IS-688: Web Mining Project Report

TEXT MINING AND SENTIMENT ANALYSIS ON IMDB USER COMMENTS

Team 6:
Amey Katkar (aak83)
Chirantan Ghosh (cg333)
Aarti Dandvate (ad484)
Pooja Shinde (ps647)
Rakesh Vijayakumar (rv269)

Contents

Objective of the project

The project has been carried out by our group in two parts -

In the first part of the project we have performed a text mining of the IMDB dataset to find out what are the most frequently occurring words and their association and/or implications. We have further used visualization techniques such as histogram and word cloud to better represent our analysis.

In the second part we have identified and categorized opinions expressed by the users in the form of reviews to determine whether the writer's attitude towards a particular movie, TV show is positive, negative or neutral and what are the most common emotions associated with it.

Introduction to Web Mining

Web Mining is the use of the data mining techniques to automatically discover and extract information from web documents/services. Web Mining can be divided into three different types – Web usage mining, Web content mining, and Web structure mining

Difference between Data Mining and Web Mining

Data Mining: Process of transforming data into knowledge

Web Mining: Process of applying data mining techniques to extract and uncover knowledge from web documents and services

1. **Web Usage Mining:**
    - To analyze and discover interesting patterns of user's usage data on the web
    - Usage data records the user's behavior when the user browses or makes transactions on the website
    - Automatically generated data stored in server access logs, referrer logs, agent logs, client-side cookies, user sessions, web server etc.
    - Ex. Google Analytics by Google

2. **Web Structure Mining**
    - Using the graph theory to analyze the node and connection structure of a website to discover the structural summary about the web site and web pages
    - Primary goal is to discover link structure of Hyperlinks
    - Discovering the structure of web document itself
    - Discovering the nature of the hierarchy or network of hyperlinks in the website
    - Ex. HITS (Hypertext Induced Topic Search) and Page Rank algorithm

3. **Web Content Mining**
    - Discovering useful information from web contents, data, documents
    - Web data could be – text, image, audio, metadata, video and hyperlinks
    - Used for developing intelligent tools for information retrieval
    - Techniques used – Classification, Clustering, and Association Rules
    - Ex. Search Engines like Google and Bing

Application of Web Mining

- E-Commerce
- Information Retrieval (Search)
- Network Management

Issues with Web Mining

- Web data is dynamic and can be very large
- Mining on single server is difficult
- Addressing security, privacy and legal issues

Challenges in Web Mining

To develop new incremental web mining algorithms and adapt traditional data mining algorithms to exploit hyper-links and access patterns

Introduction to Text Mining

Text mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation.

The difference between regular data mining and text mining is that in text mining the patterns are extracted from natural language text rather than from structured databases or facts.

Typical applications of text mining could include analyzing open-ended survey responses. For example, you may discover a certain set of words or terms that are commonly used by respondents to describe the pros and cons of a product or a service suggesting common misconceptions or confusion regarding the items in the study.

Another application includes to aid in the automatic classification of texts. E.g. filtering out junk mail based on certain keywords

Text mining algorithm consists of three steps

- Train: create a repository of words/keywords
- Filter: remove stop words such as the, as, we, but, it, is etc.
- Classify: check each document to be classified for the presence and frequency of the chosen attributes

Introduction to Sentiment Analysis

What is Sentiment?

Sentiment describes the feeling that comes from within a comment or review. Is someone for or against a product? Do they think a service was good or bad? Did they like or dislike something? While sentiment is usually described as having a binary opposition it is often more complex. There are comments and reviews that offer neither a good or bad opinion, often described as a neutral opinion.

What is Sentiment Analysis?

Sentiment Analysis aims to determine the attitude of the author of a specific piece of content with respect to the topic of interest. Comments and content can be referred to as Positive, Negative, Neutral or have no sentiment at all.

Techniques in Sentiment Analysis

- Semantic orientation and polarity of words
- Text-based sentiment classification
- Incorporating shallow linguistics

Challenges in Sentiment Analysis

- People express opinions in complex ways
- In opinion texts, lexical content alone can be misleading
- Intra-textual reversals, negation, logic change common
- Dealing with rhetorical devices/modes such as sarcasm, irony, implication etc.

About the Dataset

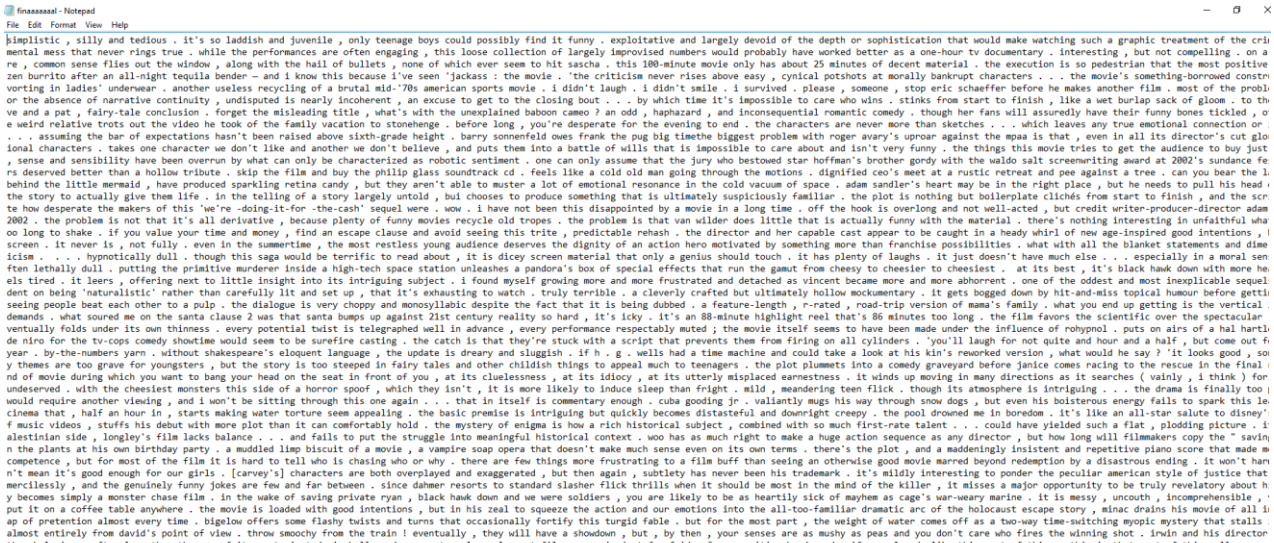
The IMDB movie review dataset was obtained from Cornell's department of Computer Science's website. The core dataset contains 1000 positive and 1000 negative movie reviews split in two different files.

In addition to our base dataset we have also made use of AFINN wordlist to analyze the sentence content which has 2477 words and phrases rated from -5 [very negative] to +5 [very positive].

Why IMDB

The primary reason for choosing IMDB as our base dataset was for the fact that we needed large amount of natural piece of text and phrases. Natural texts have the ability to express a positive, negative or neutral sentiment accompanied with wide range of human emotions.

Over the past decade there has been a substantial growth in user interactions on the IMDB review section which we thought would provide us great insights in to the most frequently occurring words as well as the sentiment associated with it (if any). We plan to use the results of our analysis to find patterns or themes in user reviews.


finaaaaaal - Notepad

File Edit Format View Help

bimplistic , silly and tedious . it's so laddish and juvenile , only teenage boys could possibly find it funny . exploitative and largely devoid of the depth or sophistication that would make watching such a graphic treatment of the cri
mental mess that never rings true . while the performances are often engaging , this loose collection of largely improvised numbers would probably have worked better as a one-hour tv documentary . interesting , but not compelling . on a
re , common sense flies out the window , along with the hail of bullets , none of which ever seem to hit sascha . this 100-minute movie only has about 25 minutes of decent material . the execution is so pedestrian that the most positive
zen burrito after an all-night tequila bender – and i know this because i've seen 'jackass : the movie . 'the criticism never rises above easy , cynical potshots at morally bankrupt characters . . . the movie's something-borrowed constr
vorting in ladies' underwear . another useless recycling of a brutal mid-'70s american sports movie . i didn't laugh . i didn't smile . i survived . please , someone , stop eric schaeffer before he makes another film . most of the probl
or the absence of narrative continuity , undisputed is nearly incoherent , an excuse to get to the closing bout . . . by which time it's impossible to care who wins . stinks from start to finish , like a wet burlap sack of gloom . to th
ve and a pat , fairy-tale conclusion . forget the misleading title , what's with the unexplained baboon cameo ? an odd , haphazard , and inconsequential romantic comedy . though her fans will assuredly have their funny bones tickled , o
e weird relative treats out the video for the family vacation to stonehenge . before long , you're desperate for the evening to end . the characters are never more than sketches . . . which leaves any true emotional connection or
. . . assuming the bar of expectations hasn't been raised above sixth-grade height . barry sonnenfeld owes frank the pug big timethe biggest problem with roger avary's uproar against the mpaa is that , even in all its director's cut glo
ional characters . takes one character we don't like and another we don't believe , and puts them into a battle of wills that is impossible to care about and isn't very funny . the things this movie tries to get the audience to buy just
, sense and sensibility have been overrun by what can only be characterized as robotic sentiment . one can only assume that the jury who bestowed star hoffman's brother gordy with the waldo salt screenwriting award at 2002's sundance fe
rs deserved better than a hollow tribute . skip the film and buy the philip glass soundtrack cd . feels like a cold old man going through the motions . dignified ceo's meet at a rustic retreat and pee against a tree . can you bear the l
behind the little mermaid , have produced sparkling retina candy , but they aren't able to muster a lot of emotional resonance in the cold vacuum of space . adam sandler's heart may be in the right place , but he needs to pull his head
the story to actually give them life . in the telling of a story largely untold , bui chooses to produce something that is ultimately suspiciously familiar . the plot is nothing but boilerplate clichés from start to finish , and the scr
te how desperate the makers of this 'we're -doing-it-for -the-cash' sequel were . . . wow . i have not been this disappointed by a movie in a long time . off the hook is overlong and not well-acted , but credit writer-producer-director adam
2002 . the problem is not that it's all derivative , because plenty of funny movies recycle old tropes . the problem is that van wilder does little that is actually funny with the material . there's nothing interesting in unfaithful wha
oo long to shake . if you value your time and money , find an escape clause and avoid seeing this trite , predictable rehash . the director and her capable cast appear to be caught in a heady whirl of new age-inspired good intentions ,
screen . it never is , not fully . even in the summertime , the most restless young audience deserves the dignity of an action hero motivated by something more than franchise possibilities . what with all the blanket statements and dime
icism . . . . hypnotically dull . though this saga would be terrific to read about , it is dicey screen material that only a genius should touch . it has plenty of laughs . it just doesn't have much else . . . especially in a moral sen
ften lethally dull . putting the primitive murderer inside a high-tech space station unleashes a pandora's box of special effects that run the gamut from cheesy to cheesier to cheesiest . at its best , it's black hawk down with more he
els tired . it leers , offering next to little insight into its intriguing subject . i found myself growing more and more frustrated and detached as vincent became more and more abhorrent . one of the oddest and most inexplicable sequel
dent on being 'naturalistic' rather than carefully lit and set up , that it's exhausting to watch . truly terrible . a cleverly crafted but ultimately hollow mockumentary . it gets bogged down by hit-and-miss topical humour before getti
seeing people beat each other to a pulp . the dialogue is very choppy and monosyllabic despite the fact that it is being dubbed . a feature-length , r-rated , road-trip version of mama's family . what you end up getting is the vertical
demands . what soured me on the santa clause 2 was that santa bumps up against 21st century reality so hard , it's icky . it's an 88-minute highlight reel that's 86 minutes too long . the film favors the scientific over the spectacular
ventually folds under its own thinness . every potential twist is telegraphed well in advance , every performance respectably muted ; the movie itself seems to have been made under the influence of rohypnol . puts on airs of a hal hartl
de niro for the tv-cops comedy showtime would seem to be surefire casting . the catch is that they're stuck with a script that prevents them from firing on all cylinders . 'you'll laugh for not quite and hour and a half , but come out f
year . by-the-numbers yarn . without shakespeare's eloquent language , the update is dreary and sluggish . if h . g . wells had a time machine and could take a look at his kin's reworked version , what would he say ? 'it looks good , so
y themes are too grave for youngsters , but the story is too steeped in fairy tales and other childish things to appeal much to teenagers . the plot plummets into a comedy graveyard before janice comes racing to the rescue in the final
nd of movie during which you want to bang your head on the seat in front of you , at its clueslessness , at its idiocy , at its utterly misplaced earnestness . it winds up moving in many directions as it searches (vainly , i think) for
undeserved . with the cheesiest monsters this side of a horror spoof , which they isn't , it is more likely to induce sleep than fright . mild , meandering teen flick . though its atmosphere is intriguing . . . the drama is finally too
would require another viewing , and i won't be sitting through this one again . . . that in itself is commentary enough . cuba gooding jr . valiantly mugs his way through snow dogs , but even his boisterous energy fails to spark this le
cinema that , half an hour in , starts making water torture seem appealing . the basic premise is intriguing but quickly becomes distasteful and downright creepy . the pool drowned me in boredom . it's like an all-star salute to disney'
f music videos , stuffs his debut with more plot than it can comfortably hold . the mystery of enigma is how a rich historical subject , combined with so much first-rate talent . . . could have yielded such a flat , plodding picture . i
alestinian side , longley's film lacks balance . . . and fails to put the struggle into meaningful historical context . woo has as much right to make a huge action sequence as any director , but how long will filmmakers copy the " savin
n the plants at his own birthday party . a muddled limp biscuit of a movie , a vampire soap opera that doesn't make much sense even on its own terms . there's the plot , and a maddeningly insistent and repetitive piano score that made m
competence , but for most of the film it is hard to tell who is chasing who or why . there are few things more frustrating to a film buff than seeing an otherwise good movie marred beyond redemption by a disastrous ending . it won't har
n't mean it's good enough for our girls . [carvey's] characters are both overplayed and exaggerated , but then again , subtlety has never been his trademark . it's mildly interesting to ponder the peculiar american style of justice that
mercilessly , and the genuinely funny jokes are few and far between . since dahmer resorts to standard slasher flick thrills when it should be most in the mind of the killer , it misses a major opportunity to be truly revelatory about h
y becomes simply a monster chase film . in the wake of saving private ryan , black hawk down and we were soldiers , you are likely to be as heartily sick of mayhem as cage's war-weary marine . it is messy , uncouth , incomprehensible ,
put it on a coffee table anywhere . the movie is loaded with good intentions , but in his zeal to squeeze the action and our emotions into the all-too-familiar dramatic arc of the holocaust escape story , minac drains his movie of all i
ap of pretention almost every time . bigelow offers some flashy twists and turns that occasionally fortify this turgid fable . but for the most part , the weight of water comes off as a two-way time-switching myopic mystery that stalls
almost entirely from david's point of view . throw smoochy from the train ! eventually , they will have a showdown , but , by then , your senses are as mushy as peas and you don't care who fires the winning shot . irwin and his director
the whole is so often less than the sum of its parts in today's hollywood . an extremely unpleasant film . a movie just for friday fans , critics be damned . if you already like this sort of thing , this is that sort of thing all over a
script is full of unhappy , two-dimensional characters who are anything but compelling . labute can't avoid a fatal mistake in the modern era : he's changed the male academic from a lower-class brit to an american , a choice that upsets
t-time writer-director neil burger follows up with are terribly convincing , which is a pity , considering barry's terrific performance . gets better after foster leaves that little room . the movie is as padded as allen's jelly belly .
, city by the sea swings from one approach to the other , but in the end , it stays in formula -- which is a waste of de niro , mcdormand and the other good actors in the cast . plotless collection of moronic stunts is by far the worst
is all too predictable and far too cliched to really work . let's issue a moratorium , effective immediately , on treacly films about inspirational prep-school professors and the children they so heartwarmingly motivate . it's the eleme
e screen in frustration . see clockstoppers if you have nothing better to do with 94 minutes . but be warned , you too may feel time has decided to stand still . or that the battery on your watch has died . suffers from over-familiarity
ers . it's not so much a movie as a joint promotion for the national basketball association and teenaged rap and adolescent poster-boy lil' bow wow . peralta's mythmaking could have used some informed , adult hindsight . amazingly dopey
atever reason , be thinking about going to see this movie is hereby given fair warning . mr . soderbergh's direction and visual style struck me as unusually and unimpressively fussy and pretentious . do you say " hi " to your lover when
y and murky points . about as enjoyable , i would imagine , as searching for a quarter in a giant pile of elephant feces . . . positively dreadful . a generic international version of a typical american horror film . . . . while certai
o-big (and not-so-hot) directorial debut . yet another iteration of what's become one of the movies' creepiest conventions , in which the developmentally disabled are portrayed with almost supernatural powers to humble , teach and ult
ctor . responsável direto pelo fracasso 'artístico' de doce lar , o roteirista c . jay cox não consegue sequer aproveitar os poucos momentos em que escapa da mediocridade . just as the lousy tarantino imitations have subsided , he
r adults . it's just a little too self-satisfied . clever but not especially compelling . mckay seems embarrassed by his own invention and tries to rush through the intermediary passages , apparently hoping that the audience will not no

IMDB Review Dataset

AFINN Word List

Approach

Text Mining Process

- Corpus Creation Stage

Preprocessing Stage

- Remove Punctuation
- Remove Numbers
- Converting to lower case
- Remove Stop Words
- Remove Whitespace
- Convert to plain text
- Stemming

Staging the data

- Creating document term matrix
- Organizing terms by frequency
- Removing sparse items
- Checking for most frequently occurring words
- Plot word frequencies (Histogram)
- Word Cloud
- Cluster Analysis

Sentiment Analysis

- Test with AFINN list
- Naives Bayes Classifier
- Confusion Matrix
- Probability of Success
- Plot Range of Emotions

Analysis

Tools Used:

- R Programming Language
- R studio
- Text Editor

PART 1: Text Mining

OBJECTIVE: To analyze nature of frequently occurring words in the review dataset

Packages Used: Text Mining in R (tm package), SnowballC (for stemming), ggplot2 (plotting), wordcloud, NLP

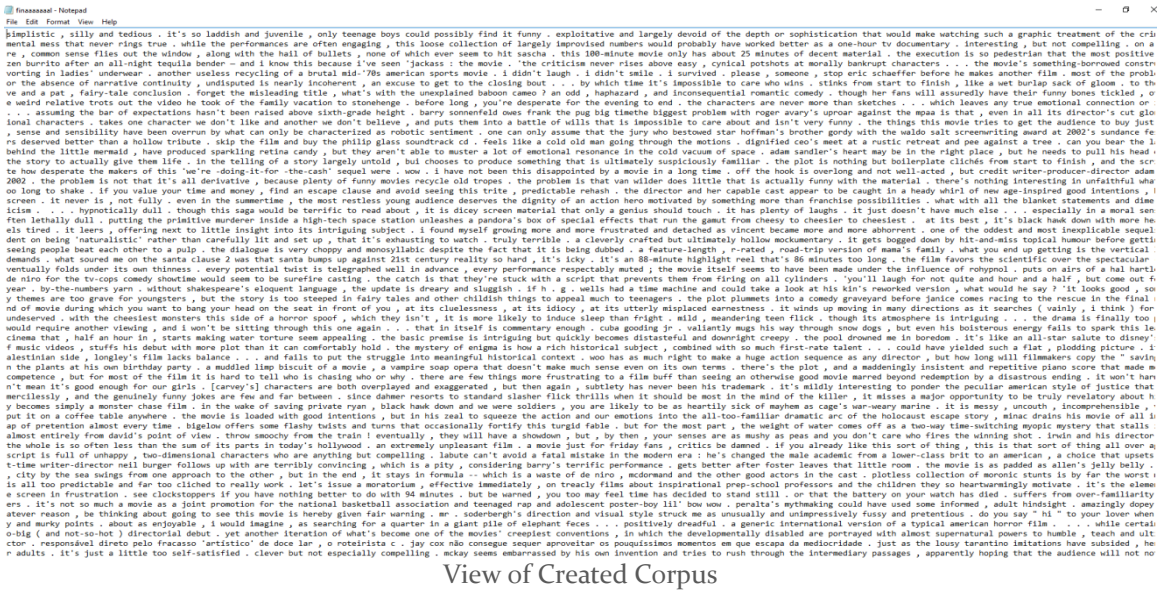
Step1: Loading libraries and data into R

```
library(tm)
```

Step2: Corpus Creation

Corpus is a collection of documents. We combined both the files (each file having 1000 reviews each) into a single document. Using the corpus() function we were able to load the dataset file into the corpus object

```
docs <- Corpus(DirSource("C:\\IS-688-Web-Mining-Final-Team-6"))
```



View of Created Corpus

Step3: Preprocessing the data

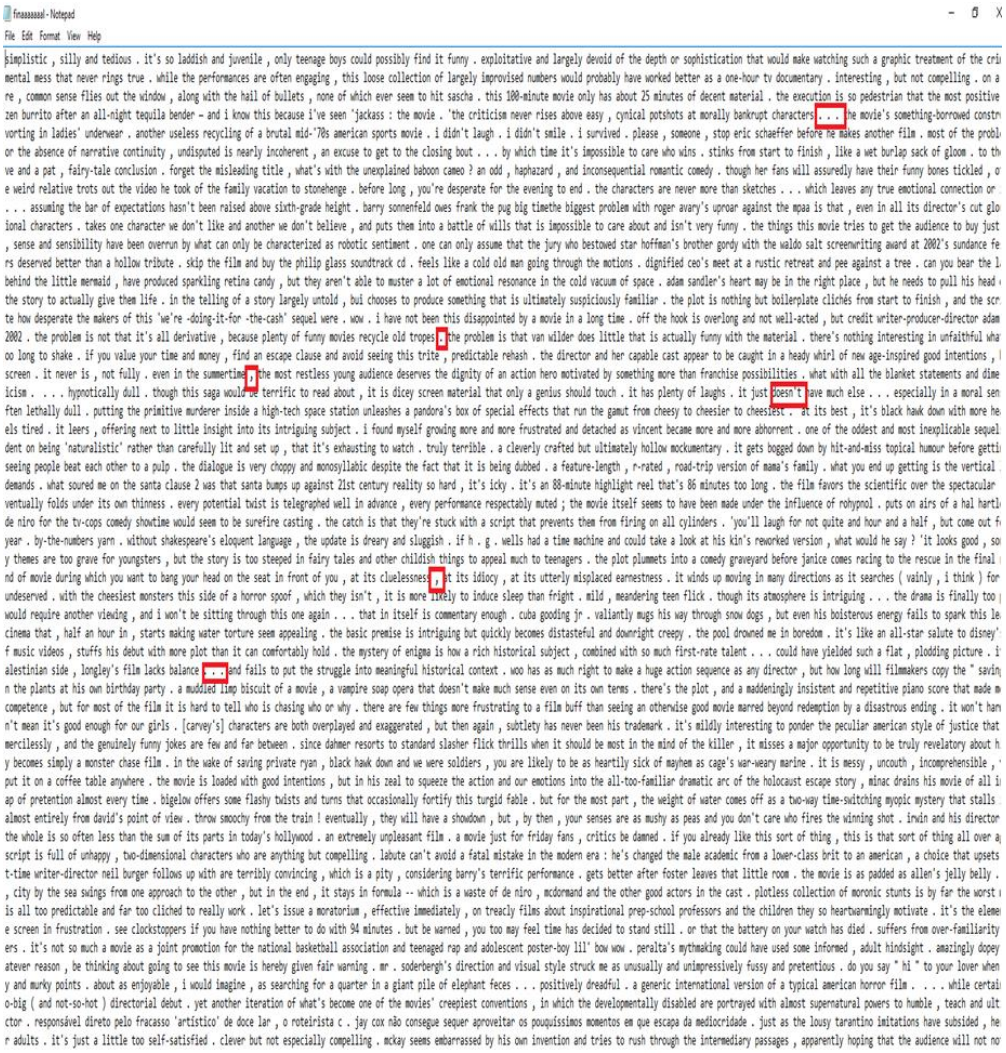
Data cleaning, is perhaps the most important step in text analysis. The tm package offers a number of transformations that make the process of data cleaning easy. Below are the supported transformations

- removeNumbers
- removePunctuation
- removeWords
- stemDocument
- stripWhitespace

Removing Punctuation

Machine does not understand punctuations and special characters and have serve no purpose in text mining. Below is the function to get rid of most punctuations.

```
docs <- tm_map(docs, removePunctuation)
```

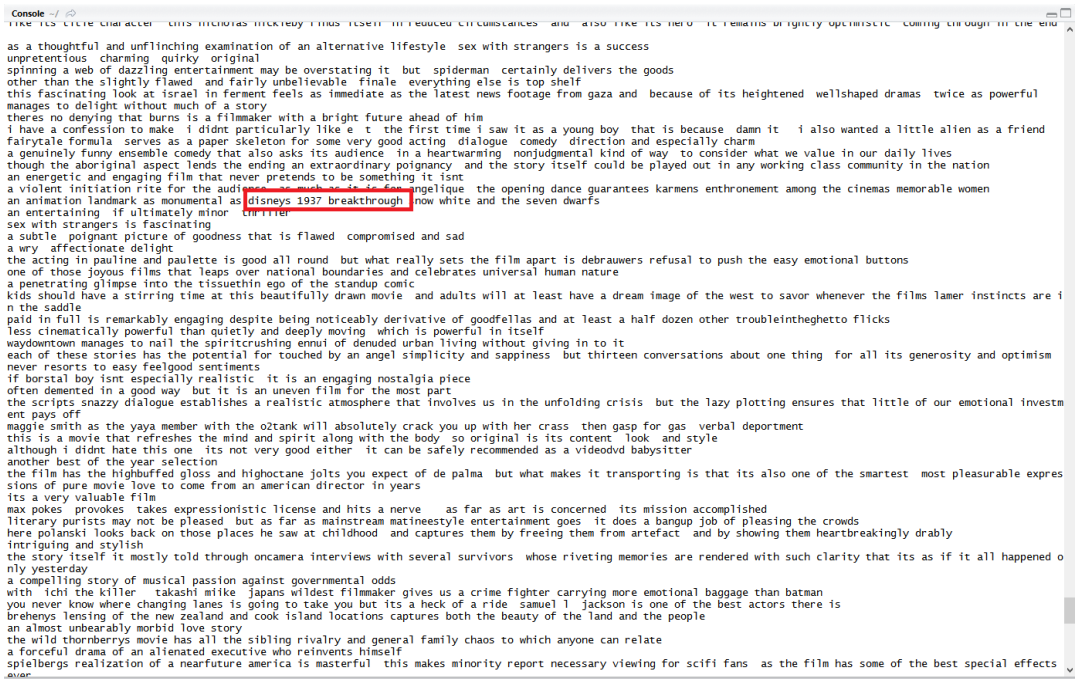


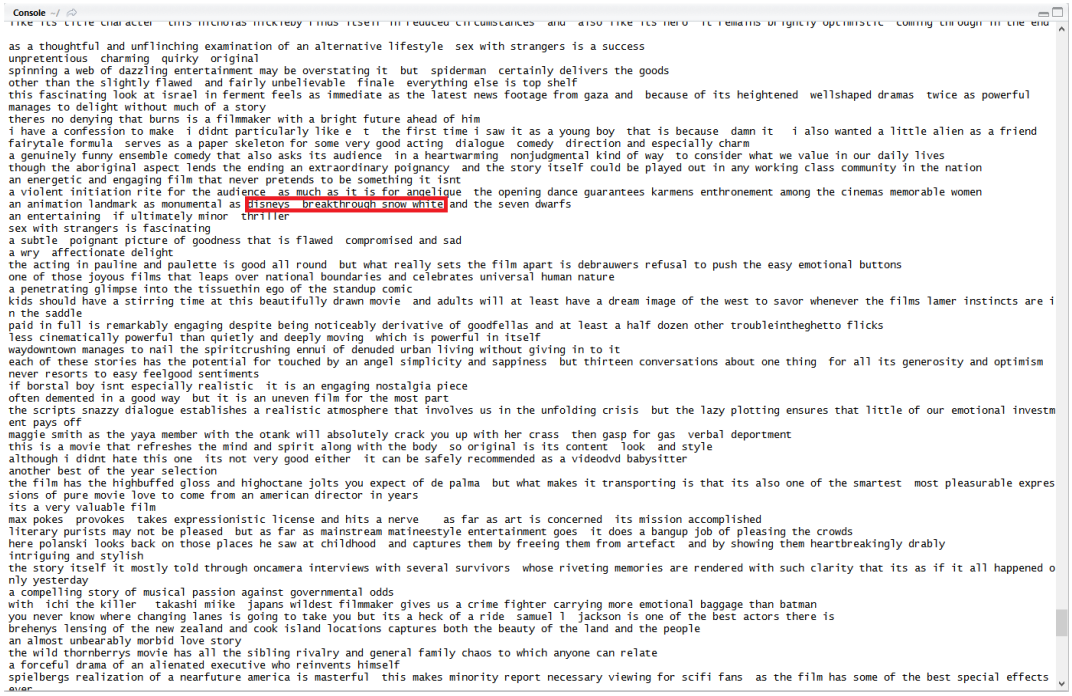

Console ~/
there is simply no doubt that this film asks the right questions at the right time in the history of our country
if youve the patience there are great rewards here
as a science fiction movie minority report astounds
watching e t now in an era dominated by cold loud specialeffectsladen extravaganzas one is struck less by its lavish grandeur than by its intimacy and precision
visually breathtaking viscerally exciting and dramatically moving its the very definition of epic adventure
chris columbus sequel is faster livelier and a good deal funnier than his original
watching this film what we feel isnt mainly suspense or excitement the dominant feeling is something like nostalgia
a great participatory spectator sport
a rather brilliant little cult item a pastiche of childrens entertainment superhero comics and japanese animation
believes so fervently in humanity that it feels almost anachronistic and it is too cute by half but arriving at a particularly dark moment in history it offers flickeri
ng reminders of the ties that bind us
adam sandler in an art film
as averse as i usually am to feelgood followyourdream hollywood fantasies this one got to me
stone seems to have a knack for wrapping the theater in a cold blanket of urban desperation
a funny yet dark and seedy clash of cultures and generations
the hook is the drama within the drama as an unsolved murder and an unresolved moral conflict jockey for the spotlight
over the years hollywood has crafted a solid formula for successful animated movies and ice age only improves on it with terrific computer graphics inventive action se
quences and a droll sense of humor
like smoke signals the film is also imbued with strong themes of familial ties and spirituality that are powerful and moving without stooping to base melodrama
one of those movies that make us pause and think of what we have given up to acquire the fastpaced contemporary society
one of the most original american productions this year youll find yourself remembering this refreshing visit to a sunshine state
melds derivative elements into something that is often quite rich and exciting and always a beauty to behold
gives everyone something to shout about
the entire movie has a truncated feeling but whats available is lovely and lovable
a thoughtful visually graceful work
admirers of director abel ferrara may be relieved that his latest feature r xmas marks a modest if encouraging return to form
the slambang superheroics are kinetic enough to engross even the most antsy youngsters
a worthy addition to the cinematic canon which at last count numbered 52 different versions
deliciously meanspirited and wryly observant
the kind of primal storytelling that george lucas can only dream of
even if the ring has a familiar ring its still unusually crafty and intelligent for hollywood horror
the sheer joy and pride they took in their work and in each other shines through every frame
a solidly constructed entertaining thriller that stops short of true inspiration
the cast keeps this pretty watchable and casting mick jagger as director of the escort service was inspired
an entertaining if somewhat standardized action movie
it has a dashing and resourceful hero a lisping reptilian villain big fights big hair lavish period scenery and a story just complicated enough to let you bask in yo
ur own cleverness as you figure it out
an enjoyable comedy of lingual and cultural differences… the château is a film full of life and small delights that has all the wiggling energy of young kitten
intriguing and downright intoxicating
an incredibly thoughtful deeply meditative picture that neatly and effectively captures the debilitating grief felt in the immediate aftermath of the terrorist attacks
with an obvious rapport with her actors and a striking style behind the camera hélène angel is definitely a director to watch
could easily be called the best korean film of 2002
full of detail about the man and his country and is well worth seeing
the banter between calvin and his fellow barbers feels like a streetwise mclaughlin group and never fails to entertain
thoroughly engrossing and ultimately tragic
peter jackson and company once again dazzle and delight us fulfilling practically every expectation either a longtime tolkien fan or a moviegoing neophyte could want
bill morrisons decasia is uncompromising difficult and unbearably beautiful
full of bland hotels highways parking lots with some glimpses of nature and family warmth time out is a discreet moan of despair about entrapment in the maze of modern
life
even with all its botches enigma offers all the pleasure of a handsome and wellmade entertainment
his work transcends the boymeetsgirl posturing of typical love stories
if the reallife story is genuinely inspirational the movie stirs us as well

Removing Numbers

Text analysts are typically not interested in numbers since these do not usually contribute to the meaning of the text. (This may not hold true in all cases though)

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

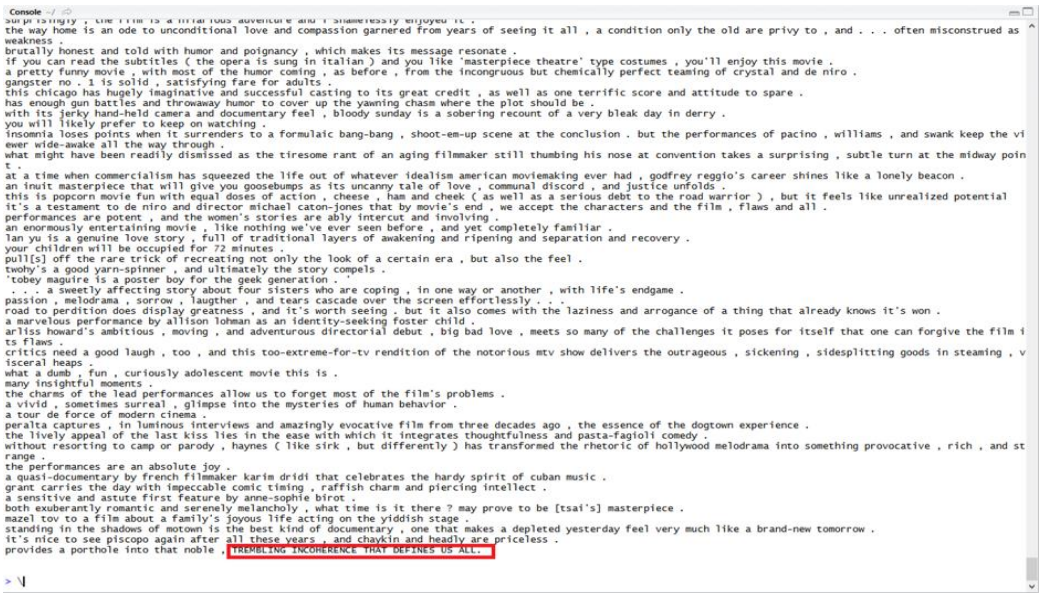


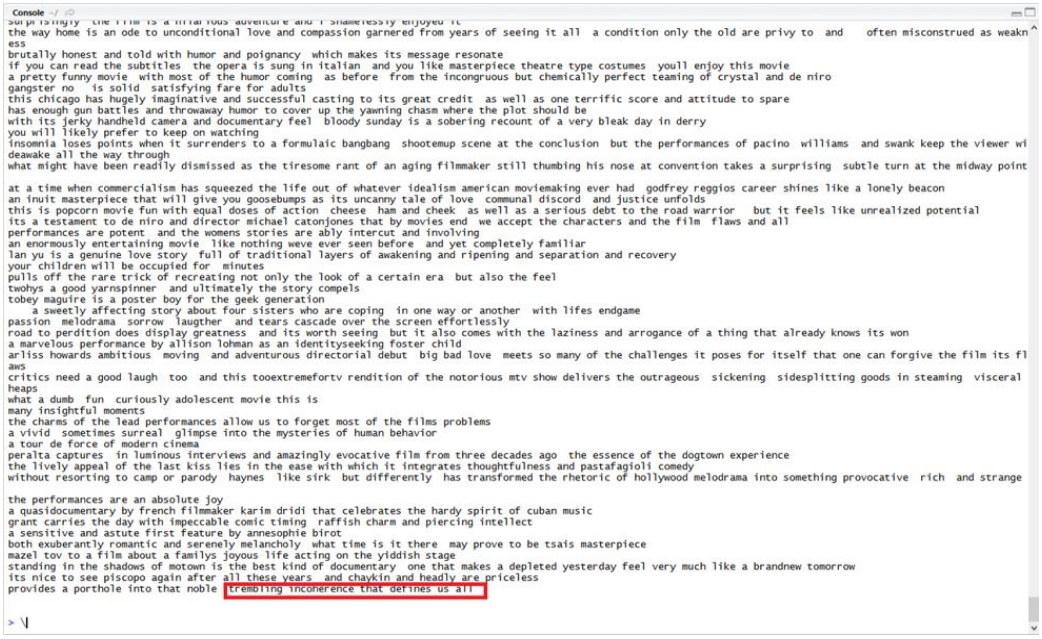


Converting the corpus to lowercase

It is important to remember that R is case sensitive, “Word” is not equal to “word” – and hence the reason for converting the corpus to lower case.

```
docs <- tm_map(docs, tolower)
```





Removing Stop Words

In computing, stop words are words which are filtered out before or after processing of natural language data. Though stop words usually refer to the most common words in a language, there is no single universe list of stop words used by all natural language processing tools, and indeed not all tools even such use a list. Stop words have no analytical value.

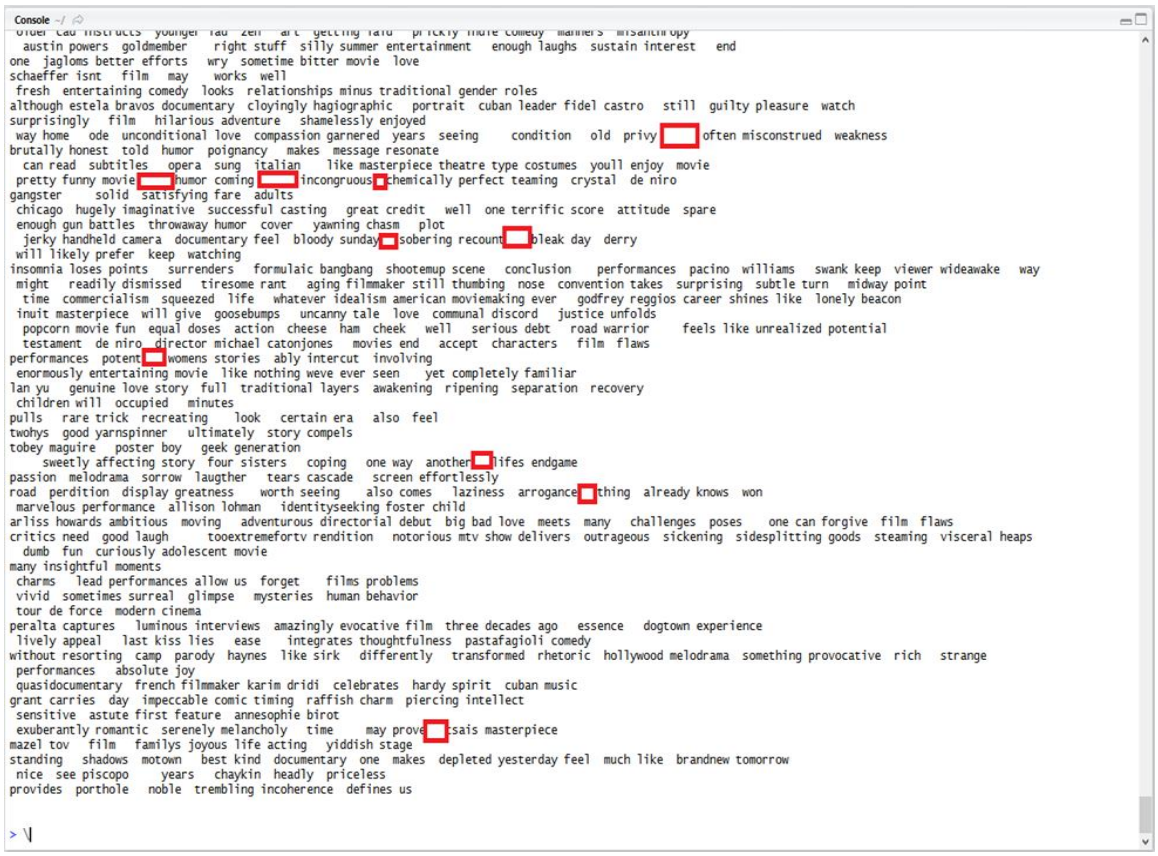
Example of stop words – a, about, out, she, he, an, all, be, at, because, the, their, then, can't, been, before, no, ours, myself etc.

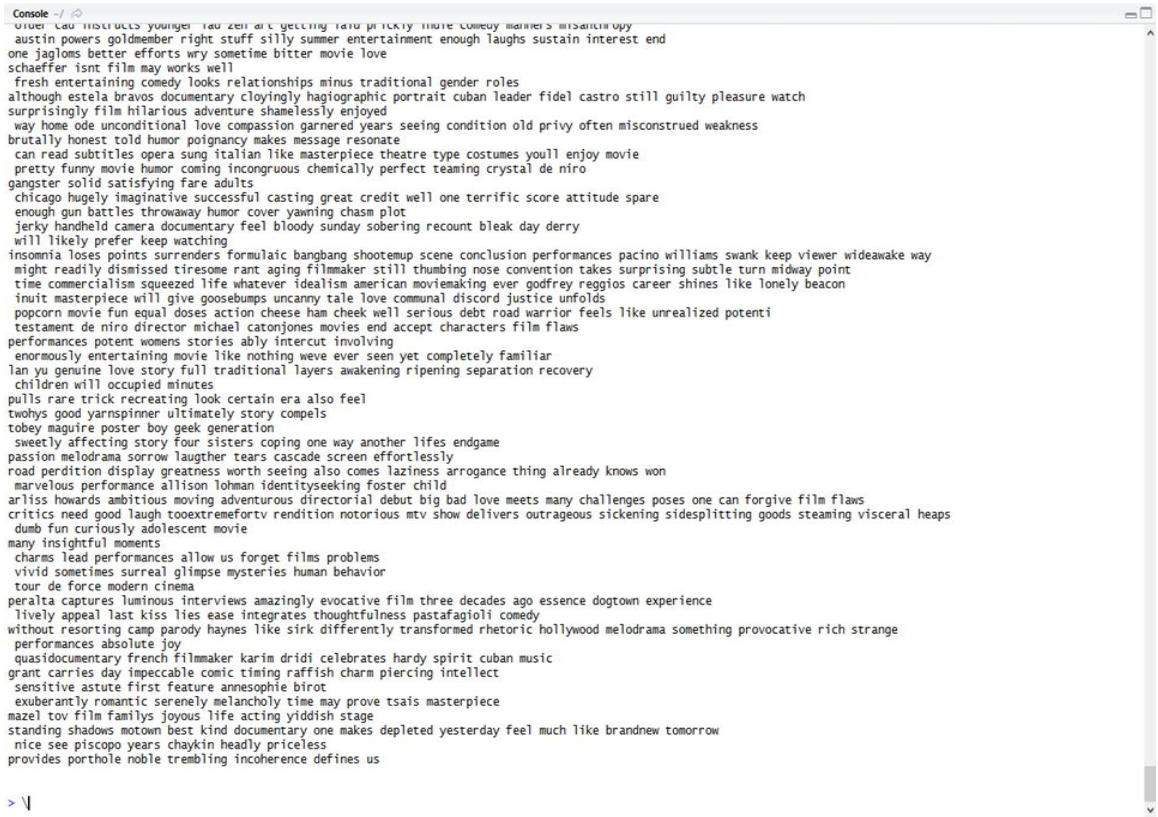
```
docs <- tm_map(docs, removeWords, stopwords("english"))
```

Removing Whitespace

When we remove words, numbers and punctuations from the corpus a lot of white space is left over which can be removed using the below functions

```
docs <- tm_map(docs, stripWhitespace)
```

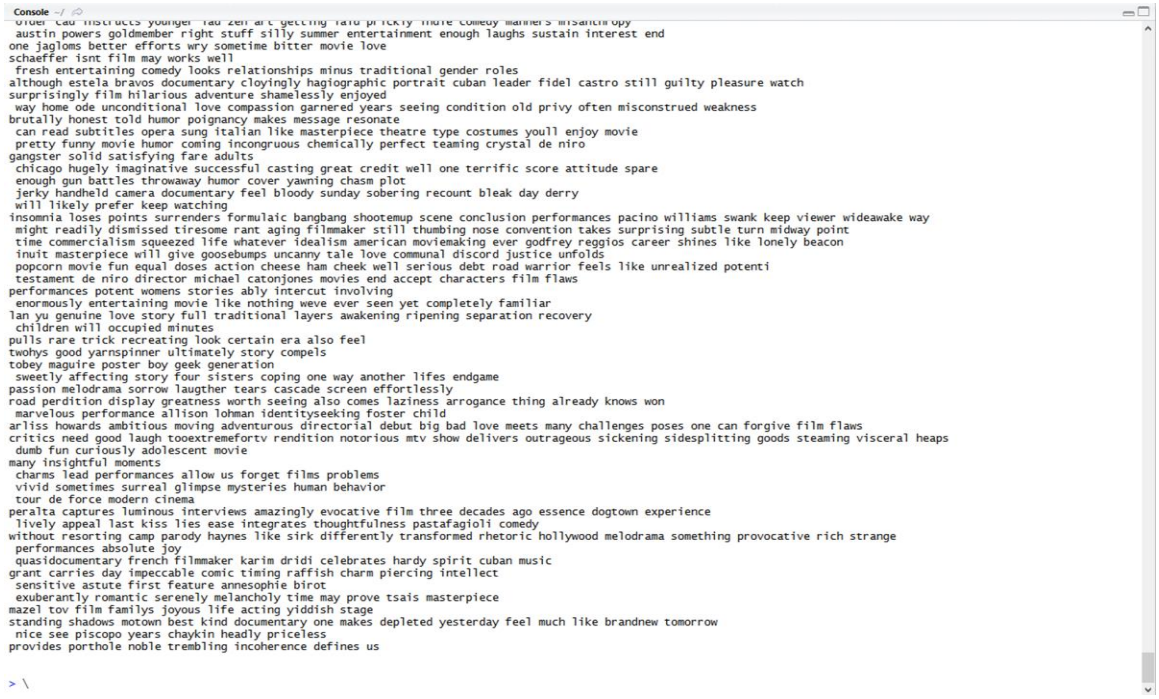

Console ~/
austin powers goldmember right stuff silly summer entertainment enough laughs sustain interest end
one jagloms better efforts wry sometime bitter movie love
schaeffer isnt film may works well
fresh entertaining comedy looks relationships minus traditional gender roles
although estela bravos documentary cloyingly hagiographic portrait cuban leader fidel castro still guilty pleasure watch
surprisingly film hilarious adventure shamelessly enjoyed
way home ode unconditional love compassion garnered years seeing condition old privy often misconstrued weakness
brutally honest told humor poignancy makes message resonate
can read subtitles opera sung italian like masterpiece theatre type costumes youll enjoy movie
pretty funny movie humor coming incongruous chemically perfect teaming crystal de niro
gangster solid satisfying fare adults
chicago hugely imaginative successful casting great credit well one terrific score attitude spare
enough gun battles throwaway humor cover yawning chasm plot
jerky handheld camera documentary feel bloody sunday sobering recount bleak day derry
will likely prefer keep watching
insomnia loses points surrenders formulaic bangbang shootemup scene conclusion performances pacino williams swank keep viewer wideawake way
might readily dismissed tiresome rant aging filmmaker still thumbing nose convention takes surprising subtle turn midway point
time commercialism squeezed life whatever idealism american moviemaking ever godfrey reggios career shines like lonely beacon
inuit masterpiece will give goosebumps uncanny tale love communal discord justice unfolds
popcorn movie fun equal doses action cheese ham cheek well serious debt road warrior feels like unrealized potential
testament de niro director michael catonjones movies end accept characters film flaws
performances potent womens stories ably intercut involving
enormously entertaining movie like nothing weve ever seen yet completely familiar
lan yu genuine love story full traditional layers awakening ripening separation recovery
children will occupied minutes
pulls rare trick recreating look certain era also feel
twohys good yarnspinner ultimately story compels
tobey maguire poster boy geek generation
sweetly affecting story four sisters coping one way another lifes endgame
passion melodrama sorrow laugther tears cascade screen effortlessly
road perdition display greatness worth seeing also comes laziness arrogance thing already knows won
marvelous performance allison lohman identityseeking foster child
arliss howards ambitious moving adventurous directorial debut big bad love meets many challenges poses one can forgive film flaws
critics need good laugh tooextremefortv rendition notorious mtv show delivers outrageous sickening sidesplitting goods steaming visceral heaps
dumb fun curiously adolescent movie
many insightful moments
charms lead performances allow us forget films problems
vivid sometimes surreal glimpse mysteries human behavior
tour de force modern cinema
peralta captures luminous interviews amazingly evocative film three decades ago essence dogtown experience
lively appeal last kiss lies ease integrates thoughtfulness pastafagioli comedy
without resorting camp parody haynes like sirk differently transformed rhetoric hollywood melodrama something provocative rich strange
performances absolute joy
quasidocumentary french filmmaker karim dridi celebrates hardy spirit cuban music
grant carries day impeccable comic timing raffish charm piercing intellect
sensitive astute first feature annesophie birot
exuberantly romantic serenely melancholy time may prove tsais masterpiece
mazel tov film familys joyous life acting yiddish stage
standing shadows motown best kind documentary one makes depleted yesterday feel much like brandnew tomorrow
nice see piscopo years chaykin headly priceless
provides porthole noble trembling incoherence defines us
> \


Console ~/
austin powers goldmember right stuff silly summer entertainment enough laughs sustain interest end
one jagloms better efforts wry sometime bitter movie love
schaeffer isnt film may works well
fresh entertaining comedy looks relationships minus traditional gender roles
although estela bravos documentary cloyingly hagiographic portrait cuban leader fidel castro still guilty pleasure watch
surprisingly film hilarious adventure shamelessly enjoyed
way home ode unconditional love compassion garnered years seeing condition old privy often misconstrued weakness
brutally honest told humor poignancy makes message resonate
can read subtitles opera sung italian like masterpiece theatre type costumes youll enjoy movie
pretty funny movie humor coming incongruous chemically perfect teaming crystal de niro
gangster solid satisfying fare adults
chicago hugely imaginative successful casting great credit well one terrific score attitude spare
enough gun battles throwaway humor cover yawning chasm plot
jerky handheld camera documentary feel bloody sunday sobering recount bleak day derry
will likely prefer keep watching
insomnia loses points surrenders formulaic bangbang shootemup scene conclusion performances pacino williams swank keep viewer wideawake way
might readily dismissed tiresome rant aging filmmaker still thumbing nose convention takes surprising subtle turn midway point
time commercialism squeezed life whatever idealism american moviemaking ever godfrey reggios career shines like lonely beacon
inuit masterpiece will give goosebumps uncanny tale love communal discord justice unfolds
popcorn movie fun equal doses action cheese ham cheek well serious debt road warrior feels like unrealized potenti
testament de niro director michael catonjones movies end accept characters film flaws
performances potent womens stories ably intercut involving
enormously entertaining movie like nothing weve ever seen yet completely familiar
lan yu genuine love story full traditional layers awakening ripening separation recovery
children will occupied minutes
pulls rare trick recreating look certain era also feel
twohys good yarnspinner ultimately story compels
tobey maguire poster boy geek generation
sweetly affecting story four sisters coping one way another lifes endgame
passion melodrama sorrow laugther tears cascade screen effortlessly
road perdition display greatness worth seeing also comes laziness arrogance thing already knows won
marvelous performance allison lohman identityseeking foster child
arliss howards ambitious moving adventurous directorial debut big bad love meets many challenges poses one can forgive film flaws
critics need good laugh tooextremefortv rendition notorious mtv show delivers outrageous sickening sidesplitting goods steaming visceral heaps
dumb fun curiously adolescent movie
many insightful moments
charms lead performances allow us forget films problems
vivid sometimes surreal glimpse mysteries human behavior
tour de force modern cinema
peralta captures luminous interviews amazingly evocative film three decades ago essence dogtown experience
lively appeal last kiss lies ease integrates thoughtfulness pastafagioli comedy
without resorting camp parody haynes like sirk differently transformed rhetoric hollywood melodrama something provocative rich strange
performances absolute joy
quasidocumentary french filmmaker karim dridi celebrates hardy spirit cuban music
grant carries day impeccable comic timing raffish charm piercing intellect
sensitive astute first feature annesophie birot
exuberantly romantic serenely melancholy time may prove tsais masterpiece
mazel tov film familys joyous life acting yiddish stage
standing shadows motown best kind documentary one makes depleted yesterday feel much like brandnew tomorrow
nice see piscopo years chaykin headly priceless
provides porthole noble trembling incoherence defines us
> \

Converting corpus to plain text document

By converting the document to plain text document (.txt) file we can get rid of any extraneous formatting.

```
docs <- tm_map(docs, PlainTextDocument)
```



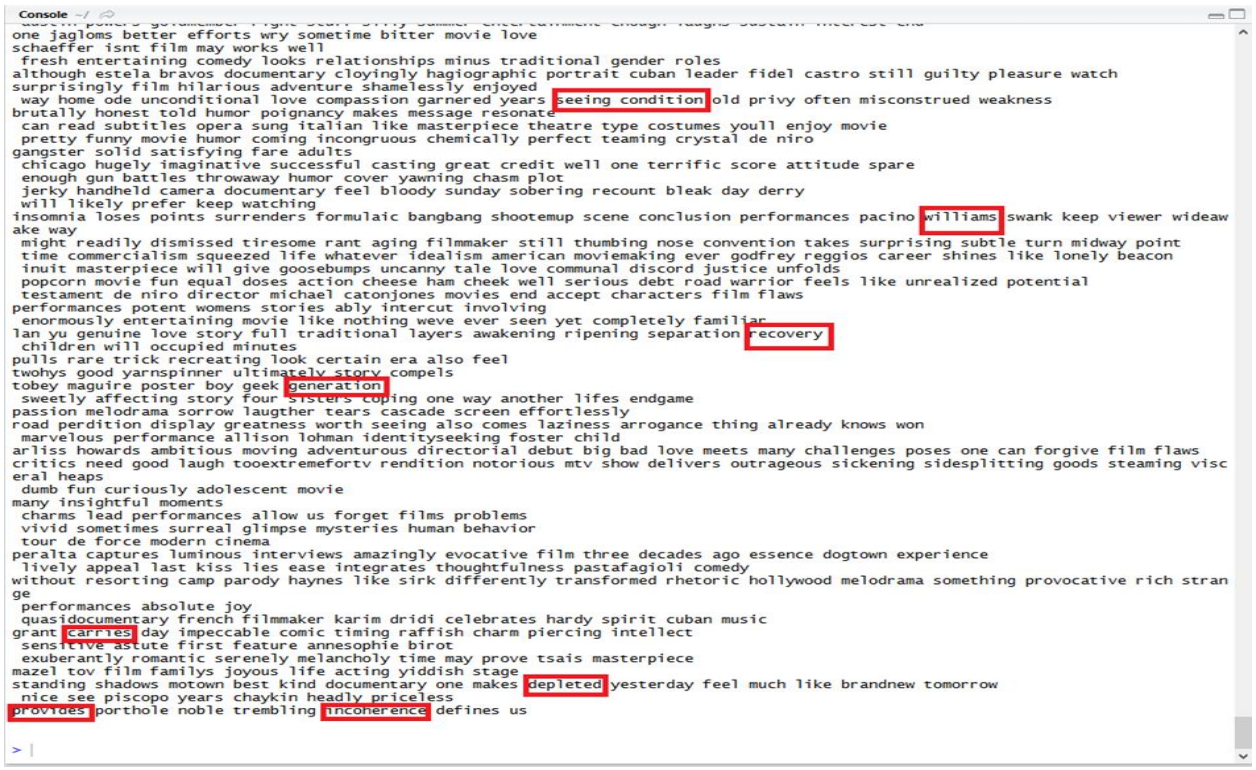
Stemming

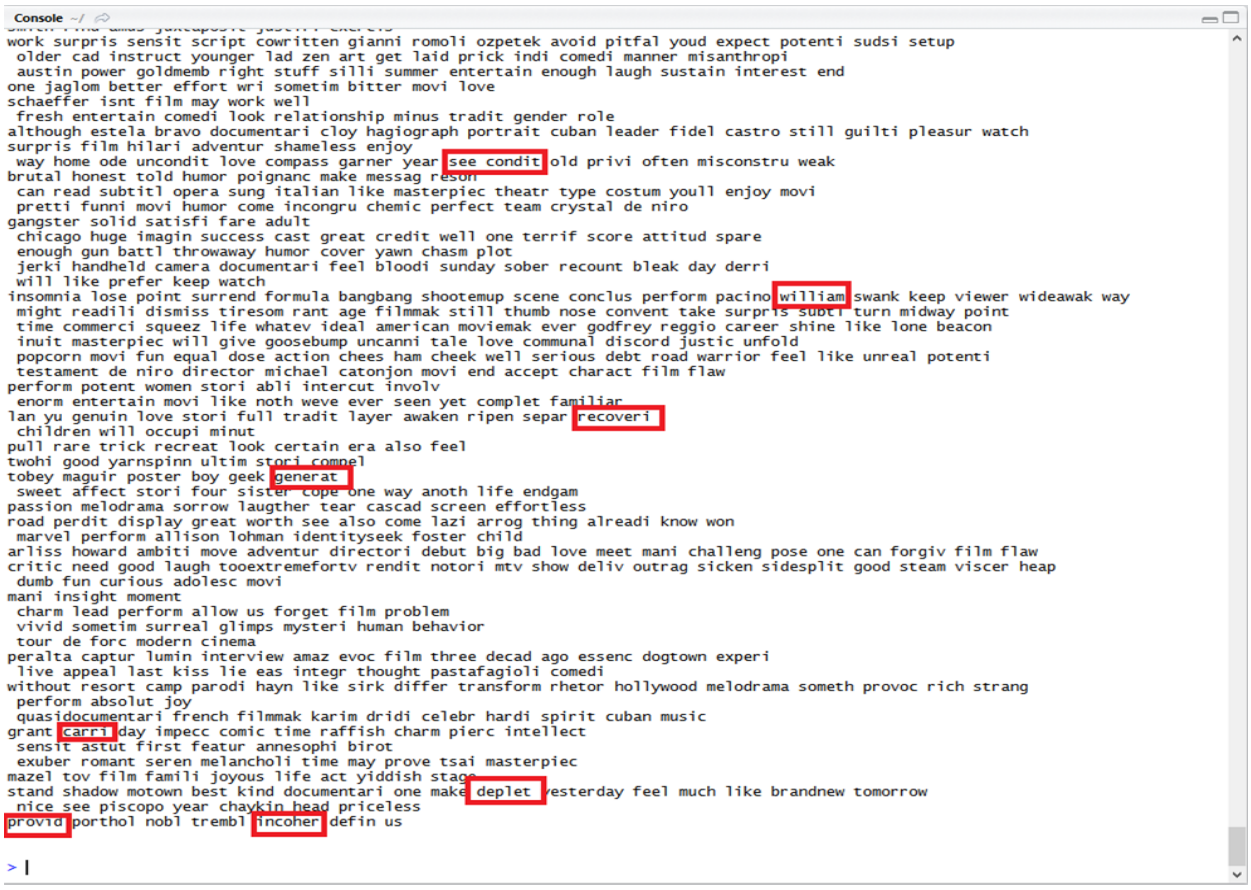
Stemming is the process of reducing words to their word stem or root. For example – fishing, fished, and fisher all relate to the root word fish.

The process we have used in our analysis although bit crude works in the following way –

Suffix-stripping: Suffix stripping algorithm is supported by tm package and do not rely on lookup table that consists of inflected forms and root forms relations. Instead, a typically smaller list of "rules" is stored which provides a path for the algorithm, given an input word form, to find its root form. Some examples of the rules include

- if the word ends in 'ed', removing the 'ed'
- if the word ends in 'ing', removing the 'ing'
- if the word ends in 'ly', removing the 'ly'


Console
one jagloms better efforts wry sometime bitter movie love
schaeffer isnt film may works well
fresh entertaining comedy looks relationships minus traditional gender roles
although estela bravos documentary cloyingly hagiographic portrait cuban leader fidel castro still guilty pleasure watch
surprisingly film hilarious adventure shamelessly enjoyed
way home ode unconditional love compassion garnered years seeing condition old privy often misconstrued weakness
brutally honest told humor poignancy makes message resonate
can read subtitles opera sung italian like masterpiece theatre type costumes youll enjoy movie
pretty funny movie humor coming incongruous chemically perfect teaming crystal de niro
gangster solid satisfying fare adults
chicago hugely imaginative successful casting great credit well one terrific score attitude spare
enough gun battles throwaway humor cover yawning chasm plot
jerky handheld camera documentary feel bloody sunday sobering recount bleak day derry
will likely prefer keep watching
insomnia loses points surrenders formulaic bangbang shootemup scene conclusion performances pacino williams swank keep viewer wideaw
ake way
might readily dismissed tiresome rant aging filmmaker still thumbing nose convention takes surprising subtle turn midway point
time commercialism squeezed life whatever idealism american moviemaking ever godfrey reggios career shines like lonely beacon
inuit masterpiece will give goosebumps uncanny tale love communal discord justice unfolds
popcorn movie fun equal doses action cheese ham cheek well serious debt road warrior feels like unrealized potential
testament de niro director michael catonjones movies end accept characters film flaws
performances potent womens stories ably intercut involving
enormously entertaining movie like nothing weve ever seen yet completely familiar
lan yu genuine love story full traditional layers awakening ripening separation recovery
children will occupied minutes
pulls rare trick recreating look certain era also feel
twohys good yarnspinner ultimately story compels
tobey maguire poster boy geek generation
sweetly affecting story four sisters coping one way another lifes endgame
passion melodrama sorrow laugther tears cascade screen effortlessly
road perdition display greatness worth seeing also comes laziness arrogance thing already knows won
marvelous performance allison lohman identityseeking foster child
arliss howards ambitious moving adventurous directorial debut big bad love meets many challenges poses one can forgive film flaws
critics need good laugh tooextremefortv rendition notorious mtv show delivers outrageous sickening sidesplitting goods steaming visc
eral heaps
dumb fun curiously adolescent movie
many insightful moments
charms lead performances allow us forget films problems
vivid sometimes surreal glimpse mysteries human behavior
tour de force modern cinema
peralta captures luminous interviews amazingly evocative film three decades ago essence dogtown experience
lively appeal last kiss lies ease integrates thoughtfulness pastafagioli comedy
without resorting camp parody haynes like sirk differently transformed rhetoric hollywood melodrama something provocative rich stran
ge
performances absolute joy
quasidocumentary french filmmaker karim dridi celebrates hardy spirit cuban music
grant carries day impeccable comic timing raffish charm piercing intellect
sensitive astute first feature annesophie birot
exuberantly romantic serenely melancholy time may prove tsais masterpiece
mazel tov film familys joyous life acting yiddish stage
standing shadows motown best kind documentary one makes depleted yesterday feel much like brandnew tomorrow
nice see piscopo years chaykin headly priceless
provides porthole noble trembling incoherence defines us
>


Console
work surpris sensit script cowritten gianni romoli ozpetek avoid pitfal youd expect potenti sudsi setup
older cad instruct younger lad zen art get laid prick indi comedi manner misanthropi
austin power goldmemb right stuff silli summer entertain enough laugh sustain interest end
one jaglom better effort wri sometim bitter movi love
schaeffer isnt film may work well
fresh entertain comedi look relationship minus tradit gender role
although estela bravo documentari cloy hagiograph portrait cuban leader fidel castro still guilti pleasur watch
surpris film hilari adventur shameless enjoy
way home ode uncondit love compass garner year see condit old privi often misconstru weak
brutal honest told humor poignanc make messag reson
can read subtitl opera sung italian like masterpiec theatr type costum youll enjoy movi
pretti funni movi humor come incongru chemic perfect team crystal de niro
gangster solid satisfi fare adult
chicago huge imagin success cast great credit well one terrif score attitud spare
enough gun battl throwaway humor cover yawn chasm plot
jerki handheld camera documentari feel bloodi sunday sober recount bleak day derri
will like prefer keep watch
insomnia lose point surrend formula bangbang shootemup scene conclus perform pacino william swank keep viewer wideawak way
might readili dismiss tiresom rant age filmmak still thumb nose convent take surpris subtl turn midway point
time commerci squeez life whatev ideal american moviemak ever godfrey reggio career shine like lone beacon
inuit masterpiec will give goosebump uncanni tale love communal discord justic unfold
popcorn movi fun equal dose action chees ham cheek well serious debt road warrior feel like unreal potenti
testament de niro director michael catonjon movi end accept charact film flaw
perform potent women stori abli intercut involv
enorm entertain movi like noth weve ever seen yet complet familiar
lan yu genuin love stori full tradit layer awaken ripen separ recoveri
children will occupi minut
pull rare trick recreat look certain era also feel
twohi good yarnspinn ultim stori compel
tobey maguir poster boy geek generat
sweet affect stori four sister cope one way anoth life endgam
passion melodrama sorrow laugther tear cascad screen effortless
road perdit display great worth see also come lazi arrog thing alreadi know won
marvel perform allison lohman identityseek foster child
arliss howard ambiti move adventur directori debut big bad love meet mani challeng pose one can forgiv film flaw
critic need good laugh tooextremefortv rendit notori mtv show deliv outrag sicken sidesplit good steam viscer heap
dumb fun curious adolesc movi
mani insight moment
charm lead perform allow us forget film problem
vivid sometim surreal glimps mysteri human behavior
tour de forc modern cinema
peralta captur lumin interview amaz evoc film three decad ago essenc dogtown experi
live appeal last kiss lie eas integr thought pastafagioli comedi
without resort camp parodi hayn like sirk differ transform rhetor hollywood melodrama someth provoc rich strang
perform absolut joy
quasidocumentari french filmmak karim dridi celebr hardi spirit cuban music
grant carri day impecc comic time raffish charm pierc intellect
sensit astut first featur annesophi birot
exuber romant seren melancholi time may prove tsai masterpiec
mazel tov film famili joyous life act yiddish stage
stand shadow motown best kind documentari one make deplet yesterday feel much like brandnew tomorrow
nice see piscopo year chaykin head priceless
provid porthol nobl trembl incoher defin us
>

Step4: Staging the data for analysis

Document-Term Matrix | Term-Document Matrix

A document-term matrix or term-document matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents and columns corresponding to the terms. For example

D1 = "I like apples"

D2 = "I hate apples"

The document-term matrix would be:

	i	**like**	**hate**	**apples**
D1	1	1	0	1
D2	1	0	1	1

The above table shows which documents contain which terms and how many times they appear

If you transpose the rows and columns we get a Term-Document Matrix in which the terms are rows and documents are columns. We could work with either DTM or TDM but in this case we have used Document-Term Matrix (DTM)

```
docs <- tm_map(docs,stemDocument)

dtm <- DocumentTermMatrix(docs)

dtm

tdm <- TermDocumentMatrix(docs)

tdm
```

```
> ### Stage the Data
> dtm <- DocumentTermMatrix(docs)
> dtm
<<DocumentTermMatrix (documents: 1, terms: 13791)>>
Non-/sparse entries: 13791/0
Sparsity           : 0%
Maximal term length: 33
Weighting          : term frequency (tf)
> tdm <- TermDocumentMatrix(docs)
> tdm
<<TermDocumentMatrix (terms: 13791, documents: 1)>>
Non-/sparse entries: 13791/0
Sparsity           : 0%
Maximal term length: 33
Weighting          : term frequency (tf)
```

As we can observe in the above image, the corpus is a matrix of the dimensions 1 x 13791 in which 0% of the rows are zero (sparsity). The maximum term length is 33 and the weighting is based on term frequency (tf).

Finding frequently occurring words

Once the corpus of text is converted into a mathematical object we can analyze it using quantitative methods.

```
findFreqTerms(dtm, lowfreq=100)
```

```
> findFreqTerms(dtm, lowfreq=100)
  [1] "act"       "action"    "actor"     "almost"    "also"      "american"  "anoth"     "audienc"
  [9] "back"      "bad"       "beauti"    "becom"     "best"      "better"    "big"       "can"
 [17] "cant"      "care"      "cast"      "charact"   "charm"     "come"      "comedi"    "compel"
 [25] "despit"    "direct"    "director"  "documentari" "doesnt"  "dont"      "drama"     "effect"
 [33] "emot"      "end"       "enjoy"     "enough"    "entertain" "even"      "ever"      "everi"
 [41] "famili"    "fan"       "far"       "feel"      "film"      "filmmak"   "find"      "first"
 [49] "full"      "fun"       "funni"     "get"       "give"      "good"      "great"     "hard"
 [57] "heart"     "high"      "hollywood" "human"     "humor"     "idea"      "imagin"    "intellig"
 [65] "interest"  "isnt"      "just"      "keep"      "kid"       "kind"      "know"      "lack"
 [73] "laugh"     "leav"      "less"      "life"      "like"      "littl"     "live"      "long"
 [81] "look"      "lot"       "love"      "made"      "make"      "man"       "manag"     "mani"
 [89] "may"       "might"     "minut"     "moment"    "move"      "movi"      "much"      "music"
 [97] "need"      "never"     "new"       "noth"      "offer"     "often"     "old"       "one"
[105] "origin"    "part"      "peopl"     "perform"   "pictur"    "piec"      "play"      "plot"
[113] "point"     "power"     "quit"      "rather"    "real"      "realli"    "right"     "romant"
[121] "say"       "scene"     "screen"    "script"    "see"       "seem"      "seen"      "sens"
[129] "set"       "show"      "someth"    "star"      "still"     "stori"     "subject"   "surpris"
[137] "take"      "tale"      "that"      "there"     "thing"     "think"     "though"    "thriller"
[145] "time"      "tri"       "turn"      "two"       "ultim"     "use"       "visual"    "want"
[153] "watch"     "way"       "well"      "will"      "without"   "wonder"    "work"      "world"
[161] "worth"     "year"      "yet"       "your"
```

We have used the findFreqTerms() function to return all terms that occur more than 100 times in the entire corpus. Also note that the result is ordered alphabetically, and not by frequency.

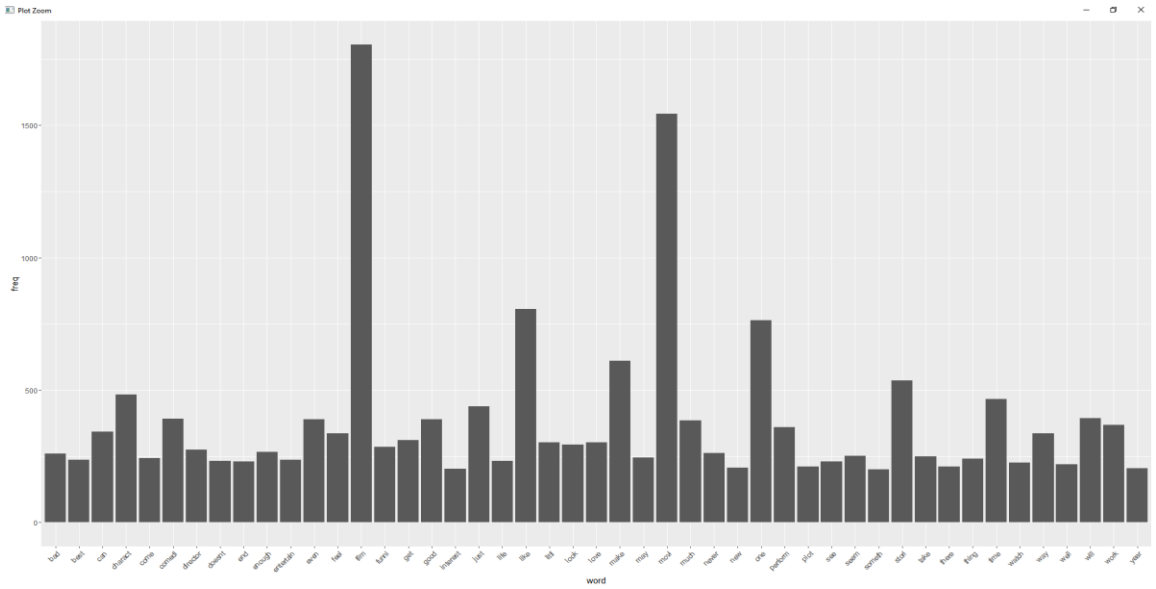
We can observe from the above image that words related to movie genres (action, thriller, humor, documentary and action etc.), user emotions in reviews (like, real, lack, compelling, enjoy, worth etc.) and positive and negative words (cant, never, like, bad, better, worth etc.) have occurred frequently.

Visualizing using Histogram

R has its own graphing capability. We tried to use the plotting function to plot the above analysis on a Histogram for better understanding of user reviews.

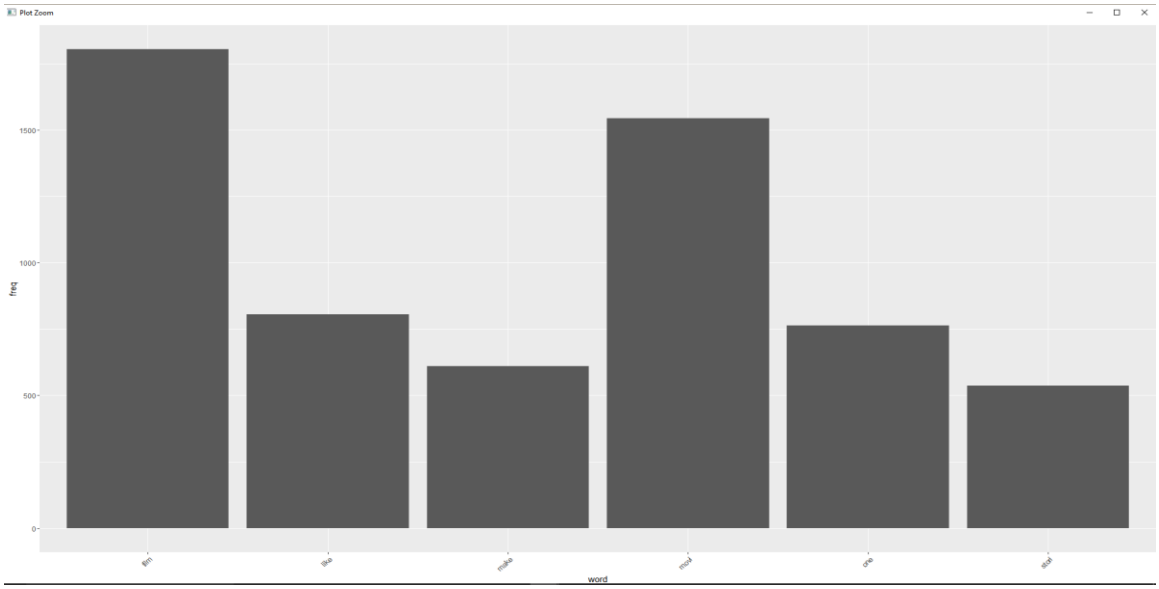
```
library(ggplot2)

wf <- data.frame(word=names(freq), freq=freq)

p <- ggplot(subset(wf, freq>500), aes(word, freq))

p <- p + geom_bar(stat="identity")

p <- p + theme(axis.text.x=element_text(angle=45, hjust=1))

p
```

The first line creates a data frame (list of columns of equal length). The data frame also contains the name of the columns – in this case these are term and occurrence respectively. Using the ggplot() function we are plotting only those terms that occur more than 500 times. The aes option in ggplot describes the plot aesthetics. The stat=”identity” option in geom_bar() ensures that the height of each bar is proportional to the data value that is mapped to the y-axis. The last line specifies that the x-axis label should be at a 45 degree angle and should be horizontally justified.



The above Histogram shows all the frequently occurring words [Freq > 200]

Through the above Histogram we realized that words related to movie genre, and user emotions were quite substantial. It was quite obvious that words like movie and film will have the highest occurrence but it was of not much help to us in our sentiment analysis. We have primarily focused on words that talk about a certain type of movie or the reviewer sentiment associated with it.



The above Histogram shows all the frequently occurring words [Freq > 500]

On changing the frequency from 200 to 500 lot of words which were under our observation were filtered out. We could conclude that most of the words associated with movie genre and reviewer emotions were in the range of 50 to 200 but not more.

Word Cloud

Word Cloud is an image composed of words used in a particular text or subject, in which the size of each word indicates its frequency or importance.

```
library(wordcloud)

dtms <- removeSparseTerms(dtm, 0.15) # Prepare the data (max 15% empty space)

freq <- colSums(as.matrix(dtm)) # Find word frequencies

dark2 <- brewer.pal(6, "Dark2")

wordcloud(names(freq), freq, max.words=100, rot.per=0.2, colors=dark2)
```

Setting a seed number ensures that you get the same look each time. For our analysis we have specified a max word limit = 100, which means it will show only the top 100 occurring words in our corpus document. We could use filter condition (set by frequency) to achieve similar results.

```
wordcloud(names(freq), freq, min.freq=100, rot.per=0.2, colors=dark2)
```

The color to the word cloud has been added using the RcolorBrewer package.

As we can see the frequency of word is directly related to the boldness of the word. Words like movi, film have a higher occurrence frequency as compared to words like drama, moment or love. We also observed that our work on cleaning up the data (in the pre-processing stage) has been as expected but not perfect. We can still notice that the stop word removal process has not done its job well, words like isn't, don't and also, need to be removed from the corpus.

To solve the above issue of stop words we recommend creating a built-in stop word list with the custom one.

Step4: Cluster Analysis

Library: cluster and fpc

Cluster Analysis or clustering is the task of grouping set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters)

For our project we wanted to cluster our corpus using

- Hierarchical Clustering
- K-Means Clustering

Hierarchical Clustering

Hierarchical Clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally falls into two types

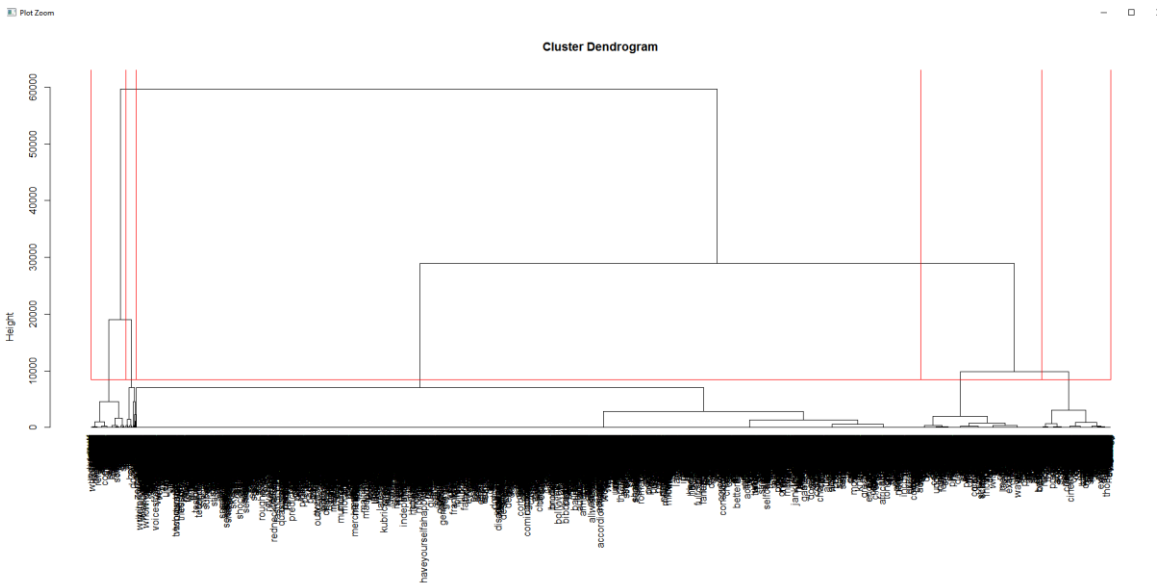
  - Agglomerative: bottom up approach where each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy
  - Divisive: top down approach where all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy

In the process we first calculated the distance between words and then clustered them accordingly to similarity.

```
dtms <- removeSparseTerms(dtm, 0.15)

library(cluster)

d <- dist(t(dtms), method="euclidian")
```

Analyzing clusters can be difficult and therefore in order to fairly read the hierarchical clustering we have highlighted the cluster groups in red. We choose to have the value of k as 5 (creation of 5 clusters)

```
fit <- hclust(d=d, method="ward")

plot.new()

plot(fit, hang=-1)

groups <- cutree(fit, k=5)

rect.hclust(fit, k=5, border="red")
```



During the hierarchical clustering execution we faced a lot of problems with respect to system memory. After several trial and error attempts we were still unable to obtain the right hierarchical cluster analysis nor did breaking down of dataset helped us. This is the output we achieved which was difficult to analyze.

The efficient way to fix this problem would be to write efficient code and usage of better sampling methods.

PART 2: Sentiment Analysis

Libraries used: plyr, stringr, e1071

The basic flow is as follows –

- Select corpus
- Analyze the corpus for positive and negative words
- Training classification
- Use of classification algorithm (Naïve Bayes)

We started the process by loading up the positive and negative sentences and linking the AFINN polarity list

```
afinn_list <- read.delim(file='C:\\IS-688-Web-Mining-Final-Team-6\\sentiment-analysis\\sentiment_analysis\\AFINN\\AFINN-111.txt ', header=FALSE, stringsAsFactors=FALSE)

posText <- read.delim(file='C:\\IS-688-Web-Mining-Final-Team-6\\sentiment-analysis\\sentiment_analysis\\polarityData\\pd\\polarity-2.txt ', header=FALSE, stringsAsFactors=FALSE)

negText <- read.delim(file='C:\\IS-688-Web-Mining-Final-Team-6\\sentiment-analysis\\sentiment_analysis\\polarityData\\pd\\polarity-1.txt ', header=FALSE, stringsAsFactors=FALSE)
```

Next we classified the AFINN words into four categories

- Very Negative (rating -5 to -4)
- Negative (rating -3,-2 or -1)
- Positive (rating 1,2 or 3)
- Very Positive (rating 4 or 5)

Next we used Naïve Bayes classifier from the e1071 package to classify the sentences as positive or negative. The basic working of the classifier is as follows – it looks at how the number of words in each of the four categories relates to whether the sentence is positive or negative. It then tries to guess whether a sentence is positive or negative by examining how many words it has in each category and relating this to the probabilities of those numbers appearing in positive and negative sentences.

Lastly, we generated a confusion matrix to visualize the results of the classification algorithm. The result was as follows

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known.

```
          actual
predicted  positive negative
  positive     2846     1546
  negative     2485     3786
>
> #run a binomial test for confidence interval of results
> binom.test(confTable[1,1] + confTable[2,2], nrow(results), p=0.5)

        Exact binomial test

data:  confTable[1, 1] + confTable[2, 2] and nrow(results)
number of successes = 6632, number of trials = 10663, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.6126810 0.6311799
sample estimates:
probability of success
             0.6219638
```

In our analysis we found out that the probability of success was around 62% which leaves room for immense improvement which can be achieved by writing better code and using efficient classification algorithms or using a different training data.

Part 3: Finding Range of Emotions

Libraries used: Syuzhet, Pander

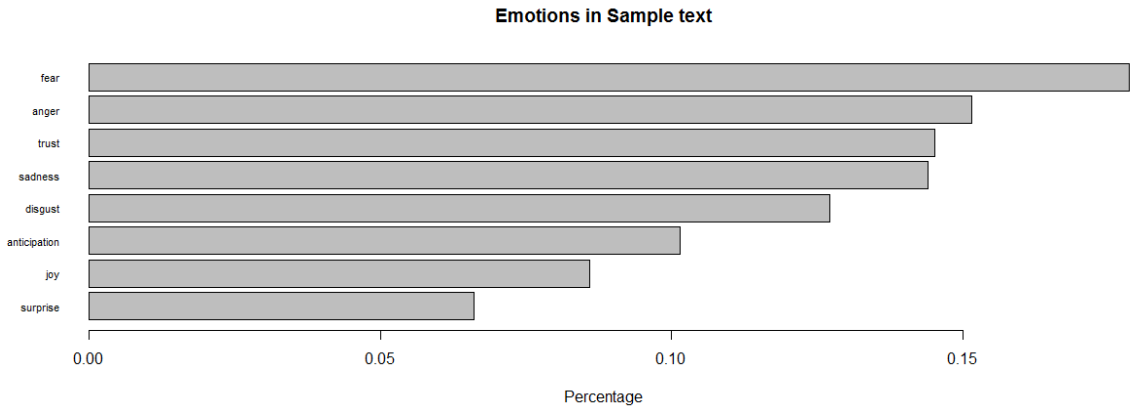
The basic flow is as:

- Reading the data from the specified Files.
- Reading it into a Class vector.
- Getting the Head.
- Getting the Sum, Mean and Summary of the Vector elements.
- Plotting the Emotions using barplot function.

```
summary(sentiment_vector)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -2146   -2146   -2146   -2146   -2146   -2146
```

In this part we are trying to figure out the number of emotional words under eight main categories. Then we are plotting it as a bar graph having the range of 0-0.15 percent.

anger	anticipation	disgust	fear	joy	sadness	surprise	trust
1173	785	984	1382	665	1114	512	1124



Limitations

Over the course of the project we have faced the following challenges/limitations –

1. Comments are not specific to a particular movie, so it becomes quite difficult to predict user sentiment related to a specific movie
2. Instead of downloading premade dataset a better way is to use data scrapping
3. System resources were not sufficient to process larger dataset (we faced this problem in clustering analysis)

Conclusion

After careful analysis of the Histogram, Word-Cloud and Bar-graph of emotions, we have concluded two main things:

- The dataset contains more reviews about Horror and Action movies.
- The word 'Fear' has the highest frequency or range among all other words in the dataset.

References

IS-688 Moodle Study Material

https://cran.r-project.org/web/packages/tm/tm.pdf

https://cran.r-project.org/web/packages/SnowballC/SnowballC.pdf

https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf

https://cran.r-project.org/web/packages/cluster/cluster.pdf

https://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf

https://cran.r-project.org/web/packages/NLP/NLP.pdf

https://cran.r-project.org/web/packages/stringr/stringr.pdf

Sentiment Analysis – Andy Bromberg

https://www.wikipedia.org/

http://www.cs.cornell.edu/people/pabo/movie-review-data/

http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010

https://www.youtube.com/