Jing Li (S5374553)

A report about Extract Keywords and Topic Modelling

In this report, I'll go through the process of extracting the top 20 keywords from a given dataset and applying topic modelling. I'll explain the methods used to get these results, what the outputs represent, and what insights I may gain from them.

Extract the Top 20 Keywords

To extract the top 20 keywords from a given dataset, I performed several processing steps. First, I cleaned the dataset and got a DataFrame containing 388 rows and 2 columns. And for the pre-processing I did lowercase, remove punctuation, stopwords removed, and Lemmatization. Then using the TfidfVectorizer to create a matrix of word frequencies and computed the tf-idf score. The tf-idf score evaluates how relevant a word is to a text by considering the frequency of the word in the text as well as the frequency of the word in the entire dataset. Finally, I can sort the words by tf-idf score and select the top 20 words.

The output shows that the top 20 keywords for the whole dataset are: 'book', 'read', 'story', 'just', 'characters', 'good', 'im', 'like', 'really', 'great', 'love', 'novel', 'reading', 'think', 'character', 'did', 'books', 'doesnt', 'interesting', 'liked'. These words can provide insights into the most important topics discussed in the text.

No.	Keywords	Score	No.		Keywords	Score
1	book	0.522423	11		love	0.111450
2	read	0.313454	12		novel	0.097519
3	story	0.243797	13		reading	0.097519
4	just	0.181107	14		think	0.097519
5	characters	0.167175	15		character	0.090553
6	good	0.139313	16		did	0.083588
7	im	0.139313	17		books	0.076622
8	like	0.132347	18		doesnt	0.076622
9	really	0.132347	19		interesting	0.076622
10	great	0.111450	20		liked	0.076622

Table 1. Top 20 keywords

Topic Modelling with Positive and Negative Texts

To create two separate topic models, I first created two new DataFrames with only positive and negative texts, and then pre-processed the texts. The positive sentences are identified by the 'sentiment' column, where a value of 1 indicates a positive review.

The text in these positive texts is pre-processed, with non-alphabetic characters removed and stopwords removed. The text is then lemmatized to reduce words to their base form. A topic model is then generated using the LDA, with 4 topics and 100 passes over the data. The resulting topics are printed using pprint and visualized by pyLDAvis. And for the negative texts that are pre-processed with the same routine,

since the corpus is less than positive, I set the chunksize as 5.

From the positive sentiment topics (Figure 1), topic 1 seems to be related to a series with the hero, and topic 2 is related to characters maybe in a mystery story. From the negative sentiment topics (Figure 2), topic 3 mentions the word “interesting” but in this context, it seems to be not that enough to keep reading.

Topic	Top 10 words in each topic
Topic 1	world, series, well, developed, could, hero, many, one, crawford, part
Topic 2	character, good, novel, great, lot, reader, mystery, liked, better, story
Topic 3	book, read, much, really, love, think, story, first, time, way
Topic 4	read, well, bit, hill, little, take,fun, voice, see, star

Figure 1 Results with positive sentiment topics

Topic	Top 10 words in each topic
Topic 1	character, felt, feel, made, author, every, scene, even, cheesier, one
Topic 2	book, first, part, like, sequel, unfortunately, overlooked,largely, ignored, flaw
Topic 3	get, interesting, much, reading, place, got, though, sort, found, lee
Topic 4	disappointed, story, time, really, got, good, something, much, wanted, finished

Figure 2 Results with negative sentiment topics

PyLDAvis provides an interactive visualization that can be used to explore the topics and their relationship to each other. The circles represent the topics, with larger circles indicating more prevalent topics. The distance between circles represents the similarity between topics. We can see from Figure 3,4 that the topics have no overlapping areas, so the topics have low similarity and are clearly separated.

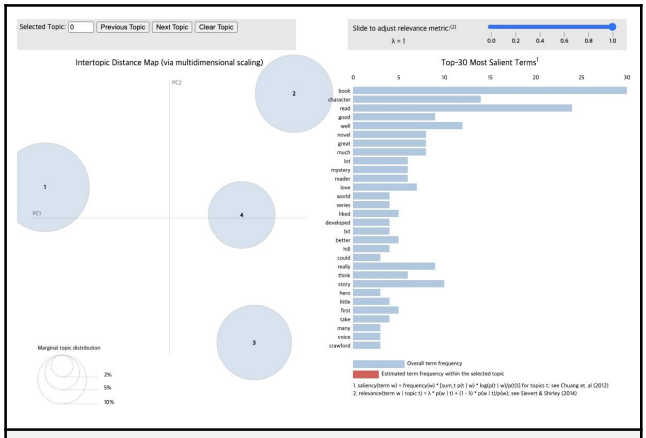

Figure 3 Topic modelling with Positive Texts

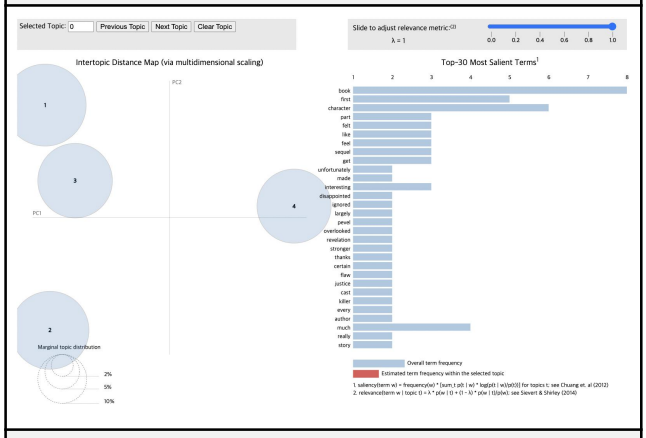

Figure 4 Topic modelling with Negative Texts

In conclusion, topic modelling and the extraction of top keywords are effective methods for analyzing large amounts of text. By using these methods, we can get an understanding of the key aspects of a certain text or corpus of texts.