



# TEXT MINING FOR IMDB REVIEWS

*Le Thanh Hoang*

*Instructor: Professor Sergiy Shevchenko*

---

Class: ALY 6040 80971 Data Mining Applications SEC 02 Spring  
2018 CPS

# Contents

---

**1. Dataset**

**2. Pre-  
processing**

**3. Insights**

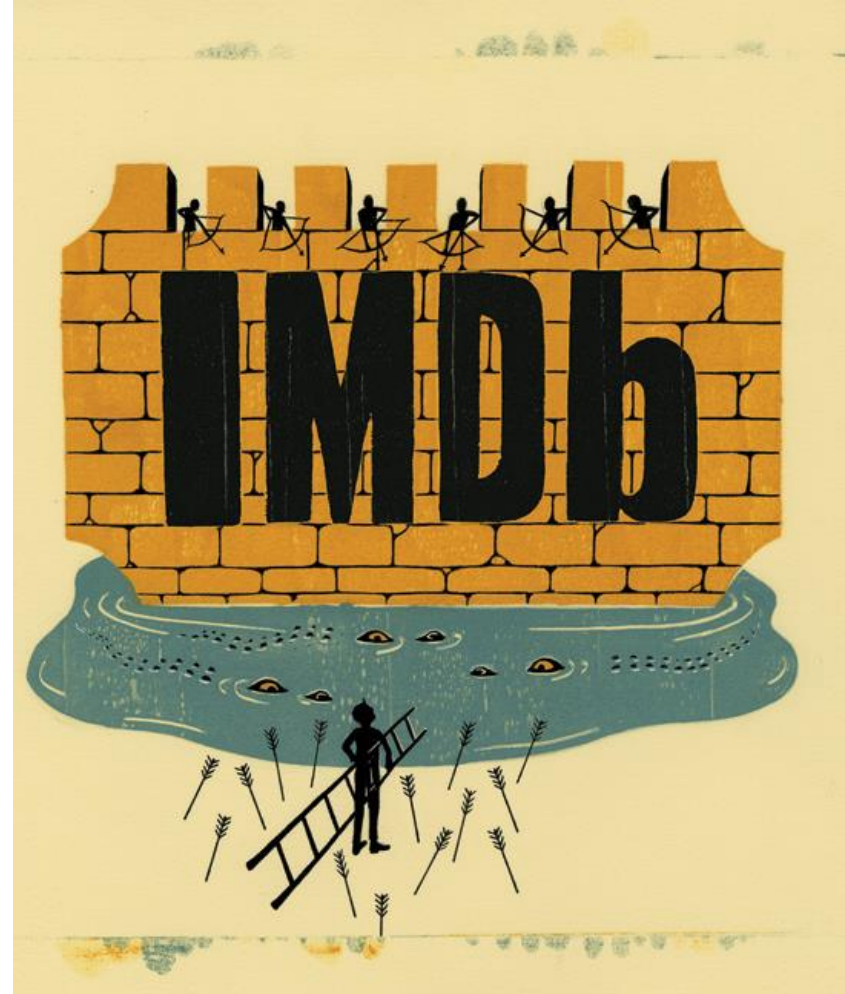
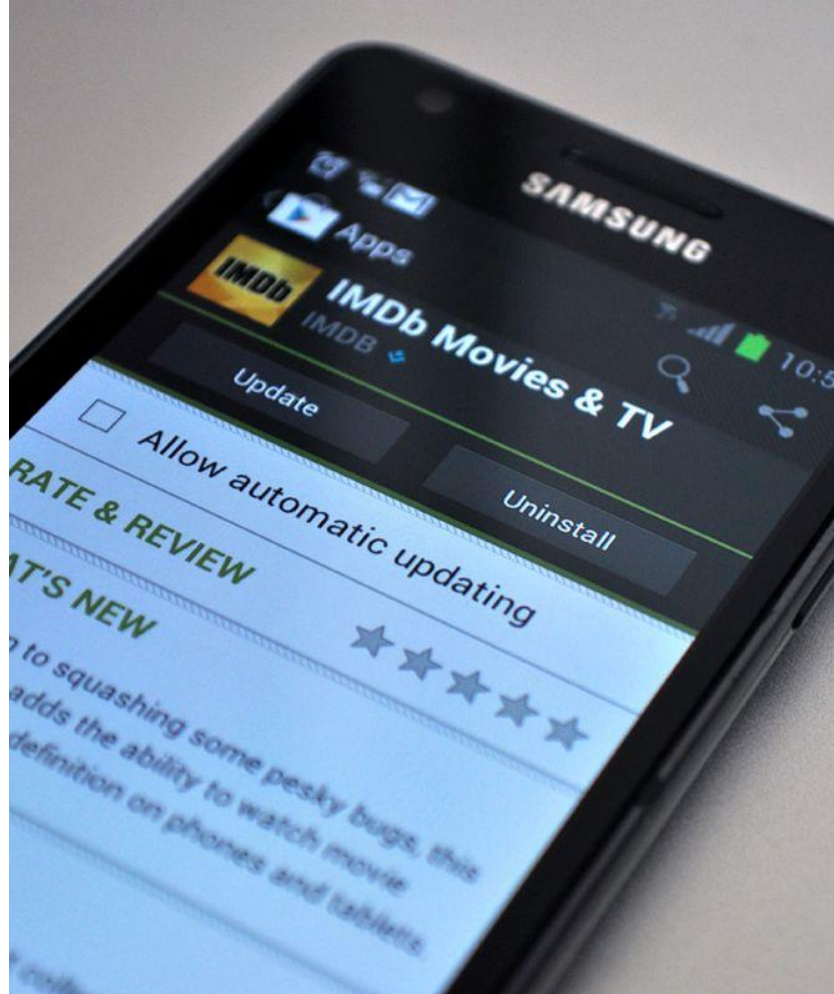
# 1. Dataset

---

IMDB, Internet Movie database is an online database of information related to word films, television programs, etc. with fan reviews and ratings. This site is extremely well-known among movie fanatics.

---

The dataset is the labeled dataset of 50,000 IMDB reviews, which has binary sentiment. It means that IMDB rating  $<5$  results in a sentiment score of 0 and IMDB rating  $\geq 7$  have a sentiment score of 1.



## 2. Pre-processing

## 2.1. Remove punctuation

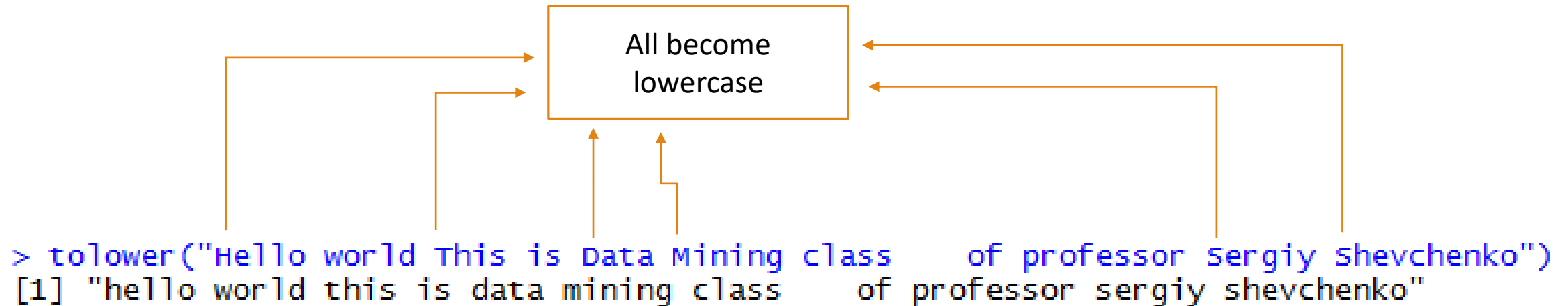
---

All punctuations  
have been removed

```
> removePunctuation("Hello @world. This is Data Mining class    of professor sergiy shevchenko!!!")  
[1] "Hello world This is Data Mining class    of professor sergiy shevchenko"
```

## 2.2. Makes all text lowercase

---





## 2.3. Remove stop words

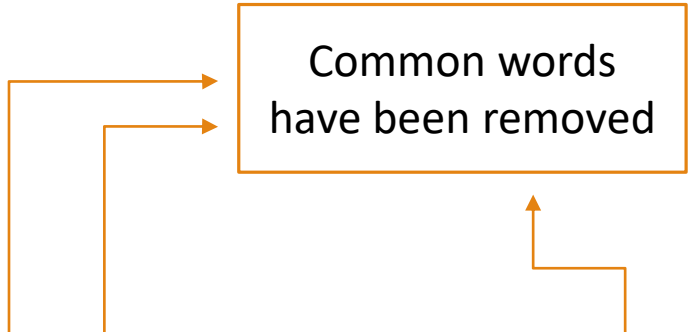
- Remove frequent words with little information
- “tm” package provides a list with 174 common words
- Users can add more words to this list based on context.

```
> stopwords("en")
[1] "i" "me" "my" "myself" "we" "our" "ours" "ourselves" "you" "your"
[11] "yours" "yourself" "yourselves" "he" "him" "his" "himself" "she" "her" "hers"
[21] "herself" "it" "its" "itself" "they" "them" "their" "theirs" "themselves" "what"
[31] "which" "who" "whom" "this" "that" "these" "those" "am" "is" "are"
[41] "was" "were" "be" "been" "being" "have" "has" "had" "having" "do"
[51] "does" "did" "doing" "would" "should" "could" "ought" "i'm" "you're" "he's"
[61] "she's" "it's" "we're" "they're" "i've" "you've" "we've" "they've" "i'd" "you'd"
[71] "he'd" "she'd" "we'd" "they'd" "i'll" "you'll" "he'll" "she'll" "we'll" "they'll"
[81] "isn't" "aren't" "wasn't" "weren't" "hasn't" "haven't" "hadn't" "doesn't" "don't" "didn't"
[91] "won't" "wouldn't" "shan't" "shouldn't" "can't" "cannot" "couldn't" "mustn't" "let's" "that's"
[101] "who's" "what's" "here's" "there's" "when's" "where's" "why's" "how's" "a" "an"
[111] "the" "and" "but" "if" "or" "because" "as" "until" "while" "of"
[121] "at" "by" "for" "with" "about" "against" "between" "into" "through" "during"
[131] "before" "after" "above" "below" "to" "from" "up" "down" "in" "out"
[141] "on" "off" "over" "under" "again" "further" "then" "once" "here" "there"
[151] "when" "where" "why" "how" "all" "any" "both" "each" "few" "more"
[161] "most" "other" "some" "such" "no" "nor" "not" "only" "own" "same"
[171] "so" "than" "too" "very"
```



## 2.3. Remove stop words

---

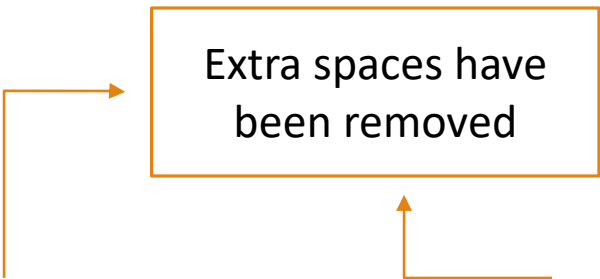


Common words  
have been removed

```
> removewords("hello world this is data mining class of professor sergiy shevchenko", stopwords("en"))  
[1] "hello world data mining class professor sergiy shevchenko"
```

## 2.4. Remove tabs and extra spaces

---



Extra spaces have been removed

```
> stripwhitespace("hello world    data mining class    professor sergiy shevchenko")  
[1] "hello world data mining class professor sergiy shevchenko"
```

The diagram illustrates the function's effect. An orange box with the text "Extra spaces have been removed" has two arrows pointing to the original string in the code. One arrow points to the four spaces between "world" and "data", and the other points to the four spaces between "class" and "professor". The output string shows these spaces reduced to single spaces.

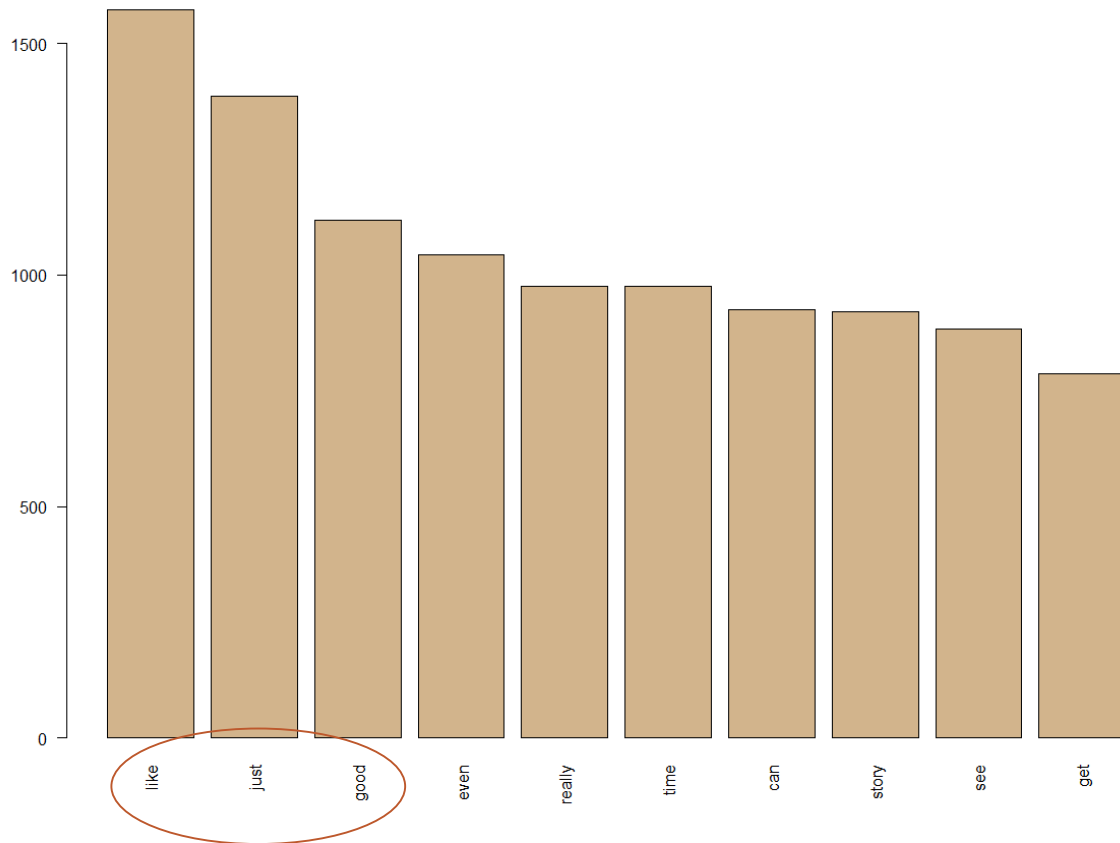
---

## 3. Insights



## 3.1. Bar plot

---



Most common words in any comment is “like”, followed by “just” and “good”.

=> Reviewers may always like the movies although it cannot be good as expected, no matter how bad it is because there will always be something to be said about the movie.

The second common word “just” may indicate that reviews are not totally disregard or be fond of movies, they are “just” somehow into movies for some aspects because flawless films are rare.

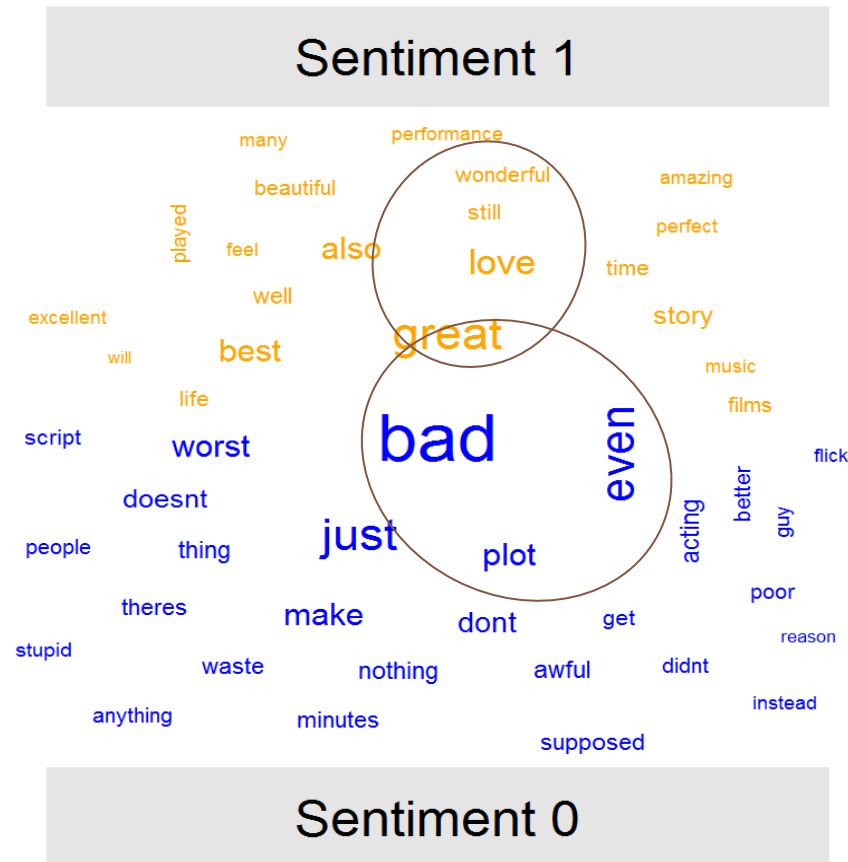
## 3.2. Commonality cloud



Commonality cloud is cloud of words in common in both sentiment.

“good” is dominant, meaning that there are many “good” in both positive and negative reviews. In other words, a bad movie can still have some good characteristics that draw a crowd.

## 3.3. Comparison cloud



Comparison cloud depicts words not in common between two sentiments.

If a review contains such words like “great”, “wonderful” or “love”, it would be categorized as a positive review.

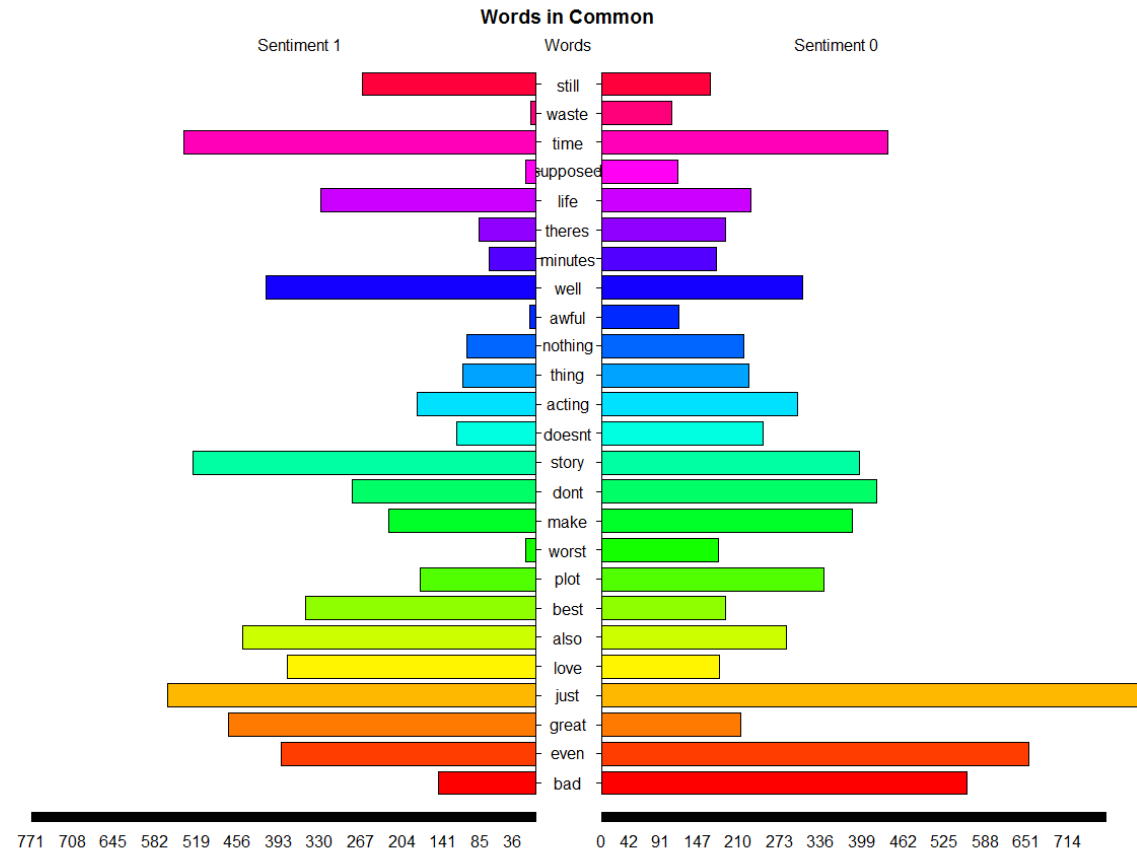
Conversely, a negative review will consist of those words such as “bad”, “plot”, “even” or “nothing”

## 3.4. Pyramid plot

Pyramid plot is the combination of commonality cloud and comparison cloud because it presents the common words in both sentiment with the highest difference in frequency

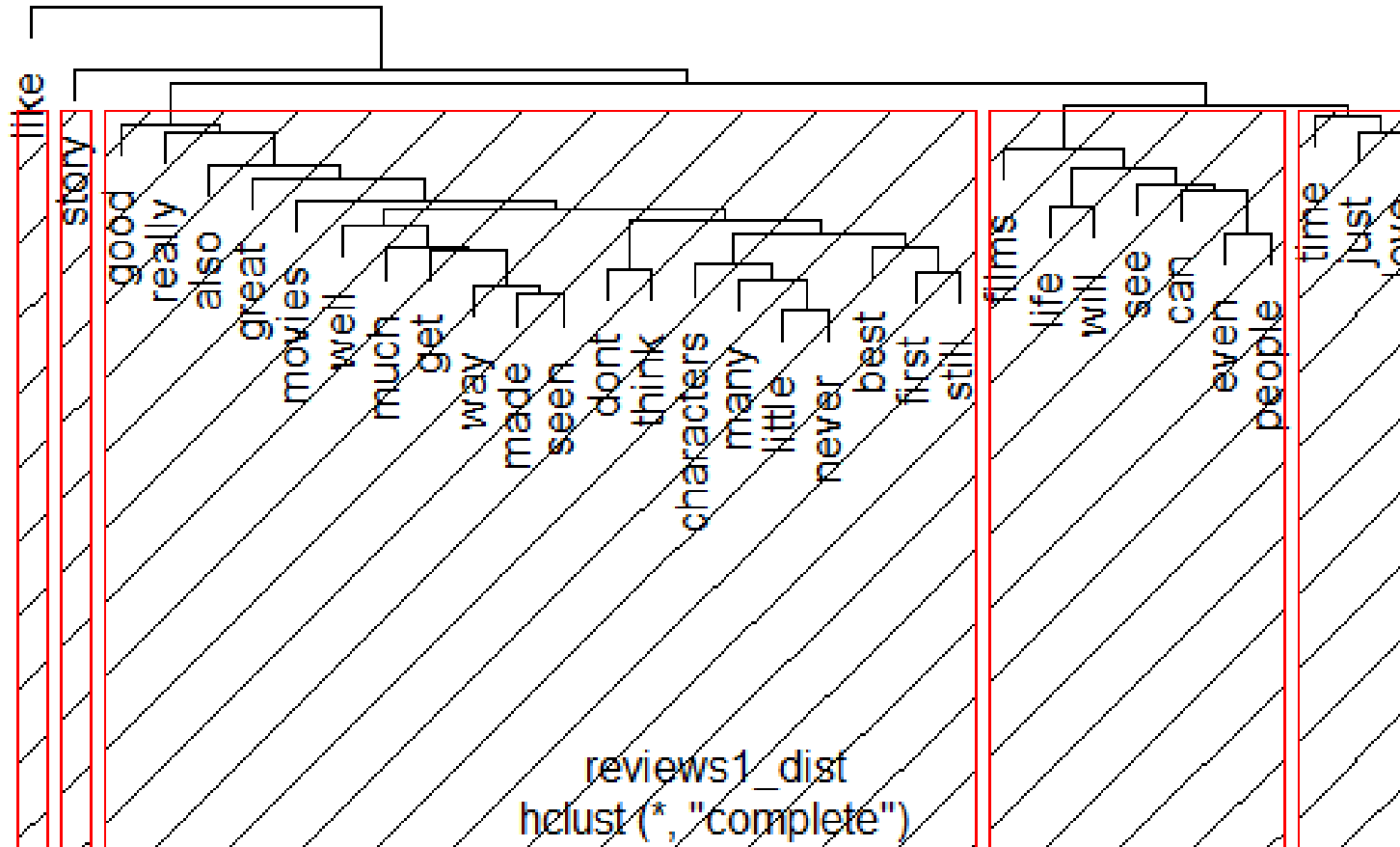
There may be a clear distinction in some common words between two sentiments. For instance, the word “bad” can be common but it is results mostly from negative sentiment.

The same pattern happens with “even” word and “just” word. The word “great”, “love”, “also”, “best” is the main driving force to make reviews become commonplace in positive reviews and in total.





## Cluster Dendrogram



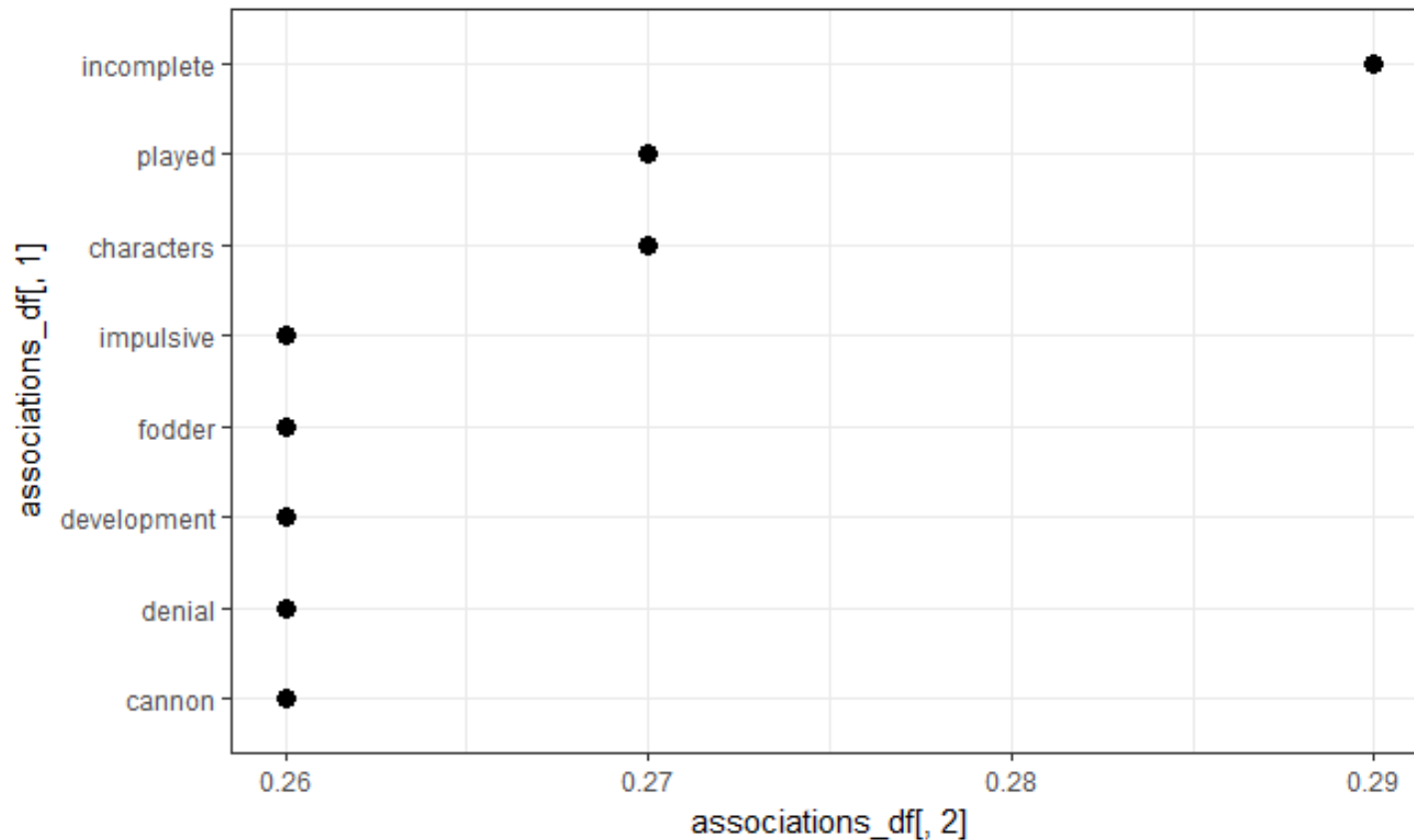
## 3.5. Dendrogram for Sentiment 1

This is a dendrogram with 5 clusters

In cluster 5, “time” is in the same cluster as “love” => for positive reviews of “love” story, the romantic and precious time could be mentioned and vice versa.

In the cluster 4, “people” goes along with “life” and “see => those movies revolving around “life” could involves the social network around “people” when they “see” each other.

“story” is in its own cluster, meaning that the term “story” doesn’t often appear with “love”, or “people” at the same time or “story” appears in almost any documents so there is negligible connection with other terms.



### 3.6. Associations for the word “character” in Sentiment 1

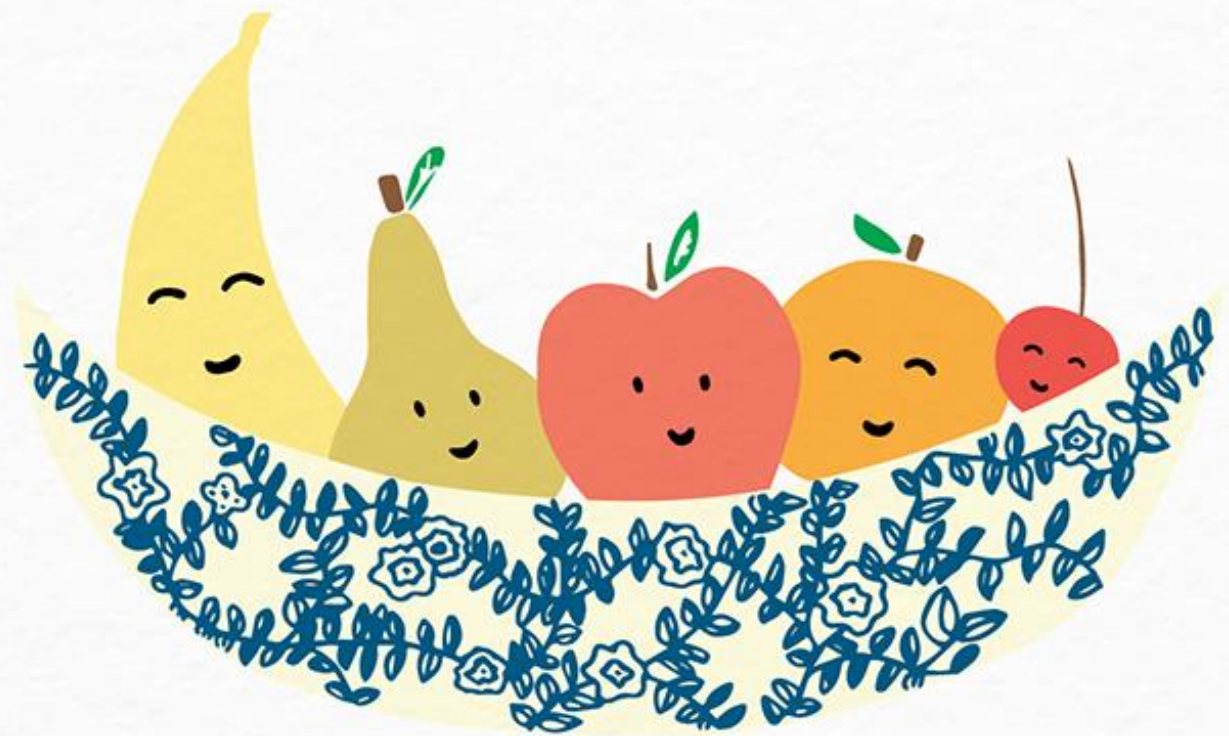
The word “character” is most related to the word “incomplete” with the correlation of 0.29, followed by “played” and “character”.

It could be inferred that for positive reviews, those “character” could be “incomplete”, “impulsive” or in “development”

# Summary

---

1. For a large majority of movies, reviewers are not completely satisfied.
2. There would always be a good aspect about a movie even if it is really bad.
3. There are different words distinguish a good and a bad review.
4. Distinguished words may not necessarily be completely different words.
5. There will always be some words happening at the same time.



T H A N K   Y O U