Applied Machine Learning Homework 5

Due 12 Dec,2022 (Monday) 11:59PM EST

Natural Language Processing

We will train a supervised model to predict if a movie has a positive or a negative review.

Dataset loading & dev/test splits

1.0) Load the movie reviews dataset from NLTK library

In [1]:

```
import nltk
nltk.download("movie_reviews")
import pandas as pd
from nltk.corpus import twitter_samples
from sklearn.model_selection import train_test_split
from nltk.corpus import stopwords
nltk.download('stopwords')
nltk.download('punkt')
stop = stopwords.words('english')
import string
import re
from nltk.stem import PorterStemmer
from nltk.tokenize import word_tokenize
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.feature_extraction.text import TfidfVectorizer
```

```
[nltk_data] Downloading package movie_reviews to
[nltk_data]     C:\Users\Clare\AppData\Roaming\nltk_data...
[nltk_data]   Package movie_reviews is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\Clare\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\Clare\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

In [2]:

```
from nltk.corpus import movie_reviews
```

In [3]:

```
len(movie_reviews.fileids())
```

Out[3]:

```
2000
```

In [4]:

```
negative_fileids = movie_reviews.fileids('neg')
positive_fileids = movie_reviews.fileids('pos')

pos_document = [(' '.join(movie_reviews.words(file_id)),category) for file_id in movie_reviews.fileids() for category in movie_re
neg_document = [(' '.join(movie_reviews.words(file_id)),category) for file_id in movie_reviews.fileids() for category in movie_re

# List of postive and negative reviews
pos_list = [pos[0] for pos in pos_document]
neg_list = [neg[0] for neg in neg_document]
```

In [5]:

```
len(pos_document)
```

Out[5]:

```
1000
```

In [6]:

```
pos_document[0]
```

Out[6]:

```
('films adapted from comic books have had plenty of success , whether they \' re about superheroes ( batman , super
man , spawn ) , or geared toward kids ( casper ) or the arthouse crowd ( ghost world ) , but there \' s never reall
y been a comic book like from hell before . for starters , it was created by alan moore ( and eddie campbell ) , wh
o brought the medium to a whole new level in the mid \' 80s with a 12 - part series called the watchmen . to say mo
ore and campbell thoroughly researched the subject of jack the ripper would be like saying michael jackson is start
ing to look a little odd . the book ( or " graphic novel , " if you will ) is over 500 pages long and includes near
ly 30 more that consist of nothing but footnotes . in other words , don \' t dismiss this film because of its sourc
e . if you can get past the whole comic book thing , you might find another stumbling block in from hell \' s direc
tors , albert and allen hughes . getting the hughes brothers to direct this seems almost as ludicrous as casting ca
rrot top in , well , anything , but riddle me this : who better to direct a film that \' s set in the ghetto and fe
atures really violent street crime than the mad geniuses behind menace ii society ? the ghetto in question is , of
course , whitechapel in 1888 london \' s east end . it \' s a filthy , sooty place where the whores ( called " unfo
rtunates " ) are starting to get a little nervous about this mysterious psychopath who has been carving through the
ir profession with surgical precision . when the first stiff turns up , copper peter godley ( robbie coltrane , the
world is not enough ) calls in inspector frederick abberline ( johnny depp , blow ) to crack the case . abberline ,
a widower , has prophetic dreams he unsuccessfully tries to quell with copious amounts of absinthe and opium . upon
arriving in whitechapel , he befriends an unfortunate named mary kelly ( heather graham , say it isn \' t so ) and
proceeds to investigate the horribly gruesome crimes that even the police surgeon can \' t stomach . i don \' t thi
nk anyone needs to be briefed on jack the ripper , so i won \' t go into the particulars here , other than to say m
oore and campbell have a unique and interesting theory about both the identity of the killer and the reasons he cho
oses to slay . in the comic , they don \' t bother cloaking the identity of the ripper , but screenwriters terry ha
yes ( vertical limit ) and rafael yglesias ( les mis ? rables ) do a good job of keeping him hidden from viewers un
til the very end . it \' s funny to watch the locals blindly point the finger of blame at jews and indians because
, after all , an englishman could never be capable of committing such ghastly acts . and from hell \' s ending had
me whistling the stonecutters song from the simpsons for days ( " who holds back the electric car / who made steve
guttenberg a star ? " ) . don \' t worry - it \' ll all make sense when you see it . now onto from hell \' s appear
ance : it \' s certainly dark and bleak enough , and it \' s surprising to see how much more it looks like a tim bu
rton film than planet of the apes did ( at times , it seems like sleepy hollow 2 ) . the print i saw wasn \' t comp
letely finished ( both color and music had not been finalized , so no comments about marilyn manson ) , but cinemat
ographer peter deming ( don \' t say a word ) ably captures the dreariness of victorian - era london and helped mak
e the flashy killing scenes remind me of the crazy flashbacks in twin peaks , even though the violence in the film
pales in comparison to that in the black - and - white comic . oscar winner martin childs \' ( shakespeare in love
) production design turns the original prague surroundings into one creepy place . even the acting in from hell is
solid , with the dreamy depp turning in a typically strong performance and deftly handling a british accent . ians
holm ( joe gould \' s secret ) and richardson ( 102 dalmatians ) log in great supporting roles , but the big surpri
se here is graham . i cringed the first time she opened her mouth , imagining her attempt at an irish accent , but
it actually wasn \' t half bad . the film , however , is all good . 2 : 00 - r for strong violence / gore , sexuali
ty , language and drug content',
 'pos')
```

In [7]:

```
pos_document[0][0]
```

Out[7]:

```
'films adapted from comic books have had plenty of success , whether they \' re about superheroes ( batman , superm
an , spawn ) , or geared toward kids ( casper ) or the arthouse crowd ( ghost world ) , but there \' s never really
been a comic book like from hell before . for starters , it was created by alan moore ( and eddie campbell ) , who
brought the medium to a whole new level in the mid \' 80s with a 12 - part series called the watchmen . to say moor
e and campbell thoroughly researched the subject of jack the ripper would be like saying michael jackson is startin
g to look a little odd . the book ( or " graphic novel , " if you will ) is over 500 pages long and includes nearly
30 more that consist of nothing but footnotes . in other words , don \' t dismiss this film because of its source .
if you can get past the whole comic book thing , you might find another stumbling block in from hell \' s directors
, albert and allen hughes . getting the hughes brothers to direct this seems almost as ludicrous as casting carrot
top in , well , anything , but riddle me this : who better to direct a film that \' s set in the ghetto and feature
s really violent street crime than the mad geniuses behind menace ii society ? the ghetto in question is , of cours
e , whitechapel in 1888 london \' s east end . it \' s a filthy , sooty place where the whores ( called " unfortuna
tes " ) are starting to get a little nervous about this mysterious psychopath who has been carving through their pr
ofession with surgical precision . when the first stiff turns up , copper peter godley ( robbie coltrane , the worl
d is not enough ) calls in inspector frederick abberline ( johnny depp , blow ) to crack the case . abberline , a w
idower , has prophetic dreams he unsuccessfully tries to quell with copious amounts of absinthe and opium . upon ar
riving in whitechapel , he befriends an unfortunate named mary kelly ( heather graham , say it isn \' t so ) and pr
oceeds to investigate the horribly gruesome crimes that even the police surgeon can \' t stomach . i don \' t think
anyone needs to be briefed on jack the ripper , so i won \' t go into the particulars here , other than to say moor
e and campbell have a unique and interesting theory about both the identity of the killer and the reasons he choose
s to slay . in the comic , they don \' t bother cloaking the identity of the ripper , but screenwriters terry hayes
( vertical limit ) and rafael yglesias ( les mis ? rables ) do a good job of keeping him hidden from viewers until
the very end . it \' s funny to watch the locals blindly point the finger of blame at jews and indians because , af
ter all , an englishman could never be capable of committing such ghastly acts . and from hell \' s ending had me w
histling the stonecutters song from the simpsons for days ( " who holds back the electric car / who made steve gutt
enberg a star ? " ) . don \' t worry - it \' ll all make sense when you see it . now onto from hell \' s appearance
: it \' s certainly dark and bleak enough , and it \' s surprising to see how much more it looks like a tim burton
film than planet of the apes did ( at times , it seems like sleepy hollow 2 ) . the print i saw wasn \' t completel
y finished ( both color and music had not been finalized , so no comments about marilyn manson ) , but cinematograp
her peter deming ( don \' t say a word ) ably captures the dreariness of victorian - era london and helped make the
flashy killing scenes remind me of the crazy flashbacks in twin peaks , even though the violence in the film pales
in comparison to that in the black - and - white comic . oscar winner martin childs \' ( shakespeare in love ) prod
uction design turns the original prague surroundings into one creepy place . even the acting in from hell is solid
, with the dreamy depp turning in a typically strong performance and deftly handling a british accent . ians holm (
joe gould \' s secret ) and richardson ( 102 dalmatians ) log in great supporting roles , but the big surprise here
is graham . i cringed the first time she opened her mouth , imagining her attempt at an irish accent , but it actua
lly wasn \' t half bad . the film , however , is all good . 2 : 00 - r for strong violence / gore , sexuality , lan
guage and drug content'
```

In [8]:

```
len(neg_document)
```

Out[8]:

```
1000
```

In [9]:

```
pos_list
```

Out[9]:

```
['films adapted from comic books have had plenty of success , whether they \' re about superheroes ( batman , su
perman , spawn ) , or geared toward kids ( casper ) or the arthouse crowd ( ghost world ) , but there \' s never
really been a comic book like from hell before . for starters , it was created by alan moore ( and eddie campbel
l ) , who brought the medium to a whole new level in the mid \' 80s with a 12 - part series called the watchmen
. to say moore and campbell thoroughly researched the subject of jack the ripper would be like saying michael ja
ckson is starting to look a little odd . the book ( or " graphic novel , " if you will ) is over 500 pages long
and includes nearly 30 more that consist of nothing but footnotes . in other words , don \' t dismiss this film
because of its source . if you can get past the whole comic book thing , you might find another stumbling block
in from hell \' s directors , albert and allen hughes . getting the hughes brothers to direct this seems almost
as ludicrous as casting carrot top in , well , anything , but riddle me this : who better to direct a film that
\' s set in the ghetto and features really violent street crime than the mad geniuses behind menace ii society ?
the ghetto in question is , of course , whitechapel in 1888 london \' s east end . it \' s a filthy , sooty plac
e where the whores ( called " unfortunates " ) are starting to get a little nervous about this mysterious psycho
path who has been carving through their profession with surgical precision . when the first stiff turns up , cop
per peter godley ( robbie coltrane , the world is not enough ) calls in inspector frederick abberline ( johnny d
epp , blow ) to crack the case . abberline , a widower , has prophetic dreams he unsuccessfully tries to quell w
ith copious amounts of absinthe and opium . upon arriving in whitechapel , he befriends an unfortunate named mar
y kelly ( heather graham , say it isn \' t so ) and proceeds to investigate the horribly gruesome crimes that ev
```

In [10]:

```
neg_list
```

Out[10]:

```
['plot : two teen couples go to a church party , drink and then drive . they get into an accident . one of the g
uys dies , but his girlfriend continues to see him in her life , and has nightmares . what \' s the deal ? watch
the movie and " sorta " find out . . . critique : a mind - fuck movie for the teen generation that touches on a
very cool idea , but presents it in a very bad package . which is what makes this review an even harder one to w
rite , since i generally applaud films which attempt to break the mold , mess with your head and such ( lost hig
hway & memento ) , but there are good and bad ways of making all types of films , and these folks just didn \' t
snag this one correctly . they seem to have taken this pretty neat concept , but executed it terribly . so what
are the problems with the movie ? well , its main problem is that it \' s simply too jumbled . it starts off " n
ormal " but then downshifts into this " fantasy " world in which you , as an audience member , have no idea what
\' s going on . there are dreams , there are characters coming back from the dead , there are others who look li
ke the dead , there are strange apparitions , there are disappearances , there are a looooot of chase scenes , t
here are tons of weird things that happen , and most of it is simply not explained . now i personally don \' t m
ind trying to unravel a film every now and then , but when all it does is give me the same clue over and over ag
ain , i get kind of fed up after a while , which is this film \' s biggest problem . it \' s obviously got this
big secret to hide , but it seems to want to hide it completely until its final five minutes . and do they make
things entertaining , thrilling or even engaging , in the meantime ? not really . the sad part is that the arrow
and i both dig on flicks like this , so we actually figured most of it out by the half - way point , so all of t
he strangeness after that did start to make a little bit of sense . but it still didn \' t the make the film all
```

1.1) Make a data frame that has reviews and its label

In [38]:

```
# code here
movies = pd.DataFrame(pos_document+neg_document, columns = ["Review", "Label"])
movies
```

Out[38]:

	Review	Label
0	films adapted from comic books have had plenty...	pos
1	every now and then a movie comes along from a ...	pos
2	you ' ve got mail works alot better than it de...	pos
3	" jaws " is a rare film that grabs your attent...	pos
4	moviemaking is a lot like being the general ma...	pos
...	...	...
1995	if anything , " stigmata " should be taken as ...	neg
1996	john boorman ' s " zardoz " is a goofy cinemat...	neg
1997	the kids in the hall are an acquired taste . i...	neg
1998	there was a time when john carpenter was a gre...	neg
1999	two party guys bob their heads to haddaway ' s...	neg

2000 rows × 2 columns

1.2 look at the class distribution of the movie reviews

In [39]:

```
# code here
movies["Label"].value_counts()
```

Out[39]:

```
pos    1000
neg    1000
Name: Label, dtype: int64
```

1.3) Create a development & test split (80/20 ratio):

In [40]:

```
# code here
x_dev, x_test, y_dev, y_test = train_test_split(movies["Review"], movies["Label"],
                                                test_size = 0.2, random_state = 42)
```

Data preprocessing

We will do some data preprocessing before we tokenize the data. We will remove `#` symbol, hyperlinks, stop words & punctuations from the data. You may use `re` package for this.

1.4) Replace the # symbol with '' in every review

In [41]:

```
# code here
x_dev = x_dev.str.replace('#', '\"')
x_test = x_test.str.replace('#', '\"')
```

In [42]:

```
x_dev[240]
```

Out[42]:

```
'seen september 13 , 1998 at 4 p . m at rotterdam square mall cinema 6 , theater " 2 , with chris wessell for free
using my sony / loews critic \' s pass . [ theater rating : * * 1 / 2 : good seats , average sound , picture unstab
le ] " rounders " is exactly the kind of movie parents don \' t want their kids to see . it \' s not that it \' s a
drunken orgy of sex and violence , but because it \' s a film that flat - out says you can make a career out of gam
bling . and to take make things " worse " it proves this through its original , fascinating story . there have been
countless crime films both past and present that evoke the " noir " mood , that is , the dark , shady atmosphere wh
ere the vices of the world become more fascinating on screen than they would in real life . this film starts off in
the tradiational noir style , introducing us to the underworld of modern gambling where the stakes are high and so
is the price for losing . matt damon stars as mike mcdermott , a 20 - something law student in present - day new yo
rk city who tells us how the game of poker is really played . damon narrates throughout the film , but the entire o
pening scene is voiced - over so perfectly to completely and totally define the setting . mike \' s about to go up
against teddy kgb ( malkovich ) , a russian gangster who looks like a serial killer . but then again , he practical
ly is one and the film does everything to convey that sense - the look in his eyes , his slow movements , his intri
cate mannerisms - all combined with the classic noir cinematography of isolated brightness within the darkness of t
he underworld ( literally ) . damon in the flesh might seem a little out of place with his expensive clothes and pe
rfectly - groomed features , but his narration is what brings it all together . he never sounds like he \' s readin
g from a script , nor that he \' s trying to embellish anything , it just comes natural to him . the screenwriters
use the right words and phrases to describe the mood , from the smell of the air , to the logic involved in reading
the other guys \' faces and cards , and all without sounding remotely trite . immediately we get the sense that pok
er isn \' t for gamblers , but for near - geniuses with nerves of steel . the game is a quiet war , with strategies
just as complex and the same sense of honor among the soldiers . the gangsters mike plays against are the same ones
that might kill someone for scratching their car , but when it comes to the game of poker , all respect is due to t
he winner because he is truly the better man . the film does an excellent job in establishing its atmosphere during
the first act . it concentrates so much in this aspect that the background and the progression of the story stumble
a bit . we learn only a little about mike , both past and present . presently we know he has a girlfriend named jo
( gretchen mol ) who he constantly argues with over his gambling . they go through a few break - up / make - up cyc
les until mike \' s childhood pal and fellow rounder " worm " ( norton ) is released from prison . it \' s not at a
ll surprising worm owes thousand of dollars to the mob , but what is surprising is how the film is able to take suc
h a predictable element and execute it the way it does . technically , the plot isn \' t unlike many children \' s
sitcoms in which the " good " kid \' s " bad " friend gets the good kid in trouble and yet the good kid remains fri
ends with the bad kid . what this film does is use a different medium to tell that story . mike and worm have been
in over their heads their entire lives , but both share a passion for out - thinking the other player who is trying
to do the same to them . where as worm prefers to go the sleazy route of cheating ( hence his jail time ) , mike al
ways takes the cards he \' s dealt and works with them . sometimes they pay off ( i . e . his ability to pay his wa
y through law school on his gambling money ) , but other times they don \' t ( i . e . the fact he takes himself ou
t of the game and works a steady job after dropping $ 30 , 000 on a single hand ) . once things start happening the
film is able to expand and develop its plot into an intricate web of detail and mood . mike and worm bob and weave
through all kinds of games at all kinds of places , from socialites \' mansions , to taking the tourists at atlanti
c city , to outwitting the gangsters that control it all . everything they come into contact with is a big poker ga
me in that everything \' s a battle against the cards destiny deals . one scene demonstrates this perfectly in whic
h mike is told by a judge ( martin landou , in a perfectly cast and performed role ) that destiny is everything and
yet nothing at the same time . matt \' s good at gambling but he \' s also got potential to be a great lawyer . he
could go professional as either , but with one he could lose everything or win big , but with the other there \' s
stability but not much risk involved . can someone who \' s gambled his entire life really cash in his chips and le
ave ? if the film had been just a subtle lesson in poker - playing , then the ending is our test . everything is to
ld from mike \' s perspective , but we \' re finally able to recognize some things on our own . this makes the fina
l , against - all - odds showdown seem like just that . it works just like the game it revolves around - showing us
some of the cards , but still evokes the element of the unknown , and the consequences thereof . what separates " r
ounders " from most other films about games is the fact the challenge and the skills are more important than winnin
g in the end .'
```

1.5) Replace hyperlinks with '' in every review

In [43]:

```
# code here
```

1.6) Remove all stop words

In [44]:

```
# code here
stop
```

Out[44]:

```
['i',
 'me',
 'my',
 'myself',
 'we',
 'our',
 'ours',
 'ourselves',
 'you',
 "you're",
 "you've",
 "you'll",
 "you'd",
 'your',
 'yours',
 'yourself',
 'yourselves',
 'he'.
```

In [46]:

```
for word in stop:
    temp_string = ' ' + word + ' '
    x_dev = x_dev.str.replace(temp_string, ' ').replace(temp_string, ' ')
    x_test = x_test.str.replace(temp_string, ' ').replace(temp_string, ' ')
```

1.7) Remove all punctuations

In [49]:

```
# code here
punctuation_list = string.punctuation
punctuation_list
```

Out[49]:

```
'!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'
```

In [50]:

```
for p in punctuation_list:
    x_dev = x_dev.str.replace(p, ' ')
    x_test = x_test.str.replace(p, ' ')
```

```
<ipython-input-50-f2fa45b7a224>:2: FutureWarning: The default value of regex will change from True to False in a fu
ture version. In addition, single character regular expressions will *not* be treated as literal strings when regex
=True.
  x_dev = x_dev.str.replace(p, ' ')
<ipython-input-50-f2fa45b7a224>:3: FutureWarning: The default value of regex will change from True to False in a fu
ture version. In addition, single character regular expressions will *not* be treated as literal strings when regex
=True.
  x_test = x_test.str.replace(p, ' ')
```

In [51]:

```
x_dev[240]
```

Out[51]:

```
'seen september 13   1998 4 p   rotterdam square mall cinema 6   theater   2   chris wessell free using sony   loew
s critic   pass     theater rating       1   2   good seats   average sound   picture unstable     rounders   exact
ly kind movie parents   want kids see       drunken orgy sex violence     film flat   says make career gambling   t
ake make things   worse   proves original   fascinating story   countless crime films past present evoke   noir   m
ood     dark   shady atmosphere vices world become fascinating screen would real life   film starts tradiational no
ir style   introducing us underworld modern gambling stakes high price losing   matt damon stars mike mcdermott   2
0   something law student present   day new york city tells us game poker really played   damon narrates throughout
film   entire opening scene voiced   perfectly completely totally define setting   mike   go teddy kgb   malkovich
russian gangster looks like serial killer     practically one film everything convey sense   look eyes   slow movem
ents   intricate mannerisms   combined classic noir cinematography isolated brightness within darkness underworld
literally     damon flesh might seem little place expensive clothes perfectly   groomed features   narration brings
together   never sounds like   reading script     trying embellish anything   comes natural   screenwriters use rig
ht words phrases describe mood   smell air   logic involved reading guys   faces cards   without sounding remotely
trite   immediately get sense poker   gamblers   near   geniuses nerves steel   game quiet war   strategies complex
sense honor among soldiers   gangsters mike plays ones might kill someone scratching car   comes game poker   respe
ct due winner truly better man   film excellent job establishing atmosphere first act   concentrates much aspect ba
ckground progression story stumble bit   learn little mike   past present   presently know girlfriend named jo   gr
etchen mol   constantly argues gambling   go break     make   cycles mike   childhood pal fellow rounder   worm
norton   released prison     surprising worm owes thousand dollars mob   surprising film able take predictable elem
ent execute way   technically   plot   unlike many children   sitcoms   good   kid     bad   friend gets good kid t
rouble yet good kid remains friends bad kid   film use different medium tell story   mike worm heads entire lives
share passion   thinking player trying   worm prefers go sleazy route cheating   hence jail time     mike always ta
kes cards   dealt works   sometimes pay     e   ability pay way law school gambling money     times       e   fact
takes game works steady job dropping   30   000 single hand     things start happening film able expand develop plo
t intricate web detail mood   mike worm bob weave kinds games kinds places   socialites   mansions   taking tourist
s atlantic city   outwitting gangsters control   everything come contact big poker game everything   battle cards d
estiny deals   one scene demonstrates perfectly mike told judge   martin landou   perfectly cast performed role   d
estiny everything yet nothing time   matt   good gambling   also got potential great lawyer   could go professional
either   one could lose everything win big     stability much risk involved   someone   gambled entire life really
cash chips leave   film subtle lesson poker   playing   ending test   everything told mike   perspective     finall
y able recognize things   makes final       odds showdown seem like   works like game revolves around   showing us
cards   still evokes element unknown   consequences thereof   separates   rounders   films games fact challenge ski
lls important winning end  '
```

1.8) Apply stemming on the development & test datasets using Porter algorithm

In [58]:

```
#code here
def stemSentence(sentence):
    porter = PorterStemmer()
    token_words = word_tokenize(sentence)
    stem_sentence = [porter.stem(word) for word in token_words]
    return " ".join(stem_sentence)
```

In [64]:

```
for index, sentence in x_dev.iteritems():
    x_dev[index] = stemSentence(sentence)

for index, sentence in x_test.iteritems():
    x_test[index] = stemSentence(sentence)
```

In [65]:

```
x_dev
```

Out[65]:

```
968     insan inspir music alferd packer first man eve...
240     seen septemb 13 1998 4 p rotterdam squar mall ...
819     one biggest clich serial killer film also one ...
692     make sequel wide belov film weighti proposit i...
420     with team 200 graphic artist anim work first f...
                              ...
1130    i never understood clich hell earth truli mean...
1294    in make 1954 japan monster film godzilla trans...
860     the verdict spine chill drama horror maestro s...
1459    scientist dr alexand mccabe bob gunton respon ...
1126    plot separ glamor hollywood coupl must pretend...
Name: Review, Length: 1600, dtype: object
```

In [66]:

```
x_test
```

Out[66]:

```
1860    i guess wild bachelor parti gone realli bad wo...
353     with abund trite recycl movi late 90 tremend d...
1333    as hot shot defens attorney kevin lomax keanu ...
905     hedwig john cameron mitchel born boy name hans...
1289    i heard call jaw claw fair summat plot though ...
                              ...
965     in mani way twotg tough guy movi la confidenti...
1284    if austin power spi shag half origin zani sill...
1739    when film produc shoestr budget coupl hardwork...
261     titan close perfect movi upset film cost 200 m...
535     it curiou thing found willi call carri whole m...
Name: Review, Length: 400, dtype: object
```

Model training

1.9) Create bag of words features for each review in the development dataset

In [87]:

```
#code here
vector = CountVectorizer(stop_words = 'english')
x_dev_transform = vector.fit_transform(x_dev)
x_dev_transform
feature_names = vector.get_feature_names()
feature_names
print(feature_names[:10])
print(feature_names[10000:10020])
print(feature_names[::3000])
```

```
['00', '000', '0009f', '007', '03', '04', '05425', '10', '100', '1000']
['imperson', 'impervi', 'impetu', 'impish', 'implant', 'implau', 'implausibilit', 'implement', 'impli', 'implic',
'implicit', 'implicitli', 'implod', 'implor', 'impo', 'impond', 'import', 'importantli', 'imposs', 'impossibilti']
['00', 'businessmen', 'drifter', 'hansel', 'loiter', 'payload', 'secretli', 'trilian']
```

1.10) Train a Logistic Regression model on the development dataset

In [88]:

```
#code here
lr = LogisticRegression().fit(x_dev_transform, y_dev)
```

```
C:\Users\Clare\anaconda3\lib\site-packages\sklearn\linear_model\_logistic.py:763: ConvergenceWarning: lbfgs failed
to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html (https://scikit-learn.org/stable/modules/preprocessi
ng.html)
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression (https://scikit-learn.org/stable/
modules/linear_model.html#logistic-regression)
  n_iter_i = _check_optimize_result(
```

In [89]:

```
x_test_transform = vector.transform(x_test)
```

1.11) Create TF-IDF features for each review in the development dataset

In [82]:

```
#code here
vector2 = TfidfVectorizer()
x_dev_transform2 = vector2.fit_transform(x_dev)
x_test_transform2 = vector2.transform(x_test)
feature_names2 = vector2.get_feature_names()
feature_names2
print(feature_names2[:10])
print(feature_names2[10000:10020])
print(feature_names2[::3000])
```

```
['00', '000', '0009f', '007', '03', '04', '05425', '10', '100', '1000']
['iliopulo', 'ilk', 'ill', 'illeana', 'illeg', 'illegitim', 'illena', 'illicit', 'illinoi', 'illit', 'illog', 'ill
u', 'illumin', 'illuminata', 'illuminati', 'illusionist', 'illusori', 'illustr', 'illustri', 'ilm']
['00', 'burnel', 'drang', 'hallstr', 'liquor', 'parasit', 'scope', 'towel']
```

1.12) Train the Logistic Regression model on the development dataset with TF-IDF features

In [84]:

```
#code here
lr2 = LogisticRegression().fit(x_dev_transform2, y_dev)
```

1.13) Compare the performance of the two models on the test dataset. Explain the difference in results obtained?

In [90]:

```
#code here
lr.score(x_test_transform, y_test)
```

Out[90]:

```
0.8
```

In [86]:

```
lr2.score(x_test_transform2, y_test)
```

Out[86]:

```
0.805
```

It is slightly better than the results obtained in the Bag of words, since it not only counts the world, but assign weight importance

In [ ]:

```

```