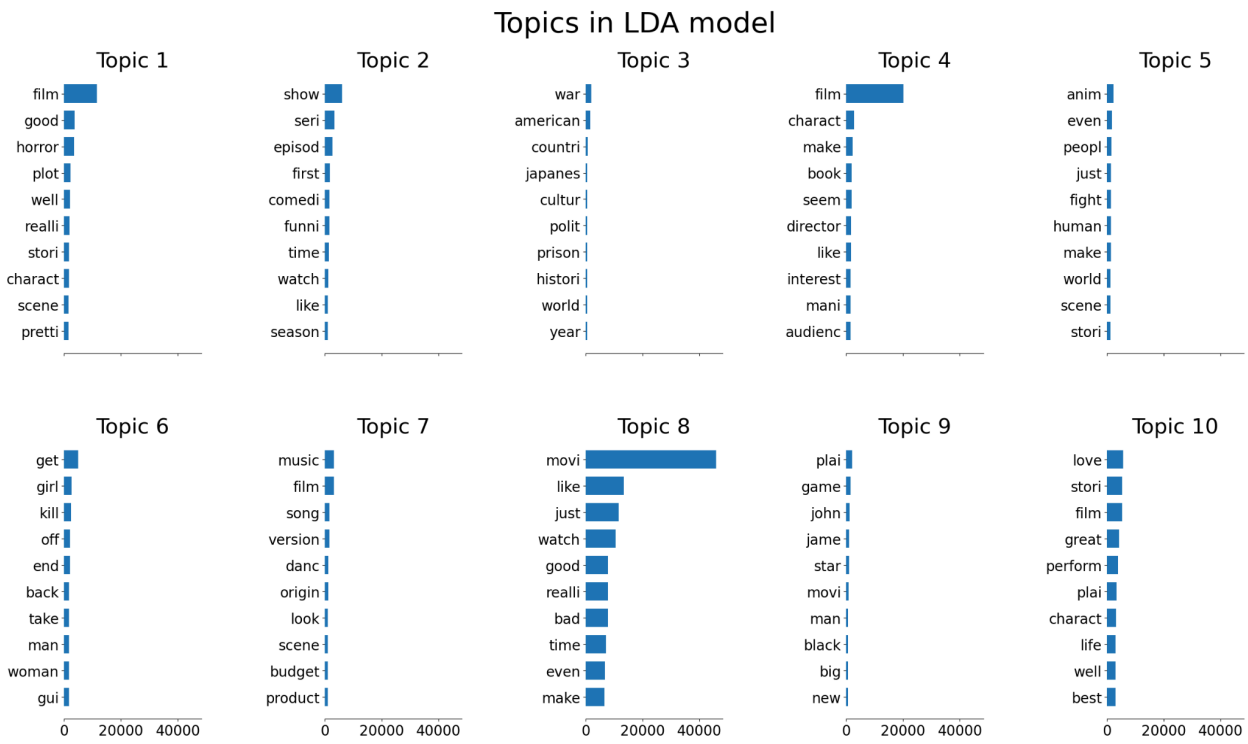

Figure 1. Top 10 topics from unfiltered bag-of-words

In this project, I wanted to analyze different movie reviews and uncover hidden information. Given a collection of different movie reviews, I wanted to find the most common topic clusters amongst them. In order to do so, I used Latent Dirichlet Allocation (LDA) in order to generate a topic model. LDA is a probabilistic model used for topic modeling that assumes that a document is a combination of topics. Each topic has its own vocabulary of which words are drawn from given a topic-specific distribution. LDA is able to find which words belong best to each topic and cluster these words. From these topics and their vocabulary, the dominant topic subjects of a document can be revealed. Figure 1 shows the topic results of performing LDA from the given data.

In order to perform LDA, I first needed to generate the sparse document-feature matrix. I was able to accomplish this by first creating a matrix with a row for each document and a column for each word within the vocabulary across the documents. Then I was able to extract the data within the ‘counts.csv’ file and load this data into the document-feature matrix I created. Once I had the document-feature matrix I was able to perform LDA using the sklearn function. I was able to avoid long run times by specifying

the model to use ‘online’ learning instead of ‘batch’ learning when fitting the large and sparse document-feature matrix.

The results of this process are illustrated in figure 1 which shows the top 10 topics found by LDA and the top 10 words which fit best. Interpreting these topics subject matters is somewhat difficult as none of the topics’ words show a strong connection, however some trends can be seen. For example topic 3 contains words such as ‘war’, ‘american’, ‘countries’, ‘japanese’, ‘world’, and ‘history’. These words appear to indicate that topic 3 is related to war movies specifically about World War II. Topic 6 could be about suspenseful mystery murder movies with words such as ‘get’, ‘girl’, ‘kill’, ‘end’, ‘take’. Topic 10 could be about character-driven romantic dramas with words such as ‘love’, ‘stories’, ‘characters’, and ‘perform’. Topic 7 could be about musicals with words such as ‘music’, ‘film’, ‘song’, and ‘dance’. Topic 5 could represent action animated movies with words such as ‘anime’, ‘people’, ‘fight’, and ‘world’. While some of these topics can be assumed some of the topics are harder to interpret. Topics such as 8 and 9 are difficult to interpret as these topics contain words that are common to most movie descriptions such as ‘movie’, ‘like’, ‘watch’, ‘man’, ‘bad’. After my initial analysis, I recognize the topics are not clear enough as they contain common words. I hope to eliminate some of these common words in order to make each topic more distinct.

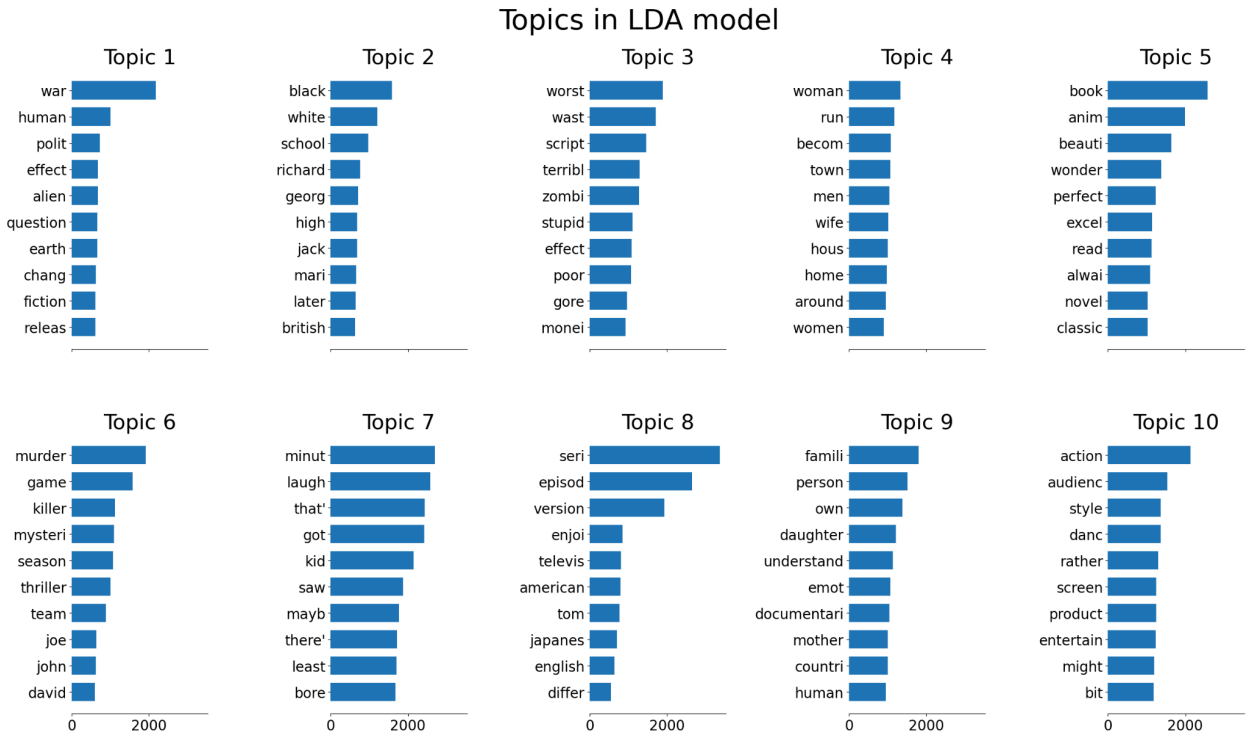

Figure 2. Top 10 topics from filtered bag-of-words

Topic #	*Subject*
1	War, interplanetary, political
2	Segregation, civil rights movement
3	apocalypse, zombies, violent
4	coming-of-age, running away from home
5	movies based on books, novels, classics
6	murder, mysteries, drama
7	comedies, children
8	tv-show, different countries
9	dramas, family
10	action, musical

Figure 3. Topics found by LDA using filtered bag-of-words

Through analyzing the results of the performing LDA on the complete bag-of-words, I noticed some of the topics were difficult to interpret. In order to improve these results, I decided to remove the 100 most frequent words with the bag-of-words. Through removing these words, LDA will not consider words that are common to all movie reviews. This will improve thes results as these common words hold no true information of each document as these words are simple words that are used across all reviews.

After identifying and removing these words, I was able to perform LDA on the filtered document-feature matrix. The topics found by LDA and their top words are shown in figure 2. The results of this experiment are much better than before and show much better correlation between words. The topics reveal more distinct subjects such as war-movies, comedies, dramas, civil rights movements. My interpretations are shown in figure 3, which were drawn from the words found in each topic. After removing the most frequent words, the topics found appear to be more distinct and specific such as the civil rights movement, zombie-apocalypse, and novel-based movies.

Topic #	*Sentiment*
1	0.002
2	0
3	0
4	0
5	0.002
6	0.002
7	0
8	0.002
9	0.002
10	0

Figure 4. Average sentiment for each each topic

After developing a better topic model from part b, I was motivated to see what other information could be found. Along with the vocabulary of the reviews, I was also given the sentiment of the review being either negative or positive. Through pairing this information with the reviews that best fit each topic, I could analyze if there was a correlation between topic and sentiment of the review.

My first step was to find the top documents that best match within each topic. This was possible as LDA finds the probability of each topic appearing within all the individual documents. Therefore I needed to find the top 500 documents for each topic that had the highest probability. After identifying these reviews, I was able to find the sentiment that corresponded with each and average the sentiment of these reviews.

The average sentiment results are shown in figure 4. These results were initially surprising as half of the topics had an average score of 0 and the other half had a score of 0.002. Therefore this would mean of all the reviews examined, only 5 had a positive review. After examining further my processes and the nature of the data and subject, I

concluded that this is acceptable results. These results make sense as they support the fact there is no correlation between the sentiment of the review and the topic of the movie. This can be explained by assuming the reviews are written by different authors that have different movie preferences that lead them to writing a positive or negative review. This information does not make it easier to interpret the topics found by LDA. This makes sense as I assume the sentiment score represents whether the review favored the movie or did not. This information does not reveal anything about the subject itself. Although some averages were different, the differences are too small to believe there to be a statistically significant correlation.

	Vocab
Top 20 words with largest TF-IDF score difference	bad, worst, great, waste, love, movie, bore, just, terrible, even, excel, best, stupid, noth, horrible, worst, plot, minut, why, act
Bottom 20 words with smallest TF-IDF score difference	level, suspense, frustrate, brief, circumstance, jone, break, bear, luck, camp, woman, offer, order, forget, common, soft, expose, punch, everi, heroin

Figure 5. Top and bottom words with TF-IDF average difference

After clustering the topics, I was still motivated to explore the sentiment feature of each review. Figure 4 revealed the topics found by LDA to be a poor predictor of the sentiment review. In hopes to find a better predictor, I decided to explore which words within the bag-of-words were most unique and important to each sentiment class of reviews.

I was able to accomplish this by first separating the documents into positive and negative sentiment groups. For each group, I calculated the TF-IDF matrix. The TF-IDF scores would be a good estimator of importance as it identifies words within a goldilocks' zone. By down-weighting frequent words, words that appear with an average frequency are more pronounced. These words that appear in the goldilocks' zone are of most interest as they capture information that appears frequent while removing common

words that capture no distinct information. Therefore, averaging the TF-IDF score for each word across the documents within each sentiment class shows how important that word is to that particular class. After doing this for both classes, I found the absolute difference between each word. Doing this and sorting the differences reveal which unique words have the highest difference in frequency.

Figure 5. shows the words found during this process. These results align with words I expected to see. The top 2 words appear to be more expressive words that express one's opinion towards the movie being reviewed, These words would better identify one's sentiment towards a movie as compared to the bottom 20 words. The bottom 20 words appear to be common words that describe movie plots rather than one's opinion.