", "great", "bad", "best", "movie", "notable", "full", "ba
sed", "nudity", "oscar", stopwords("english")))</pre>
```

transformation drops documents

Hide

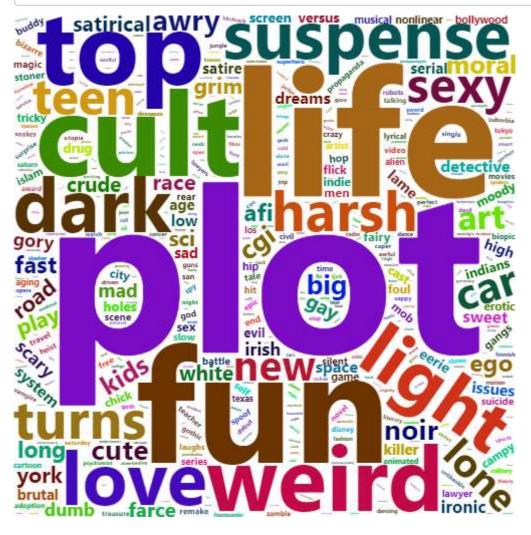
```
dtm<-TermDocumentMatrix(docs)
matrix<-as.matrix(dtm)
words<-sort(rowSums(matrix), decreasing=TRUE)

df<-data.frame(word=names(words), freq=words)</pre>
```

Prepare word clouds

Hide

```
set.seed(1)
wordcloud2(data = df, size = 1.5, color = 'random-dark')
```



Hide

NA

Do Sentiment Analysis

Hide

```
library("syuzhet")

text<-gsub("[][!#$%()*,.:;<=>@^_`|~.{}]", "", text)

t<-as.vector(text)

mysentiment<-get_nrc_sentiment((t))
```

Plot

2020/5/12 NLP Over Movie Data

Sentiments of Tags of the Top Rated Movies

