Week3

March 31, 2023

0.1 Part 1: Using the TextBlob Sentiment Analyzer

0.1.1 1. Import the movie review data as a data frame and ensure that the data is loaded properly.

[1]:
```
import pandas as pd
from textblob import TextBlob
```

[2]:
```
#Load the dataset as a Pandas data frame.
labeled_train_data_df = pd.read_csv('labeledTrainData.tsv', sep='\t')

print(labeled_train_data_df.shape)
labeled_train_data_df.head()
```

```
(25000, 3)
```

[2]:

	id	sentiment	review
0	5814_8	1	With all this stuff going down at the moment w…
1	2381_9	1	\The Classic War of the Worlds\" by Timothy Hi…
2	7759_3	0	The film starts with a manager (Nicholas Bell)…
3	3630_4	0	It must be assumed that those who praised this…
4	9495_8	1	Superbly trashy and wondrously unpretentious 8…

0.1.2 2. How many of each positive and negative reviews are there?

[3]:
```
labeled_train_data_df.groupby(['sentiment'])['sentiment'].count().sort_index()
# 50% are positive and 50% are negative reviews.
```

[3]:
```
sentiment
0    12500
1    12500
Name: sentiment, dtype: int64
```

0.1.3 3. Use TextBlob to classify each movie review as positive or negative.

[4]:
```
#Assume that a polarity score greater than or equal to zero is a positive␣
 ↪sentiment and
#less than 0 is a negative sentiment.
```

```
#labeled_train_data_df['TextBlob_sentiment'] = labeled_train_data_df['review'].
 ↪apply(lambda review:
#                                                                          ␣
 ↪TextBlob(review).sentiment)
labeled_train_data_df['subjectivity'] = labeled_train_data_df['review'].
 ↪apply(lambda review:
                                                                          ␣
 ↪TextBlob(review).sentiment.subjectivity)

labeled_train_data_df['polarity'] = labeled_train_data_df['review'].
 ↪apply(lambda review:
                                                                          ␣
 ↪TextBlob(review).sentiment.polarity)

labeled_train_data_df['analysis'] = labeled_train_data_df['polarity'].
 ↪apply(lambda x: 1 if x >=0 else 0)

labeled_train_data_df.groupby(['analysis'])['analysis'].count().sort_index()
```

[4]:
```
analysis
0     5983
1    19017
Name: analysis, dtype: int64
```

0.1.4 4. Check the accuracy of this model. Is this model better than random guessing?

[5]:
```
#Comparing sentiment and analysis for accuracy
from sklearn.metrics import accuracy_score

accuracy_score(labeled_train_data_df['sentiment'],labeled_train_data_df['analysis'])

#It appears the accuracy of this model is better than the random guessing based␣
 ↪on existing sentiment data.
```

[5]:
```
0.68524
```

0.1.5 5. For up to five points extra credit, use another prebuilt text sentiment analyzer, e.g., VADER, and repeat steps (3) and (4).

[6]:
```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer

analyzer = SentimentIntensityAnalyzer()
labeled_train_data_df['Vader_polarity'] = labeled_train_data_df['review'].
 ↪apply(lambda review: analyzer.polarity_scores(review)['compound'])
```

[7]:
```
labeled_train_data_df['Vader_analysis'] =␣
↪labeled_train_data_df['Vader_polarity'].apply(lambda x: 1 if x >= 0.05 else␣
↪0)

labeled_train_data_df.groupby(['Vader_analysis'])['Vader_analysis'].count().
↪sort_index()
```

[7]:
```
Vader_analysis
0     8493
1    16507
Name: Vader_analysis, dtype: int64
```

[8]:
```
accuracy_score(labeled_train_data_df['sentiment'],labeled_train_data_df['Vader_analysis'])
```

[8]:
```
0.69556
```

Accuracy from TextBlob is 68.5% while that from VADER is 69.6%

0.2 Part 2: Prepping Text for a Custom Model

[9]:
```
#Load the dataset as a Pandas data frame.
labeled_train_data_df = pd.read_csv('labeledTrainData.tsv', sep='\t')
```

0.2.1 1. Convert all text to lowercase letters.

[10]:
```
labeled_train_data_df  = labeled_train_data_df.applymap(lambda x: x.lower() if␣
↪type(x)==str else x)
```

0.2.2 2. Remove punctuation and special characters from the text.

[11]:
```
# Remove punctuations
import string
labeled_train_data_df.review = labeled_train_data_df.review.apply(lambda review:
↪ review.translate(str.maketrans('', '', string.punctuation)))
```

[12]:
```
#Remove special characters
labeled_train_data_df.review = labeled_train_data_df.review.str.
↪replace(r"[^a-zA-Z0-9]+", " " , regex=True)
labeled_train_data_df.head()
```

[12]:

	id	sentiment	review
0	5814_8	1	with all this stuff going down at the moment w…
1	2381_9	1	the classic war of the worlds by timothy hines…
2	7759_3	0	the film starts with a manager nicholas bell g…
3	3630_4	0	it must be assumed that those who praised this…
4	9495_8	1	superbly trashy and wondrously unpretentious 8…

0.2.3 3. Remove stop words.

[13]:
```
# Load library
from nltk.corpus import stopwords
stopwords= stopwords.words('english')

# Remove stop words
labeled_train_data_df.review = labeled_train_data_df.review.apply(
    lambda review: ' '.join([word for word in review.split() if word not in␣
 ↪(stopwords)]))
labeled_train_data_df.head()
```

[13]:

	id	sentiment	review
0	5814_8	1	stuff going moment mj ive started listening mu…
1	2381_9	1	classic war worlds timothy hines entertaining …
2	7759_3	0	film starts manager nicholas bell giving welco…
3	3630_4	0	must assumed praised film greatest filmed oper…
4	9495_8	1	superbly trashy wondrously unpretentious 80s e…

0.2.4 4. Apply NLTK's PorterStemmer.

[14]:
```
from nltk.stem.porter import PorterStemmer

# Create stemmer
porter = PorterStemmer()
# Apply stemmer
labeled_train_data_df.review = labeled_train_data_df.review.apply(lambda review:
 ↪' '.join([porter.stem(word) for word in review.split()]))
labeled_train_data_df.head()
```

[14]:

	id	sentiment	review
0	5814_8	1	stuff go moment mj ive start listen music watc…
1	2381_9	1	classic war world timothi hine entertain film …
2	7759_3	0	film start manag nichola bell give welcom inve…
3	3630_4	0	must assum prais film greatest film opera ever…
4	9495_8	1	superbl trashi wondrous unpretenti 80 exploit …

0.2.5 5. Create a bag-of-words matrix from your stemmed text (output from (4)), where each row is a word-count vector for a single movie review (see sections 5.3 & 6.8 in the Machine Learning with Python Cookbook). Display the dimensions of your bag-of-words matrix. The number of rows in this matrix should be the same as the number of rows in your original data frame.

[15]:
```
import numpy as np
from sklearn.feature_extraction.text import CountVectorizer

# Create the bag of words feature matrix
```

```
count = CountVectorizer()
bag_of_words = count.fit_transform(labeled_train_data_df.review)
# Show feature matrix
bag_of_words
```

[15]:
```
<25000x91908 sparse matrix of type '<class 'numpy.int64'>'
        with 2439277 stored elements in Compressed Sparse Row format>
```

[16]:
```
bag_of_words.shape
```

[16]:
```
(25000, 91908)
```

0.2.6 6. Create a term frequency-inverse document frequency (tf-idf) matrix from your stemmed text, for your movie reviews (see section 6.9 in the Machine Learning with Python Cookbook). Display the dimensions of your tf-idf matrix. These dimensions should be the same as your bag-of-words matrix.

[17]:
```
from sklearn.feature_extraction.text import TfidfVectorizer

# Create the tf-idf feature matrix
tfidf = TfidfVectorizer()
feature_matrix = tfidf.fit_transform(labeled_train_data_df.review)
# Show tf-idf feature matrix
tfidf.vocabulary_
```

[17]:
```
{'stuff': 77888,
 'go': 33711,
 'moment': 52832,
 'mj': 52606,
 'ive': 42173,
 'start': 76735,
 'listen': 47296,
 'music': 54427,
 'watch': 88063,
 'odd': 57487,
 'documentari': 23379,
 'wiz': 89827,
 'moonwalk': 53151,
 'mayb': 50395,
 'want': 87812,
 'get': 33084,
 'certain': 14835,
 'insight': 41100,
 'guy': 35396,
 'thought': 81392,
 'realli': 66251,
 'cool': 18413,
```

```
'eighti': 25631,
'make': 49128,
'mind': 52047,
'whether': 89047,
'guilti': 35198,
'innoc': 41023,
'part': 60219,
'biographi': 10152,
'featur': 29070,
'film': 29583,
'rememb': 67058,
'see': 71248,
'cinema': 16169,
'origin': 58683,
'releas': 66940,
'subtl': 78171,
'messag': 51411,
'feel': 29122,
'toward': 82790,
'press': 63711,
'also': 4576,
'obviou': 57383,
'drug': 24490,
'bad': 7840,
'mkaybr': 52611,
'br': 11731,
'visual': 87252,
'impress': 40339,
'cours': 18939,
'michael': 51627,
'jackson': 42257,
'unless': 85363,
'remot': 67096,
'like': 46993,
'anyway': 5816,
'hate': 36549,
'find': 29899,
'bore': 11387,
'may': 50390,
'call': 13488,
'egotist': 25579,
'consent': 18087,
'movi': 53642,
'fan': 28510,
'would': 90379,
'say': 70154,
'made': 48815,
```

```
'true': 83551,
'nice': 55735,
'himbr': 37873,
'actual': 3091,
'bit': 10242,
'final': 29874,
'20': 1020,
'minut': 52238,
'exclud': 27566,
'smooth': 74672,
'crimin': 19496,
'sequenc': 71938,
'joe': 42960,
'pesci': 61255,
'convinc': 18372,
'psychopath': 64637,
'power': 63278,
'lord': 47904,
'dead': 20904,
'beyond': 9826,
'overheard': 59291,
'plan': 62065,
'nah': 54735,
'charact': 15111,
'rant': 65802,
'peopl': 60883,
'know': 44739,
'suppli': 78752,
'etc': 27041,
'dunno': 24840,
'musicbr': 54453,
'lot': 47971,
'thing': 81107,
'turn': 83837,
'car': 13923,
'robot': 68324,
'whole': 89226,
'speed': 75892,
'demon': 21581,
'director': 22706,
'must': 54510,
'patienc': 60447,
'saint': 69540,
'came': 13567,
'kiddi': 44209,
'usual': 86123,
'work': 90161,
```

```
'one': 58118,
'kid': 44200,
'let': 46610,
'alon': 4492,
'bunch': 12873,
'perform': 61027,
'complex': 17682,
'danc': 20447,
'scenebr': 70372,
'bottom': 11526,
'line': 47150,
'level': 46659,
'anoth': 5443,
'think': 81161,
'stay': 76851,
'away': 7477,
'tri': 83286,
'give': 33497,
'wholesom': 89238,
'iron': 41698,
'bestest': 9668,
'buddi': 12635,
'girl': 33377,
'truli': 83588,
'talent': 79585,
'ever': 27236,
'grace': 34382,
'planet': 62077,
'well': 88515,
'attent': 7086,
'gave': 32670,
'subjecthmmm': 78073,
'dont': 23699,
'differ': 22427,
'behind': 9237,
'close': 16704,
'door': 23755,
'fact': 28232,
'either': 25680,
'extrem': 28009,
'stupid': 77939,
'sickest': 73410,
'liar': 46739,
'hope': 38647,
'latter': 45862,
'classic': 16440,
'war': 87833,
```

```
'world': 90243,
'timothi': 81984,
'hine': 37956,
'entertain': 26534,
'obvious': 57384,
'goe': 33820,
'great': 34632,
'effort': 25525,
'length': 46437,
'faith': 28360,
'recreat': 66524,
'book': 11246,
'mr': 53959,
'succe': 78199,
'appreci': 6070,
'standard': 76578,
'predict': 63472,
'hollywood': 38370,
'fare': 28671,
'come': 17318,
'everi': 27276,
'year': 90994,
'eg': 25538,
'spielberg': 75981,
'version': 86695,
'tom': 82301,
'cruis': 19767,
'slightest': 74358,
'resembl': 67359,
'everyon': 27329,
'look': 47805,
'envis': 26649,
'amateur': 4718,
'critic': 19559,
'everyth': 27339,
'other': 58860,
'rate': 65940,
'import': 40311,
'baseslik': 8573,
'never': 55548,
'agre': 3747,
'enjoy': 26412,
'put': 64977,
'hg': 37568,
'novel': 56903,
'found': 31117,
'easi': 25189,
```

```
'overlook': 59323,
'perceiv': 60962,
'shortcom': 73078,
'manag': 49344,
'nichola': 55768,
'bell': 9344,
'welcom': 88500,
'investor': 41574,
'robert': 68303,
'carradin': 14163,
'primal': 63856,
'park': 60153,
'secret': 71191,
'project': 64179,
'mutat': 54542,
'anim': 5284,
'use': 86065,
'fossil': 31094,
'dna': 23315,
'jurassik': 43389,
'scientist': 70672,
'resurrect': 67515,
'natur': 55012,
'fearsom': 29053,
'predat': 63456,
'sabretooth': 69353,
'tiger': 81738,
'smilodon': 74620,
'scientif': 70669,
'ambit': 4780,
'deadli': 20932,
'howev': 39097,
'high': 37658,
'voltag': 87402,
'fenc': 29242,
'open': 58435,
'creatur': 19321,
'escap': 26893,
'begin': 9187,
'savag': 70089,
'stalk': 76540,
'prey': 63806,
'human': 39340,
'visitor': 87244,
'tourist': 82772,
'scientificmeanwhil': 70670,
'youngster': 91335,
```

```
'enter': 26524,
'restrict': 67493,
'area': 6248,
'secur': 71227,
'center': 14790,
'attack': 7053,
'pack': 59612,
'larg': 45669,
'prehistor': 63544,
'deadlier': 20933,
'bigger': 10000,
'addit': 3159,
'agent': 3654,
'staci': 76449,
'haiduk': 35643,
'mate': 50204,
'brian': 12071,
'wimmer': 89534,
'fight': 29507,
'hardli': 36318,
'carnivor': 14130,
'real': 66187,
'star': 76632,
'astound': 6895,
'terrifyingli': 80518,
'though': 81377,
'giant': 33217,
'group': 34980,
'run': 69135,
'afoul': 3497,
'furthermor': 32072,
'third': 81207,
'danger': 20499,
'slow': 74449,
'victimsbr': 86846,
'deliv': 21485,
'good': 33998,
'blood': 10656,
'gore': 34187,
'behead': 9232,
'hairrais': 35683,
'chillsful': 15753,
'scare': 70270,
'appear': 6003,
'mediocr': 50891,
'special': 75830,
'effectsth': 25508,
```

'stori': 77354,
'provid': 64424,
'excit': 27545,
'stir': 77197,
'result': 67502,
'quit': 65352,
'major': 49115,
'comput': 17741,
'gener': 32824,
'seem': 71285,
'total': 82673,
'lousi': 48044,
'middl': 51731,
'player': 62173,
'react': 66137,
'appropri': 6090,
'becom': 9040,
'foodactor': 30743,
'vigor': 87000,
'physic': 61588,
'dodg': 23419,
'beast': 8897,
'runningbound': 69163,
'leap': 46134,
'dangl': 20511,
'wall': 87711,
'ridicul': 67935,
'scene': 70367,
'small': 74542,
'realisticgori': 66210,
'violent': 87140,
'follow': 30698,
'sabretooth2002': 69354,
'jame': 42358,
'hickox': 37612,
'vanessa': 86387,
'angel': 5196,
'david': 20760,
'keith': 43954,
'john': 42994,
'rhi': 67782,
'davi': 20758,
'much': 54044,
'better': 9739,
'10000': 62,
'bc2006': 8813,
'roland': 68490,

```
'emmerich': 26055,
'steven': 77045,
'strait': 77533,
'cliff': 16609,
'curti': 20077,
'camilla': 13619,
'motion': 53504,
'pictur': 61661,
'fill': 29570,
'bloodi': 10681,
'badli': 7899,
'direct': 22669,
'georg': 32979,
'miller': 51979,
'take': 79528,
'mani': 49462,
'element': 25783,
'previou': 63796,
'australian': 7257,
'televis': 80254,
'tidal': 81707,
'wave': 88190,
'journey': 43164,
'earth': 25155,
'occasion': 57403,
'man': 49340,
'snowi': 74860,
'river': 68195,
'zeu': 91668,
'roxannerobinson': 68949,
'cruso': 19796,
'averag': 7392,
'barrel': 8478,
'assum': 6868,
'prais': 63358,
'greatest': 34646,
'opera': 58461,
'didnt': 22363,
'read': 66152,
'somewher': 75272,
'care': 13987,
'wagner': 87582,
'anyth': 5805,
'except': 27518,
'desir': 21903,
'cultur': 19968,
'represent': 67273,
```

'swansong': 79060,
'strike': 77694,
'unmitig': 85415,
'disast': 22850,
'leaden': 46091,
'score': 70734,
'match': 50190,
'tricksi': 83334,
'lugubri': 48389,
'realis': 66199,
'textbr': 80634,
'question': 65242,
'idea': 39842,
'matter': 50296,
'play': 62153,
'especi': 26948,
'shakespear': 72417,
'allow': 4372,
'anywher': 5839,
'near': 55141,
'theatr': 80769,
'studio': 77868,
'syberberg': 79258,
'fashion': 28787,
'without': 89788,
'smallest': 74550,
'justif': 43422,
'text': 80632,
'decid': 21169,
'parsif': 60214,
'bisexu': 10223,
'integr': 41239,
'titl': 82085,
'stage': 76471,
'transmut': 83045,
'kind': 44399,
'beatnik': 8930,
'babe': 7657,
'continu': 18259,
'sing': 73741,
'tenor': 80396,
'actor': 2980,
'singer': 73750,
'doubl': 23894,
'dose': 23864,
'armin': 6394,
'jordan': 43111,

'conductor': 17898,
'seen': 71302,
'face': 28182,
'heard': 36838,
'voic': 87344,
'amforta': 4869,
'monstrous': 53053,
'exposur': 27883,
'batonzilla': 8705,
'ate': 6956,
'monsalvat': 53014,
'friday': 31565,
'way': 88207,
'transcend': 83007,
'loveli': 48109,
'repres': 67271,
'scatter': 70342,
'shopworn': 73060,
'flaccid': 30195,
'crocus': 19619,
'stuck': 77833,
'illlaid': 40061,
'turf': 83812,
'expedi': 27761,
'baffl': 7967,
'sometim': 75250,
'piec': 61694,
'imperfect': 40274,
'cant': 13809,
'couldnt': 18809,
'splice': 76102,
'gurnemanz': 35352,
'mountain': 53557,
'pastur': 60399,
'lush': 48490,
'juli': 43296,
'andrew': 5139,
'sound': 75544,
'hard': 36273,
'endur': 26293,
'trumpet': 83598,
'particular': 60252,
'possess': 63015,
'aural': 7233,
'glare': 33552,
'add': 3140,
'sort': 75475,

'fatigu': 28907,
'impati': 40259,
'uninspir': 85268,
'conduct': 17896,
'paralyt': 60042,
'unfold': 85120,
'ritual': 68178,
'someon': 75211,
'review': 67670,
'mention': 51247,
'1951': 732,
'bayreuth': 8778,
'record': 66509,
'knappertsbusch': 44663,
'tempi': 80309,
'often': 57716,
'altogeth': 4638,
'lack': 45301,
'sens': 71816,
'puls': 64783,
'ebb': 25266,
'flow': 30534,
'half': 35719,
'centuri': 14812,
'orchestr': 58600,
'set': 72105,
'modern': 52709,
'still': 77125,
'superior': 78637,
'superbl': 78556,
'trashi': 83094,
'wondrous': 90011,
'unpretenti': 85512,
'80': 2110,
'exploit': 27841,
'hooray': 38625,
'precredit': 63452,
'somewhat': 75265,
'fals': 28425,
'deal': 20963,
'seriou': 72034,
'harrow': 36450,
'drama': 24158,
'need': 55266,
'fear': 29032,
'bare': 8384,
'ten': 80336,

```
'later': 45803,
'neck': 55235,
'nonsens': 56470,
'chainsaw': 14923,
'battl': 8721,
'rough': 68866,
'fistfight': 30118,
'lurid': 48484,
'dialog': 22222,
'gratuit': 34572,
'nuditi': 57067,
'bo': 10862,
'ingrid': 40931,
'two': 84092,
'orphan': 58744,
'sibl': 73389,
'unusu': 85794,
'even': 27196,
'slightli': 74361,
'pervert': 61248,
'relationship': 66913,
'imagin': 40129,
'rip': 68107,
'towel': 82794,
'cover': 19004,
'sister': 73875,
'nake': 54780,
'bodi': 10929,
'stare': 76656,
'unshaven': 85664,
'genit': 32869,
'sever': 72189,
'judg': 43251,
'dub': 24587,
'laughter': 45937,
'doesnt': 23449,
'sick': 73402,
'dude': 24651,
'fled': 30342,
'russia': 69216,
'parent': 60100,
'nasti': 54951,
'soldier': 75110,
'brutal': 12507,
'slaughter': 74223,
'mommi': 52861,
'daddi': 20270,
```

'friendli': 31600,
'smuggler': 74695,
'took': 82432,
'custodi': 20110,
'rais': 65627,
'train': 82944,
'expert': 27804,
'plot': 62302,
'lift': 46921,
'theyr': 81052,
'ultim': 84412,
'quest': 65240,
'mythic': 54666,
'incred': 40581,
'valuabl': 86306,
'white': 89128,
'fire': 29989,
'diamond': 22260,
'coincident': 17072,
'mine': 52111,
'life': 46825,
'littl': 47362,
'narr': 54905,
'structur': 77789,
'sure': 78796,
'fun': 31918,
'time': 81813,
'clue': 16802,
'who': 89202,
'beat': 8907,
'caus': 14586,
'bet': 9705,
'understood': 84926,
'less': 46551,
'whatev': 88940,
'violenc': 87121,
'magnific': 48985,
'grotesqu': 34953,
'singl': 73779,
'twist': 84066,
'pleasingli': 62252,
'retard': 67528,
'script': 70923,
'bonker': 11187,
'repair': 67196,
'suddenli': 78280,
'wont': 90021,

```
'reveal': 67620,
'reason': 66315,
'replac': 67233,
'fred': 31397,
'williamson': 89489,
'big': 9971,
'cigar': 16134,
'mouth': 53599,
'sleazi': 74271,
'black': 10323,
'finger': 29939,
'local': 47528,
'prostitut': 64349,
'princip': 63888,
'oppon': 58520,
'italian': 41947,
'chick': 15628,
'breast': 11966,
'hideou': 37634,
'accent': 2648,
'preposter': 63642,
'catchi': 14473,
'theme': 80820,
'song': 75310,
'least': 46155,
'dozen': 24090,
'throughout': 81569,
'there': 80968,
'obligatori': 57312,
'werefallinginlov': 88801,
'montag': 53055,
'load': 47499,
'attract': 7124,
'god': 33747,
'brilliant': 12167,
'experi': 27780,
'french': 31484,
'translat': 83034,
'surviv': 78896,
'uniqu': 85302,
'rest': 67460,
'none': 56292,
'got': 34276,
'pretti': 63771,
'action': 2901,
'chang': 15026,
'locat': 47538,
```

'harri': 36427,
'hurt': 39545,
'offens': 57592,
'eastwood': 25222,
'form': 30973,
'dirti': 22786,
'pat': 60401,
'hingl': 37958,
'town': 82800,
'cop': 18463,
'pool': 62770,
'45': 1632,
'could': 18802,
'short': 73066,
'cheesi': 15465,
'effect': 25472,
'soso': 75495,
'act': 2825,
'past': 60373,
'wasnt': 88019,
'background': 7757,
'around': 6458,
'evil': 27379,
'druid': 24518,
'witch': 89736,
'link': 47205,
'woman': 89920,
'migrain': 51849,
'drag': 24115,
'clearli': 16542,
'explain': 27815,
'keep': 43924,
'plod': 62296,
'christoph': 16012,
'walken': 87683,
'complet': 17665,
'senseless': 71833,
'potenti': 63197,
'tv': 83921,
'avoid': 7430,
'video': 86872,
'friend': 31588,
'hous': 39015,
'im': 40115,
'glad': 33523,
'wast': 88032,
'money': 52901,

'buy': 13199,
'1975': 839,
'capricorn': 13892,
'clip': 16661,
'rocket': 68372,
'blowup': 10768,
'relat': 66909,
'flight': 30414,
'sibrel': 73395,
'smoke': 74648,
'gun': 35271,
'astronaut': 6917,
'prepar': 63628,
'broadcast': 12267,
'edit': 25385,
'voiceov': 87354,
'instead': 41171,
'us': 86034,
'crew': 19437,
'curious': 20034,
'end': 26188,
'show': 73171,
'zaprud': 91551,
'claim': 16359,
'radiat': 65499,
'shield': 72798,
'photographi': 61547,
'lead': 46087,
'believ': 9320,
'ignor': 39962,
'ax': 7563,
'grind': 34857,
'nasa': 54929,
'american': 4829,
'scienc': 70659,
'bought': 11543,
'grossli': 34944,
'overpr': 59375,
'despit': 21949,
'name': 54799,
'adam': 3121,
'sandler': 69785,
'billi': 10084,
'bob': 10894,
'thornton': 81357,
'burt': 13013,
'young': 91311,

'funni': 31977,
'chisel': 15817,
'hammer': 35961,
'straight': 77493,
'earhol': 25105,
'tire': 82044,
'comed': 17326,
'techniqu': 80094,
'consist': 18117,
'break': 11929,
'fourth': 31156,
'talk': 79619,
'audienc': 7160,
'seemingli': 71294,
'pointless': 62528,
'hot': 38924,
'girlsbr': 33444,
'waiter': 87624,
'ship': 72863,
'success': 78208,
'comedian': 17328,
'order': 58612,
'women': 89955,
'resid': 67385,
'shamelessli': 72471,
'dicki': 22319,
'due': 24670,
'unfathom': 85093,
'opposit': 58531,
'gender': 32809,
'presum': 63733,
'lost': 47956,
'sea': 71032,
'shecker': 72641,
'he': 36733,
'rather': 65953,
'lock': 47548,
'bathroom': 8680,
'sickbr': 73403,
'perhap': 61077,
'vomit': 87418,
'worst': 90346,
'full': 31858,
'refer': 66689,
'mad': 48790,
'max': 50373,
'ii': 39979,

'wild': 89417,
'ladybug': 45357,
'clear': 16530,
'tribut': 83314,
'peter': 61274,
'lorr': 47932,
'masterpiec': 50133,
'futur': 32095,
'happen': 36184,
'armi': 6393,
'wetback': 88881,
'towelhead': 82795,
'godless': 33788,
'eastern': 25212,
'european': 27144,
'commi': 17510,
'gather': 32631,
'forc': 30832,
'south': 75606,
'border': 11380,
'gari': 32539,
'busey': 13037,
'kick': 44183,
'butt': 13151,
'laughabl': 45885,
'exampl': 27464,
'reaganera': 66180,
'fallout': 28417,
'bulletproof': 12802,
'decent': 21150,
'support': 78757,
'cast': 14359,
'head': 36734,
'jone': 43083,
'thalmu': 80661,
'rasulala': 65923,
'although': 4622,
'remak': 67024,
'far': 28650,
'comment': 17477,
'opinion': 58490,
'pure': 64894,
'comparisonbr': 17602,
'written': 90561,
'capot': 13873,
'wordsbr': 90144,
'anthoni': 5506,

```
'edward': 25424,
'eric': 26811,
'superb': 78543,
'case': 14273,
'alway': 4666,
'wonder': 89981,
'number': 57112,
'famou': 28503,
'brother': 12386,
'certainti': 14839,
'top': 82496,
'professionbr': 64114,
'recommend': 66479,
'50': 1735,
'dvd': 24945,
'shelv': 72720,
'harvey': 36489,
'light': 46930,
'candl': 13731,
'anchor': 5055,
'spallbr': 75728,
'titular': 82110,
'moros': 53339,
'tight': 81749,
'teacher': 80002,
'catharsi': 14507,
'base': 8545,
'deep': 21245,
'emot': 26066,
'unveil': 85803,
'surpris': 78854,
'spall': 75727,
'rang': 65775,
'convey': 18359,
'move': 53616,
'portray': 62970,
'mike': 51868,
'leigh': 46363,
'repertorybr': 67222,
'expect': 27739,
'school': 70565,
'bu': 12568,
'trip': 83395,
'comic': 17416,
'purpos': 64927,
'simpson': 73696,
'central': 14804,
```

```
'situat': 73926,
'visit': 87239,
'salisburi': 69625,
'cathedr': 14513,
'rhidian': 67783,
'brook': 12355,
'wellcontain': 88566,
'dramat': 24190,
'almost': 4461,
'formal': 30975,
'divid': 23244,
'actsbr': 3087,
'introduc': 41506,
'urban': 85985,
'british': 12236,
'racial': 65454,
'religi': 66983,
'divers': 23236,
'uniform': 85244,
'tell': 80273,
'privat': 63939,
'public': 64681,
'rap': 65812,
'asian': 6723,
'muslim': 54495,
'bulli': 12814,
'mean': 50735,
'individu': 40719,
'recogn': 66462,
'shameless': 72470,
'exuber': 28031,
'junior': 43363,
'social': 74956,
'pressur': 63721,
'celia': 14720,
'imri': 40386,
'warmth': 87912,
'supervisor': 78738,
'role': 68494,
'martinet': 49967,
'playbr': 62163,
'transform': 83018,
'crisi': 19531,
'remain': 67016,
'amusingli': 4986,
'oblivi': 57317,
'ben': 9415,
```

```
 'mile': 51906,
 'coupl': 18918,
 'spoon': 76205,
 'fed': 29097,
 'didact': 22344,
 'lesson': 46576,
 'toler': 82277,
 'bbc': 8794,
 'england': 26357,
 'easter': 25209,
 'america': 4819,
 'christmasbr': 15995,
 'nathali': 54978,
 'summer': 78429,
 'love': 48061,
 'key': 44087,
 'redempt': 66566,
 'movieoftheweek': 53781,
 'preach': 63397,
 'touch': 82704,
 'reach': 66133,
 'unexpect': 85068,
 'unfortun': 85138,
 'saw': 70124,
 'intens': 41271,
 'interrupt': 41422,
 'commercialsbr': 17504,
 'heavyhand': 36961,
 'pointedli': 62518,
 'road': 68250,
 'pilgrimag': 61767,
 'quiet': 65305,
 'best': 9663,
 'evoc': 27403,
 'men': 51163,
 'holidaythem': 38335,
 'submerg': 78094,
 'wouldnt': 90390,
 …}
```

[18]:
```
feature_matrix.shape
```

[18]:
```
(25000, 91908)
```