Movie Reviews: Clustering, Topic Modeling, and Sentiment Analysis

W. Andy Holst

School of Professional Studies, Northwestern University

MSDS 453: Natural Language Processing

Dr. Syamala Srinivasan

July 25, 2021

Movie Reviews: Clustering, Topic Modeling, and Sentiment Analysis

Let's face it, most humans *love* to watch movies. It is a uniquely human experience to set aside a few hours of time to immerse ourselves into a visual, audial, and often emotional display of imagination. With this initial perspective in mind, executives at movie theaters may desire to better connect the movies they are promoting with the viewers who ultimately decide upon which movie to fork over $25+ to go see in a theater. Moreover, a recent study by Deloitte revealed that some movie genres -- such as Action and Sci-Fi movies -- perform better at the box office (Arkenberg et al., 2020). Figure 1 below shows the multiple return of revenue (by genre) for every dollar spent on marketing in the film industry. What if it were possible for a movie theater operator to automatically scan the world wide web to find all the movie reviews about Action and Sci-Fi movies; then promote these reviews to movie viewers (on social media or through targeted ads)? This might be a fantastic way for movie theater operators to successfully encourage people after the Covid-19 pandemic to return to movie theaters once again.

This business opportunity for movie theaters provides the backdrop for the results that I present in this second paper on Natural Language Processing (NLP). I previously analyzed the language contents of a corpus (200 documents) of movie reviews collected by students. Various clustering and topic modeling techniques are explored, with varying results. After recognizing a distinct trend in the topic models produced across this 200 document corpus, I then turned attention to developing a method of differentiating Action and Sci-Fi from other types of movie reviews (Comedy and Horror). The final model developed is based upon a relatively simple technique known as the Support Vector Machine (SVM), which identifies clear decision boundaries in high dimensional spaces (Srinivasan, 2021). This SVM model has been shown to achieve **100% Accuracy** in identifying Action and Sci-Fi movie from a Test Set of 30 reviews.

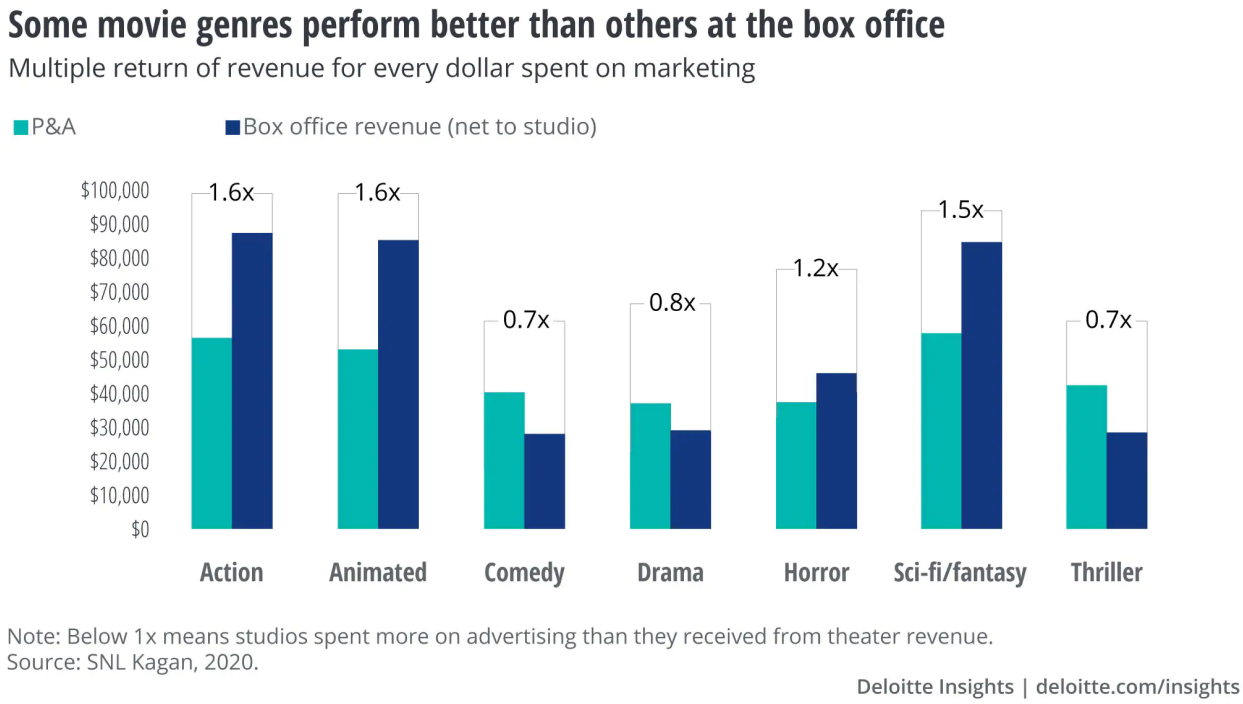

Figure 1 - Deloitte's *Digital media trends*: Action, Animated, and Sci-Fi Movies perform better

Data

The data utilized in this research consist of 200 movie reviews that have been collected and compiled by students in this NLP course. Each movie review (the "text") is at least 500 words long and includes the following metadata associated with each text:

- **Document ID** (numbering 1 to 200)
- **Unique Document Name**
- **Student Name** (20 students total, each student contributed 10 movie reviews)
- **Movie Title Name** (20 different movies included, based upon the above distribution)
- **Genre of Movie** (4 genres were selected, and movies are equally distributed across each genre: Action, Comedy, Horror, and Sci-Fi)

- **Review Type** (positive or negative, each student contributed 5 of each, so there are equal numbers of positive reviews and negative reviews)

The data preparation, exploratory data analysis and visualization steps undertaken to prepare were presented in the previous research paper entitled “Movie Reviews: A First Vectorized Representation”, dated July 11, 2021. The general specifications for research, analysis, and modeling were as follows:

- A Google Colaboratory Notebook (Python 3.7.10) was created.
- Python Packages imported as necessary for pre-processing and modeling, including: Numpy (1.19.5), Pandas (1.1.5), Sci-kit Learn (0.22.2), Matplotlib (3.2.2), NLTK (3.2.5), Gensim (3.6.0), and Re (2.2.1).
- Much of the exploratory data analysis involves printing strings of characters from each text. These strings are compiled into lists and dictionaries for review and analysis.
- Further exploratory data analysis in this research paper include: developing clustering results using a K-Means algorithm, available as a package from Sci-Kit Learn; developing topic models using statistical techniques known as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (packages available from Gensim).

One facet of our dataset that was not analyzed in previous research is document length. There was a naive assumption made that using techniques like TF-IDF would provide results not susceptible to document length. But, some consideration was provided in this week’s research to look specifically at this naive assumption. I determined that the length of documents in the corpus ranged from approximately 500 words to over 2,000 words. After text cleaning the number of tokens per document still ranged from 200-800 tokens. So, to analyze whether

document length had any impact on analysis, I pre-processed a secondary “truncated” corpus which only considered the first 500 words of each document. After text cleaning the range of token length of each document was thus reduced from 200-800 down to 160-260 tokens.

Research Design and Modeling Method(s)

Previous Research

In my previous research paper presented on July 11, 2021, results of initial Natural Language Processing (NLP) studies of vector representations were presented. By studying techniques such as Term-Document Matrices using TF-IDF (Term Frequency combined with Inverse Document Frequency) it was possible to begin to see some themes and topics emerge from our 200 movie review corpus. As a summary, the following were top TF-IDF terms identified, with strong elements of all four genres (Action, Sci-Fi, Comedy, Horror) expressed:

```
[('film',57),('movie',49),('horror',23),('family',22),('action',22),
('alien',19),('character',19),('cruise',18),('planet',17),('time',17)
]
```

Step 1: Clustering

To further the analysis of this corpus, I looked to clustering techniques -- specifically the K-Means clustering -- in order to ascertain whether any natural clustering was present in the corpus. The following K-means clustering analysis on the “truncated” corpus’ TF-IDF Term-Document Matrix was performed:

1. 20 clusters, using unigrams (single word values)
2. 20 clusters, using unigrams and bigrams (word combinations, eg. “Emma Stone”)
3. 20 clusters, using unigrams, bigrams, and trigrams (eg. “Edge of Tomorrow”)
4. 4 clusters, using bigrams

5. Clustering ranging from 2 to 25, using Method #1 above, to view silhouette scores
6. Repeating Methods #1 through #3 and #5 above, but instead using the corpus that was not truncated to a 500 word length
7. Clustering ranging from 2 to 25, using the Doc2Vec Method, to view silhouette scores

Step 2: Topic Modeling using Latent Semantic Analysis and LDA

After exploring the K-Means clustering techniques described above, I turned to topic modeling techniques known as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). The LSA algorithm has became known by NLP experimenters as an effective way to reveal meaning of entire documents, and it uses singular value decomposition (SVD) to break TF-IDF term-document matrices down into three components, then reduces the dimensionality of these components to reveal the most influential information within the data (Lane et al., 2019). Latent Dirichlet Allocation (LDA) is also a topic modeling technique like LSA, but it instead uses probabilities to achieve its goal. LDA is based on the premise that “documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words” (Blei et al., 2003). By computing the statistics of these two distributions, topics “emerge” from the corpus. The goal of LSA and LDA are both to extract “meaning” and “semantics” out of the mathematical vector space created by TF-IDF matrices.

During my research, the following LSA and LDA topic modeling was conducted on the corpus’ TF-IDF Term-Document Matrix:

1. LSA on the “truncated” corpus, with the number of topics ranging from 2 to 24, with coherence scores collected for each analysis.
2. LSA on the original length corpus, with the number of topics ranging from 2 to 24, with coherence scores collected for each analysis.

3. LDA on the “truncated” corpus, with the number of topics ranging from 2 to 24, with coherence scores collected for each analysis.
4. LDA on the original length corpus, with the number of topics ranging from 2 to 24, with coherence scores collected for each analysis.
5. A deeper dive into LDA with 4 topics, with visualizations created showing coherence scores in a corpus organized by the four genres of movies (Action, Comedy, Horror, and Sci-Fi).

Step 3: Sentiment Analysis

At the heart of research this week was to nominate appropriate classes for labeling documents in the corpus, with the ultimate goal of facilitating a multi-class classification analysis of the corpus, called a “sentiment analysis”. This sentiment analysis could take any number of forms, so I initiated my research by analyzing whether or not a sentiment analysis could accurately predict the reviews as either positive or negative. This initial methodology was chosen simply because this metadata already exists in the corpus. Predictive models were developed by using Random Forest, Gradient Boosting, and Support Vector Machine, a relatively simple algorithm which identifies clear decision boundaries in high dimensional spaces (Srinivasan, 2021). The final “sentiment analysis” that I conducted was inspired by the visualizations produced in previous Step 2. In recognizing a very high coherence produced in my LSA analysis with 2 Topics (**Coherence = 0.50**). All 200 reviews were randomly split into a Training Set of 170 reviews and a Test Set of 30 reviews.

Results

Clustering Results

With the K-means clustering work described above, the results presented below show Silhouette Scores which are not impressive at all. All Silhouette Scores were below 0.09 and the peak occurred at approximately 20 topics, which is to be expected from this data set of 20 movie reviews. The fact that there was low performance on all range of topics tells me that K-means clusters are not necessarily effective in providing insight into this particular dataset.

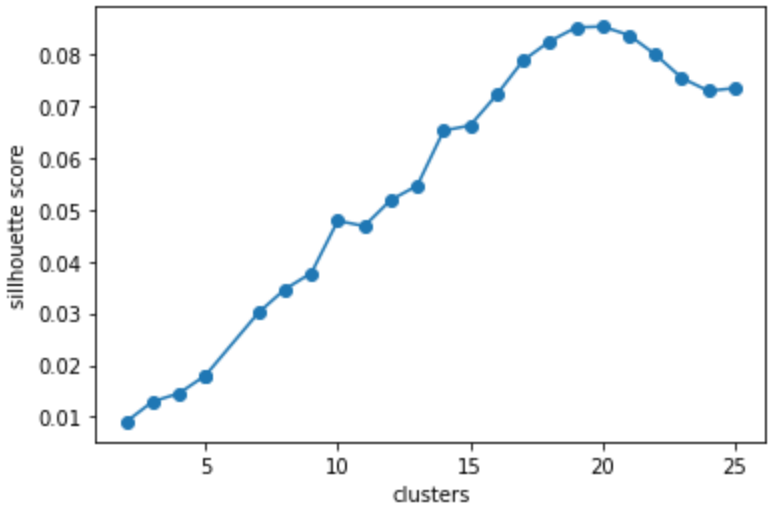

Figure 2 - K-Means Clustering Silhouette Scores reveal poor performance (all results < 0.09)

With an understanding that 20 Topics provides the “best” performing K-Means clustering, I did analyze which of the 20 Topic models performed the best. The following results are reported for the dataset when it’s lengths were limited to 500 words total:

- TF-IDF matrix with **unigrams**: mis-classified two movies - **Superbad** and **The Martian.**
- TF-IDF matrix with **bigrams**: mis-classified three movies - **John Wick, Pirates of the Caribbean, Superbad** and **The Martian.**

- TF-IDF matrix with **trigrams**: mis-classified two movies - **Superbad** and **The Martian, *but only misclassified one out of the 10 Superbad reviews*.**

The following results are reported for the dataset with the full length of each review remained intact (i.e. NOT shortened to 500 words):

- TF-IDF matrix with **unigrams**: mis-classified two movies - **Mission Impossible** and **Poltergeist, *but only misclassified one Mission Impossible and two Poltergeist*.**
- TF-IDF matrix with **trigrams**: mis-classified two movies - **Mission Impossible** and **Us.**

While not entirely conclusive, the K-Means Clustering on the full length reviews, using **TF-IDF matrix with unigrams only** did appear to provide the “cleanest” clustering of any K-means cluster analyzed. These results are listed in Table 1 below:

K-Means Group #	Correctly Classified Movie	Correct Count	Incorrect Count	Incorrectly Classified Movie
1	Pirates of the Caribbean	10	0	
2	Frozen II	10	0	
3	Us	10	0	
4	The Conjuring	10	0	
5	Taken	10	0	
6	James Bond Casino Royale	10	0	
7	The King of Staten Island	10	0	
8	The Ring	10	0	
9	Superbad	10	0	
10	Arrival	10	0	
11	Edge of Tomorrow	10	***1***	***Mission Impossible***
12	Pacific Rim	10	0	
13	Cruella	10	0	
14	Hereditary	10	0	
15	The Martian	10	***2***	***Poltergeist***
16	John Wick	10	0	
17	Mission Impossible	***9***	0	
18	Poltergeist	***8***	0	
19	Home Alone	10	0	
20	Interstellar	10	0	

Table 1 - Results of K-Means Clustering of Movies (20 topics)

Topic Modeling Results

Coherence Scores for the LSA and LDA topic models are presented in Table 2 below, using both the “truncated” corpus (limited to 500 words per movie review) and the original length movie reviews.

	Coherence Scores			
Number of Topics	LSA “Truncated” 10 words	LSA Orig. Length 20 words	LDA “Truncated”	LDA Orig. Length
2	0.234	***0.502***	0.252	0.255
4	0.376	0.325	0.258	0.232
6	0.386	0.431	0.242	0.241
8	0.401	0.302	0.255	0.248
10	0.392	0.375	0.246	0.248
12	0.397	0.441	0.243	0.235
14	0.472	0.404	0.247	0.237
16	0.422	0.400	0.237	0.246
18	0.423	0.406	0.240	0.245
20	0.436	0.340	0.248	0.240
22	0.436	0.341	0.250	0.241
24	0.412	0.404	0.258	0.243

Table 2 - Coherence Scores: Latent Semantic Analysis (LSA) & Latent Dirichlet Allocation (LDA)

In addition to the Coherence Scores, the top ten words for the ***most coherent*** LDA Topic Model (LSA Topic Model, Original Movie Review Lengths, Coherence = 0.502) is displayed below:

```
Group 0: 0.373*"movie" + 0.181*"first" + 0.174*"movies" + 0.165*"story" +
         0.118*"world" + 0.118*"films" + 0.118*"family" + 0.116*"never" +
         0.114*"little" + 0.109*"character"

Group 1: 0.446*"conjuring" + 0.269*"movie" + 0.255*"devil" +
         0.228*"warrens" + 0.196*"horror" + -0.148*"casino" +
        -0.144*"royale" + -0.138*"world" + 0.119*"farmiga" +
         0.114*"wilson"')]
```

While conducting a deeper dive into LDA, I created the visualization shown in Figure 3 below, depicting coherence scores (yellow have high coherent, green have low coherence), but this time around I organized the movie reviews **by the four genres of movies** (Action, Comedy, Horror, and Sci-Fi). What I immediately noticed was there is a high coherence between Action movies and Sci-Fi Movies. Three of the 5 comedy movies reviews, on the other hand, have some partial coherence with Action/Sci-Fi. At the other end of the spectrum, Horror movies have relatively low coherence of topics with any other genre.

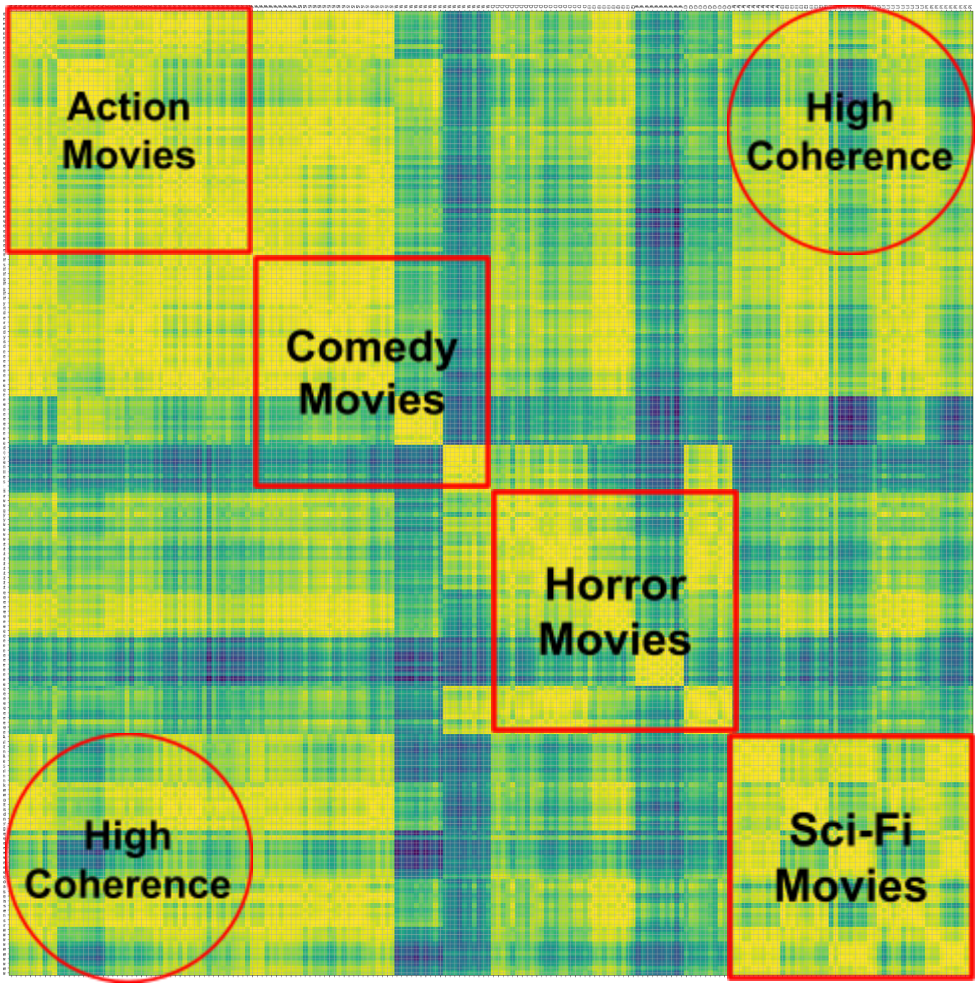

Figure 3 - Coherence Scores for LDA Model (Original Length) Reviews, ***Organized by Genre***

This may be a bit over-simplistic, but what really started to emerge for me during this analysis was that may be two different “themes”, and I propose one group of reviews largely correlate to “Action/Sci-Fi” movies (since they share the most correlation on topics). The other “group” is not really Comedy and Horror per se, because these types of movies have completely different semantic . So I decided to call this second group “Not Action/Sci-Fi”.

Semantic Analysis Results

Following on the hunch that I developed while reviewing the Topic Modeling results described above, I proceeded to nominate appropriate classes of “**Action / Sci-Fi**” and “**Not Action / Sci-Fi**” for labeling my documents in the corpus. Following this labeling, I used Random Forest, Gradient Boosting and Support Vector Machine models to predict the labels on the movie review data set *NOTE: it is also highly possible that labeling the movies by Rating (i.e. G, PG, PG-13, R, etc) might also be a good way to classify these movie reviews and make predictions. This should be considered for future research.*

Results for the Random Forest, Gradient Boosting, and Support Vector Machine models are published in Tables 3, 4 and 5 below:

		Predicted Classes		
		Action/Sci-Fi	Not Action Sci-Fi	
Actual Results	Action/Sci-Fi	12	1	**Recall: 0.92**
	Not Action/Sci-Fi	8	9	
		Precision: 0.60		**Accuracy: 0.70** **F1-score: 0.73**

Table 3 - Random Forest Confusion Matrix, Precision, Recall, Accuracy, and F-statistic

		Predicted Classes		
		Action/Sci-Fi	Not Action Sci-Fi	
Actual Results	Action/Sci-Fi	12	1	**Recall: 0.92**
	Not Action/Sci-Fi	0	17	
		Precision: 1.00		**Accuracy: 0.97** **F1-score: 0.96**

Table 4 - Gradient Boosting Confusion Matrix, Precision, Recall, Accuracy, and F-statistic

		Predicted Classes		
		Action/Sci-Fi	Not Action Sci-Fi	
Actual Results	Action/Sci-Fi	13	0	**Recall: 1.00**
	Not Action/Sci-Fi	0	17	
		Precision: 1.00		**Accuracy: 1.00** **F1-score: 1.00**

Table 4 - Support Vector Machine Confusion Matrix, Precision, Recall, Accuracy, and F-statistic

The best model developed was based upon a relatively simple technique known as the Support Vector Machine (SVM), which identifies clear decision boundaries in high dimensional spaces (Srinivasan, 2021). This SVM model achieved 100% Accuracy in predicting "Action/Sci-Fi" movie reviews within the Test Set of 30 movie reviews. There may still be much to be studied further, especially considering that the most prominent terms which appear are simply the names of the movies themselves, thus clearly biasing the results. So future work

might be considered to better filter out terms and find deeper semantic meaning from the corpus.

Analysis and Interpretation

At the heart of research this week was to nominate appropriate classes for labeling documents in the corpus, with the ultimate goal of facilitating a multi-class classification analysis. I nominated two classes within this corpus as "**Action / Sci-Fi**" and "**Not Action / Sci-Fi**" for purposes of labeling the documents. In recognizing a very high coherence produced in my LSA analysis with 2 Topics (Coherence = 0.502), the final classification model achieved a 100% Accuracy in identifying Action/Sci-Fi movie reviews on the movie review dataset. There may still be much that needs improvement with this analysis especially considering that the most prominent terms which appear are simply the names of the movies themselves, thus clearly biasing the results. So future work might be considered to better filter out terms and find deeper semantic meaning from the corpus.

A final note regarding implementation of this model into production: Because the data needed for predicting the type of movie based upon the language of a movie review, and the evolution of this type of data is expected to be constantly changing, it is expected that this model will require routine updating after it has been placed into production and as new movies are released and new reviews become available.

Conclusions

A business opportunity exists for movie theater operators to automatically scan the world wide web to find all the movie reviews about Action and Sci-Fi movies which will provide them with the highest margins for their theaters, then promote these reviews to movie viewers (on social media or through targeted ads). I previously analyzed the language contents of a corpus (200 documents) of movie reviews collected by students. Various clustering and topic modeling

techniques were explored, with varying results. After recognizing a distinct trend in the topic models produced across this 200 document corpus, I developed a method of differentiating Action and Sci-Fi from other types of movie reviews (Comedy and Horror). The final model developed is based upon a relatively simple technique known as the Support Vector Machine (SVM)_. This SVM model has been shown to achieve **100% Accuracy** in identifying "Action/Sci-Fi" movies from a Test Set of 30 reviews.

References

Arkenberg, C., Cutbill, D., Loucks, J., & Westcott, K. (2020, December 10). *Digital media trends*. Deloitte Insights. https://www2.deloitte.com/us/en/insights/industry/technology/future-of-the-movie-industry.html.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003, January). *Latent Dirichlet Allocation*. Journal of Machine Learning Research. https://www.jmlr.org/papers/v3/blei03a.html.

Lane, H., Howard, C., Hapke, H. M., & Griffioen, A. (2019). Chapter 4: Finding Meaning in Word Counts (Semantic Analysis). In *Natural language processing in action understanding, analyzing, and generating text with Python* (pp. 97–152). essay, Manning Publications Company.

Srinivasan, S. (2021, July 19). *MSDS 453 Week 5 & 6 Lecture – Classification and Clustering, Knowledge Graphs* [Lecture notes, PowerPoint slides]. Northwestern University. https://canvas.northwestern.edu/courses/142977/modules/items/1976556.