DS 5001 Final Report

Nick Kalenichenko (nhk6up@virginia.edu) DS 5001 Spring 2023

1. Introduction

The corpus contains each book in JK Rowling's Harry Potter books. My analysis is coming from someone who has never read the books nor seen the movies, who is trying to extract as much information about the books as possible through text analytics. To do this, I will use dispersion plots, similarity analysis, hierarchical clustering, correlation plots, and tsne plots.

2. Source Data

1. Provenance

The data came from:
https://github.com/prakhar21/whiteboard/tree/master/nbviewer/notebooks/data/harrypotter

This is a github repo containing text files for all of the Harry Potter books.

2. Location

https://virginia.box.com/s/uqzk085yu73wog9o0ych8uc4n8ael3kk

3. Description

The general subject matter of the corpus is Harry Potter, which is a fictional book series about magic and spells. There are 7 books (observations) in the corpus, and the average book length is 189055 tokens.

4. Format

The format of the source file formats is plaintext (.txt) files.

3. Data Model

All tables can be found in the `tables` folder in the github directory

- **LIB**: https://github.com/nick-kalen/DS5001-final-data-project/blob/main/tables/LIB.csv
  - book_id: ID of book
  - title: title of book

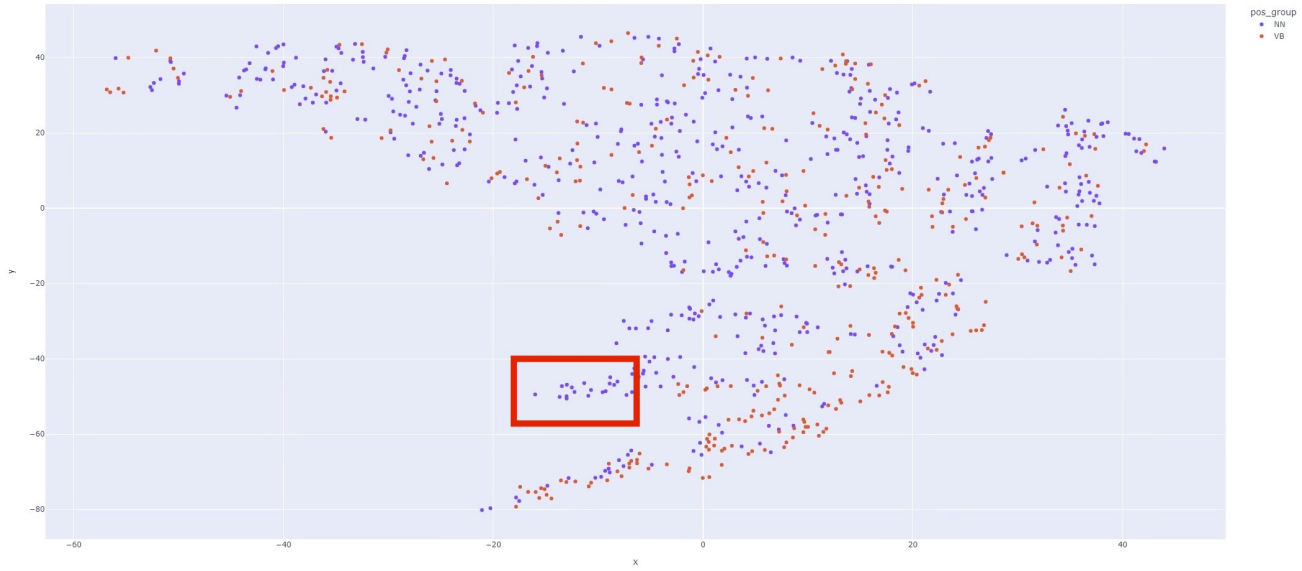
    - chapter_regex: regex used for chapter in book
    - book_len: number of tokens in book
    - nb_chaps: # of chapters
    - kendall_sum: kendall_sum metric used for correlation heatmap
- **CORPUS**: https://github.com/nick-kalen/DS5001-final-data-project/blob/main/tables/CORPUS.csv.zip
    - INDEX: book_id, chap_num, para_num, sent_num, token_num
    - pos_tuple: token string with part of speech
    - pos: part of speech
    - token_str: token string
    - term_str: token string (no formatting)
- **VOCAB**: https://github.com/nick-kalen/DS5001-final-data-project/blob/main/tables/VOCAB.csv
    - n: number of occurences
    - p: probability of term occuring
    - i: information for each term
    - n_chars: number of characters in word
    - max_pos: highest part of speech for word
    - n_pos: how many parts of speech word has
    - cat_pos: set of parts of speech for the word
    - stop: boolean of if it is a stop word
    - stem_porter: porter method of stem for word
    - stem_snowball: snowball method of stem for word
    - stem_lancaster: lancaster method of stem for word
    - dfidf: global boolean term entropy
    - mean_tfidf: average tfidf for the word in a document

These are our core tables, and in the github repo in the same `tables` folder you can see the document and components, and loadings table that were used for PCA, as well as our document and topics, and topics and term counts tables used for LDA. I have also included the Word2Vec csv file, as well as the sentiment and emotion values as features of vocab table, and the sentiment polarity and emotions table for each document. While PCA and LDA are not talked about in this report, the code that was used to generate these models are in the exploratory data analytics notebook. All of these tables follow the standards we learned in class, and follow the examples learned in the example code provided in the class GitHub.

In the helper_files folder, there is HarryPotterETA.py containing t

4. Exploration

To kick off my exploration, I decided to look at the t-distributed stochastic neighbor embedding to glean information from the different clusters. To create this plot, I used the TSNE function with parameters: perplexity=20, n_components=2, init='random', n_iter=1000, random_state=42, learning_rate = 200



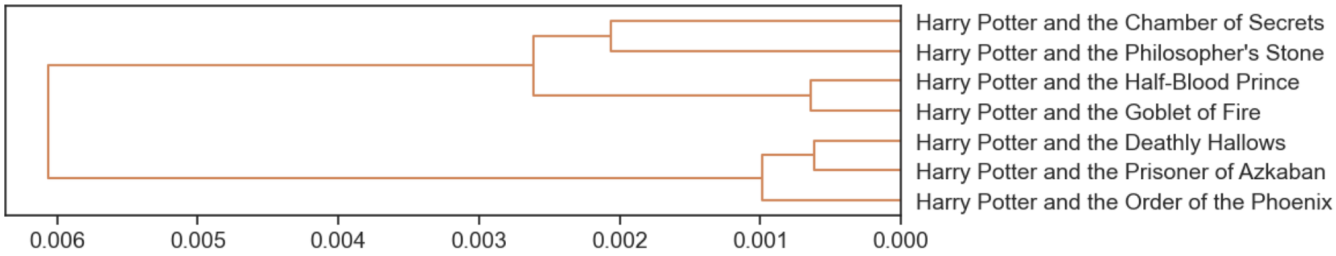
While the words do not appear, I noticed that this cluster contained many proper nouns. Harry was the leftmost dot, and next to him were the words: **professor, Herimione, Ron, Snape, Dumbledore, Neville, Percy, Jon, and Fred**. Knowing that Harry is the main character, my first impression is that these are his friends and family, and people who are for him. What also leads me in this direction is that in the cluster right above the red box, there is another cluster of proper nouns that contains Voldemort (who I know is evil). The names around Voldemort are: **Parvati, Crookshanks, Mundungus, Fleur, and Parvati**. These names also sound different than the first group that contain some more traditional names, like Jon, whereas these names are all very unique.

For further exploration into the characters, I decided to use the Word2Vec model with window=2, vector_size=256, and min*count=80. I used the analogy `Harry: light, Voldemort: __`* which gave me the results: dark, stone, witches, hidden, and against. This shows that Voldemort is indeed a bad guy in this series, which is one of the only things I know about Harry Potter. Instead of analogies, I got the most similar words for Ron using the code below:

```
pd.DataFrame(w2v_model.wv.most_similar('ron'), columns=['term', 'sim'])
```

	term	sim
0	hermione	0.871872
1	ginny	0.838262
2	lavender	0.772547
3	seamus	0.769468
4	hagrid	0.757866
5	luna	0.752300
6	neville	0.750800
7	cho	0.741290
8	ernie	0.726092
9	grinning	0.724408

His associated words are all weird words unlike the words that were returned with Voldemort, making me think that Ron either is a part of a group with really weird names, or maybe he uses a lot of spells. My inclination is that he uses spells because most of these words are not even English. But then again, I do not know the language used in Harry Potter spells, so this can be an interesting area of further investigation. Next, we will take a look at how similar each of the books are by using hierarchical clustering.



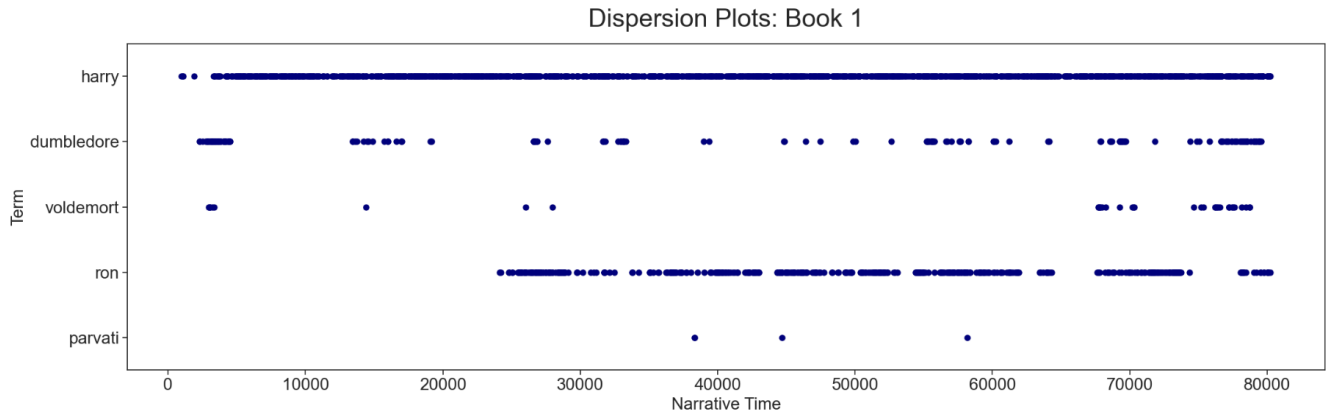
Correlation Heatmap

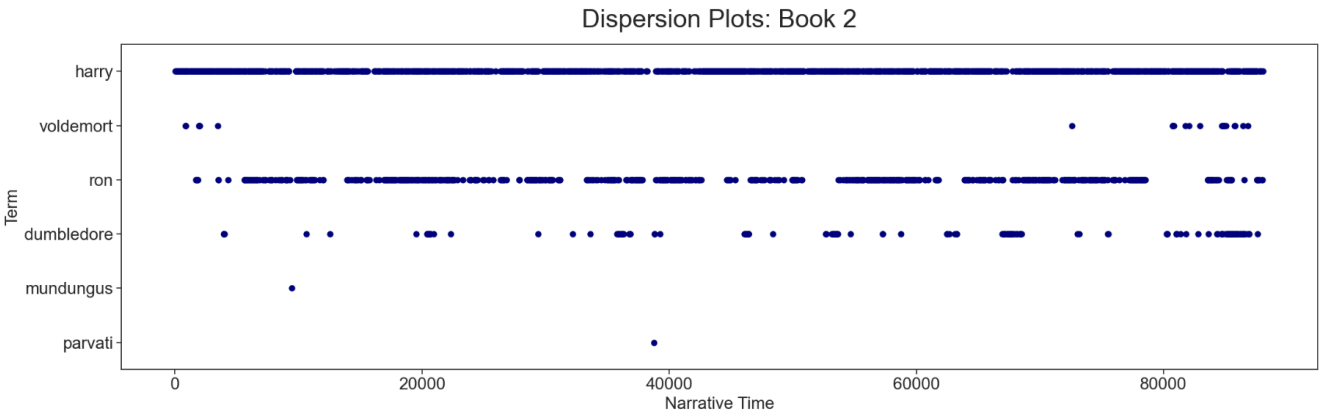
book_id	1	2	3	4	5	6	7
book_id							
1	1.000000	0.209466	0.142343	0.114144	0.069146	0.005537	-0.027950
2	0.209466	1.000000	0.176609	0.109546	0.071958	-0.030763	-0.066422
3	0.142343	0.176609	1.000000	0.099792	0.073522	0.012533	-0.043966
4	0.114144	0.109546	0.099792	1.000000	0.129329	0.026931	-0.038104
5	0.069146	0.071958	0.073522	0.129329	1.000000	0.104032	-0.055005
6	0.005537	-0.030763	0.012533	0.026931	0.104032	1.000000	0.034809
7	-0.027950	-0.066422	-0.043966	-0.038104	-0.055005	0.034809	1.000000

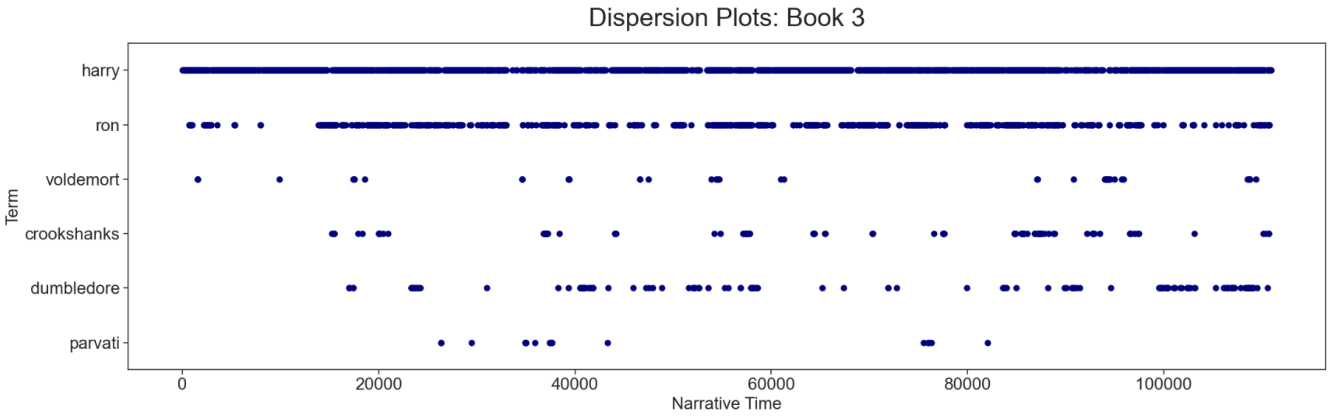
From this dendogram, we can see that the third and seventh book are the most similar, and they form a cluster with the 5th book as well. This is interesting because our hierarchical model found that every other book was similar for this cluster. Up at the top of the graph, we can see that the first 2 books also form a cluster. This intuitively makes sense, because JK Rowling likely wrote in the same style during her first few books, where she can then switch up the tone or bring in new characters as the plot develops. This model also found that books 4 and 6 were similar, which can be an area for further investigation. What is interesting is that when I looked up the order of the books online to interpret the dendrogram, I found the books in order with their movie pictures on the front. The first two movie covers seemed to have many more characters in the background, whereas in the later movies there seems to be only the 3 main characters. This matches with our dendograms, so maybe the Harry Potter series follows something along the lines of Lord of the Rings where at first there are many characters, but then by the end a majority of the story follows Frodo and Sam.

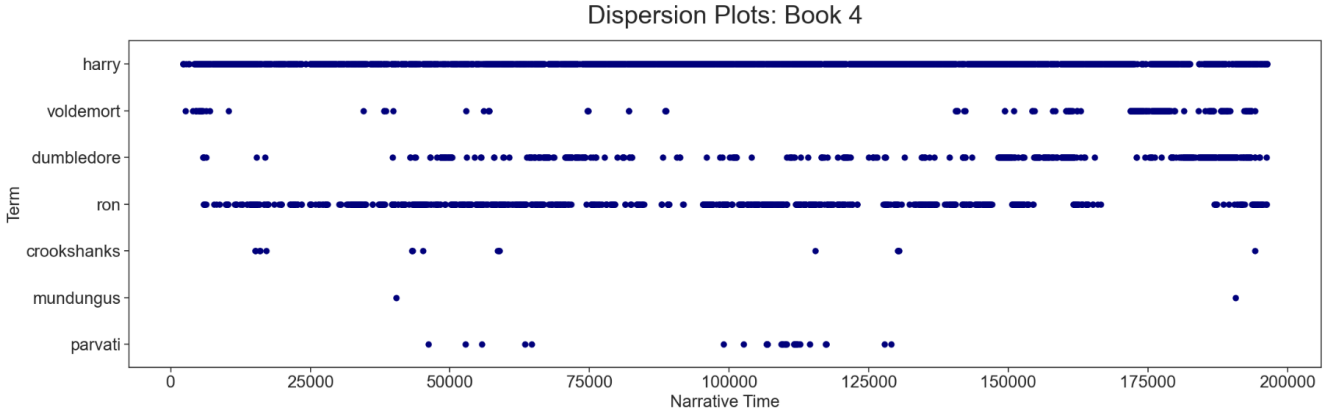
Our correlation heatmap on the books agrees with the dendrograms that book 1 and 2 are related, but does not agree well with the other observations. This is likely because this heatmap was calculated using the top 1000 terms from the tfidf, whereas the HAC is calculated using our bag of words.

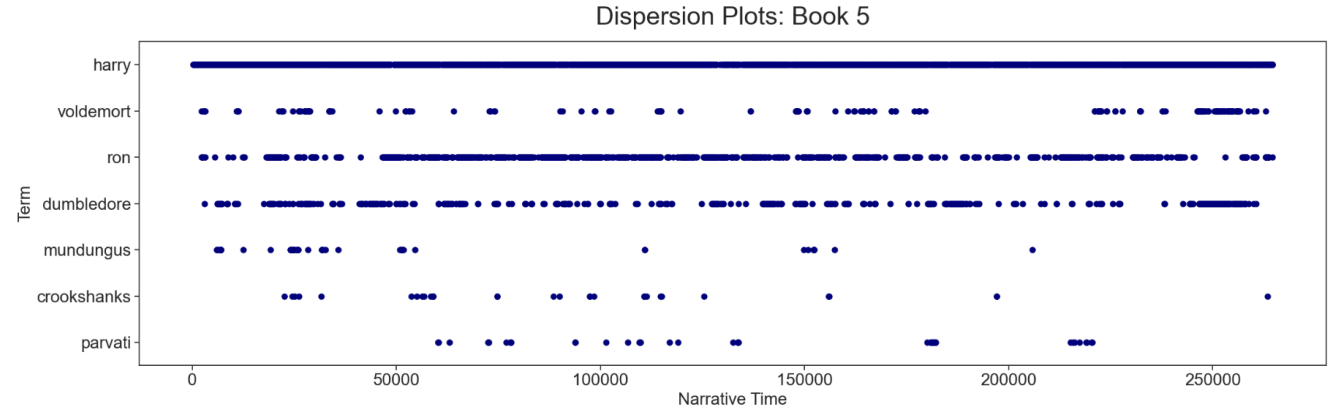
Next I will explore how some characters from our first exploration (voldemort, dumbledore, ron, and parvati) plays into the sentiment of each book.

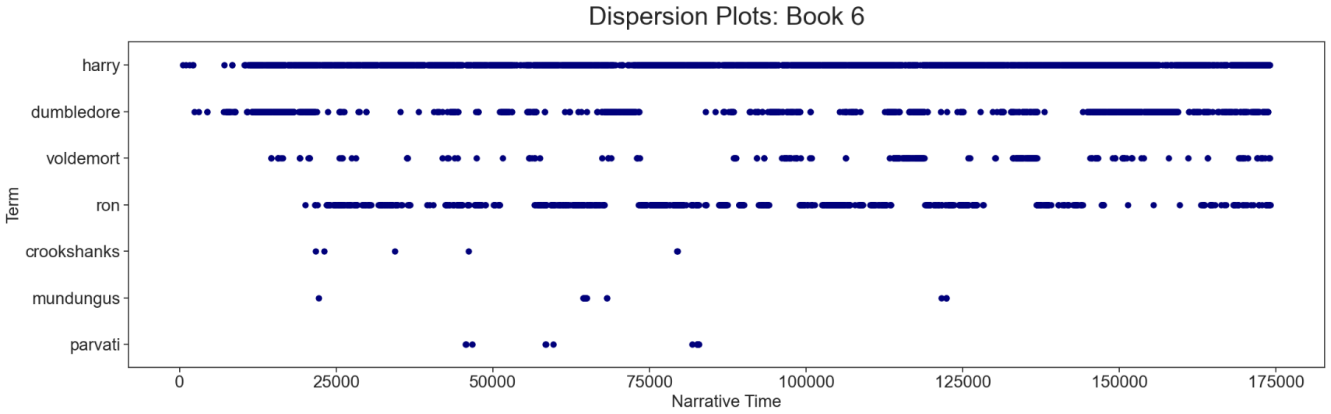

Dispersion Plots: Book 1
harry
dumbledore
voldemort
ron
parvati
Term
0
10000
20000
30000
40000
50000
60000
70000
80000
Narrative Time

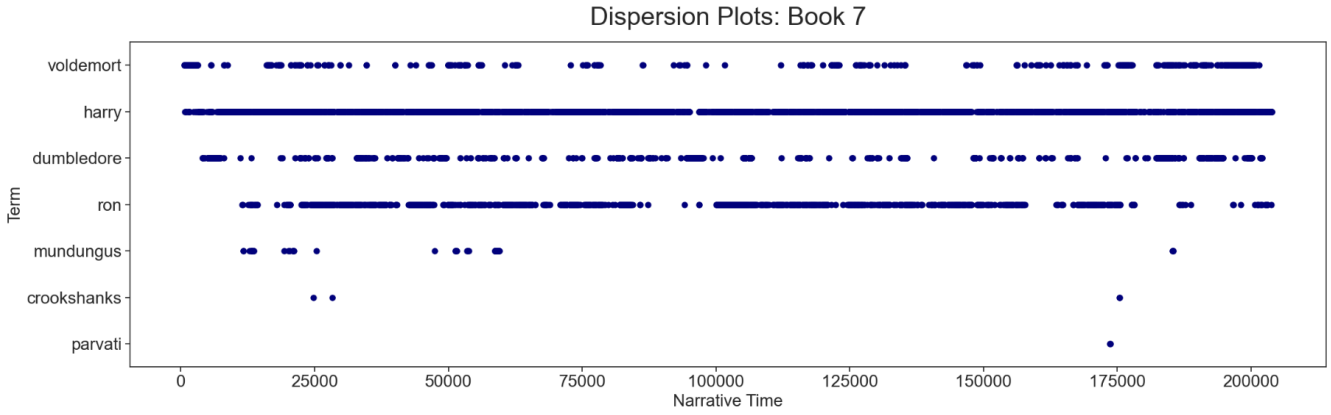

Dispersion Plots: Book 2
harry
voldemort
ron
dumbledore
mundungus
parvati
Term
0
20000
40000
60000
80000
Narrative Time


Dispersion Plots: Book 3
harry
ron
voldemort
crookshanks
dumbledore
parvati
Term
0
20000
40000
60000
80000
100000
Narrative Time


Dispersion Plots: Book 4
harry
voldemort
dumbledore
ron
crookshanks
mundungus
parvati
Term
0
25000
50000
75000
100000
125000
150000
175000
200000
Narrative Time







It is apparent that Harry is the main character of all the books from the dispersion plots. Ron and Dumbledore seem to also be a somewhat main character as they are alongside Harry in parts that he is mentioned. The Crookshanks appear in book 3, but then are mentioned scarcely, alongside the mundungus who are introduced in book 2 but then make their appearence for a second time in book 4. More interesting, Voldemort is mentioned rarely in book 1, but then appears at only the beginning and end of book 2. This indicates that there was probably a cliffhanger in book 2, as in book 3 he starts to appear more regularly. He appears in somewhat regular time intervals in book 5 and 6, making me think that JK rowling probably cuts to his plot somewhat frequently throughout these books. In the last book, he is mentioned very often most likely due to the rising action and climax of the plot.

5. Interpretation

Although I have never read a Harry Potter book or watched the movies, I feel as though I have learned about the books through this project. The correlation heatmap and hierarchical clustering helped lead me to see that there is a divide in JK Rowlings books, as the first 2 books are similar in language compared to the rest, according to the models. From the t-SNE plot, I was able to get an idea of what the character names are, which I was then able to put into the `most_similar()` function to learn more about each character. I confirmed that Voldemort is bad based on this, and then expanded to other characters like Harry or Ron to see what words are most similar to theirs based on our model. Finally, I used dispersion plots to see the occurence of characters across each book. This shows me how the plot in Harry Potter works, with the story being about him and some other main characters (Ron and Dumbledore), and the evil Voldemort appearing in regular intervals in most of the books, with his presence increasing until the last book. I found my analytics to be interesting and insightful, and am happy with what I have learned about a novel I have never read. Now that I am done, I might have to go watch the film with my project group or read the books this summer.

Use `jupyter nbconvert FINAL_REPORT.ipynb --no-input --to html` to convert to html