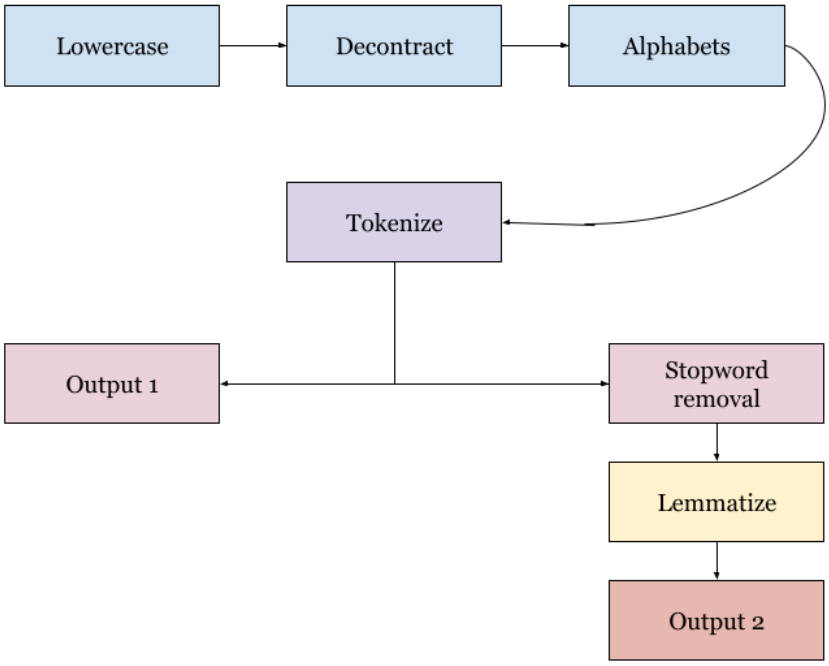
Assignment - 1 Report

S.No	Name	Roll Number
1	Anurag Saraswat	M20CS066
2	Parsa Revanth	M20CS058

Module - 1
(Statistics)

Pipeline



- **Lowercase block:** Convert the input text into lower case
- **Decontract block:** Convert the selected short forms to full forms like ‘re -> are, won't -> will not, can’t -> can not, ‘d -> would, ‘s -> is, ‘t -> not, ‘ll -> will, ‘m -> am, ‘ve -> have.
- **Alphabets block:** It only retains alphabets in the decontracted text.
- **Tokenize block:** To tokenize the words which will be further used for processing of text
- **Stopword removal block:** Remove the selected stop words from the text.
- **Lemmatize block:** To find the root of the words using the meaning and context in the text
- **Output 1 block:** It is used for calculating the word frequency to plot log-log scale Zirf’s law and for calculating word frequency for unigram, bigram and trigram
- **Output 2 block:** It is used to get an idea on the type of the dataset based on the top 50 most frequent words

Vocabulary size with word frequencies:

	Words	Frequency	POS
64	the	191516	determiner
11	and	103920	coordinating conjunction
17	is	95506	verb, 3rd person
31	a	88917	determiner
24	to	83886	to
...	...	...	...
37895	mephistophelian	1	adjective
20224	unduly	1	adverb
20226	semantic	1	adjective
20228	raros	1	noun
53304	grotesqueness	1	noun

Bi-gram:

	Words	Frequency
119	(of, the)	21013
33	(it, is)	17183
137	(in, the)	11684
34	(is, a)	11119
103	(this, show)	8800
...	...	...
373089	(former, stars)	1
373090	(stars, tainted)	1
373091	(tainted, it)	1
373093	(greatly, boston)	1
816318	(past, each)	1

Tri-gram:

	Words	Frequency
32	(it, is, a)	2361
1009	(this, is, a)	2339
117	(one, of, the)	2283
1153	(i, do, not)	2184
3663	(a, lot, of)	1849
...	...	...
805041	(show, which, imho)	1
805040	(cute, show, which)	1
805039	(reviewer, this, is)	1
805038	(another, reviewer, this)	1
2066660	(watch, the, remaining)	1

POS collections:

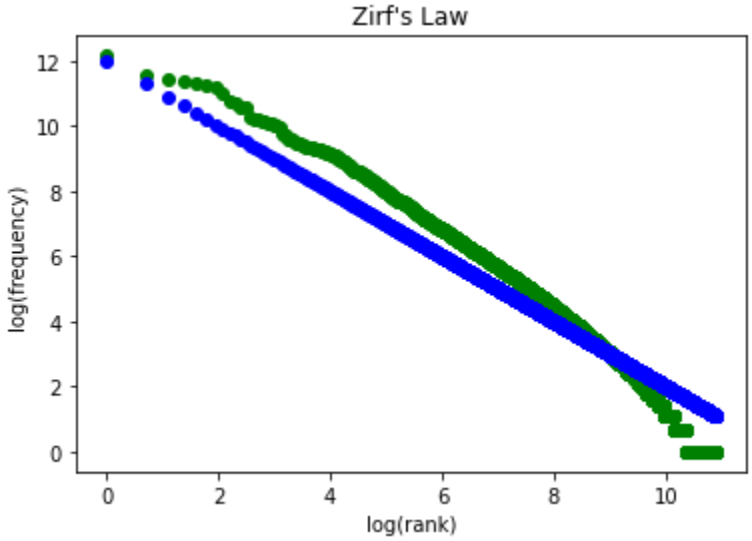
The below is the table for parts of speech collection

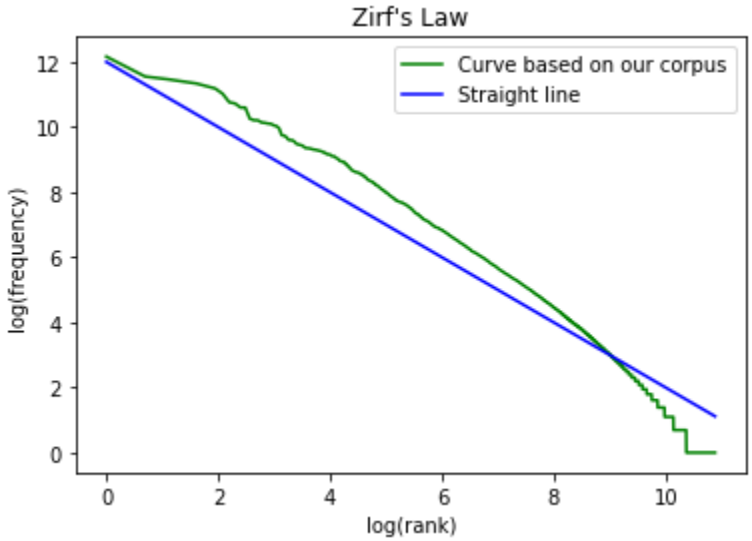
	POS	POS_Count_freq
0	noun	26586
2	adjective	10881
16	verb, present, not 3rd person	3851
8	verb, present participle	3404
7	adverb	2590
1	verb, past tense	2577
6	verb, 3rd person	1111
14	verb, past participle	1052
3	preposition	438
10	verb, base	235

18	adjective,	188
24	foreign word	102
21	proper noun	75
9	personal pronoun	30
15	particle	28
5	determiner	26
17	cardinal digit	25
13	wh-pronoun	25
12	modal	22
4	coordinating conjunction	20
22	wh-adverb	13
20	possessive pronoun	10
19	wh-determiner	8
23	existential there	3
11	to	1
25	possessive wh-pronoun	1
26	None	1
27	interjection	1
28	possessive	1

- The above table gives the parts of speech collection and its corresponding frequency in the corpus.

Verify Zipf's law:





- The above two graphs show the Zipf's law curve with logarithm of frequency of the words on the y-axis and logarithm of rank on the x-axis.
- The blue line is the straight line passing through (12,0) and (0,12), which has a slope have -1
- The green curve is the curve drawn with words in the corpus
- The above fit is the best fit for our corpus because we generated the words from the amazon instant video review dataset and it is close to the straight line even in the scatter plot also

Set of terms best describe our corpus:

The below table contains the top 50 frequency words in the corpus.

S.No	Words	Frequency
1	not	40031
2	show	30595
3	season	20127
4	one	15527
5	like	15093
6	character	13913
7	episode	12958
8	series	12903
9	good	12696
10	great	10606
11	would	10043
12	story	9949
13	love	9665
14	movie	9331
15	get	9208
16	watch	9127
17	really	9094
18	time	9048
19	well	8298
20	film	7918
21	see	7809
22	make	7126
23	much	6468

24	first	6451
25	watching	5927
26	no	5849
27	people	5562
28	even	5417
29	way	5325
30	thing	5259
31	think	5074
32	could	4967
33	new	4893
34	also	4891
35	know	4677
36	two	4615
37	go	4551
38	life	4514
39	interesting	4361
40	better	4251
41	tv	4171
42	still	4107
43	little	4086
44	many	4077
45	plot	4066
46	actor	4013
47	end	3997
48	best	3909
49	acting	3904
50	lot	3865

- Based on the above table and by seeing the following words,
    1. Show
    2. Season
    3. Character
    4. Episode
    5. Series
    6. Story
    7. Love
    8. Movie
    9. Watch
    10. Film
    11. Watching
    12. Tv
    13. Plot
    14. Actor
    15. Acting
- These words describe my corpus as a movie/tv shows review dataset.
- We arrived at this conclusion by checking the top 50 words based on the frequency of their occurrence in the corpus.

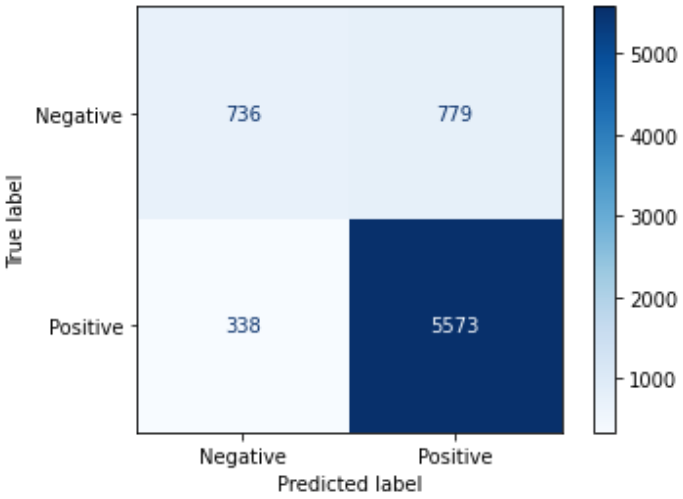
Module - 2
(Sentiment Analysis using statistical NLP)

S.NO	ML Techniques	Vector Space Models	Accuracy	F1 Score
1	Naive Bayes Model	CountVectorizer	84.96 %	0.91
2	Naive Bayes Model	TF-IDF	84.23 %	0.91
3	Naive Bayes Model	HashingVectorizer	68.72 %	0.79
4	Decision Tree	CountVectorizer	77.65 %	0.86
5	Decision Tree	TF-IDF	76.88 %	0.85
6	Decision Tree	HashingVectorizer	69.77 %	0.81
7	Logistic Regression	CountVectorizer	86.14 %	0.91
8	Logistic Regression	TF-IDF	87.06 %	0.92
9	Logistic Regression	HashingVectorizer	79.83 %	0.89

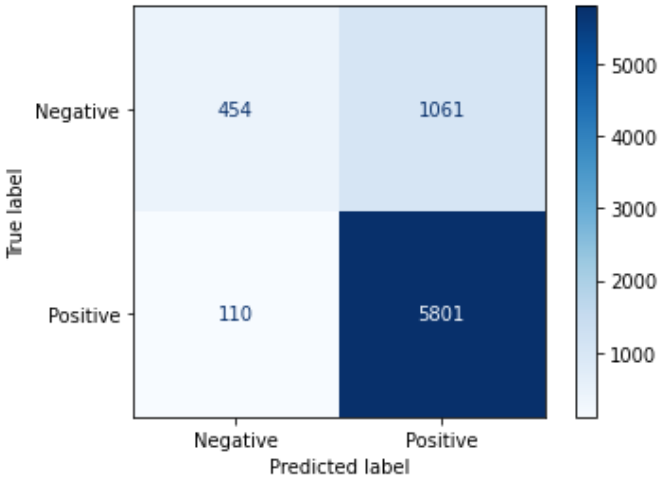
- Based on the above table we can clearly see that CountVectorizer has more accuracy and F1 score in both Naive Bayes Model and Decision Tree compared to other vector space models
- In case of Logistic Regression TF-IDF has more accuracy and F1 score compared to other vector space models
- Overall we have the best accuracy of 87.06 % and highest F1 score of 0.91 for Logistic Regression with TF-IDF vector space model
- We have sparse vector representation which will not be compatible with Gaussian Naive Bayes Model, so we used Multinomial Naive Bayes Model.

Confusion Matrix for all the possible combinations:

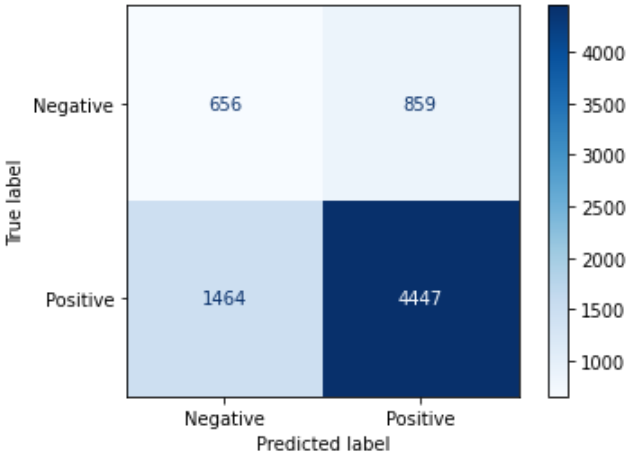
1. Naive Bayes Model with Countvectorizer:



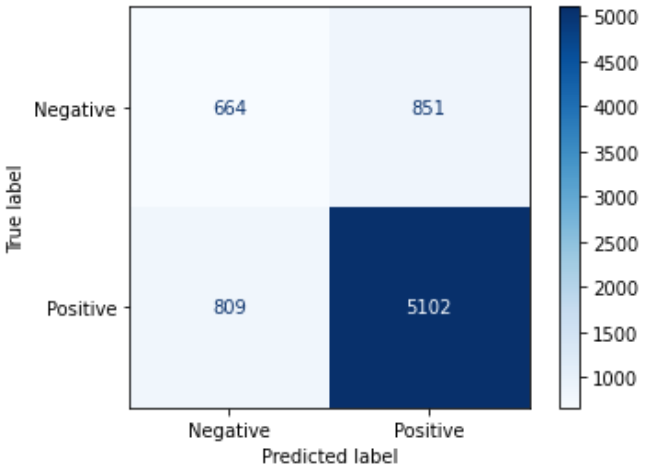
2. Naive Bayes Model with TF-IDF:



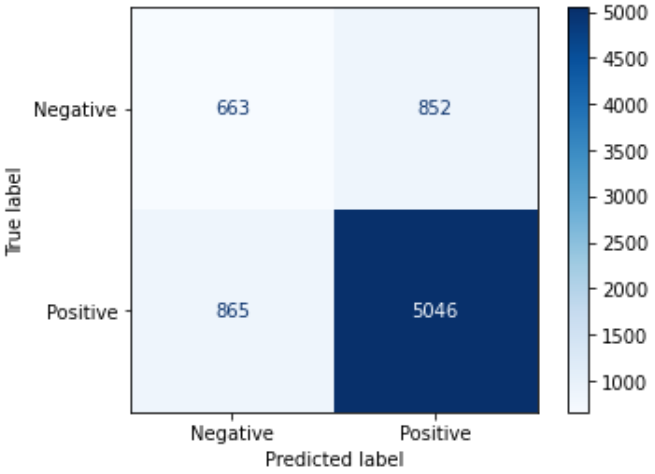
3. Naive Bayes Model with HashingVectorizer:



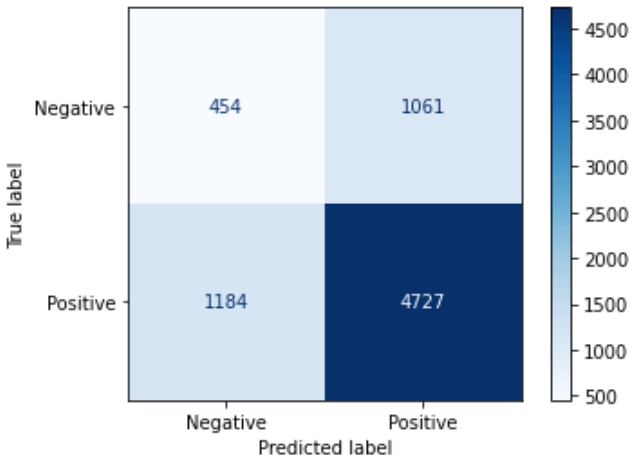
4. Decision Tree with Countvectorizer:



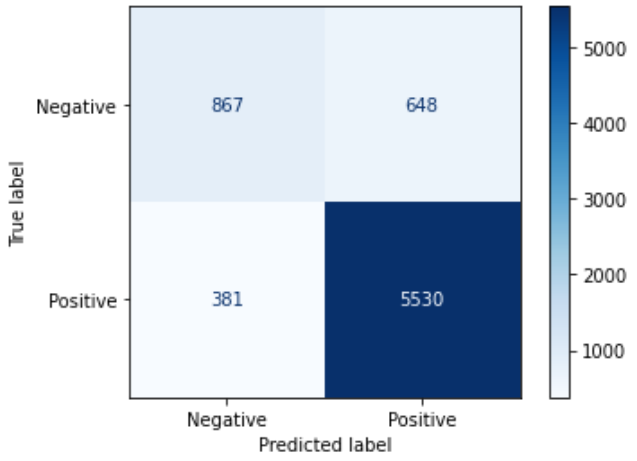
5. Decision Tree with TF-IDF:



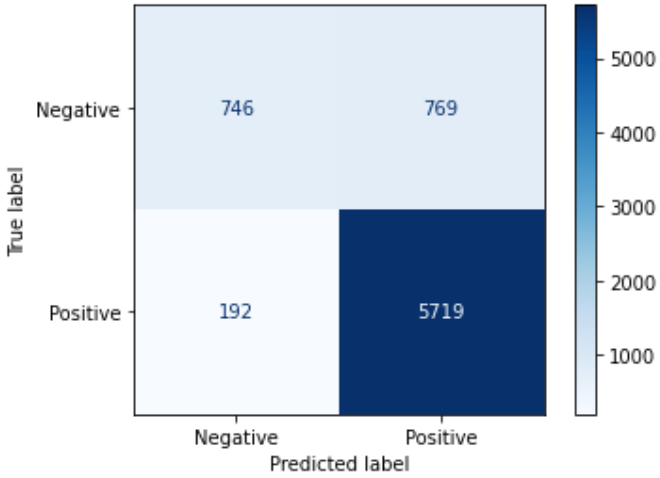
6. Decision Tree with HashingVectorizer:



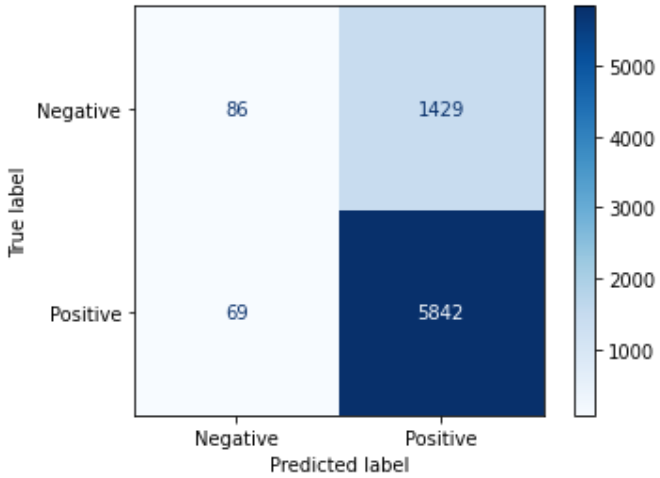
7. Logistic Regression with Countvectorizer:



8. Logistic Regression with TF-IDF:



9. Logistic Regression with HashingVectorizer:



Module - 3
(Topic analysis and topic (attribute) wise sentiment analysis)

- We used Latent Dirichlet Allocation (LDA) for the topic extraction from amazon instant video review dataset
- Topics and corresponding top 10 words in probability which are extracted from amazon instant video review dataset are
    1. Science related tv show
    2. Crime or detective film/tv show
    3. Mystery Drama film
    4. Comedy tv show
    5. Horror/Action film
    6. Drama film
    7. General topic
    8. Play (theatre) or Musical-Drama film/tv show

- Top 10 words in probability corresponding to each topic are:

    1. Science related tv show

S.No	Words	Probability
1	doctor	0.006
2	science	0.004
3	fan	0.004
4	dvd	0.004
5	tv	0.004
6	end	0.004
7	set	0.003
8	special	0.003
9	u	0.003
10	find	0.003

2. Crime or detective film/tv show

S.No	Words	Probability
1	family	0.008
2	crime	0.007
3	murder	0.005
4	life	0.005
5	case	0.005
6	detective	0.004
7	mystery	0.004
8	man	0.004
9	police	0.004
10	plot	0.004

3. Mystery Drama film

S.No	Words	Probability
1	film	0.018
2	life	0.007
3	woman	0.004
4	man	0.004
5	find	0.003
6	world	0.003
7	scene	0.003
8	work	0.003
9	never	0.003
10	look	0.003

4. Comedy tv show

S.No	Words	Probability
1	funny	0.016
2	tv	0.010
3	comedy	0.010
4	laugh	0.006
5	fun	0.006
6	actor	0.005
7	humor	0.005
8	hope	0.004
9	writing	0.004
10	family	0.004

5. Horror/Action film

S.No	Words	Probability
1	film	0.015
2	acting	0.009
3	plot	0.008
4	horror	0.006
5	scene	0.006
6	action	0.006
7	end	0.005
8	lot	0.005
9	pretty	0.004
10	little	0.004

6. Drama film

S.No	Words	Probability
1	film	0.009
2	role	0.004
3	scene	0.004
4	actor	0.004
5	play	0.003
6	city	0.003
7	life	0.003
8	real	0.003
9	game	0.003
10	guy	0.003

7. General topic

S.No	Words	Probability
1	end	0.008
2	want	0.008
3	hope	0.006
4	acting	0.006
5	got	0.006
6	never	0.006
7	forward	0.006
8	enjoyed	0.006
9	something	0.005
10	start	0.005

8. Play (theatre) or Musical-Drama film/tv show

S.No	Words	Probability
1	kid	0.010
2	drama	0.008
3	music	0.007
4	life	0.007
5	little	0.007
6	enjoy	0.006
7	actor	0.006
8	child	0.006
9	old	0.006
10	bosch	0.006

Sentences/reviews under each topic:

- Science related tv show
    1. I love the variety of comics. Great for dinner TV entertainment because of the length of each episode. Many of the featured comics have gone on to even bigger TV specials so it's great to see some of their earlier material.
    2. This show is like a runaway train ride. Every episode has poor Jack in some death defying crisis. Really enjoyed the ride.
    3. I particularly like the fact that the show (which I missed when broadcast) subtly introduces the moral dilemmas of the situations.
- Crime or detective film/tv show
    1. This series is OK, but it really lacks the substance that Criminal Minds, MI-5, CSI or Alias has. The list goes on. When is the second season of Bones and Criminal Minds coming out?
    2. "The Friends of Eddie Coyle" is a 1970's crime movie, with great dialogue, acting, and intrigue. The movie reminded me of "The Sopranos" more than classic film noir. "Eddie Coyle" is more realistic and subdued than a lot of noir. It also has a lot of criminals interacting: pulling capers, making deals, and distrusting each other, like some of the criminals in "The Sopranos". It's centered more on the interaction of the characters and the story than on action. There are several great scenes, where the dialogue and the

acting are engaging, particularly when Robert Mitchum speaks. I've never been a fan of Robert Mitchum, but here he's excellent, bringing a lot of emotion and weariness to Eddie Coyle. The story is also very good, as we see Eddie struggle to stay out of jail. The DVD is from The Criterion Collection, and looks good. There is a new commentary track from director Peter Yates. I don't think this film will appeal to a wide audience, because it's completely a crime film and nothing else, but fans of crime films might thoroughly enjoy it. Reviewed 8/23/2009 after watching on DVD.

- Mystery Drama film
    1. NCIS changed in the third season..and got strongerAt the end of season two, Kate was shot and died. Season three handles her death in the opening two parter "Kill Ari" which also introduces the audience to Ziva David (Cote de Pablo) and the new NCIS director (Lauren Holly).The change comes in many ways, Gibbs seems harder after Kate's death and it shows during the season, which builds to the two part season finale "Hiatus"...a cliffhanger to open season four. all the character grew from Abby, Ducky, Mcgee and especially Tony.Most say change is bad, for this show is GOOD and we the audience reap the benefits Bennet Pomerantz AUDIOWORLD
    2. Season 3 is much easier to use because CBS Video has splurged on the Play All function and subtitles. I started to like NCIS with the introduction of Cote de Pablo who looks like a cross between Winona Ryder and Salma Hayek. This is the season with the "homos on the train" line.
- Comedy tv show
    1. This is the best of the best comedy Stand-up. The fact that I was able to just continuously watch one comedian after another was great. I had the best laughter I have had in a long time.
    2. Watched it for Kevin Hart and only Kevin Hart! He makes me laugh. The best comedy comes from pain and Kevin does his comedy with a huge heart.
    3. It is nice to see some of the more popular comedians when they were first starting out. You can tell why some made it big and some didn't. If you like stand up comedy, I recommend giving it a try.
- Horror/Action film
    1. Non stop action with edge of your seat thrilling plots with a new turn every minute. Cast of great stars keep all events seeming like weeks but actually are in 24 hours
    2. Well if my two year old had a say she would give it ten stars. I think the show has gotten better since it's introduction. I don't mind her watching it at all and the fact that she surprised me by counting to 5 in Spanish says a lot.

3. This film was a minor letdown after reading all the five star reviews. There's nothing intrinsically wrong with the film. My theory is that after years of being out of circulation it's reputation exceeds it's actual artistic worth. Everything here is aces, though. Good direction, atmosphere and writing. The acting is superb with the stolid Robert Mitchum leading the cast. I particularly liked Steven Keats' gun dealer and Richard Jordan's duplicitous feed. I think we've been spoiled by the high quality of films in this genre with Sidney Lumet's "Prince of the City" and Martin Scorsese's "The Departed". Recommended without reservation but not as enthusiastically as some.

- Drama film
    1. This show is fantastic! I love how they wind ancient culture and Myths into futuristic events! It's a lot of fun to watch, especially if you're into ancient myths and such.
    2. An essay by David Mamet turned me on to George V. Higgins, and, completist that I am, I devoured all his novels. He is the smartest dialogue writer I have ever read. Last year's "Killing Them Softly#34; got the wit, zing, and poetry just right, and the delivery by Brad Pitt and James Gandolfini was amazing.Robert Mitchum is more realistic in his approach, so, while the dialogue is not delivered with the great timing of Pitt and Gandolfini, his physical presence is completely believable as the low-level mobster, Eddie Coyle.Otherwise, kind of plodding.
    3. if for no other reason, select this season and admire oscar in all his glory! this man's smile will break hearts around the world...
- General topic
    1. I highly recommend this series. It is a must for anyone who is yearning to watch "grown up" television. Complex characters and plots to keep one totally involved. Thank you Amazon Prime.
    2. It was a disappointment. I gave it a few episodes to get into it, but it didn't get any better.
    3. I love this show and I don't think that will ever change. I hate all of the characters and I mean that in the best way. Hilarious.
- Play (theatre) or Musical-Drama film/tv show
    1. I had big expectations because I love English TV, in particular Investigative and detective stuff but this guy is really boring. It didn't appeal to me at all.
    2. This one is a real snoozer. Don't believe anything you read or hear, it's awful. I had no idea what the title meant. Neither will you.
    3. Mysteries are interesting. The tension between Robson and the tall blond is good but not always believable. She often seemed uncomfortable.

Sentiment Distribution Table

S.No	Topics	Positive sentiment	Negative sentiment
1	Science related tv show	3225	294
2	Crime or detective film/tv show	2088	163
3	Mystery Drama film	1526	365
4	Comedy tv show	2594	338
5	Horror/Action film	2149	1033
6	Drama film	1429	243
7	General topic	6898	1255
8	Play (theatre) or Musical-Drama film/tv show	12199	1327

- The above sentiment distribution is generated based on the best combination of vector space model and ML technique that is Linear Regression with TF-IDF

Colab links

- Module 1 link: https://colab.research.google.com/drive/13iWWzsUKxDV_R8beqmUyHQQIdFaFMt8c?usp=sharing
- Module 2 link: https://colab.research.google.com/drive/1njWfMhtQv4V3fcBaw5VmXH3Uu-1CeAhc?usp=sharing
- Module 3 link: https://colab.research.google.com/drive/10G589Imv1hoHPy353mSP-wfZt8x1lTsR?usp=sharing

References

1. NLP: Extracting the main topics from your dataset using LDA
   https://towardsdatascience.com/nlp-extracting-the-main-topics-from-your-dataset-using-lda-in-minutes-21486f5aa925
2. Text Preprocessing in Natural Language Processing
   https://towardsdatascience.com/text-preprocessing-in-natural-language-processing-using-python-6113ff5decd8
3. Counting POS Tags, Frequency Distribution & Collocations in NLTK
   https://www.guru99.com/counting-pos-tags-nltk.html
4. Expanding English language contractions in Python
   https://stackoverflow.com/questions/19790188/expanding-english-language-contractions-in-python
5. HashingVectorizer (used in Module 2 as external vectorizer)
   https://github.com/scikit-learn/scikit-learn/blob/95119c13a/sklearn/feature_extraction/text.py#L513