

# Movies and Video Series recommendation using emotional information.

ISEP supervisor

Professor: Raja CHIKY

By: Rakesh NUVVULA

61088

## Table of Contents

1. Objective: .....	2
2. About datasets .....	2
2.1 Emotional Dataset:.....	2
2.2 IMDB Dataset: .....	2
3. Cleaning Dataset .....	3
3.1 Removing unwanted data .....	3
3.2 Adding IMDB ids to Emotional data .....	3
3.3 Cleaning TV-series data of Emotional dataset .....	4
4. Merging Emotional dataset and IMDB dataset.....	5
4.1 Merge with <i>name.basics.tsv</i> .....	5
4.2 Merge with <i>title.crew.tsv</i> .....	6
4.3 Merge with <i>title.crew.tsv</i> .....	6
5. Recommended System: .....	7
5.1 Recommendations Based on Emotion type.....	7
5.2 Recommendations based on Emotion and Time period.....	8
5.3 Recommendations based on Director Name:.....	9
5.4 Recommendations Based on Movie name .....	9

## 1. Objective:

Creating dataset using given Emotional dataset and IMDB dataset to building recommended system.

## 2. About datasets

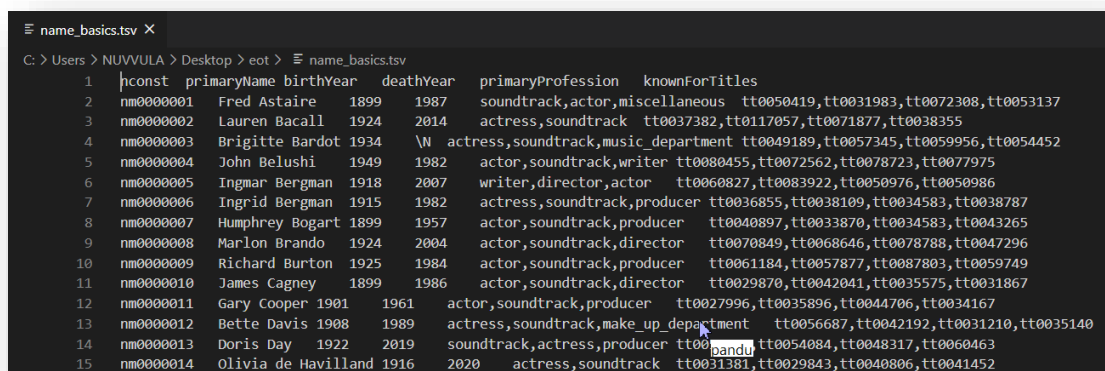
### 2.1 Emotional Dataset:

Emotional dataset contains 3659 records of data. It contains emotional scores for the movies and series episodes. It contains scores for 20 different emotions (Aggressiveness, Anger, Anticipation, Awe, Contempt, Disgust, Disapproval, Fear, Joy, Love, Negative, Optimism, Positive, Remorse, Sadness, Submission, Surprise, Trust, AFINN (-4&-5), Ero).

### 2.2 IMDB Dataset:

IMDB dataset contains data about every movie, Tv-series, episodes, Tv-shows etc, it contains all information about crew, production, writer, director, actress, rating, start and end year etc.

There were 7 files in IMDB dataset. They were name.basics.tsv, title.akas.tsv, title.basics.tsv, title.crew.tsv, title.episode.tsv, title.principals.tsv, title.ratings.tsv. Each file contains huge data. For sample 'names.basics.tsv' file is opened in virtual code studio and shown below.



id	primaryName	birthYear	deathYear	primaryProfession	knownForTitles
nm0000001	Fred Astaire	1899	1987	soundtrack,actor,miscellaneous	tt0050419,tt0031983,tt0072308,tt0053137
nm0000002	Lauren Bacall	1924	2014	actress,soundtrack	tt0037382,tt0117057,tt0071877,tt0038355
nm0000003	Brigitte Bardot	1934	\N	actress,soundtrack,music_department	tt0049189,tt0057345,tt0059956,tt0054452
nm0000004	John Belushi	1949	1982	actor,soundtrack,writer	tt0080455,tt0072562,tt0078723,tt0077975
nm0000005	Ingmar Bergman	1918	2007	writer,director,actor	tt0060827,tt0083922,tt0050976,tt0050986
nm0000006	Ingrid Bergman	1915	1982	actress,soundtrack,producer	tt0036855,tt0038109,tt0034583,tt0038787
nm0000007	Humphrey Bogart	1899	1957	actor,soundtrack,producer	tt0040897,tt0033870,tt0034583,tt0043265
nm0000008	Marlon Brando	1924	2004	actor,soundtrack,director	tt0070849,tt0068646,tt0078788,tt0047296
nm0000009	Richard Burton	1925	1984	actor,soundtrack,producer	tt0061184,tt0057877,tt0087803,tt0059749
nm0000010	James Cagney	1899	1986	actor,soundtrack,director	tt0029870,tt0042041,tt0035575,tt0031867
nm0000011	Gary Cooper	1901	1961	actor,soundtrack,producer	tt0027996,tt0035896,tt0044706,tt0034167
nm0000012	Bette Davis	1908	1989	actress,soundtrack,make_up_department	tt0056687,tt0042192,tt0031210,tt0035140
nm0000013	Doris Day	1922	2019	soundtrack,actress,producer	tt0054084,tt0048317,tt0060463
nm0000014	Olivia de Havilland	1916	2020	actress,soundtrack	tt0031381,tt0029843,tt0040806,tt0041452

Figure 1: Sample IMDB Data

For our project we were using 3 files from IMDB

1. name.basics.tsv for IMDB id(tconst), titles and titleType
2. title.crew.tsv for directors Id(nconst)
3. title.crew.tsv for directors' names.

### 3. Cleaning Dataset

#### 3.1 Removing unwanted data.

The emotional dataset contains many unwanted columns which were not required for the project, so they were removed manually. The removed data columns were as below LoveA, OptimismA, AggressivenessA, ContemptA, SubmissionA, AweA, DissapprovalA, RemorseA, erowords and nrc words.

#### 3.2 Adding IMDB ids to Emotional data.

Titles of the emotional data movie and tv-series are not same as titles in IMBD. Emotional data titles contain -, .srt for every movie and some titles are separated by “.”, “white space” as shown in the below image.

0	-12_Angry_Men_1957.en.srt	movie name seperated with _ and followed bt year
1	-1954. Seven Samurai.srt	year followed by moviename and seperated by space
2	-1984_Once_Upon_a_Time_in_America-Il_etait_une_fois_en_Amerique_ext_HD_VOSTI	year followed by movie name in english and french. seperated by _ and some extra text
3	-2001.A.Space.Odyssey.1968.REMASTERED.720p.BluRay.X264-AMIABLE.srt	
4	-3 Idiots 2009 Hindi 1080p Blu-Ray x264 DD 5.1 MSubs-HDSector.eng.srt	
5	-A Clockwork Orange (1971) 1080p H.264 Multi (moviesbyrizzo).srt	

*Figure 2* About Emotional data titles

So, titles were not used to merge two datasets and there is no other way to merge. So IMDB ID's (tconst) are assigned to emotional data titles manually.

	IMBD-tconst	0	Score	Normalise	Year
0	tt12389600	12 Angry Men 1957 en	8.9	87.23404	1957
1	tt0047478	1954 Seven Samurai	8.6	80.85106	1954
2	tt12409982	1984 Once Upon a Time in America Il etait une fois en Ameriq	8.3	74.46809	1984
3	tt0062622	2001 A Space Odyssey 1968 REMASTERED 720p BluRay X264	8.3	74.46809	1968
4	tt1187043	3 Idiots 2009 Hindi 1080p Blu Ray x264 DD 5 1 MSubs HDSe	8.3	74.46809	2009
5	tt0066921	A Clockwork Orange (1971) 1080p H 264 Multi (moviesbyrizzo	8.3	74.46809	1971

Figure 3 Emotional dataset with IMDB id's

### 3.3 Cleaning TV-series data of Emotional dataset

Emotional data contains data of Tv-series episodes but there is only one IMDB id for complete series and for recommended system we cannot give recommendations by episode. To overcome this problem, the emotional scores of each Tv-series episode were calculated manually

Boardwalk Em	8.6	80.8511	2010	2148.241	840.017241	50.8930586	63.306953	13.21022296	22.9929	34.914888	13.589	22.92337
Boston Med S	8.6	80.8511	2010	3164	1083	37.6941607	60.903251	4.059095106	15.8834	52.965668	23.818	7.9841509
Boston Med S	8.6	80.8511	2010	2980	1000	35.8562612	71.5881	3.297	15.5658	49.44	8.5985	13.363291
Boston Med S	8.6	80.8511	2010	3170	1110	39.8471547	62.561953	3.96036036	17.7613	44.438984	10.328	25.494355
Boston Med S	8.6	80.8511	2010	3114	1108	41.3942191	68.149225	4.959386282	12.8067	52.924188	12.934	16.317461
Boston Med S	8.6	80.8511	2010	3162	1131	41.9065255	64.043421	3.886825818	18.037	46.666667	12.671	19.460766
Boston Med S	8.6	80.8511	2010	2689	971	42.8410383	64.078342	9.054582904	15.7257	55.315045	13.775	20.238825
Boston Med S	8.6	80.8511	2010	2673	990	45.3770831	60.704819	2.22020202	16.1207	44.565657	5.7902	19.850403
Boston Med S	8.6	80.8511	2010	2603	1033	52.6215241	68.11875	6.383349468	8.92136	51.42128	7.3989	8.3706055
Boston Med S	8.6	80.8511	2010	2944.375	1053.25	42.1922458	65.018483	4.727600245	15.1028	49.717186	11.914	16.384982

Figure 4 Average of emotional scores for Tv-series

The average scores of tv-series, movies data were taken into another file to build a clean and reduced dataset.

## 4. Merging Emotional dataset and IMDB dataset

### 4.1 Merge with *name.basics.tsv*

'Titles' and 'title-type' were added and 'titles' were removed from emotional dataset in name.basics.tsv using the below code.

```
data_with_titles = pd.merge(ed,title, on ='tconst')
```

Figure 5. Data Merge.

The results were as below.

	Trust	AFINN(-4&-5)	Ero	titleType	primaryTitle
	34.706143	0.040480	10.181415	tvEpisode	12 Angry Men (1957)
	30.772156	19.278291	20.043469	movie	Seven Samurai
	41.001855	14.087598	18.451172	tvEpisode	Once Upon A Time In America (1984)
	52.073334	2.011218	4.221965	movie	2001: A Space Odyssey
	62.270554	3.627473	14.311616	movie	3 Idiots

Figure 6. emotional data with IMDB titles

#### 4.2 Merge with *title.crew.tsv*

Using tconst data was merged in emotional data with title.crew.tsv to get director id's (nconst) for respective movies and tv series.

titleType	primaryTitle	nconst
tvEpisode	12 Angry Men (1957)	nm6020528
movie	Seven Samurai	nm0000041
tvEpisode	Once Upon A Time In America (1984)	nm6020528

*Figure 7. emotional data with nconst*

#### 4.3 Merge with *title.crew.tsv*

Merging emotional data with title.crew.tsv on nconst to get director names for respective movies and tv-series.

Trust	AFINN(-4&-5)	Ero	directors
34.706143	0.040480	10.181415	Rob Boor
30.772156	19.278291	20.043469	Akira Kurosawa
41.001855	14.087598	18.451172	Rob Boor

*Figure 8. emotional data with director names*

## 5. Recommended System:

Considering the dataset, 4 different types of recommended systems were built as below.

1. Recommendations Based on Emotion type
2. Recommendations based on Emotion and Time period
3. Recommendations based on Director Name.
4. Recommendations Based on Movie name.

### 5.1 Recommendations Based on Emotion type

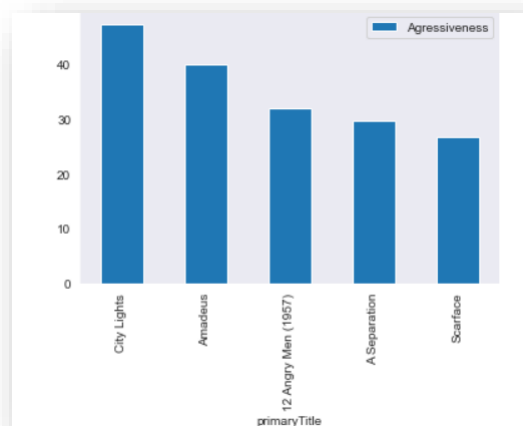
The aim of this recommendation type is movies, tv-series suggestions based on emotion and count.

```
sort_by_emotion("Agressiveness", 5)
```

	Score	primaryTitle	titleType	Agressiveness	director
23	8.5	City Lights	movie	47.370690	['Charles Chaplin']
10	8.3	Amadeus	movie	40.008172	['Milos Forman']
0	8.9	12 Angry Men (1957)	tvEpisode	32.022145	['Rob Boor']
6	8.2	A Separation	short	29.772674	['Yalan Hu']
68	8.2	Scarface	movie	26.732432	['Brian De Palma']

*Figure 9. Recommendations Based on Emotion type*

All the emotions were captured in a graphical representation as below, which displays 5 movies, tv-series and it shows 'Agressiveness' as the main emotion for the taken data.



*Figure 10. Bar graph for Recommendations Based on Emotion type*



## 5.2 Recommendations based on Emotion and Time period

The aim of this recommendation type is movies, tv-series suggestions based on emotion, time period (like 2010 - 2015) and count.

```
sort_by_year_emotion("Anger", 1985,1990,5)
```

	Score	Year	primaryTitle	titleType	Anger \
36	8.3	1987	Full Metal Jacket	movie	33.475033
39	8.6	1990	Goodfellas	movie	31.188208
40	8.4	1988	Grave of the Fireflies	movie	31.141707
21	8.4	1988	Cinema Paradiso	movie	22.406215
9	8.3	1986	Aliens	movie	20.884473

	director
36	['Stanley Kubrick']
39	['Martin Scorsese']
40	['Isao Takahata']
21	['Giuseppe Tornatore']
9	['James Cameron']

Figure 11. Recommendations based on Emotion and Time period

All the emotions were captured in a graphical representation as below, which displays 5 movies, tv-series, time period is between 1985 – 1990 and it shows 'Anger' as the main emotion for the taken data.

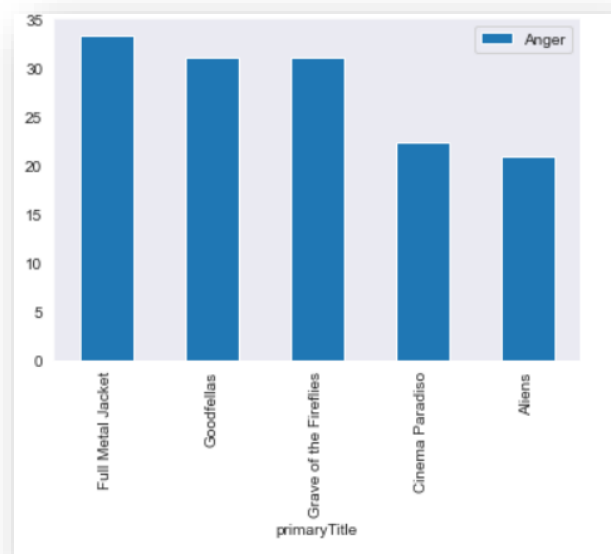


Figure 12. Bar graph for Recommendations based on Emotion and Time period

### 5.3 Recommendations based on Director Name:

This recommendation type is based on director and the sample is as below.

```
director(['\Rob Boor\'])
```

	Score	Year	primaryTitle	titleType	director
0	8.9	1957	12 Angry Men (1957)	tvEpisode	['Rob Boor']
2	8.3	1984	Once Upon A Time In America (1984)	tvEpisode	['Rob Boor']

*Figure 13. Recommendations based on Director name.*

Here we passed “Rob Boor” as director name and we got result of movies director by Rob boor.

### 5.4 Recommendations Based on Movie name

This recommendation type is based on the specific emotion which has highest percentage in the given movie.

```
by_title('Scarface')
```

	Score	Year	primaryTitle	titleType	Trust
60	8.4	1957	Paths of Glory	movie	95.731117
48	8.2	1962	Lawrence of Arabia	movie	91.630259
5	8.3	1971	A Clockwork Orange	movie	86.796997
37	8.5	2000	Gladiator	movie	78.046007
23	8.5	1931	City Lights	movie	74.755836

*Figure 14. Recommendations Based on Movie name*

As we seen in the above picture, the movie name was given, the code helps to finds the highest emotion score among all emotion scores and give results based on the emotion.