Text Mining

- The first 5 rows and a three columns of the data

title	release_year	description
3%	2020	In a future where the elite inhabit an island paradise far from the crowded slums, you get one chance to join the 3% saved from squalor.
7:19	2016	After a devastating earthquake hits Mexico City, trapped survivors from all walks of life wait to be rescued while trying desperately to stay alive.
23:59	2011	When an army recruit is found dead, his fellow soldiers are forced to confront a terrifying secret that's haunting their jungle island training camp.
9	2009	In a postapocalyptic world, rag-doll robots hide in fear from dangerous machines out to exterminate them, until a brave newcomer joins the group.
21	2008	A brilliant group of students become card-counting experts with the intent of swindling millions out of Las Vegas casinos by playing blackjack.

Token

- A token is a meaningful unit of text.
- One row of text will be converted to multiple rows of tokens.

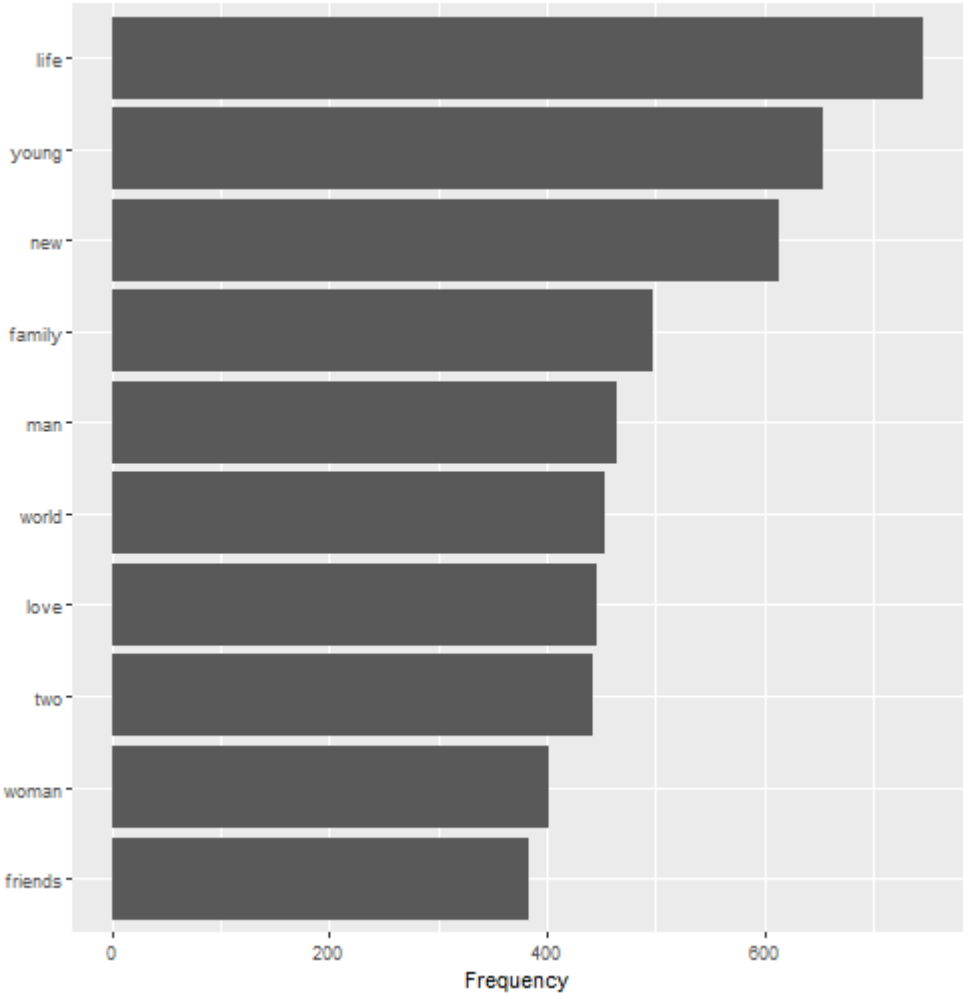
title	description
Avengers: Infinity War	Superheroes amass to stop intergalactic sociopath Thanos from acquiring a full set of Infinity Stones and wiping out half of all life in the universe.

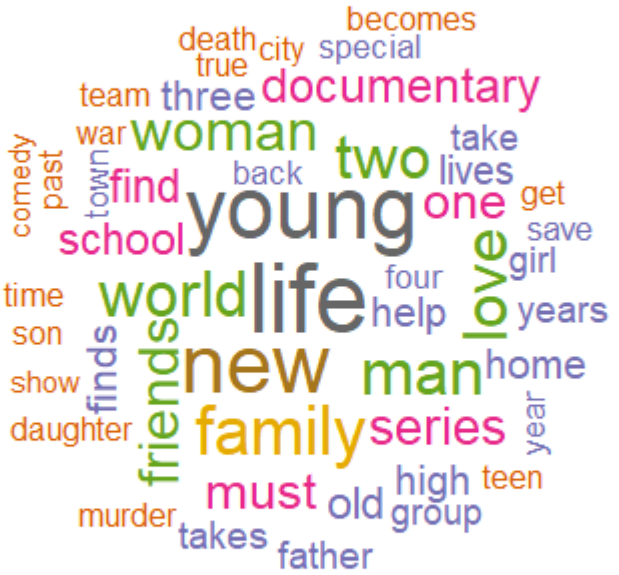
title	word
Avengers: Infinity War	superheroes
Avengers: Infinity War	amass
Avengers: Infinity War	to
Avengers: Infinity War	stop
Avengers: Infinity War	intergalactic
Avengers: Infinity War	sociopath
Avengers: Infinity War	thanos
Avengers: Infinity War	from
Avengers: Infinity War	acquiring
Avengers: Infinity War	a
Avengers: Infinity War	full
Avengers: Infinity War	set
Avengers: Infinity War	of
Avengers: Infinity War	infinity
Avengers: Infinity War	stones
Avengers: Infinity War	and
Avengers: Infinity War	wiping
Avengers: Infinity War	out
Avengers: Infinity War	half
Avengers: Infinity War	of
Avengers: Infinity War	all
Avengers: Infinity War	life
Avengers: Infinity War	in
Avengers: Infinity War	the

title	word
Avengers: Infinity War	superheroes
Avengers: Infinity War	amass
Avengers: Infinity War	stop
Avengers: Infinity War	intergalactic
Avengers: Infinity War	sociopath
Avengers: Infinity War	thanos
Avengers: Infinity War	acquiring
Avengers: Infinity War	full
Avengers: Infinity War	set
Avengers: Infinity War	infinity
Avengers: Infinity War	stones
Avengers: Infinity War	wiping
Avengers: Infinity War	half
Avengers: Infinity War	life
Avengers: Infinity War	universe

title	word
Avengers: Infinity War	superheroes
Avengers: Infinity War	amass
Avengers: Infinity War	stop
Avengers: Infinity War	intergalactic
Avengers: Infinity War	sociopath
Avengers: Infinity War	thanos
Avengers: Infinity War	acquiring
Avengers: Infinity War	full
Avengers: Infinity War	set
Avengers: Infinity War	infinity
Avengers: Infinity War	stones
Avengers: Infinity War	wiping
Avengers: Infinity War	half
Avengers: Infinity War	life
Avengers: Infinity War	universe
Spider-Man 3	seemingly
Spider-Man 3	invincible
Spider-Man 3	spider
Spider-Man 3	man
Spider-Man 3	goes
Spider-Man 3	new
Spider-Man 3	crop
Spider-Man 3	villains
Spider-Man 3	third
Spider-Man 3	installment
Spider-Man 3	blockbuster
Spider-Man 3	adventure

word	n
life	746
young	655
new	613
family	497
man	464
world	453
love	447
two	443
woman	402
friends	383


life
young
new
family
man
world
love
two
woman
friends
0
200
400
600
Frequency


becomes
death
city
special
true
documentary
team
three
woman
war
take
two
lives
comedy
past
town
back
find
young
one
get
school
save
love
girl
four
time
world
life
help
years
son
friends
new
man
home
finds
show
family
series
year
daughter
must
high
teen
old
group
murder
takes
father

Sentiment Analysis

title	description
Avengers: Infinity War	Superheroes amass to stop intergalactic sociopath Thanos from acquiring a full set of Infinity Stones and wiping out half of all life in the universe.

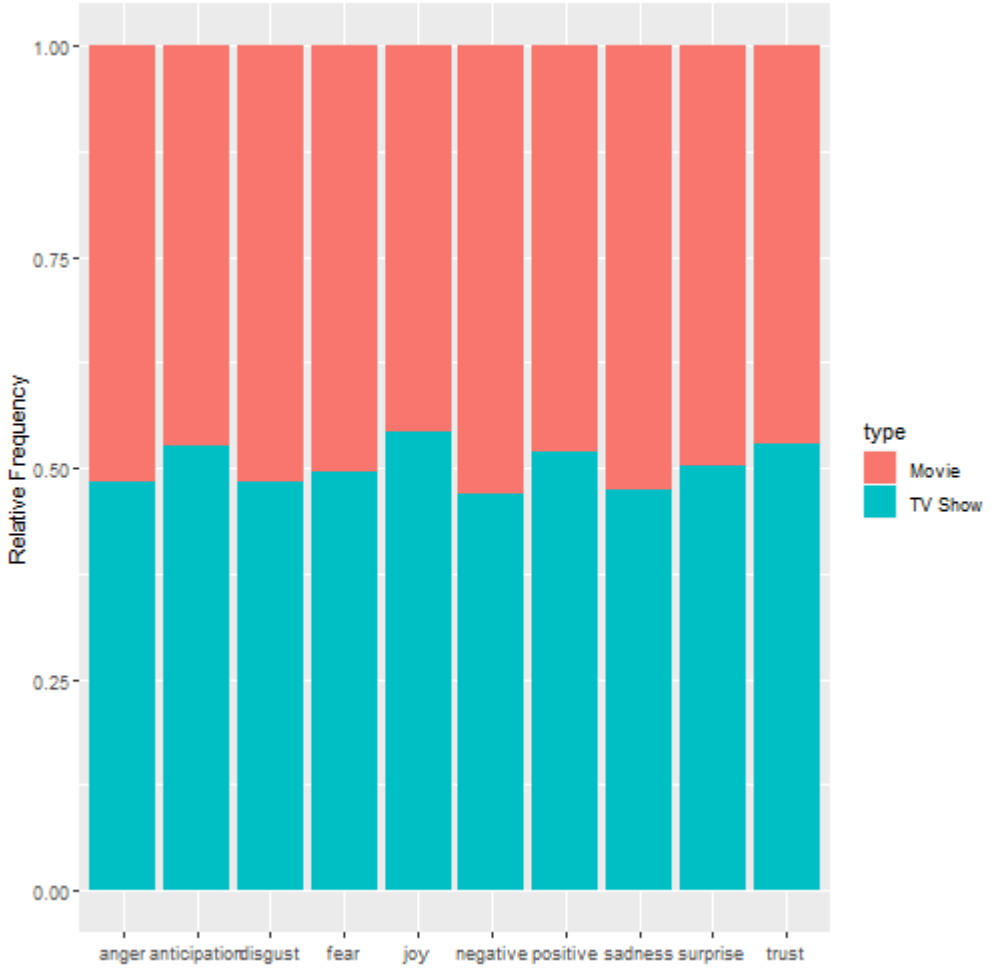
title	word
Avengers: Infinity War	superheroes
Avengers: Infinity War	amass
Avengers: Infinity War	to
Avengers: Infinity War	stop
Avengers: Infinity War	intergalactic
Avengers: Infinity War	sociopath
Avengers: Infinity War	thanos
Avengers: Infinity War	from
Avengers: Infinity War	acquiring
Avengers: Infinity War	a
Avengers: Infinity War	full
Avengers: Infinity War	set
Avengers: Infinity War	of
Avengers: Infinity War	infinity
Avengers: Infinity War	stones
Avengers: Infinity War	and
Avengers: Infinity War	wiping
Avengers: Infinity War	out
Avengers: Infinity War	half
Avengers: Infinity War	of
Avengers: Infinity War	all
Avengers: Infinity War	life
Avengers: Infinity War	in
Avengers: Infinity War	the

title	word
Avengers: Infinity War	superheroes
Avengers: Infinity War	amass
Avengers: Infinity War	stop
Avengers: Infinity War	intergalactic
Avengers: Infinity War	sociopath
Avengers: Infinity War	thanos
Avengers: Infinity War	acquiring
Avengers: Infinity War	full
Avengers: Infinity War	set
Avengers: Infinity War	infinity
Avengers: Infinity War	stones
Avengers: Infinity War	wiping
Avengers: Infinity War	half
Avengers: Infinity War	life
Avengers: Infinity War	universe

Using `nrc`

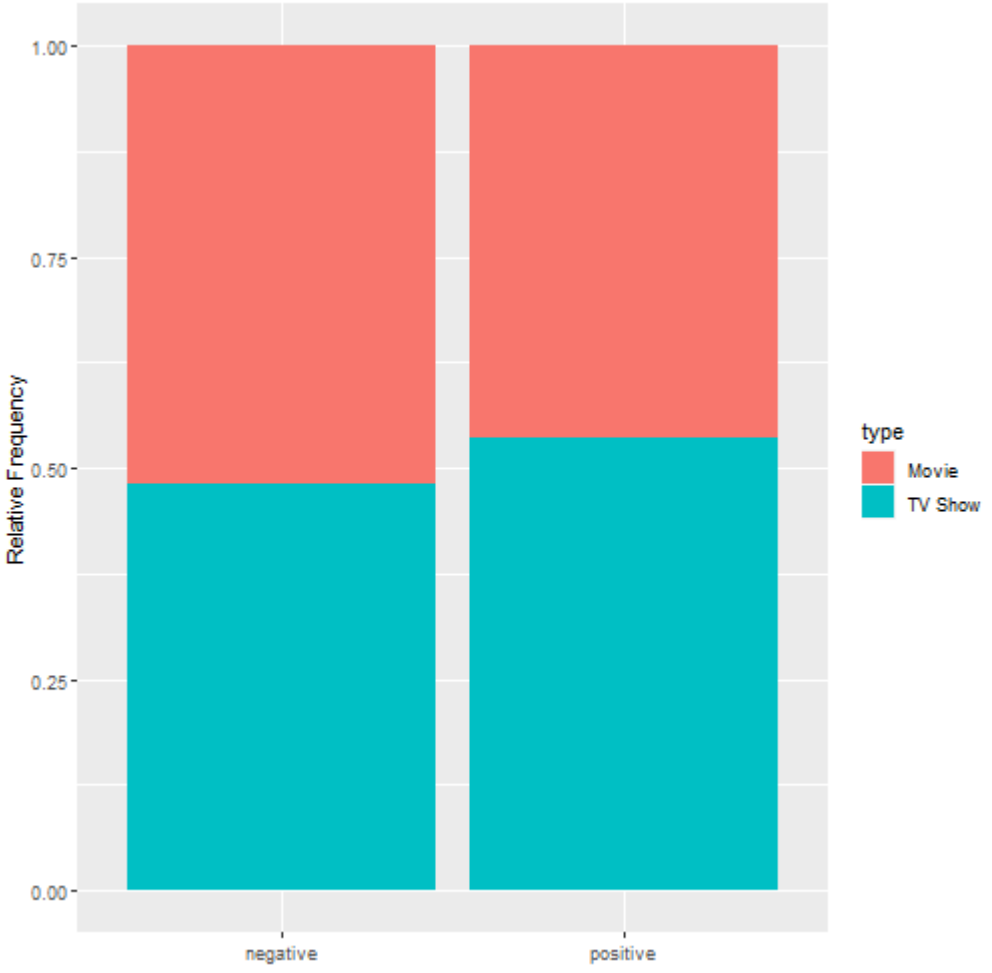
title	word	sentiment
Avengers: Infinity War	acquiring	anticipation
Avengers: Infinity War	acquiring	positive
Avengers: Infinity War	full	positive
Avengers: Infinity War	infinity	anticipation
Avengers: Infinity War	infinity	joy
Avengers: Infinity War	infinity	positive
Avengers: Infinity War	infinity	trust

- Some words are missing
- Some words have more than one sentiment


1.00
0.75
0.50
0.25
0.00
Relative Frequency
anger
anticipation
disgust
fear
joy
negative
positive
sadness
surprise
trust
type
Movie
TV Show

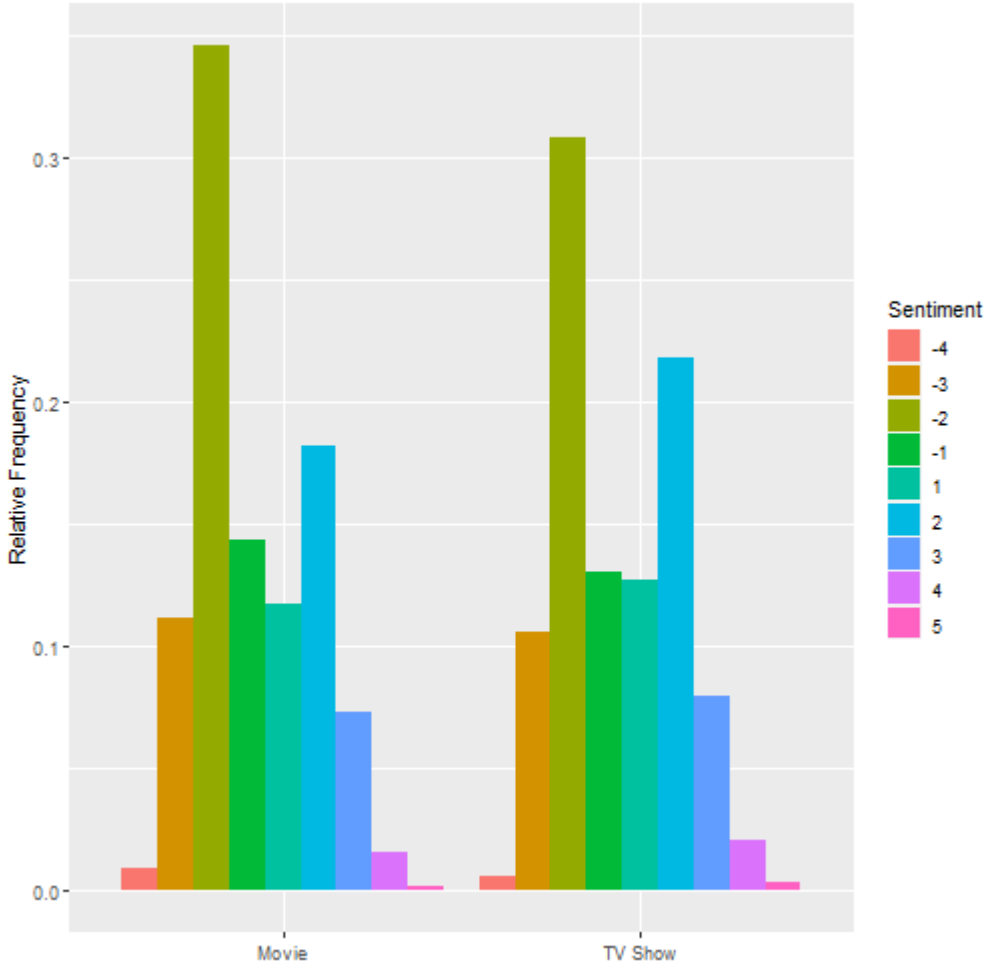
Using `bing`

title	word	sentiment
Spider-Man 3	invincible	positive
Spider-Man 3	villains	negative
Spider-Man 3	blockbuster	positive


1.00
0.75
0.50
0.25
0.00
Relative Frequency
negative
positive
type
Movie
TV Show

Using `afinn`

title	word	value
Avengers: Infinity War	stop	-1
Spider-Man 3	invincible	2
Spider-Man 3	blockbuster	3
Spider-Man 3	adventure	2


Relative Frequency
0.3
0.2
0.1
0.0
Movie
TV Show
Sentiment
-4
-3
-2
-1
1
2
3
4
5

Modeling

- TF (Term Frequency) = (Number of times term `t` appears in a text) / (Total number of terms in the text).
- IDF (Inverse Document Frequency) = log_e(Total number of texts / Number of texts with term `t` in it).
- TF_IDF = TF * IDF

Example:

Consider a text containing 100 words wherein the word `cat` appears 3 times.

Then, TF(Cat) = 3/100 = 0.03.

Now, assume we have 10 million texts and the word `cat` appears in one thousand of these.

Then, IDF(Cat) = log(10,000,000/1,000) = 4.

Thus, the Tf_idf(Cat) 0.03 * 4 = 0.12.

```
##  [1] "tfidf_description_a"    "tfidf_description_an"   "tfidf_description_
##  [4] "tfidf_description_her"  "tfidf_description_his"  "tfidf_description_
##  [7] "tfidf_description_of"   "tfidf_description_the"  "tfidf_description_
## [10] "tfidf_description_with" "target"
```