STUDENT NAME: Nilay Anand

ASSIGNMENT: EAI6010 Final Assignment - Topic Modelling

Topic Modelling with LDA

Summary

The three books provided for analysis: namely Twenty Thousand Leagues under the Sea, The War of the Worlds and Wuthering Heights. All three belong to different genres. Twenty Thousand Leagues under the Sea is an adventure, science fiction novel with a set of characters exploring under the ocean. War of the Worlds is also science fiction novel but more oriented towards mystery and survival, it deals with a family trying to survive the destruction brought on by the arrival of Martians on planet earth. Finally, Wuthering Heights is a gothic tragedy dealing with social interactions between individuals and family, it has themes of revenge, love, and betrayal. All three books have different topics and themes compared to each other.

After importing the books and dividing them by chapters, I built a document term matrix to be used for LDA. Using a 3 topic LDA model, the chapters were assigned to different topics. LDA was able to differentiate the topics well, which was to be expected as the themes of the books is very different. After plotting top 5 words in each topic (see Appendix Figure 1), I concluded that topic 1 was assigned to chapters from Twenty Thousand Leagues, topic 2 to Wuthering Heights and topic 3 to War of the Worlds. The top words in each book are distinct, the only common word was “one” in topic 1 and topic 3. It has highest frequency in topic 3 and 5th highest in topic 1.

On further inspection using the “gamma” parameter and visualising how the chapters of each book was assigned (see Appendix Figure 2), it was clear that chapters were assigned their respective topics for each book except for one chapter from The War of the Worlds being assigned to Wuthering Heights with a probability close to 50%. This is explainable as The War of the Worlds also deals with interaction between individuals within and without a family. Specifically, the word “brother” has a high frequency in The War of the Worlds and Wuthering Height which relates to themes of family in each book.

Finally, after observing the Confusion Matrix (see Appendix Figure 3), it could be concluded that the model was able to assign chapters with a high accuracy to each Topic. There are less than 10% assignations in the wrong Topic. A few chapters from War of the Worlds have been assigned to Wuthering Heights and an even fewer from Twenty Thousand Leagues under the Sea have been assigned to War of the Worlds. LDA was able to easily come to a consensus on the topics in each chapter and assign the correct one to topics associated with each book.

Ideally the model should have been able to assign the chapter to each topic with a hundred percent accuracy as the themes, settings and characters are very different from each other. LDA is a confirmatory topic model, we select the number of topics in which the documents are to be divided and then check if the topics are correctly assigned based on the content of the documents.

Some improvement can be made to this approach. Another way to index is using Term Frequency (TF) * Inverse Document Frequency (IDF). The problem with LDA is that it is unable to classify high frequency between documents. TF-IDF calculates the term frequency in each document and calculate the occurrence of term per document. Using TF-IDF gives a lower ranking to words that are common between documents which will highlight the unique words that are useful for assigning topics for the documents.

Detailed Findings

Creating the document term matrix (DTM)

Using cast_dtm(), I made the document term matrix for the chapters in each book

```
<<DocumentTermMatrix (documents: 107, terms: 16255)>>
Non-/sparse entries: 86935/1652350
Sparsity           : 95%
Maximal term length: 18
Weighting          : term frequency (tf)
```

Using LDA to create a 3-topic model

Creating a 3-topic model using LDA, we get topics, words in topic and the beta value which gives the degree association of each word with the topic.

topic <int>	term <chr>	beta <dbl>
1	brother	3.636958e-53
2	brother	5.880853e-04
3	brother	2.993568e-03
1	said	6.542973e-03
2	said	6.919754e-03
3	said	4.895905e-03
1	heathcliff	6.424428e-58
2	heathcliff	7.272050e-03
3	heathcliff	6.047084e-36
1	captain	1.222551e-02

Using dplyr's top_n() function to find the top 5 terms within each topic and visualizing

The output gives the top 5 words in each topic with their respective beta value. We use this to visualize the words in each topic (see Appendix Figure 1).

topic <int>	term <chr>	beta <dbl>
1	captain	0.012225506
1	nautilus	0.010480290
1	nemo	0.006993578
1	sea	0.006989581
1	one	0.006782637
2	heathcliff	0.007272050
2	said	0.006919754
2	linton	0.005976554
2	catherine	0.005803821
2	mr	0.005389259

Determine which topics are closest in their association with each document using the "gamma" parameter

Using gamma parameter of output from LDA, we see the association of each document with the topics.

Using the gamma values, I plot the topics and document to visualize the association of each chapter with the topic (see Appendix Figure 2). It is observed that all chapter from all three books have been assigned correctly apart from one chapter from War of the worlds being

assigned to topic 2 with a gamma ~0.5. Some topics in Twenty Thousand Leagues under the Sea are in topic 3 but have very low gamma so have not been assigned to the topic

document <chr>	topic <int>	gamma <dbl>
The War of the Worlds_16	1	9.882932e-06
The War of the Worlds_24	1	8.238822e-06
Wuthering Heights_10	1	7.284734e-06
Twenty Thousand Leagues under the Sea_26	1	9.999721e-01
Twenty Thousand Leagues under the Sea_36	1	9.999706e-01
Wuthering Heights_21	1	7.406582e-06
Twenty Thousand Leagues under the Sea_27	1	9.999703e-01
Wuthering Heights_27	1	1.012667e-05
Twenty Thousand Leagues under the Sea_21	1	9.999704e-01
Twenty Thousand Leagues under the Sea_29	1	9.999669e-01

Developing the "consensus" topic for each book

Grouping by title we develop a consensus for each topic and the results show that topic 1 has been assigned to Twenty Thousand Leagues under the Sea, topic 2 to Wuthering Heights and topic 3 to The War of the Worlds.

consensus <chr>	topic <int>
The War of the Worlds	3
Twenty Thousand Leagues under the Sea	1
Wuthering Heights	2

Words assignment for each topic

The augment function gives us word assignment in each topic, we see that "brother" is frequently used in The War of the Worlds and Wuthering Heights. This could have been a reason for incorrect classification of few chapters.

title <chr>	chapter <int>	term <chr>	count <dbl>	.topic <dbl>
The War of the Worlds	16	brother	50	3
Wuthering Heights	10	brother	5	2
Wuthering Heights	17	brother	5	2
The War of the Worlds	14	brother	26	3
Wuthering Heights	9	brother	4	2
Wuthering Heights	12	brother	2	2
Wuthering Heights	23	brother	1	2
The War of the Worlds	15	brother	1	3
Wuthering Heights	8	brother	1	2
Wuthering Heights	3	brother	2	2

Confusion matrix for all the topics

The confusion matrix (see Appendix Figure 3) gave us a clear picture of the assignment of chapter to each topic. Wuthering Height had to most accuracy with all chapters of the book being assigned to its topic. A few chapters (about 10%) of The War of the Worlds were assigned incorrectly to Wuthering Heights and about 5% chapters from Twenty Thousand Leagues under the Sea were incorrectly classified to The War of the Worlds

Appendix

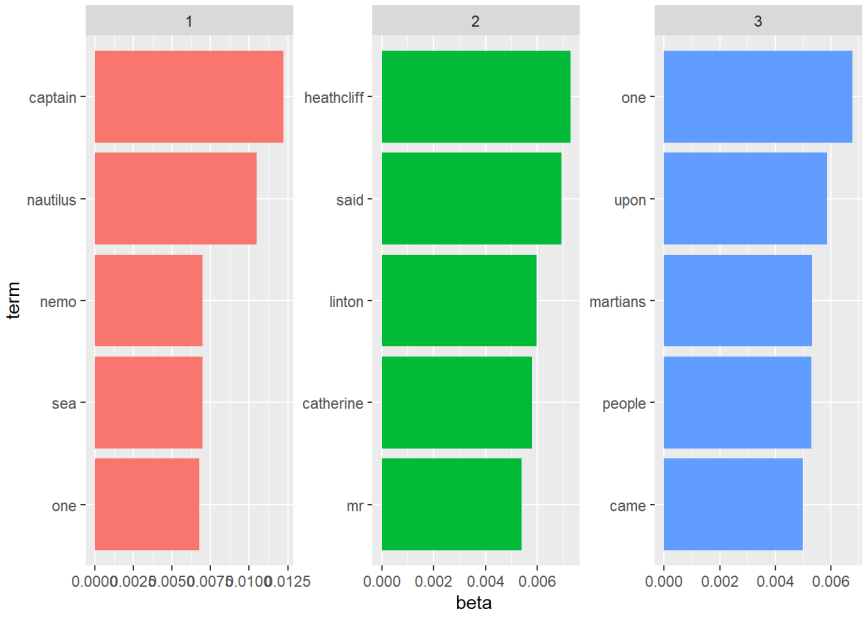

Figure 1: top 5 terms in each topic

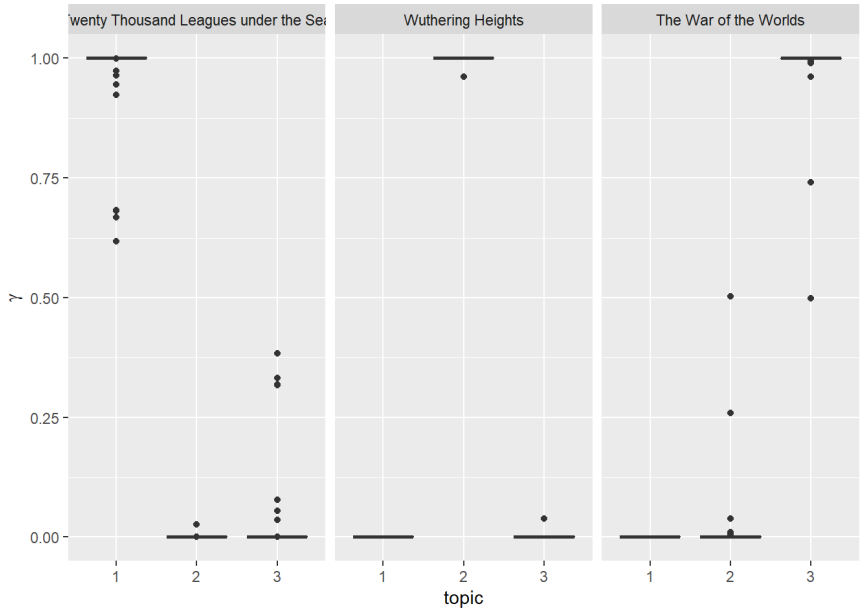

Figure 2: Per Topic Probability

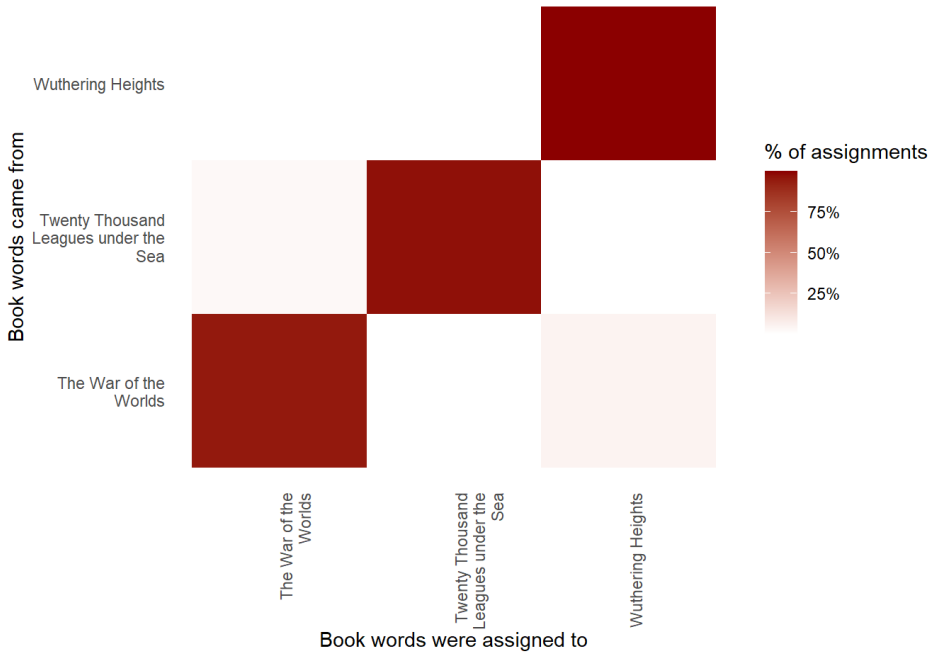

Figure 3: Confusion Matrix

References

[1] Wang, Yi (2008, August). *Distributed Gibbs Sampling of Latent Topic Models: The Gritty Details.* https://cxwangyi.files.wordpress.com/2012/01/llt.pdf

[2] Scott, W. (2021, September 26). *TF-IDF for Document Ranking from Scratch In Python On Real World Dataset. Medium.* https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089.