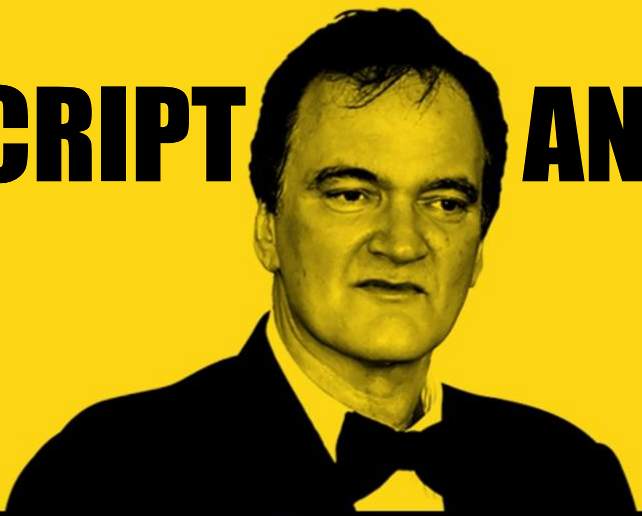
SCRIPT ANALYSIS



Ross Pingatore &
Ai Xiyao

The Data

	movie	Year	Genres	Script	Rating
0	Four Rooms	1995	comedy,drama	four rooms screenplay allison anders alexandre...	8.63
1	Inglourious Basterds	2008	action,adventure,war	jppinglourious basterds written quentin tarant...	7.44
2	From Dusk Till Down	1996	action,comedy,horror,thriller	rkfrom dusk till dawn screenplay quentin taran...	7.25
3	Natural Born killers	1995	action,romance,thriller,crime	mewsijjnatural born killers written quentin ta...	8.52
4	Django Unchained	2012	adventure,drama,western	unchained written quentin tarantino ext countr...	7.82
5	Pulp Fiction	1993	action,crime,drama,thriller	vbw pulp fiction quentin tarantino roger avary...	9.39
6	True Romance	1993	action,romance,thriller,crime	cdameddlucy laughs well enough king bout bout ...	9.88
7	Reservoir Dogs	1990	action,crime,thriller	quentin tarantino r e e r v r g october dedica...	8.89
8	Jackie Brown	1997	comedy,crime	ldajackie screenplay quentin tarantino opening...	8.27
9	Kill Bill	2003	action,comedy,crime,drama,thriller	hodsover black hear labored breathing black fr...	8.73

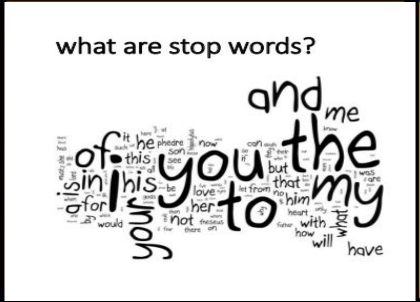
The Internet Movie Script Database (IMSDb) is the largest free online repository for movie scripts.

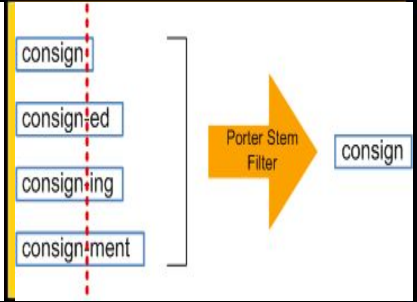
Movie scripts were chosen by producer based on the movie rating. The online repository includes the: **movie name, the year, the genre, the script and the movie rating(Rotten Tomatoes)** within the text data of each movie script.

Initially, we investigated the ten highest rated movies from Tarantino and Disney but latter expanded our analysis with Guy Ritchie.

Preprocessing

- We created two corpuses, one for the ten Disney movies and one for the ten Tarantino movies.
- We used stemming as opposed to lemmatizing to gain for familiarity with the process.
- In addition to removing NLTK stop words, we also removed a list of stop words specific to each producer. This included character names, movie names, and titles.





Sentiment Analysis

Hypothesis test 1:

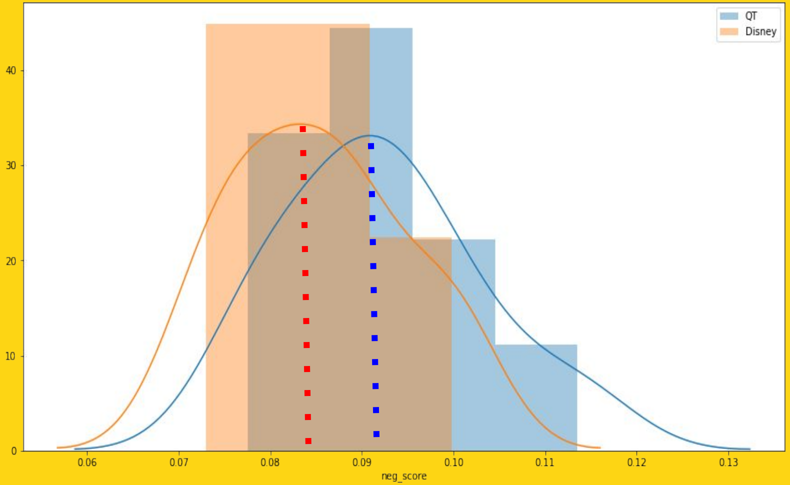
Tarantino’s movies will have a higher negative sentiment rating than that of Disney.

Null hypothesis: $\mu 1 - \mu 2 = 0$

Alternative hypothesis: $\mu 1 - \mu 2 > 0$ @ $\alpha = 0.05$

Where $\mu 1$ stands for the mean negative score for QT; $\mu 2$ stands for the mean negative for Disney.

Distribution of Negative Sentiment



Statistical Evidence

	statistic	pvalue
Levene	0.00362	0.9528
Ttest	1.4901	0.1535

Levene test: Equal variance

Two-sample t-test:

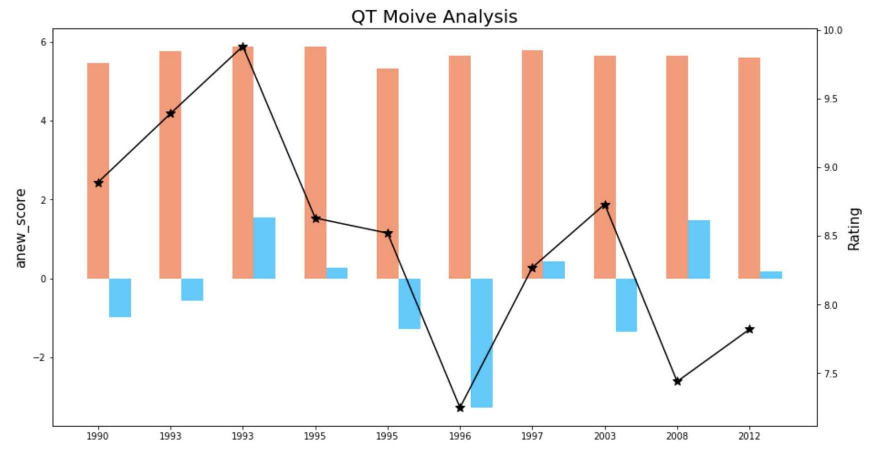
Fail to reject null hypothesis

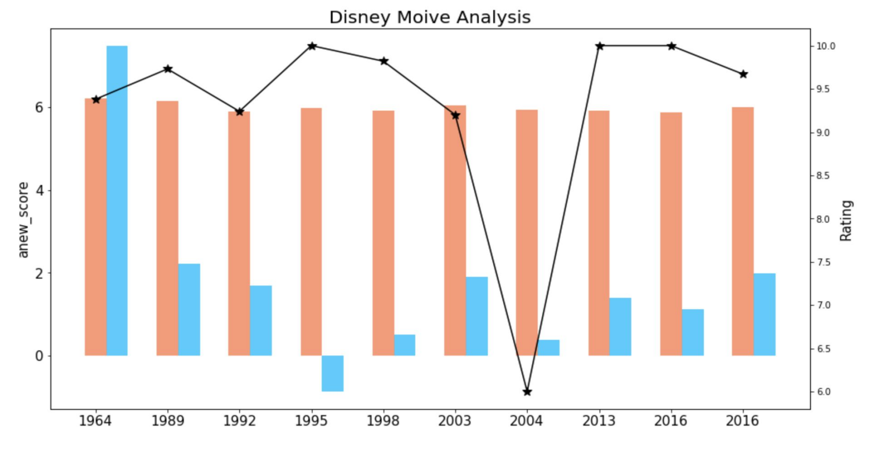
Hypothesis test 2:

Overtime, the total sentiment score rating has increased for both movie categories.

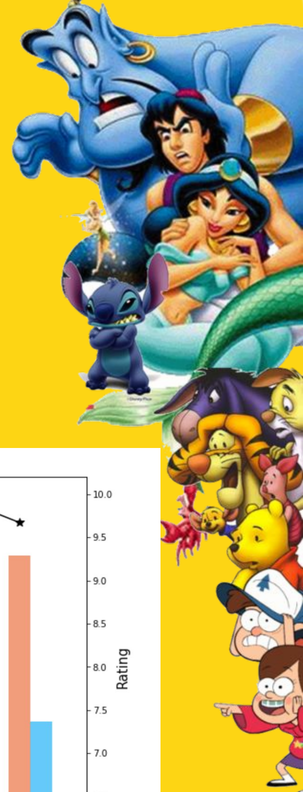
Hypothesis test 3:

The sentiment score calculated by "Harvard IV" and "ANEW" has the same trend.

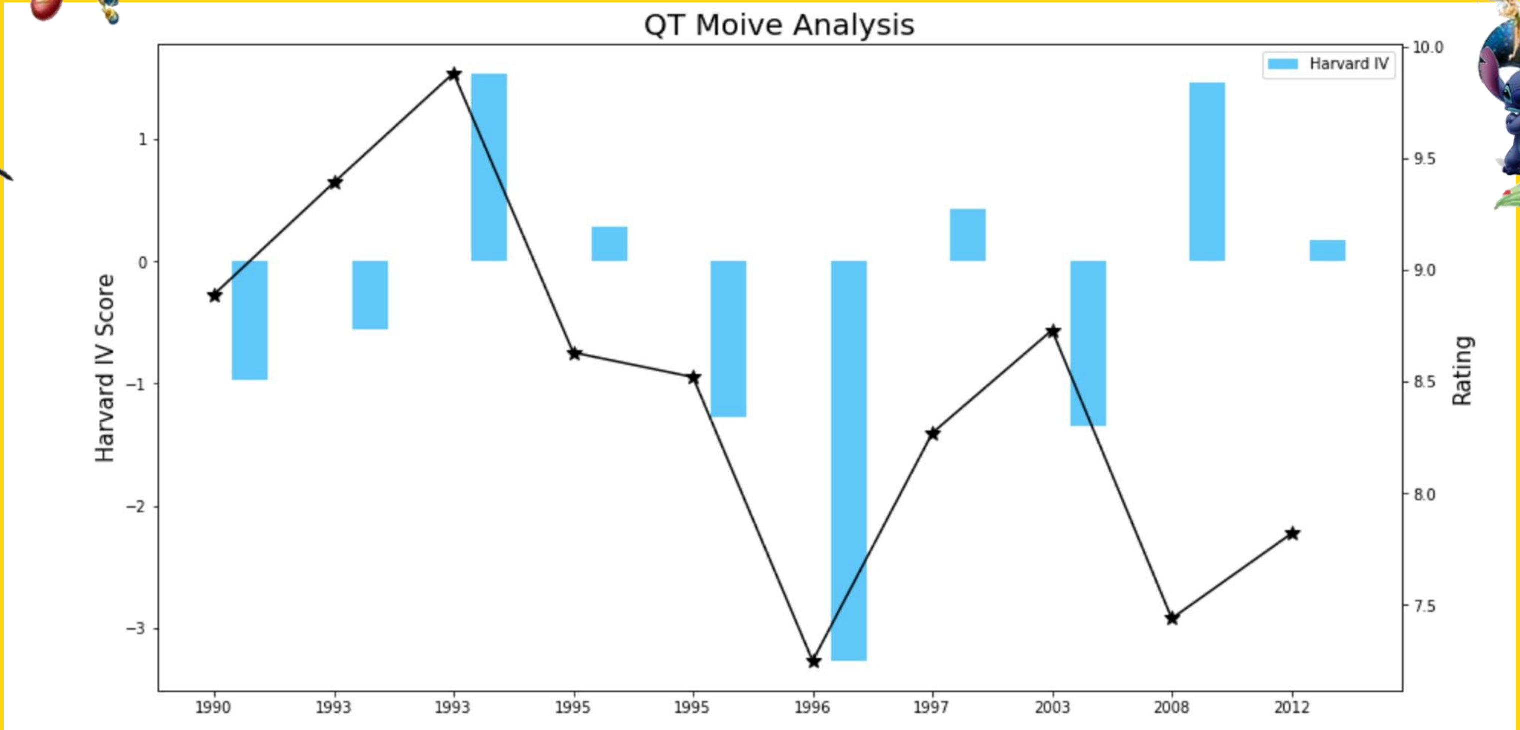




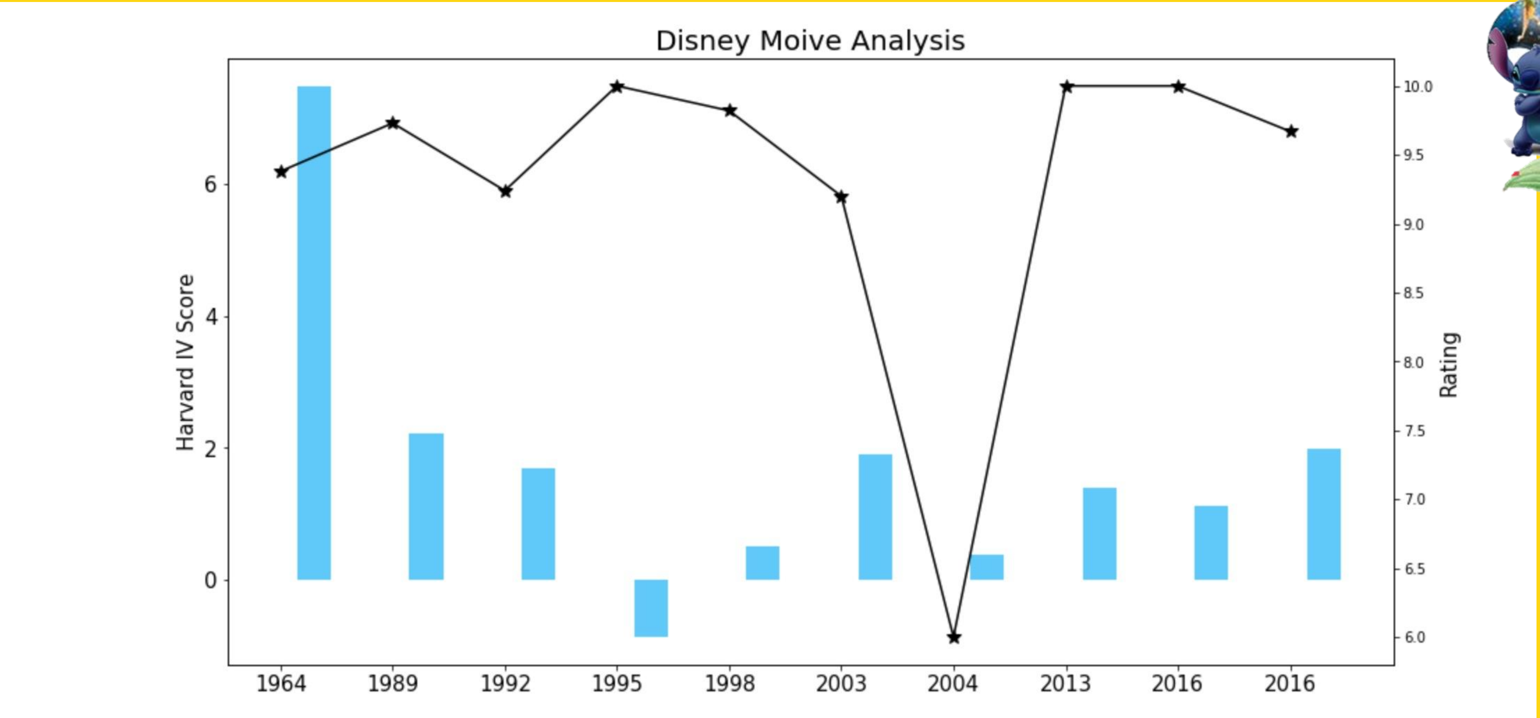
ANEW = (sum(mean(word_valance) * count) / total word count
Harvard IV = words categorized into positive and negative. Total sentiment was derived from positive - negative.



Sentiment Conclusion



Sentiment Conclusion



Topic Modeling for Tarantino

In [66]:
```
get_topics(model, 10)
```

Out[66]:

	Topic 01	Topic 02	Topic 03	Topic 04	Topic 05	Topic 06	Topic 07	Topic 08	Topic 09	Topic 10
0	like	back	mall	cu	mall	collanda	chet	theodore	virgil	wurlitzer
1	youre	like	contd	oren	contd	ltaldo	vampires	jezebel	nicholson	knox
2	back	two	bail	yuki	winston	lthicox	border	champagne	boris	deputies
3	one	one	del	back	bail	nazi	richie	witches	mustang	grace
4	know	see	ismay	two	ismay	donny	emilio	bellboy	floyd	duncan
5	get	would	robinson	black	del	colonel	vamp	altar	monty	roger
6	door	get	nigga	hanzo	robinson	hellstrom	twister	diana	worley	gayle
7	go	well	bonds	face	os	basterds	vamps	jacuzzi	wilshire	interview
8	right	right	bond	room	nigga	francesca	sex	cart	krinkle	cell
9	see	take	amo	floor	amo	hirschberg	stake	eve	sawedoff	cu

Topic Modeling for Disney

In [76]:
```
get_topics(model, 10)
```

Out[76]:

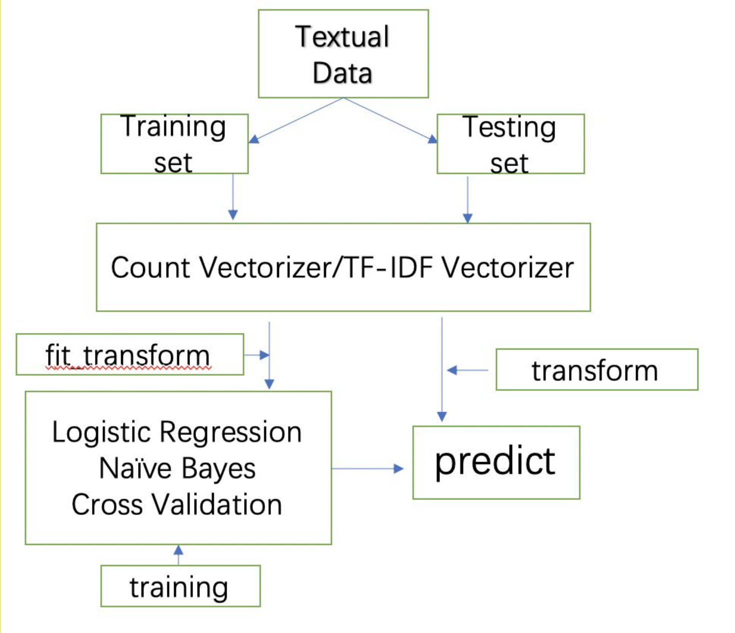
	Topic 01	Topic 02	Topic 03	Topic 04	Topic 05	Topic 06	Topic 07	Topic 08	Topic 09	Topic 10
0	contd	clawhauser	thou	shanyu	sebastion	dentist	michael	carpet	bill	shanyu
1	int	zootopia	cont	crikee	scuttle	sharkbait	chim	ali	plummer	crikee
2	phone	yeah	laurence	back	max	hey	diddle	lamp	tyler	chienpo
3	cop	elephant	gloria	chienpo	flotsam	sydney	snr	rajah	contd	khan
4	car	bunny	thy	khan	la	ha	tuppence	princess	brownies	tent
5	looks	zpd	balthasar	looks	carlotta	dad	george	turban	principal	troops
6	ext	savage	dave	face	sea	moonfish	chiminy	al	int	ping
7	hey	manchas	montague	away	vanessa	okay	medicine	back	continuous	chirp
8	little	big	car	around	err	sherman	expialidocious	zaps	scott	helmet
9	back	oh	thee	head	humans	swim	supercalifragilistic	bee	minivan	cannon

Machine Learning Analysis:

Predict:

1. If a movie (a sentence) is a Tarantino movie or a Disney movie.
(Naive Bayes & Logistic Regression)

2. If a movie (a sentence) is an Action movie or a Musical movie
(Cross Validation)

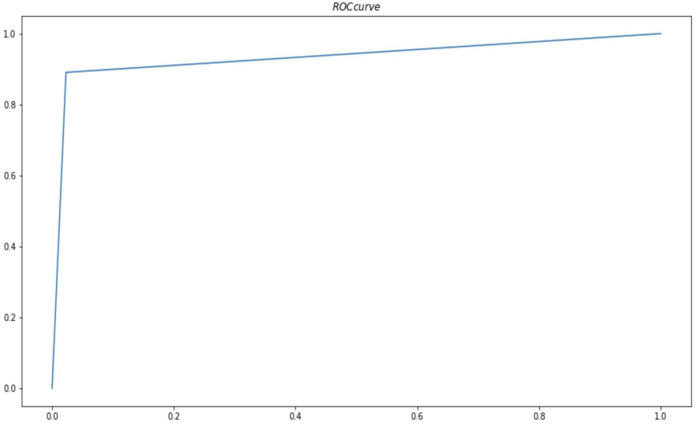


Predict:

If a movie (a sentence) is a Tarantino movie or a Disney movie.

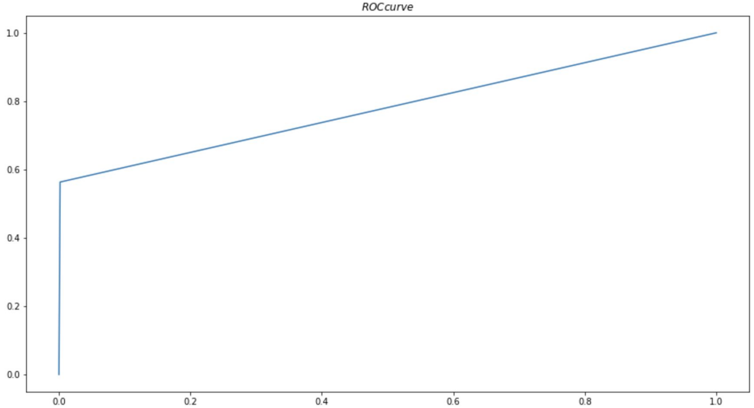
Naive Bayes
With CountVectorizer(AUC:0.93)

	precision	recall	f1-score	support
0	0.93	0.98	0.96	2318
1	0.96	0.88	0.92	1408
accuracy			0.94	3726
macro avg	0.95	0.93	0.94	3726
weighted avg	0.94	0.94	0.94	3726



Naive Bayes
With TF-IDF Vectorizer(AUC: 0.78)

	precision	recall	f1-score	support
0	0.79	1.00	0.88	2318
1	0.99	0.56	0.72	1408
accuracy			0.83	3726
macro avg	0.89	0.78	0.80	3726
weighted avg	0.87	0.83	0.82	3726

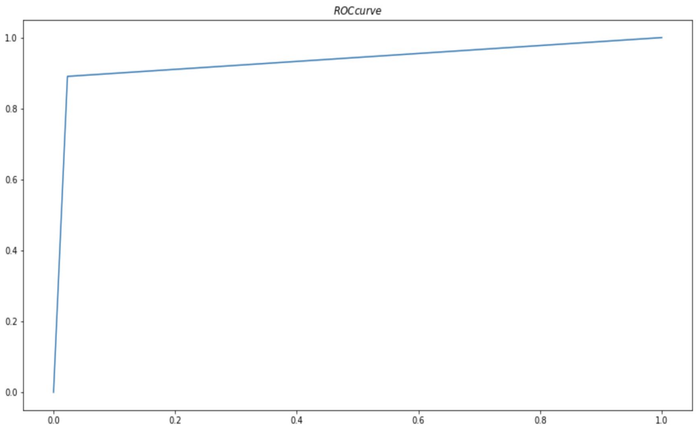


Predict:

If a movie (a sentence) is a Tarantino movie or a Disney movie.

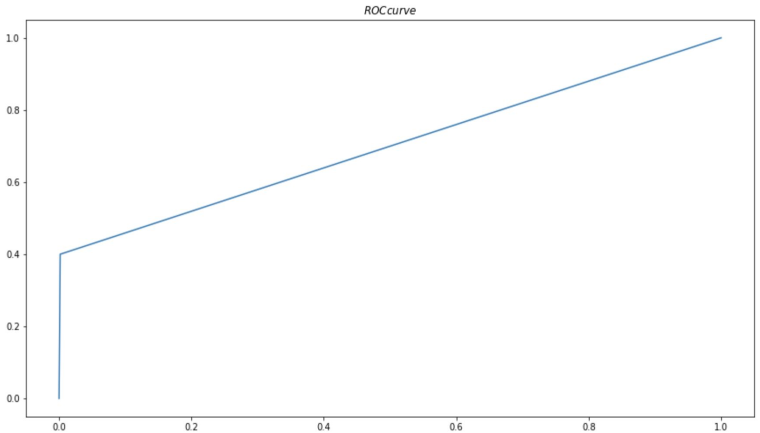
Naive Bayes
With CountVectorizer(AUC:0.93)

	precision	recall	f1-score	support
0	0.93	0.98	0.96	2318
1	0.96	0.88	0.92	1408
accuracy			0.94	3726
macro avg	0.95	0.93	0.94	3726
weighted avg	0.94	0.94	0.94	3726



Logistic Regression
With CountVectorizer(AUC:0.70)

	precision	recall	f1-score	support
0	0.73	1.00	0.84	2318
1	0.99	0.40	0.57	1408
accuracy			0.77	3726
macro avg	0.86	0.70	0.71	3726
weighted avg	0.83	0.77	0.74	3726

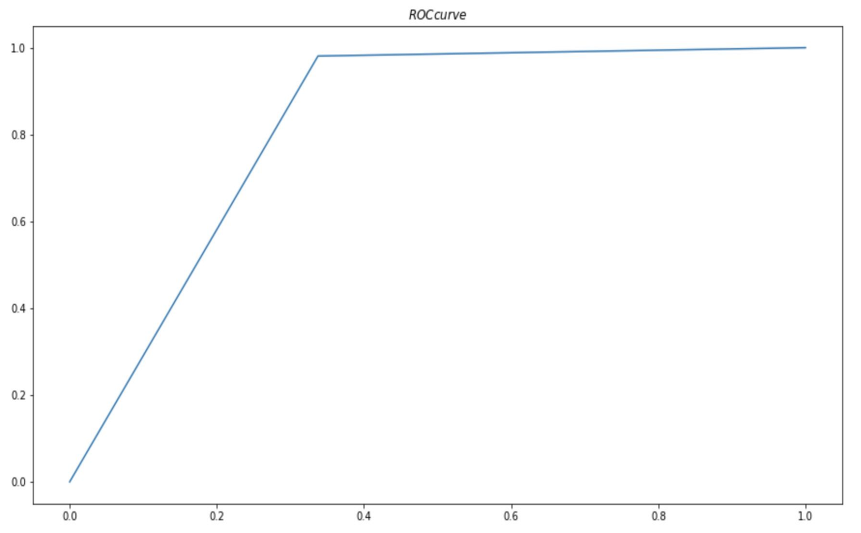


Predict:

If a movie (a sentence) is an Action movie or a Musical movie.

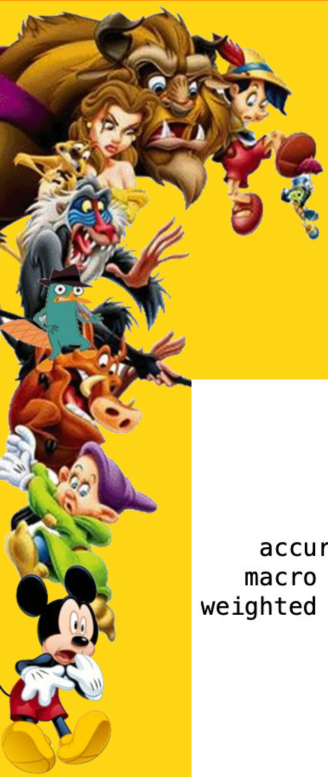
Cross Validation (k = 5)
with CountVectorizer(AUC: 0.82)

	precision	recall	f1-score	support
0	0.89	0.97	$\boxed{0.93}$	684
1	0.89	0.65	0.75	233
accuracy			0.89	917
macro avg	0.89	0.81	0.84	917
weighted avg	0.89	0.89	0.88	917



High Model Accuracy Solution

Due to our high model accuracy when comparing Disney and Tarantino movies we thought it would be useful to push our model and compared scripts that are more similar in genre to Tarantino.

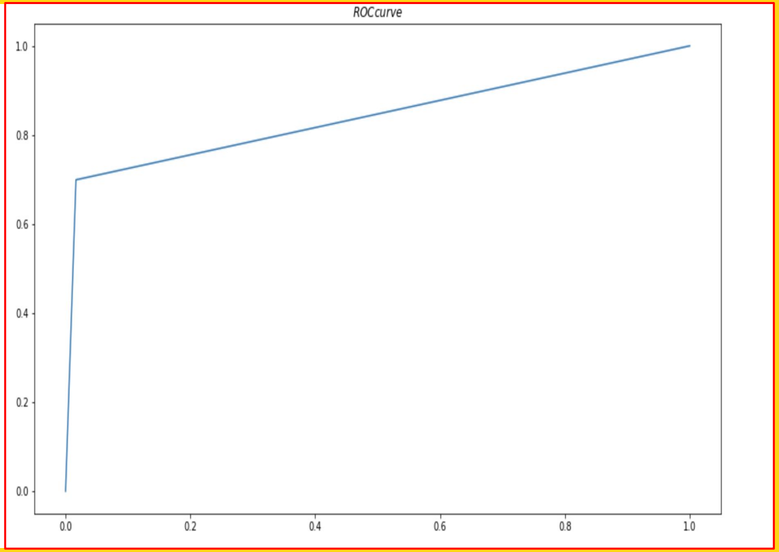


Predict:

Test if a movie (a sentence) is a action movie.
If an action movie (a sentence) belongs to QT's or Guy Ritchie's

	precision	recall	f1-score	support
0	0.58	0.92	$\boxed{0.71}$	809
1	0.36	0.06	0.11	582
accuracy			0.56	1391
macro avg	0.47	0.49	0.41	1391
weighted avg	0.48	0.56	0.46	1391

	precision	recall	f1-score	support
0	0.89	0.98	$\boxed{0.94}$	661
1	0.94	0.70	0.80	256
accuracy			0.90	917
macro avg	0.92	0.84	0.87	917
weighted avg	0.91	0.90	0.90	917



Cross Validation (k = 5)
with CountVectorizer(AUC: 0.84)

Disadvantages:

- **DATA:**
  It is hard to split the data by chapters or characters.
- **ANEW:**
  Don’t have the newest word list version, “ALL.csv” is not a complete list.
- **Machine Learning:**
  Based on the result we got, it might has a overfitting problem.
  Due to “QT has a totally different style of Disney” and “Action movie has a totally different style of Musical movie”, it is fairly easy for human to detect, so it is also easy for programming language to detect.

Improvement:

- **DATA:**
  Try to figure out a better way to split the scripts by characters or chapters
- **Machine Learning:**
  Try to deal with the overfitting problem, and improve the feature extraction.
  Find more similarity scripts to train and test the model