Problem 3 – Bag of words (30%)

Included in the assignment are four CSV files containing information about movie reviews:

- □ reviews.tsv:
- □ 'Counts.csv':
- □ 'vocabulary.csv':
- □ 'sentiment.csv':

Problem 3 – Bag of words (30%)..1

5% credit. Use Latent Dirichlet Allocation (LDA) to generate a topic model with 10 topics. 2

Display the top-10 words for each topic as a 10x10 matrix (each column being a topic) and guess what each of the 10 underlying topics may be.........2

Figure 1: 10x10 Matrix of Top-Ten Words for Each Topic Model......................2

Discussion of findings:.. 4

5% credit. Remove the 100 most frequent words from the bag-of-words, regardless of sentiment.. 5

Then, repeat part a. Display the top-10 words for each topic and discuss what each of the 10 underlying topics may be, compared to those you obtained in part a...5

Figure 3: Matrix of Top-Ten Words for Each Topic Model Following Removal of 100 most Frequent Words..5

Discussion of findings:.. 7

10% credit. Find the top 500 movie reviews for each topic in part b1, then compute the average sentiment of those 500 reviews... 8

Does the average sentiment of each topic help you interpret those topics? Please elaborate... 8

Figure 5: Pandas Dataframe Showing Statistics of the Corpus Using the Previous Model for Classification of Documents...8

Figure 6: Average Sentiment for Each Topic..8

Figure 7. Distribution of Documents Following LDA classification of Topic......9

Discussion of findings:.. 9

10% credit. Split the corpus into two sets: X+ containing only positive reviews, and X containing only negative reviews... 10

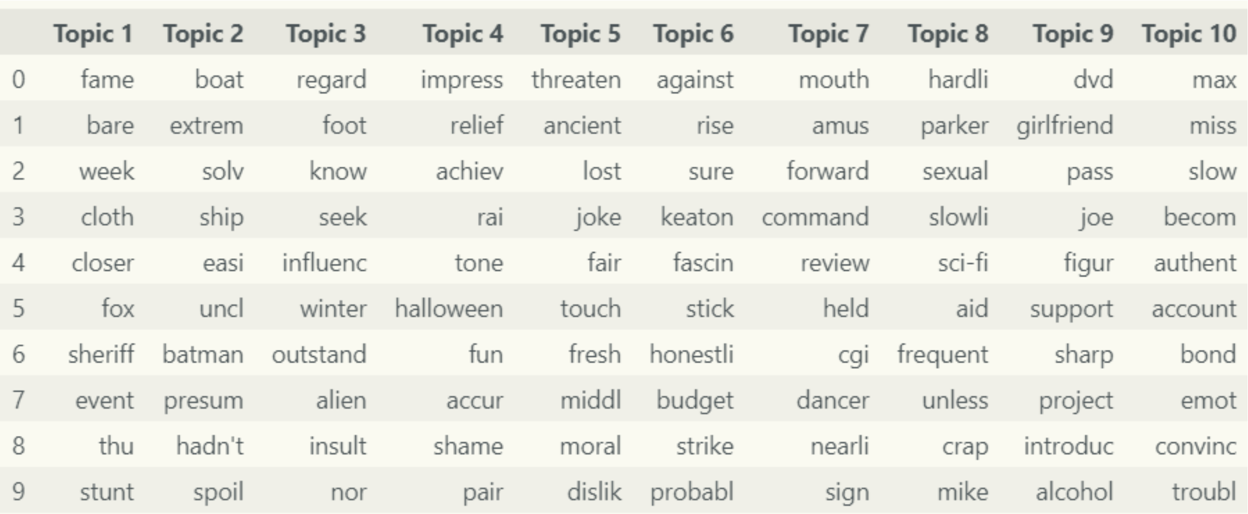
Figure 8: Top 20 and Bottom 20 Words in the Absolute Value Vector............10

Discussion of findings:..11

References:..11

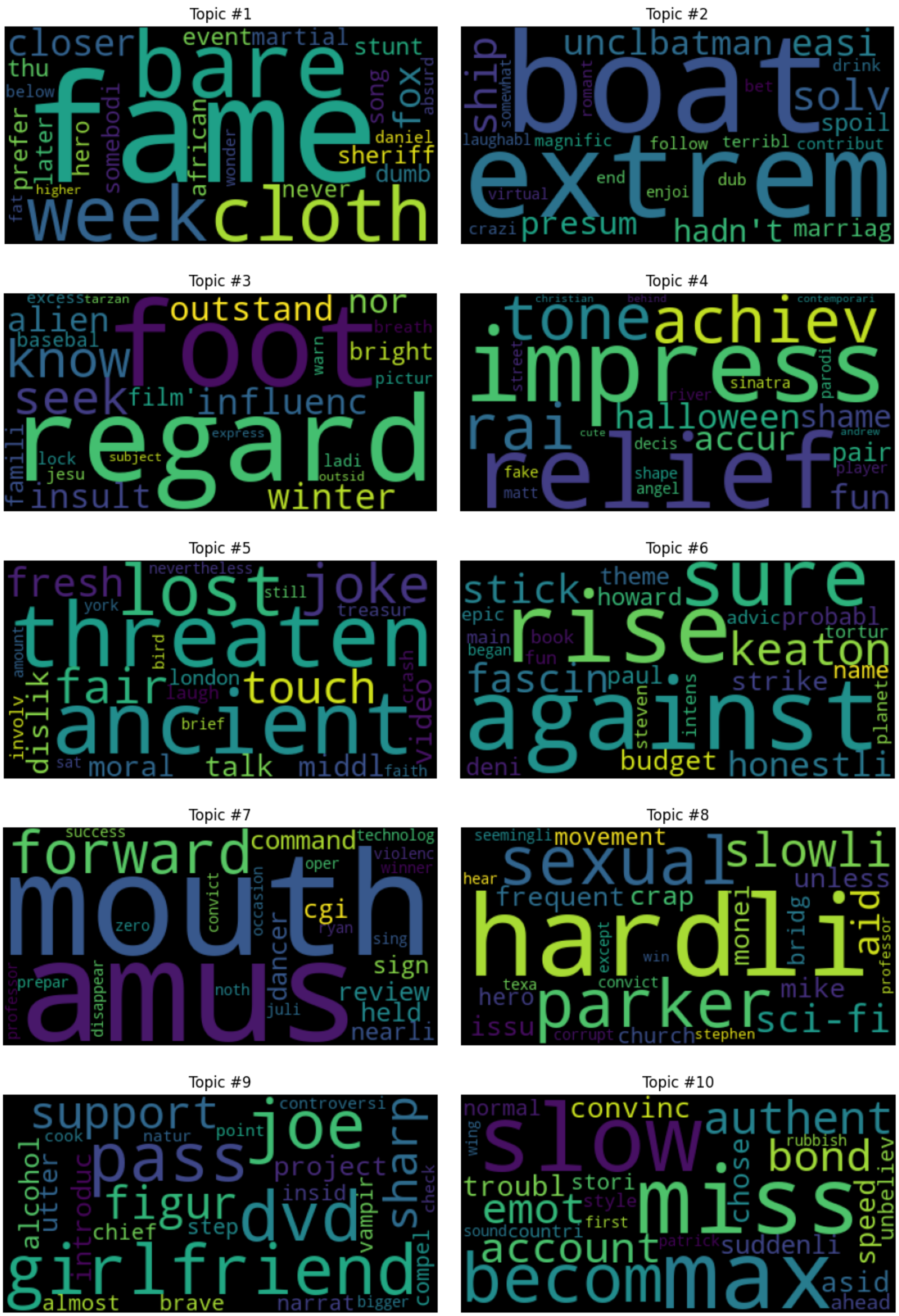
5% credit. Use Latent Dirichlet Allocation (LDA) to generate a topic model with 10 topics.

Display the top-10 words for each topic as a 10x10 matrix (each column being a topic) and guess what each of the 10 underlying topics may be.



	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
0	fame	boat	regard	impress	threaten	against	mouth	hardli	dvd	max
1	bare	extrem	foot	relief	ancient	rise	amus	parker	girlfriend	miss
2	week	solv	know	achiev	lost	sure	forward	sexual	pass	slow
3	cloth	ship	seek	rai	joke	keaton	command	slowli	joe	becom
4	closer	easi	influenc	tone	fair	fascin	review	sci-fi	figur	authent
5	fox	uncl	winter	halloween	touch	stick	held	aid	support	account
6	sheriff	batman	outstand	fun	fresh	honestli	cgi	frequent	sharp	bond
7	event	presum	alien	accur	middl	budget	dancer	unless	project	emot
8	thu	hadn't	insult	shame	moral	strike	nearli	crap	introduc	convinc
9	stunt	spoil	nor	pair	dislik	probabl	sign	mike	alcohol	troubl

Figure 1: 10x10 Matrix of Top-Ten Words for Each Topic Model


Topic #1

Topic #2

Topic #3

Topic #4

Topic #5

Topic #6

Topic #7

Topic #8

Topic #9

Topic #10

Figure 2. WordCloud Visualizations of the Top Words in Each Topic

Discussion of findings:

After recreating the sparse matrix, and preparing the data for use in the model using Latent Dirichlet Allocation, we get the above 10x10 matrix seen in Figure 1 that displays the top 10 words. For the sake of personal visualization, I created WordCloud images for each topic in order to more easily interpret the information in Figure 2.

Based on what I know about the dataset, which details the reviews of films, I can thus assume that each “topic” displayed in the results of the model details a film, with the top ten words describing said contents of a film.

Based on the following top 10 words, and their according distributions, I interpreted the topics as the following:

Topic 1:

The first topic makes me think that it is a cartoon animated film about a **fox** who seeks out **fame** and has a **week** to become a **sheriff.** The film covers the silly **stunts** that his family engage in and they experience major **events** that change their lives for the better. Maybe they work in a **clothing** business and try to get **closer** to the ugly truths in their neighborhood. Overall, I feel like this topic encapsulates animated movies about characters that start from nothing and become **heroes**. Or, perhaps this could be the *Karate Kid* or other martial arts movies, and the main character’s name is **Daniel**. He wants to be **famous**, and has a **week** to be the *Karate Kid*. Overall, this topic seems to capture a training montage sort of film.

Topic 2:

This was a very interesting topic to interpret, especially because of the word **Batman** in it. I know that Batman has a boat, so perhaps this topic encapsulates the contents of **Batman** movies, and the **extreme** stunts that Batman does. Based on my research, I think that this topic probably had a lot of reviews about the Batman movie, “The Dark Knight” because a lot of the words somewhat coincide with the plot. Maybe Batman has a frequently appearing **Uncle** in the story, and he tries to **solve** a case. Overall, the tone for this topic is somewhat grim, and perhaps captures Batman Films.

Topic 3:

I feel like this topic has a very strange set of top 10 words, but I feel like this topic covers supernatural stories about **aliens** who **seek** to **influence** the people around them as they come down to earth. This topic may have been received well, as **outstanding** appears to be a popular top word. Overall, I feel like the topic is about supernatural occurrences.

Topic 4:

I feel like this topic is probably **halloween** themed and encapsulates movies that have a **fun** theme. The words **accur** and **shame** makes me think that there are some negative connotations, and some reviews may have not thought it was accurate, or liked the **tone** of the film, but overall, I feel like the topic includes halloween themed films that are fun and lighthearted.

Topic 5:
I thought that this topic was about an **ancient** artifact that threatens the life of explorers, who get **lost** in a **fair** and have to learn their **morality** while struggling over their **dislike** of each other. Perhaps this is an Indiana Jones Topic film, as the topic appears to be a historical adventurous film about morality and has a **fresh** outlook on previous iterations of the same topic.

Topic 6:
I feel like this topic probably encapsulates movies with the actor **Keaton** within them, and often covers topics such as **budget strikes against** a big corporation, and features how the protestors **rise** up against the people who take advantage of their work. I feel like reviews generally found this topic **fascinating**, and that perhaps these involve a subset of films about money, strikes, and budgeting.

Topic 7:
I feel like this topic mainly encapsulates **amusing** films that include **cgi** aspects. Perhaps the films in this topic include **dancers** who have a **forward** attitude about life, and have a **commanding** presence on stage. Perhaps this could be a subset of films that involve **dancers** and their behind the scenes work to be the best at what they do.

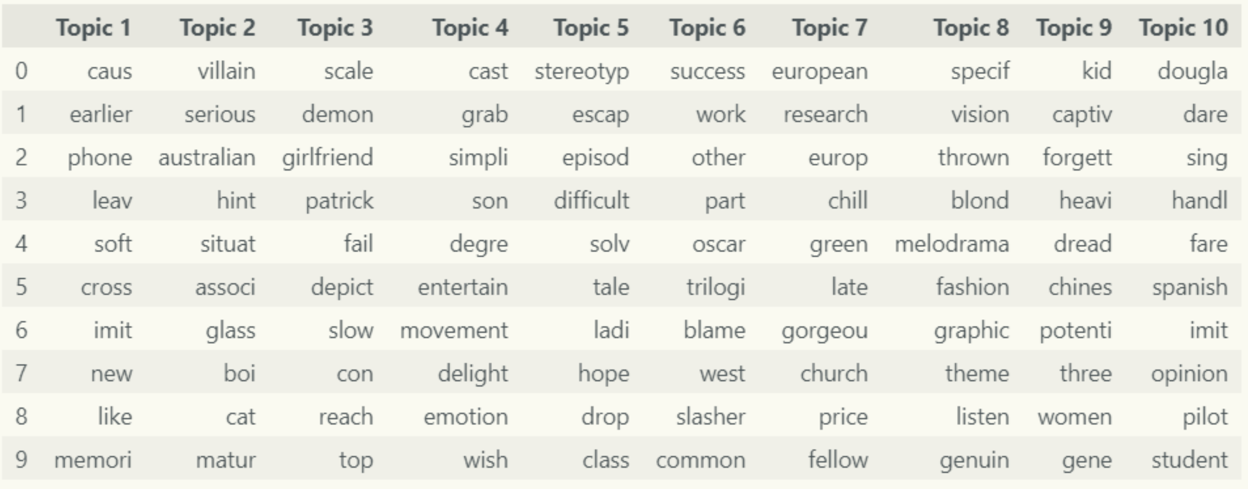
Topic 8:
I feel like this topic encapsulates films that have actors by the name of **Parker** and **Mike**, and have a general **sci-fi** theme. Perhaps the themes of these movies cover the space travels of astronauts who **frequently** seek **aid** from other planets. Maybe reviewers felt like films from this topic were **crappy** and **sexual.** Overall, I felt like this topic covered sci-fi films about providing aid to those in need.

Topic 9:
I feel like this topic encapsulates films that include either a character or an actor by the name of **Joe**, and his adventures in college. Maybe these films cover characters who aim to get a **girlfriend** by the end of the semester, and find solace in the group **project** that they work on with their crush. Perhaps they rent out a **dvd** together, get some **alcohol** and **pass** the time by getting to know more about each other. In short, maybe this topic describes college themed romance films.

Topic 10:
Based on the top ten words, I feel like this topic is about the **slow** relationship between two **accountants** who **bond** over shared **troubles** and feel **authentic emotions** for the first time. Of course, this is all speculation, but I feel like overall, this topic includes movies with lots of reflection, bonding, and emotion.

5% credit. Remove the 100 most frequent words from the bag-of-words, regardless of sentiment.

Then, repeat part a. Display the top-10 words for each topic and discuss what each of the 10 underlying topics may be, compared to those you obtained in part a.



	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
0	caus	villain	scale	cast	stereotyp	success	european	specif	kid	dougla
1	earlier	serious	demon	grab	escap	work	research	vision	captiv	dare
2	phone	australian	girlfriend	simpli	episod	other	europ	thrown	forgett	sing
3	leav	hint	patrick	son	difficult	part	chill	blond	heavi	handl
4	soft	situat	fail	degre	solv	oscar	green	melodrama	dread	fare
5	cross	associ	depict	entertain	tale	trilogi	late	fashion	chines	spanish
6	imit	glass	slow	movement	ladi	blame	gorgeou	graphic	potenti	imit
7	new	boi	con	delight	hope	west	church	theme	three	opinion
8	like	cat	reach	emotion	drop	slasher	price	listen	women	pilot
9	memori	matur	top	wish	class	common	fellow	genuin	gene	student

Figure 3: Matrix of Top-Ten Words for Each Topic Model Following Removal of 100 most Frequent Words

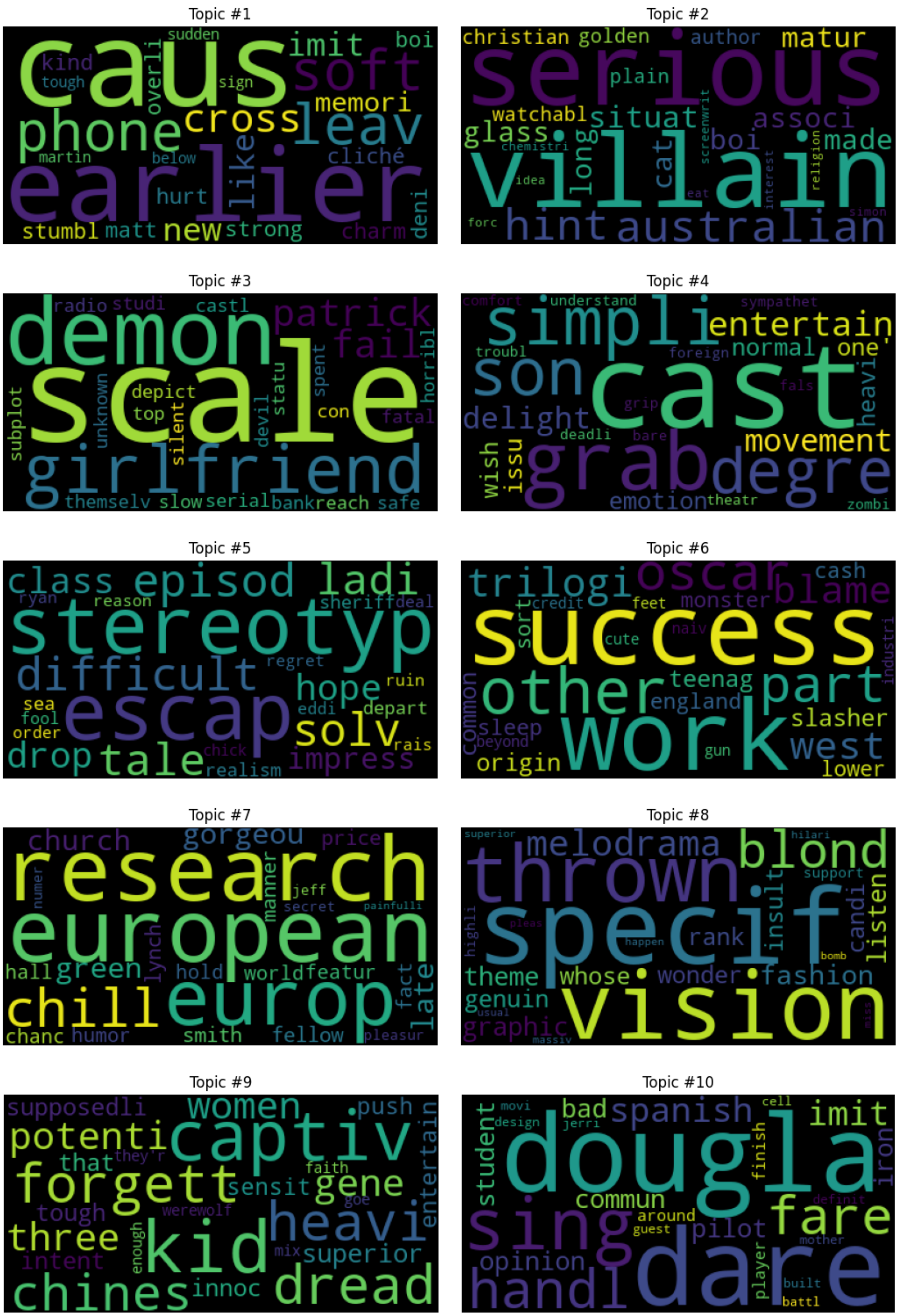


Figure 4: WordCloud Visualizations of the Top Words in Each Topic following Corpus Reduction

Discussion of findings:

Topic 1:

The contents of this topic seem to cover films that have a somewhat somber tone, as the words **memories** and **soft** were used. I feel like this topic encapsulates movies that revolve around morbid **phone** calls and **leaving** the people who matter most to you. Overall, the feeling of this topic evokes a sense of nostalgia, dream-like sequences, and reflection. This differs from the action-packed feeling that the previous Topic 1 evoked.

Topic 2:

The contents topic 2 appears to encapsulate films that involve **villains** and **serious situations.** Perhaps this topic encapsulates stories about **Australian** political issues, and perceived political **villains** based on party affiliation. Overall, this topic appears to evoke a sense of villainy and seriousness. This is very different from that of the previous Topic 2, which seemed to capture Batman Films.

Topic 3:

The contents of topic 3 appear to feature actors or a character by the name of **Patrick** and his **reach** to the **top** in some sort of competition? Perhaps it covers films that have an energy of training and self improvement, as the film can highlight the **failures** the main character experiences along the way and has him face his inner **demons**. Overall, this topic evokes a feeling of improvement, struggle, and determination. This differs from the supernatural elements and feelings that the previous topic 3 held.

Topic 4:

The contents of topic 4 appear to encapsulate films that are lighthearted, **emotional**, and **entertaining**, which makes me think that perhaps this topic covers children's movies that are a **delight** to watch. Perhaps these films were about a **son** who makes a **wish** on a star for something grand. This is very different from the previous Topic 4, which had a more halloween theme. Both Topics do appear though, to have lighthearted, fun themes.

Topic 5:

I feel that this topic encapsulates films that explore the **stereotypes** of **ladies** who undergo almost **episode** types of mishaps. The words **escape**, **difficult,** and **solve** also makes me think that this could be a murder mystery sort of theme, with strong female protagonists. This is very different from that of the previous Topic 5, which had a more archeological, adventurous theme.

Topic 6:

I feel like this topic encapsulates films that have **succeeded**, as they have received **oscars** before for the films. The films probably involve a **trilogy** and could be potentially a **slasher** film, that potentially could be a **western** film as well.Compared to the other topic six shown in Figure 1, this Topic appears to have less of a "financial" aspect, and more of an adventure packed feeling to it.

Topic 7:

This topic felt like it encapsulates films that cover the **European** countryside as a **green, chill,** and **gorgeous** landscape. Maybe it is a series of documentaries that cover European churches, and research into its history (i.e. like Rick Steves's documentary videos). Overall, this differs from that of the previous Topic 7 in Figure 1 as that topic appears to be more focused on **cgi** movies.

Topic 8:

This topic feels like it encapsulates the **melodrama** of the fashion industry, similar to the movie *The Devil Wears Prada*. The reason why I think this is because of the words **vision, blond, fashion, graphic,** and **blonde,** which evokes a sense of early 2000 movies with strong female leads who don't compromise on their femininity to succeed in their fields (i.e. *Clueless* or *Legally Blonde)*. This differs from that of the previous Topic 8, which held a more sci-fi theme.

Topic 9:

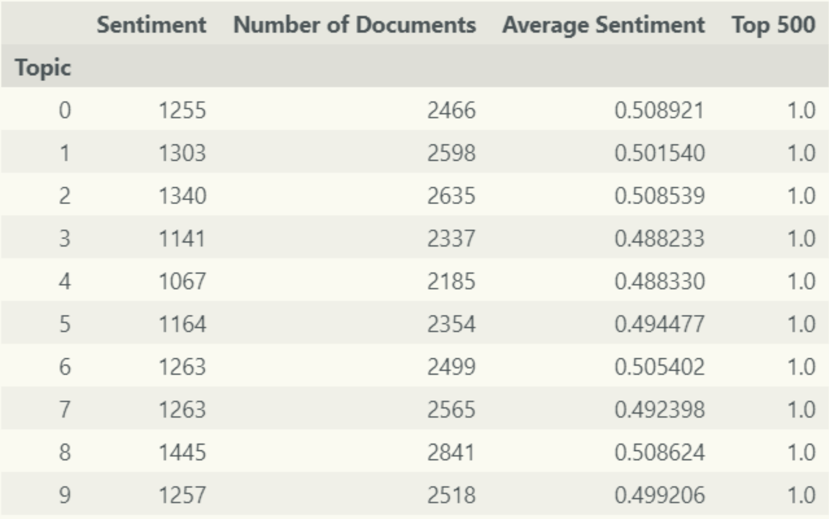
I feel like this topic encapsulates horror films, as many of the words evoke a feeling of captivity, fear, and ominousness. The reason why I think this is because some frequent words were **captive, dread,** and **forgetting.** Perhaps these movies feature **three women** who hold **kids captive** for their evil experiments. This differs a lot from the contents of the previous Topic 9, which had a more college-centric romance theme.

Topic 10:

I feel like this topic encapsulates musicals and films that involve airport themed settings. The reason why I think this is because the words **sing, fare,** and **pilot** were frequently used in this Topic. Perhaps these films feature a character or actor by the name of **Douglas** as well. Overall, this topic differs from that of the previous Topic 10 as it had a more emotional and financial based theme.

10% credit. Find the top 500 movie reviews for each topic in part b1, then compute the average sentiment of those 500 reviews.

Does the average sentiment of each topic help you interpret those topics? Please elaborate.



	Sentiment	Number of Documents	Average Sentiment	Top 500
Topic				
0	1255	2466	0.508921	1.0
1	1303	2598	0.501540	1.0
2	1340	2635	0.508539	1.0
3	1141	2337	0.488233	1.0
4	1067	2185	0.488330	1.0
5	1164	2354	0.494477	1.0
6	1263	2499	0.505402	1.0
7	1263	2565	0.492398	1.0
8	1445	2841	0.508624	1.0
9	1257	2518	0.499206	1.0

Figure 5: Pandas Dataframe Showing Statistics of the Corpus Using the Previous Model for Classification of Documents

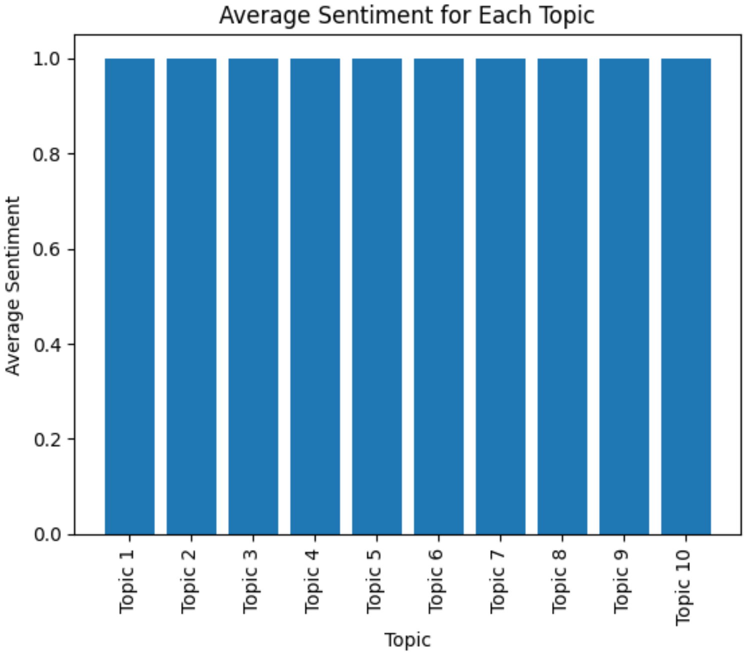


Figure 6: Average Sentiment for Each Topic

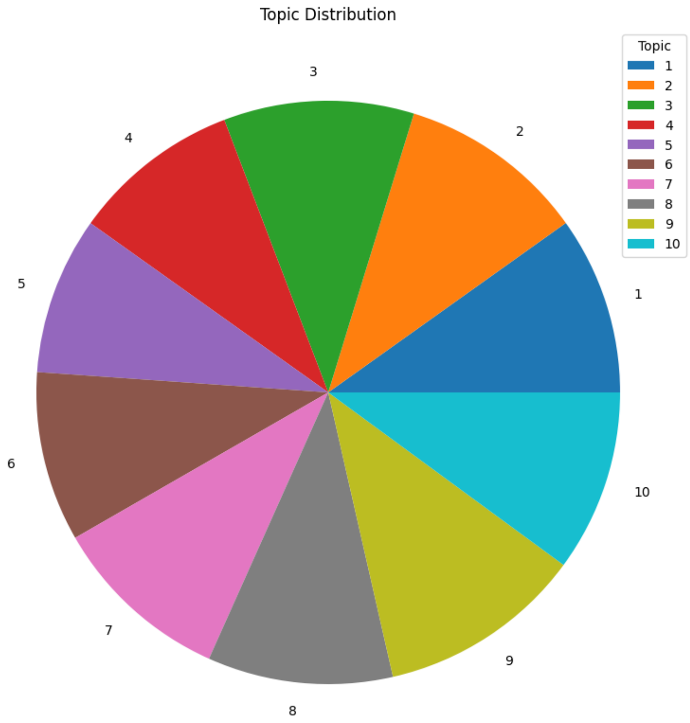


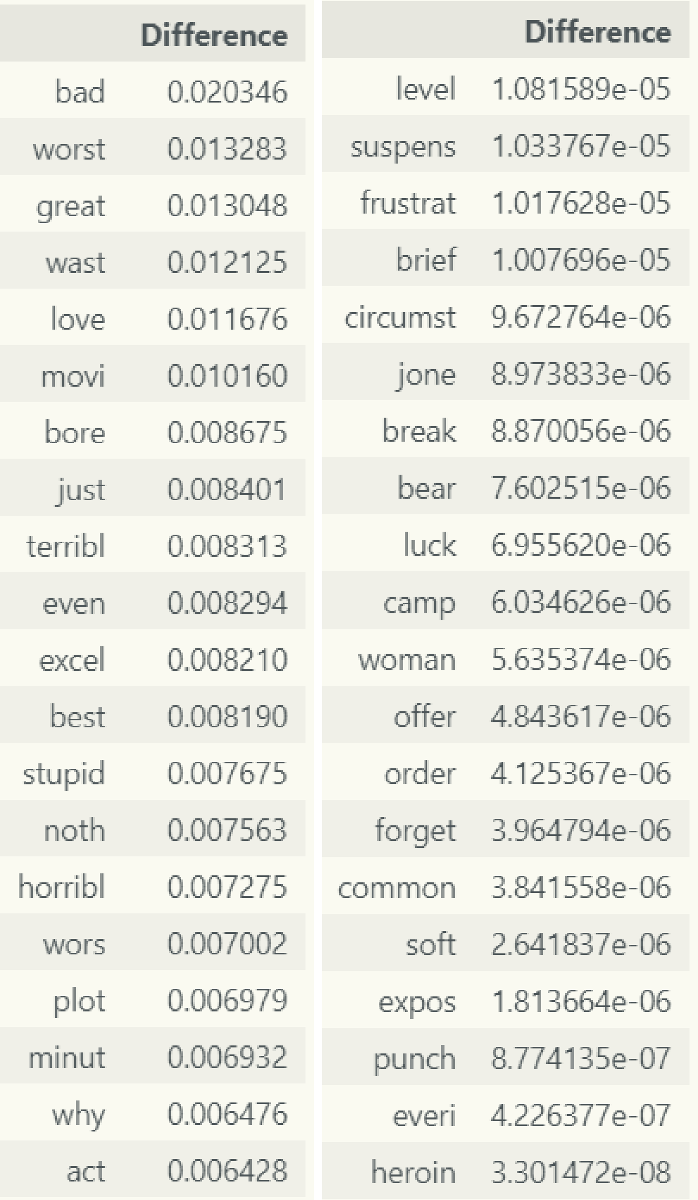
Figure 7. Distribution of Documents Following LDA classification of Topic

Discussion of findings:

- Initially, when computing the average sentiment for each topic, I was surprised to see that all the topics had the same average sentiment of 1 (i.e. Figure 6). This told me that the top 500 reviews for each topic all had positive sentiments. Thus, I generated Figure 5 following this observation, which Figure 6 was derived from.
- Figure 5 gives us the descriptive statistics regarding each document, as after assigning the 24998 documents to their respective topics (by taking the index of the max score that the model generates for the document).What surprised me was that overall, there appeared to be generally even distribution of documents in each Topic. Though some topics had more documents classified as such then other topics, overall, the ratio of positive sentiment to negative sentiment was pretty evenly split (i.e. Figure 7, which shows distribution of Documents following classification).
- While the initial Figure 6 initially drew alarm, upon further investigation of the data, I was relieved to find that this result aligned with observation of the dataset following LDA classification of Topics. Overall, it appears that within each Topic, distribution of negative and positive sentiment was generally evenly split, and thus, taking the top 500 reviews would result in an average sentiment score of 1, and likewise, taking the bottom 500 reviews would result in an average sentiment score of 0.
- This tells us within each topic, movies had generally equal numbers of people who enjoyed the topic and disliked the topic.

10% credit. Split the corpus into two sets: X+ containing only positive reviews, and X containing only negative reviews.

Then compute the tf-idf matrix for each set: M+ from X+ (roughly a 12,500✖2073 matrix), and M- from X- (~12,500✖2073). Next, compute the average tf-idf of each word in X+ (i.e., a 1✖2073 vector), and the average tf-idf of each word in X- (i.e., another 1✖2073 vector). Finally, compute the absolute value of the difference between the two vectors, and sort words by decreasing order. Which words come at the top and at the bottom of the list? Can you interpret the result?



	Difference
bad	0.020346
worst	0.013283
great	0.013048
wast	0.012125
love	0.011676
movi	0.010160
bore	0.008675
just	0.008401
terribl	0.008313
even	0.008294
excel	0.008210
best	0.008190
stupid	0.007675
noth	0.007563
horribl	0.007275
wors	0.007002
plot	0.006979
minut	0.006932
why	0.006476
act	0.006428

	Difference
level	1.081589e-05
suspens	1.033767e-05
frustrat	1.017628e-05
brief	1.007696e-05
circumst	9.672764e-06
jone	8.973833e-06
break	8.870056e-06
bear	7.602515e-06
luck	6.955620e-06
camp	6.034626e-06
woman	5.635374e-06
offer	4.843617e-06
order	4.125367e-06
forget	3.964794e-06
common	3.841558e-06
soft	2.641837e-06
expos	1.813664e-06
punch	8.774135e-07
everi	4.226377e-07
heroin	3.301472e-08

Figure 8: Top 20 and Bottom 20 Words in the Absolute Value Vector

Discussion of findings:

Following separation of the corpus into documents containing positive reviews, and documents containing negative reviews, computation of the tf-idf matrix for each set was performed, scaling each word, and averaging, we end up with two 1 by 2073 vectors that provide an overview and summary of the weighted frequency of terms in the overall document set. Thus, before we take the absolute value difference of the two vectors, we currently have two vectors that give us an idea of what words were frequently used in documents that had good sentiment, and documents that had bad sentiment.

The difference between these two values are somewhat likened to the Manhattan/City Block distance measure. Thus, before we interpret the values of the vector and the top and bottom words following sorting, larger absolute values for a term tells us that with regards to the two document sets X+ and X-, that particular term is dissimilar (i.e. that term had high frequency of use in one set, and low frequency of use in another). Likewise smaller absolute value indicates similarity, as the averages of the tf-idf values are more close in value, and the two documents used the term with the same weighted frequency.

Analysis of the top 20 words:

When looking at the words at the top of the list, many of the words are adjectives that describe emotions and opinions that people may have with regards to the movie. Some positive words that had high absolute value difference values included **great, love,** and **excel**, which shows that these terms are more indicative of one sentiment than the other. These high values most likely are the result of frequent use of these terms in the positive sentiment set, and infrequent use of these terms in the negative sentiment set.

There were a lot more negative connotated words within the top 20 words in the vector, which also aligns with understanding of how reviews and document bias can occur. For example, reviewers are more inclined to write negative reviews due to heightened emotional response. Negative emotions often are processed more thoroughly by individuals who experience it, and thus, it makes sense that many of the top words in the absolute value difference matrix are negative. Examples of these words include **bad, worst, waste, bore, terrible, stupid, horrible,** and **worst.** This does not indicate that there are more bad sentiment documents, but rather, there is a large discrepancy between usage of these words in positive and negative document sets. The above words are most likely more used in the negative set, and thus, since they have higher weighted frequency, and the positive set has lower frequency of these words, there is a large level of dissimilarity between these terms being used in the sentiment sets.

Analysis of the bottom 20 words:

Many of the bottom words appeared to be mostly nouns and actions, rather than adjectives that describe emotion. Words such as **level, bear, camp, woman, offer,** and **suspense** are some examples that are within the set. Because these words have a low absolute value difference measure, this tells us that these words were equally used with similar average frequency rates in both the X+ and X- sets. These terms are most likely common and equally used in both positive and negative documents. The lowest word with average frequency was the word **heroin** which somewhat makes sense, as it doesn't appear to be an emotionally charged word that would be used in different proportions with regards to positive and negative sentiment.

References:

- Topic Modeling in Python: Latent Dirichlet Allocation (LDA) | by Shashank Kapadia | Towards Data Science
- Interesting visualizations, LDA, Word2vec | Kaggle
- Understanding TF-IDF (Term Frequency-Inverse Document Frequency) - GeeksforGeeks