



## ASSIGNMENT SUBMISSION FORM

**This will be the first page of your assignment**

Course Name: Digital Marketing Analytics  
Assignment Title: Individual Assignment  
Submitted by:

(Student name or group names)

Student Name	PG ID
Yash Srivastava	12010060

### ISB Honour Code

- I will represent myself in a truthful manner.
- I will not fabricate or plagiarise any information with regard to the curriculum.
- I will not seek, receive or obtain an unfair advantage over other students.
- I will not be a party to any violation of the ISB Honour Code.
- I will personally uphold and abide, in theory and practice, the values, purpose and rules of the ISB Honour Code.
- I will report all violations of the ISB Honour Code by members of the ISB community.
- I will respect the rights and property of all in the ISB community.
- I will abide by all the rules and regulations that are prescribed by ISB.

**Note:** Lack of awareness of the ISB Honour Code is never an excuse for a violation. Please go through the Honour Code in the student handbook, understand it completely. Please also pay attention to the following points:

- Please do not share your assignment with your fellow students under any circumstances if the Honour Code scheme prohibits it. The HCC considers both parties to be guilty of an Honour Code violation in such circumstances.
- If the assignment allows you to refer to external sources, please make sure that you cite all your sources. Any material that is taken verbatim from an external source (website, news article etc.) must be in quotations. A much better practice is to paraphrase the source material (it still must be cited).

*(Please start writing your assignment below)*

---

## Digital Media Analytics

By Yash Srivastava

PGID: 12010060

**Data Collection:** I have extracted the movies which had either the highest grossing on box office or were the lowest.

### Movies Extracted:

Avatar (2009)  
Avengers: Endgame (2019)  
Avengers: Infinity War (2018)  
Conan the Barbarian (2011)  
Disaster Movie (2008)  
Dragonball Evolution (2009)  
Fantastic Beasts: The Crimes of Grindelwald (2018)  
Frozen (I) (2013)  
Furious 7 (2015)  
Gods of Egypt (2016)  
Green Zone (2010)  
Incredibles 2 (2018)  
Jurassic World (2015)  
Ram Gopal Varma Ki Aag (2007)  
Star Wars: Episode VII - The Force Awakens (2015)  
The Adventures of Pluto Nash (2002)  
The Lion King (2019)  
Thugs of Hindostan (2018)  
Timeline (2003)  
Titanic (1997)

### Snapshot of Data:

	Movie_Title	Rating	Review_Title	Reviews
0	Avengers: Endgame (2019)	[No rating, 10/10, 7/10, No rating, 6/10, 8/10...	[Plot holes and other blockbuster weaknesses, ...	[Much like I did for Infinity War, I was caught...
1	Avatar (2009)	[9/10, 6/10, No rating, 7/10, 8/10, 9/10, No r...	[Gorgeous and 100% otherworldly--but the story...	[As of today, there are 2675 reviews for "Avat...
2	Titanic (1997)	[9/10, 8/10, 10/10, No rating, 5/10, 9/10, 9/1...	[My review is this film's 2400th....so what mo...	[“Titanic” won a bazillion Oscars and is consi...
3	Star Wars: Episode VII - The Force Awakens (2015)	[10/10, 9/10, No rating, 5/10, 8/10, 7/10, No ...	[A slightly more adult Star Wars that shows Lu...	[I have never been a huge fan of the Star Wars...
4	Avengers: Infinity War (2018)	[9/10, 9/10, No rating, 9/10, No rating, 10/10...	[Extravagant clash of the titans, The Marvel p...	[Have found myself liking or loving a lot of M...

**Keyword Extraction:** For Keyword extraction I have used Keybert package. So for each movie the code has generated 5 keywords for low diversity score and 5 keywords for a high diversity score based on cosine similarity between candidates and the documents. The

assumption made is that the most similar candidates to the document are good keywords or phrases. I have also added generated bi-grams and tri-grams for both diversity scores.

**Maximal Marginal Relevance** – This is used for diversifying the result as written above. It tries to minimize redundancy and maximize diversity of results in text summarization.

	Movie	Ngram	Low_diversity	High_diversity
0	Avatar (2009)	(1, 1)	[(blockbuster, 0.4033), (avatar, 0.3443), (ji [(blockbuster, 0.4033), (element, -0.0719), (average, 0.121), (blatantly, -0.0071), (enraptured, 0.1273)]	
1	Avatar (2009)	(1, 2)	[(reviews avatar, 0.5623), (jurassic park, 0. [(reviews avatar, 0.5623), (create facsimiles, -0.1042), (visit theater, 0.256), (spy betray, -0.0667), (unreservedly, 0.0	
3	Avatar (2009)	(2, 2)	[(reviews avatar, 0.5623), (jurassic park, 0. [(reviews avatar, 0.5623), (create facsimiles, -0.1042), (visit theater, 0.256), (spy betray, -0.0667), (population differ	
4	Avengers: Endgame (2019)	(1, 1)	[(avengers, 0.3995), (summary, 0.3408), (s [(avengers, 0.3995), (celluloid, -0.13), (comprehensively, 0.2181), (plotting, 0.2678), (average, 0.055)]	
5	Avengers: Endgame (2019)	(1, 2)	[(summaries avengers, 0.5462), (infinity w [(summaries avengers, 0.5462), (celluloid, -0.13), (criticisms treating, 0.1658), (values charts, 0.136), (defining corne	
7	Avengers: Endgame (2019)	(2, 2)	[(summaries avengers, 0.5462), (infinity w [(summaries avengers, 0.5462), (underplayed organic, 0.0259), (values charts, 0.136), (possibly hoped, 0.1053), (clo	
8	Avengers: Infinity War (2018)	(1, 1)	[(marvel, 0.561), (reviews, 0.3316), (ultror [(marvel, 0.561), (excel, -0.0611), (saddened, 0.0472), (risks, 0.0617), (antiseptic, 0.0535)]	

Keeping a low score will give keywords which have similar cosine similarities hence a low diversity.

A high score will give keywords which have very different cosine scores hence a high diversity.

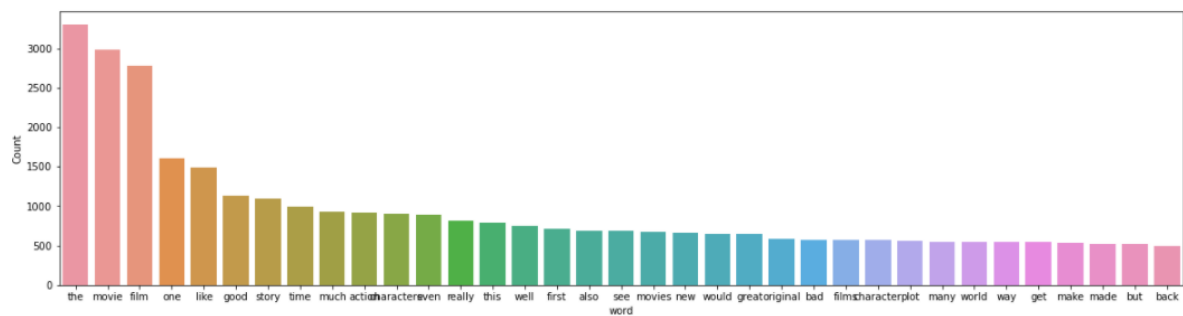
## Sentiment Analysis:

	neg	neu	pos	compound
0	0.104	0.663	0.234	1.0000
1	0.127	0.593	0.280	1.0000
2	0.152	0.591	0.257	1.0000
3	0.182	0.601	0.217	0.9999
4	0.232	0.559	0.210	-1.0000
5	0.184	0.597	0.219	1.0000
6	0.124	0.624	0.252	1.0000
7	0.099	0.600	0.301	1.0000
8	0.168	0.613	0.220	1.0000
9	0.160	0.589	0.251	1.0000
10	0.184	0.620	0.197	0.9985
11	0.108	0.594	0.298	1.0000
12	0.120	0.622	0.257	1.0000
13	0.182	0.604	0.213	0.9997
14	0.152	0.583	0.265	1.0000
15	0.155	0.591	0.254	1.0000
16	0.115	0.609	0.275	1.0000
17	0.149	0.598	0.253	1.0000
18	0.165	0.627	0.208	1.0000
19	0.133	0.583	0.284	1.0000

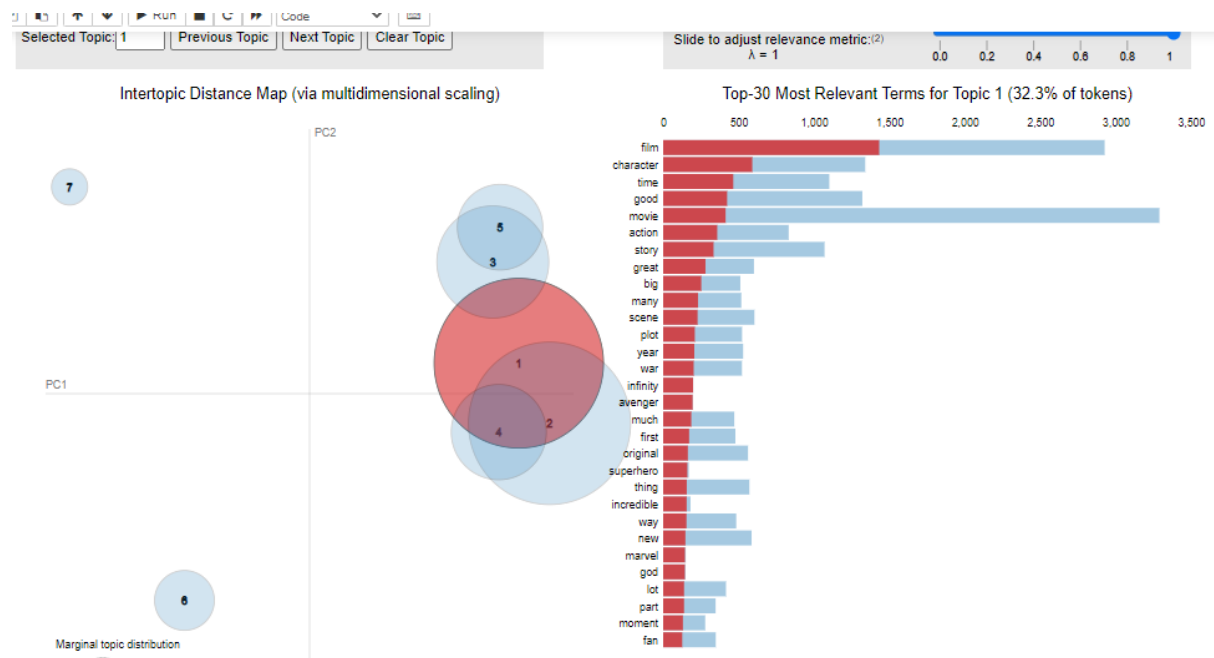
From the sentiment scores we can the overall sentiment is mostly neutral, followed by positive and then negative.

## LDA

Firstly I have checked the most occurring words:



From LDA we have made 7 topics:



As we can see Topic 1, 4, 2 have lot of common words.

Topic 1 & 3 and 3 & 5 has overlaps.

Topic 6 and 7 are very different and has no overlaps from any other.

Topics 1, 4, 2 - These topics mostly surround around superhero action movies which are really good. Most occurring terms for these are: action, war, superhero, good, great.

Topics 3 and 5: They center on good fantasy movies. These have most occurring terms like: fantasy, prequel, good and great.

Topic 6 and 7: These topics are usually for dramas like Titanic and some action movies which were not so okay.

### Associated Keywords for top 10 Keywords:

**Keywords:** The associated keywords are taken from Keywordtools.io

Keyword	Associated Keyword											
blockbuster	blockbuster movie, blockbuster meaning in hindi, blockbuster vs netflix, blockbuster action movies											
avengers	avengers endgame, avengers infinity war, avengers wallpaper, avengers series											
overrated	overrated meaning, overrated meaning in hindi, overrated movies,											
comedies	comedy action, comedy acchi acchi, comedy app, comedy actor											
disney	disney+ hotstar, disneyland, disney+ hotstar plans, c disneynow, disneyland c											
dragonball	dragon ball z, dragon ball af, dragon ball all movies, dragon ball anime											
grindlewald	grindelwald in harry potter, grindelwald movie, grindelwald vs dumbledore											
inconsistencies	inconsistency hindi meaning, consistency hobgoblin little minds, inconsistencies-plaza,											
obnoxiousness	obnoxiousness meaning, obnoxiousness meaning in hindi,											

	obnoxiousness synonym, is obnoxiousness a word, obnoxious blog											

## References:

<https://towardsdatascience.com/topic-modeling-with-nlp-on-amazon-reviews-an-application-of-latent-dirichlet-allocation-lda-ae42a4c8b369>

<https://www.analyticsvidhya.com/blog/2018/10/mining-online-reviews-topic-modeling-lda/>

[Topic Modelling with LSA and LDA | Kaggle](#)

<https://github.com/MaartenGr/KeyBERT>

<https://towardsdatascience.com/keyword-extraction-with-bert-724efca412ea>

<https://www.analyticsvidhya.com/blog/2021/06/sentiment-analysis-using-nltk-a-practical-approach/>

<https://www.analyticsvidhya.com/blog/2021/06/vader-for-sentiment-analysis/#:~:text=VADER%20is%20a%20lexicon%20and,the%20other%20positive%20or%20negative.>