Project2 - Text Mining

Dipesh Poudel

10/9/2021

Text Mining

A huge portion of information is stored in form of text in medium such as news articles, papers, books, email, blogs, websites etc. It becomes important that we can extract valuable information and intelligence from the text. The process of extracting the meaningful information and intelligence from the text data by analyzing relations, patterns and rules in text data is called text mining.

Text Mining and Natural Language Processing(NLP)

Natural Language Processing (NLP) is a sub field of artificial intelligence which deals with making machines understand human language (natural language). It includes both verbal and written aspect of natural language.

Text mining is used for extracting meaningful information from unstructured and structured text. In text mining, interesting patterns in the text data and identified and analyzed but the semantics in the text is not analyzed where as in the NLP we need to understand the semantics of the text as well.

We can say that NLP is a component of text mining that performs linguistic analysis that helps machine understand the text.

Text Mining and Machine Learning

Machine Learning provides computer (system) the ability to learn from the data without explicitly programming them.

In the process of text mining we use machine learning techniques like clustering for segmenting texts into several clusters depending on the substantial relevance.

Text Mining and Artifical Intelligence

Turing Test designed by Alan Turing in 1950 is a very famous test of intelligence of a agent(machine). A computer passes the test if a human integrator, after posing written questions, cannot tell whether the response is from human or machine. For the machine to understand the question it needs to be able to store the text information, analyze it and give out the

appropriate response. For this to be possible we need to use text mining and natural language processing.

Text Mining With Large Movie Review Dataset

We will perform text mining using the Large Movie Review Dataset.

This is a data set that contains both negative and positive movie reviews. Since we will not be doing sentiment analysis, For this project I have only used few files as I got Memory Error as the number of files go up.

Step 1: Loading the Dataset and building a corpus

```
library("tm")

## Loading required package: NLP

file_source<-DirSource("movie_reviews/")
reviewCorpus<-Corpus(file_source,readerControl = list(language="lat"))
```

Inspecting the First Five Elements of the text

```
inspect(reviewCorpus[1:5])

## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 5
##
##
0_10.txt
##
I went and saw this movie last night after being coaxed to by a few
friends of mine. I'll admit that I was reluctant to see it because
from what I knew of Ashton Kutcher he was only able to do comedy. I
was wrong. Kutcher played the character of Jake Fischer very well, and
Kevin Costner played Ben Randall with such professionalism. The sign
of a good movie is that it can toy with our emotions. This one did
exactly that. The entire theater (which was sold out) was overcome by
laughter during the first half of the movie, and were moved to tears
during the second half. While exiting the theater I not only saw many
women in tears, but many full grown men as well, trying desperately
not to let anyone see them crying. This movie was great, and I suggest
that you go see it before you judge.
##
0_2.txt
##
Once again Mr. Costner has dragged out a movie for far longer than
necessary. Aside from the terrific sea rescue sequences, of which
there are very few I just did not care about any of the characters.
Most of us have ghosts in the closet, and Costner's character are
realized early on, and then forgotten until much later, by which time
```

I did not care. The character we should really care about is a very cocky, overconfident Ashton Kutcher. The problem is he comes off as kid who thinks he's better than anyone else around him and shows no signs of a cluttered closet. His only obstacle appears to be winning over Costner. Finally when we are well past the half way point of this stinker, Costner tells us all about Kutcher's ghosts. We are told why Kutcher is driven to be the best with no prior inkling or foreshadowing. No magic here, it was all I could do to keep from turning it off an hour in.
##
1_10.txt
##
My boyfriend and I went to watch The Guardian.At first I didn't want to watch it, but I loved the movie- It was definitely the best movie I have seen in sometime.They portrayed the USCG very well, it really showed me what they do and I think they should really be appreciated more.Not only did it teach but it was a really good movie. The movie shows what the really do and how hard the job is.I think being a USCG would be challenging and very scary. It was a great movie all around. I would suggest this movie for anyone to see.The ending broke my heart but I know why he did it. The storyline was great I give it 2 thumbs up. I cried it was very emotional, I would give it a 20 if I could!
##
1_3.txt
##
This is a pale imitation of 'Officer and a Gentleman.' There is NO chemistry between Kutcher and the unknown woman who plays his love interest. The dialog is wooden, the situations hackneyed. It's too long and the climax is anti-climactic(!). I love the USCG, its men and women are fearless and tough. The action scenes are awesome, but this movie doesn't do much for recruiting, I fear. The script is formulaic, but confusing. Kutcher's character is trying to redeem himself for an accident that wasn't his fault? Costner's is raging against the dying of the light, but why? His 'conflict' with his wife is about as deep as a mud puddle. I saw this sneak preview for free and certainly felt I got my money's worth.
##
2_3.txt
It seems ever since 1982, about every two or three years we get a movie that claims to be "The Next Officer and a Gentleman." There has yet to be one movie that has lived up to this claim and this movie is no different.

We get the usual ripped off scenes from OAAG ("I want you DOR," the instructor gives the Richard Gere character his overdose of drills in hopes he'll quit, the Gere character comes back for the girl, the Gere character realizes the instructor is great, etc.) and this movie is as predictable as the sun rising in the East and is horribly miscast on top. Costner plays his usual "wise teacher" character, the only character he can play, and you really get a sense of his limited acting abilities here. Kutcher is terrible in the Richard Gere character, just miscast with acting skills barely a notch

```
above Keanu Reeves.<br /><br />The main problem with this OAAG wannabe
is the two main characters are so amazingly one-dimensional, you never
care for either in the least and when Kutcher's character finally
turns around (just like Gere did in OAAG) you just go "so what? The
movie leaves no plot point unturned and seems to never end as if to
say "oh wait, we forgot to close out the girlfriend story, or the what
happens after he graduates story, or the other six plot points in the
movie..." What's more baffling is the great "reviews" I see here. The
general public's opinions never cease to amaze me.
```

Step 2: Preprocessing

Before analyzing the text we need to preprocess the data. The text data contains punctuation, white spaces, numbers, words that have no meaning but is used very frequently like is, the, an etc these words are called stop words and we need to remove them.

Removing Punctuation

```
reviewCorpus <- tm_map(reviewCorpus, removePunctuation)
inspect(reviewCorpus[1:3])

## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 3
##
##
0_10.txt
##
I went and saw this movie last night after being coaxed to by a few
friends of mine Ill admit that I was reluctant to see it because from
what I knew of Ashton Kutcher he was only able to do comedy I was
wrong Kutcher played the character of Jake Fischer very well and Kevin
Costner played Ben Randall with such professionalism The sign of a
good movie is that it can toy with our emotions This one did exactly
that The entire theater which was sold out was overcome by laughter
during the first half of the movie and were moved to tears during the
second half While exiting the theater I not only saw many women in
tears but many full grown men as well trying desperately not to let
anyone see them crying This movie was great and I suggest that you go
see it before you judge
##
0_2.txt
## Once again Mr Costner has dragged out a movie for far longer than
necessary Aside from the terrific sea rescue sequences of which there
are very few I just did not care about any of the characters Most of
us have ghosts in the closet and Costners character are realized early
on and then forgotten until much later by which time I did not care
The character we should really care about is a very cocky
overconfident Ashton Kutcher The problem is he comes off as kid who
thinks hes better than anyone else around him and shows no signs of a
```

```
cluttered closet His only obstacle appears to be winning over Costner
Finally when we are well past the half way point of this stinker
Costner tells us all about Kutchers ghosts We are told why Kutcher is
driven to be the best with no prior inkling or foreshadowing No magic
here it was all I could do to keep from turning it off an hour in
##
1_10.txt
##
My boyfriend and I went to watch The GuardianAt first I didnt want to
watch it but I loved the movie It was definitely the best movie I have
seen in sometimeThey portrayed the USCG very well it really showed me
what they do and I think they should really be appreciated moreNot
only did it teach but it was a really good movie The movie shows what
the really do and how hard the job isI think being a USCG would be
challenging and very scary It was a great movie all around I would
suggest this movie for anyone to seeThe ending broke my heart but I
know why he did it The storyline was great I give it 2 thumbs up I
cried it was very emotional I would give it a 20 if I could
```

Removing Numbers

```
reviewCorpus<-tm_map(reviewCorpus,removeNumbers)
```

Changing the Text to Lower case

Since the case of the word do not have much meaning we convert all the words into lower case. For example word `Aside` and `aside` are same but `R` is case sensitive so it reads them as different word. So to deal with this issue we are converting the words to lower case.

```
reviewCorpus<-tm_map(reviewCorpus,tolower)
inspect(reviewCorpus[1:3])

## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 3
##
##
0_10.txt
##
i went and saw this movie last night after being coaxed to by a few
friends of mine ill admit that i was reluctant to see it because from
what i knew of ashton kutcher he was only able to do comedy i was
wrong kutcher played the character of jake fischer very well and kevin
costner played ben randall with such professionalism the sign of a
good movie is that it can toy with our emotions this one did exactly
that the entire theater which was sold out was overcome by laughter
during the first half of the movie and were moved to tears during the
second half while exiting the theater i not only saw many women in
tears but many full grown men as well trying desperately not to let
anyone see them crying this movie was great and i suggest that you go
see it before you judge
```

```
##
0_2.txt
## once again mr costner has dragged out a movie for far longer than
necessary aside from the terrific sea rescue sequences of which there
are very few i just did not care about any of the characters most of
us have ghosts in the closet and costners character are realized early
on and then forgotten until much later by which time i did not care
the character we should really care about is a very cocky
overconfident ashton kutcher the problem is he comes off as kid who
thinks hes better than anyone else around him and shows no signs of a
cluttered closet his only obstacle appears to be winning over costner
finally when we are well past the half way point of this stinker
costner tells us all about kutchers ghosts we are told why kutcher is
driven to be the best with no prior inkling or foreshadowing no magic
here it was all i could do to keep from turning it off an hour in
##
1_10.txt
##
my boyfriend and i went to watch the guardianat first i didnt want to
watch it but i loved the movie it was definitely the best movie i have
seen in sometimethey portrayed the uscg very well it really showed me
what they do and i think they should really be appreciated morenot
only did it teach but it was a really good movie the movie shows what
the really do and how hard the job isi think being a uscg would be
challenging and very scary it was a great movie all around i would
suggest this movie for anyone to seethe ending broke my heart but i
know why he did it the storyline was great i give it  thumbs up i
cried it was very emotional i would give it a  if i could
```

Removing the Stop Words

Before removing the stop words lets take a look at them. These are words that are used very often but do not carry much meaning.

```
stopwords("english")
```

```
##   [1] "i"          "me"         "my"         "myself"     "we"

##   [6] "our"        "ours"       "ourselves"  "you"        "your"

##  [11] "yours"      "yourself"   "yourselves" "he"         "him"

##  [16] "his"        "himself"    "she"        "her"        "hers"

##  [21] "herself"    "it"         "its"        "itself"     "they"

##  [26] "them"       "their"      "theirs"     "themselves" "what"

##  [31] "which"      "who"        "whom"       "this"       "that"
```

```
##  [36] "these"      "those"      "am"         "is"         "are"
##  [41] "was"        "were"       "be"         "been"       "being"
##  [46] "have"       "has"        "had"        "having"     "do"
##  [51] "does"       "did"        "doing"      "would"      "should"
##  [56] "could"      "ought"      "i'm"        "you're"     "he's"
##  [61] "she's"      "it's"       "we're"      "they're"    "i've"
##  [66] "you've"     "we've"      "they've"    "i'd"        "you'd"
##  [71] "he'd"       "she'd"      "we'd"       "they'd"     "i'll"
##  [76] "you'll"     "he'll"      "she'll"     "we'll"      "they'll"
##  [81] "isn't"      "aren't"     "wasn't"     "weren't"    "hasn't"
##  [86] "haven't"    "hadn't"     "doesn't"    "don't"      "didn't"
##  [91] "won't"      "wouldn't"   "shan't"     "shouldn't"  "can't"
##  [96] "cannot"     "couldn't"   "mustn't"    "let's"      "that's"
## [101] "who's"      "what's"     "here's"     "there's"    "when's"
## [106] "where's"    "why's"      "how's"      "a"          "an"
## [111] "the"        "and"        "but"        "if"         "or"
## [116] "because"    "as"         "until"      "while"      "of"
## [121] "at"         "by"         "for"        "with"       "about"
## [126] "against"    "between"    "into"       "through"    "during"
## [131] "before"     "after"      "above"      "below"      "to"
## [136] "from"       "up"         "down"       "in"         "out"
## [141] "on"         "off"        "over"       "under"      "again"
## [146] "further"    "then"       "once"       "here"       "there"
## [151] "when"       "where"      "why"        "how"        "all"
```

```
## [156] "any"        "both"       "each"       "few"        "more"

## [161] "most"       "other"      "some"       "such"       "no"

## [166] "nor"        "not"        "only"       "own"        "same"

## [171] "so"         "than"       "too"        "very"
```

```
reviewStopWords<-c(stopwords("english"),"<br />", "br","<br>")
reviewCorpus<-tm_map(reviewCorpus,removeWords,reviewStopWords)
inspect(reviewCorpus[1:3])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 3
##
##
0_10.txt
##
went  saw  movie last night   coaxed     friends  mine ill admit
reluctant  see      knew  ashton kutcher    able   comedy   wrong
kutcher played  character  jake fischer  well  kevin costner played
ben randall   professionalism  sign   good movie    can toy   emotions
one  exactly   entire theater   sold   overcome  laughter   first half
movie   moved  tears   second half  exiting  theater    saw many women
tears  many full grown men  well trying desperately   let anyone see
crying  movie  great   suggest   go see    judge
##
0_2.txt
##   mr costner  dragged   movie  far longer  necessary aside
terrific sea rescue sequences        just   care     characters   us
ghosts   closet  costners character  realized early    forgotten  much
later   time    care  character   really care     cocky overconfident
ashton kutcher  problem   comes   kid  thinks hes better  anyone else
around   shows  signs   cluttered closet   obstacle appears   winning
costner finally    well past  half way point   stinker costner tells
us   kutchers ghosts   told  kutcher  driven    best   prior inkling
foreshadowing  magic         keep  turning    hour
##
1_10.txt
##
boyfriend   went  watch  guardianat first  didnt want  watch    loved
movie   definitely  best movie   seen  sometimethey portrayed  uscg
well  really showed       think   really  appreciated morenot    teach
really good movie  movie shows   really    hard  job isi think   uscg
challenging   scary    great movie  around   suggest  movie  anyone
seethe ending broke  heart   know      storyline  great  give   thumbs
cried    emotional   give
```

Removing Extra Whitespaces

```
reviewCorpus <- tm_map(reviewCorpus, stripWhitespace)
inspect(reviewCorpus[1:3])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 3
##
##
0_10.txt
##                                                          went saw movie
last night coaxed friends mine ill admit reluctant see knew ashton
kutcher able comedy wrong kutcher played character jake fischer well
kevin costner played ben randall professionalism sign good movie can
toy emotions one exactly entire theater sold overcome laughter first
half movie moved tears second half exiting theater saw many women
tears many full grown men well trying desperately let anyone see
crying movie great suggest go see judge
##
0_2.txt
##  mr costner dragged movie far longer necessary aside terrific sea
rescue sequences just care characters us ghosts closet costners
character realized early forgotten much later time care character
really care cocky overconfident ashton kutcher problem comes kid
thinks hes better anyone else around shows signs cluttered closet
obstacle appears winning costner finally well past half way point
stinker costner tells us kutchers ghosts told kutcher driven best
prior inkling foreshadowing magic keep turning hour
##
1_10.txt
##
boyfriend went watch guardianat first didnt want watch loved movie
definitely best movie seen sometimethey portrayed uscg well really
showed think really appreciated morenot teach really good movie movie
shows really hard job isi think uscg challenging scary great movie
around suggest movie anyone seethe ending broke heart know storyline
great give thumbs cried emotional give
```

Stemming

Steaming is the process of converting words to their base or root form. For example, eat, eating, eats will be converted to eat.

```
# Making a copy of the corpus
reviewCorpusCP<-reviewCorpus

# stem words
reviewCorpus<-tm_map(reviewCorpus,stemDocument)
inspect(reviewCorpus[1:3])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 3
##
##
0_10.txt
##                                                 went saw movi last
night coax friend mine ill admit reluct see knew ashton kutcher abl
comedi wrong kutcher play charact jake fischer well kevin costner play
ben randal profession sign good movi can toy emot one exact entir
theater sold overcom laughter first half movi move tear second half
exit theater saw mani women tear mani full grown men well tri desper
let anyon see cri movi great suggest go see judg
##
0_2.txt
## mr costner drag movi far longer necessari asid terrif sea rescu
sequenc just care charact us ghost closet costner charact realiz earli
forgotten much later time care charact realli care cocki overconfid
ashton kutcher problem come kid think hes better anyon els around show
sign clutter closet obstacl appear win costner final well past half
way point stinker costner tell us kutcher ghost told kutcher driven
best prior inkl foreshadow magic keep turn hour
##
1_10.txt
##
boyfriend went watch guardianat first didnt want watch love movi
definit best movi seen sometimethey portray uscg well realli show
think realli appreci morenot teach realli good movi movi show realli
hard job isi think uscg challeng scari great movi around suggest movi
anyon seeth end broke heart know storylin great give thumb cri emot
give
```

Steam Completion

```
tm_map(reviewCorpus, stemCompletion, dictionary = reviewCorpusCP)

## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 10

inspect(reviewCorpus[1:3])

## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 3
##
##
0_10.txt
##                                                 went saw movi last
night coax friend mine ill admit reluct see knew ashton kutcher abl
comedi wrong kutcher play charact jake fischer well kevin costner play
ben randal profession sign good movi can toy emot one exact entir
```

```
theater sold overcom laughter first half movi move tear second half
exit theater saw mani women tear mani full grown men well tri desper
let anyon see cri movi great suggest go see judg
##
0_2.txt
## mr costner drag movi far longer necessari asid terrif sea rescu
sequenc just care charact us ghost closet costner charact realiz earli
forgotten much later time care charact realli care cocki overconfid
ashton kutcher problem come kid think hes better anyon els around show
sign clutter closet obstacl appear win costner final well past half
way point stinker costner tell us kutcher ghost told kutcher driven
best prior inkl foreshadow magic keep turn hour
##
1_10.txt
##
boyfriend went watch guardianat first didnt want watch love movi
definit best movi seen sometimethey portray uscg well realli show
think realli appreci morenot teach realli good movi movi show realli
hard job isi think uscg challeng scari great movi around suggest movi
anyon seeth end broke heart know storylin great give thumb cri emot
give
```

Fixing some seen issues caused by steamining

```
# Changing Movi to movie
reviewCorpus <- tm_map(reviewCorpus, gsub, pattern="movi",
replacement="movie")

# Changing comedi to comedy
reviewCorpus <- tm_map(reviewCorpus, gsub, pattern="comedi",
replacement="comdey")

# Removing other Issues seen after Inspecting
reviewCorpus <- tm_map(reviewCorpus, gsub, pattern="necessari",
replacement="necessary")
reviewCorpus <- tm_map(reviewCorpus, gsub, pattern="earli",
replacement="early")
reviewCorpus <- tm_map(reviewCorpus, gsub, pattern="realiz",
replacement="realize")
reviewCorpus <- tm_map(reviewCorpus, gsub, pattern="realli",
replacement="really")
reviewCorpus <- tm_map(reviewCorpus, gsub, pattern="overconfid",
replacement="overconfidence")
reviewCorpus <- tm_map(reviewCorpus, gsub, pattern="cocki",
replacement="cocky")
reviewCorpus <- tm_map(reviewCorpus, gsub, pattern="asid",
replacement="aside")
reviewCorpus <- tm_map(reviewCorpus, gsub, pattern="appreci",
replacement="appreciate")
reviewCorpus <- tm_map(reviewCorpus, gsub, pattern="definit",
replacement="definitive")
```

```
reviewCorpus <- tm_map(reviewCorpus, gsub, pattern="emot",
replacement="emotion")
reviewCorpus <- tm_map(reviewCorpus, gsub, pattern="mani",
replacement="many")
reviewCorpus <- tm_map(reviewCorpus, gsub, pattern="charact",
replacement="character")
reviewCorpus <- tm_map(reviewCorpus, gsub, pattern="stori",
replacement="story")
reviewCorpus <- tm_map(reviewCorpus, gsub, pattern="entir",
replacement="entire")
reviewCorpus <- tm_map(reviewCorpus, gsub, pattern="overcom",
replacement="overcome")
reviewCorpus <- tm_map(reviewCorpus, gsub, pattern="cri",
replacement="cry")

inspect(reviewCorpus[1:3])

## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 3
##
##
0_10.txt
##                                                    went saw movie
last night coax friend mine ill admit reluct see knew ashton kutcher
abl comdey wrong kutcher play character jake fischer well kevin
costner play ben randal profession sign good movie can toy emotion one
exact entire theater sold overcome laughter first half movie move tear
second half exit theater saw many women tear many full grown men well
tri desper let anyon see cry movie great suggest go see judg
##
0_2.txt
## mr costner drag movie far longer necessary aside terrif sea rescu
sequenc just care character us ghost closet costner character realize
early forgotten much later time care character really care cocky
overconfidence ashton kutcher problem come kid think hes better anyon
els around show sign clutter closet obstacl appear win costner final
well past half way point stinker costner tell us kutcher ghost told
kutcher driven best prior inkl foreshadow magic keep turn hour
##
1_10.txt
##
boyfriend went watch guardianat first didnt want watch love movie
definitive best movie seen sometimethey portray uscg well really show
think really appreciate morenot teach really good movie movie show
really hard job isi think uscg challeng scari great movie around
suggest movie anyon seeth end broke heart know storylin great give
thumb cry emotion give
```

Term Document Matrix

A term documnet matrix is a matrix that gives the frequency of the terms that occurs in the corpus.

```
reviewTdm <-
TermDocumentMatrix(reviewCorpus,control=list(wordLengths=c(1,Inf)))
print(reviewTdm)

## <<TermDocumentMatrix (terms: 502, documents: 10)>>
## Non-/sparse entries: 719/4301
## Sparsity           : 86%
## Maximal term length: 15
## Weighting          : term frequency (tf)

(freq.terms <- findFreqTerms(reviewTdm, lowfreq=10))

## [1] "character" "costner"   "kutcher"   "movie"     "one"
"see"
## [7] "well"
```

As we can see from above result, there are seven terms that occur more that 10 times in the document corpus.

Finding the Associtation with word movie

```
findAssocs(reviewTdm, "movie", 0.5)

## $movie
##      like    almost      good       bad    batten     becam
becom       big
##      0.82      0.79      0.70      0.69      0.69      0.69
0.69      0.69
##    borrow     butbr      came      cant     carri      cast
clever     count
##      0.69      0.69      0.69      0.69      0.69      0.69
0.69      0.69
##  distract   downfal    enough    expect      fact       fan
fashion      find
##      0.69      0.69      0.69      0.69      0.69      0.69
0.69      0.69
##       goe  guardian     hatch      head      hint     maker
memor       mrs
##      0.69      0.69      0.69      0.69      0.69      0.69
0.69      0.69
##    muchbr      near   nitpick     often overboard   probabl
remind      ride
##      0.69      0.69      0.69      0.69      0.69      0.69
0.69      0.69
##    ridebr    riskbr  robinson      save    smooth     storm
successbr    surfac
##      0.69      0.69      0.69      0.69      0.69      0.69
```

```
0.69      0.69
##      swim  tomorrow     water      wear      work      abil
seem      line
##      0.69      0.69      0.69      0.69      0.69      0.63
0.61      0.61
##      dont      make       see      play      give
##      0.58      0.58      0.54      0.52      0.51
```

As we can see from the result above the word movie is associated with term like, almost, good, bad and so on. This is natural as this is a dataset of movie reviews.

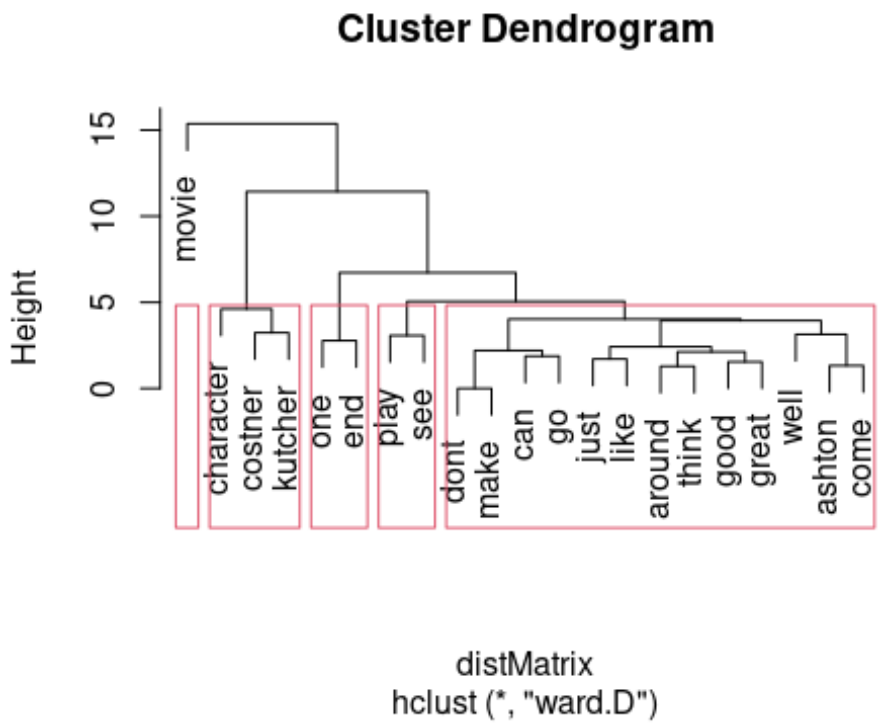
Word Cloud

```
library(wordcloud)

## Loading required package: RColorBrewer

m <- as.matrix(reviewTdm)
freq <- sort(rowSums(m), decreasing=T)
wordcloud(words=names(freq), freq=freq, min.freq=5,
random.order=F)
```

The word cloud shows the most used term in the corpus is movie and second one is character and so on. From the word cloud we can see that the data is about a movie.

Clustering of Words

```
# remove sparse terms
reviewTdm2 <- removeSparseTerms(reviewTdm, sparse=0.70)
m2 <- as.matrix(reviewTdm2)
# cluster terms
distMatrix <- dist(scale(m2))
fit <- hclust(distMatrix, method="ward.D")
plot(fit)
# cut tree into 5 clusters
rect.hclust(fit, k=5)
```




```
(groups <- cutree(fit, k=5))

##    ashton       can character   costner        go      good
great   kutcher
##         1         1         2         2         1         1
1         2
##     movie       one      play       see      well    around
come      just
##         3         4         5         5         1         1
1         1
##     think       end      like      dont      make
##         1         4         1         1         1

m3<-t(m2)
set.seed(10)
k <- 4
```

```
kmeansResult <- kmeans(m3, k)
round(kmeansResult$centers, digits=3)
```

```
##   ashton   can character costner    go good great kutcher movie one
play see
## 1  0.333 0.333     1.667       2 0.333  0.0     0   1.667   1.0 1.0
0.333 0.0
## 2  0.000 1.000     8.000       1 1.000  0.0     2   2.000   6.0 1.0
2.000 1.0
## 3  0.000 2.000     2.000       3 0.000  1.0     0   2.000  11.0 2.0
2.000 2.0
## 4  0.600 0.200     0.800       1 0.600  0.6     1   0.800   5.2 0.8
0.600 1.6
##    well around  come  just think end like  dont  make
## 1 0.333  0.333 0.333 0.333 0.333 1.0  0.0 0.333 0.333
## 2 0.000  1.000 1.000 3.000 0.000 1.0  1.0 0.000 0.000
## 3 3.000  1.000 0.000 1.000 2.000 1.0  2.0 2.000 2.000
## 4 1.200  0.200 0.600 0.200 0.600 0.2  0.6 0.400 0.400
```

Topic Modeling

Topic modeling is a technique through which we can cluster a group of words(topic). The words grouped in a topic has somekind of hidden relationship with one another.

```
library(topicmodels)
set.seed(123)
myLda <- LDA(as.DocumentTermMatrix(reviewTdm), k=4)
terms(myLda, 4)
```

```
##      Topic 1   Topic 2     Topic 3     Topic 4
## [1,] "movie"   "character" "movie"     "movie"
## [2,] "see"     "movie"     "kutcher"   "film"
## [3,] "well"    "gere"      "character" "costner"
## [4,] "costner" "one"       "costner"   "know"
```