Introduction

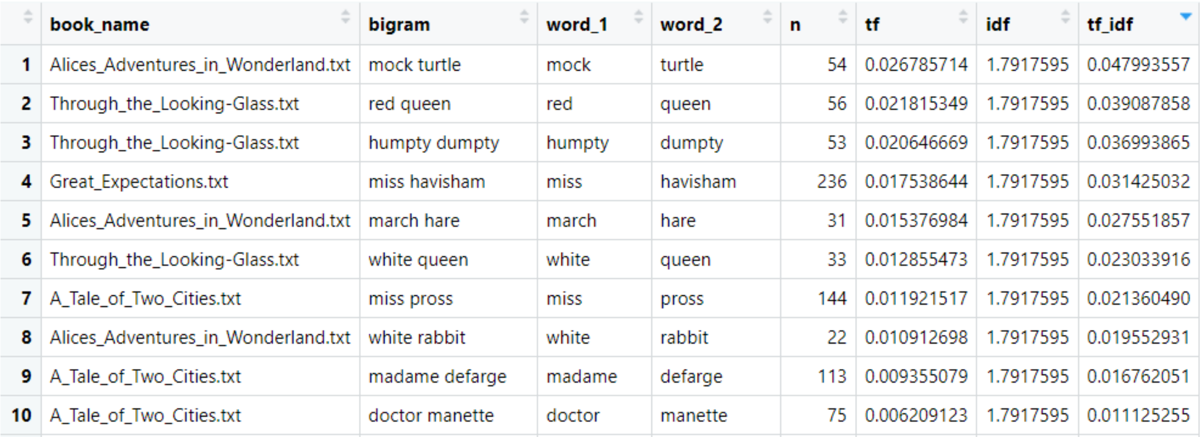
Latent Dirichlet Allocation (LDA) is a statistical model or form of topic model which is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modelled as Dirichlet distributions. Each document is a mixture of various topics and each topic is considered as a mixture of various words. LDA assumes documents are produced from a mixture of topics, which are in turn generated from a distribution of words.

Provided are 6 different books each having multiple chapters. The task is to perform a LDA model on the books and provide evaluation and analysis based on it.

Data preprocessing

The unwanted texts from the text files of the books were removed manually and the files were cleaned just by keeping the chapter wises content. After separating the lines book-chapter wise all the blank (NA) lines were filtered out.

1. Term Frequency – Inverse Document frequency



	book_name	bigram	word_1	word_2	n	tf	idf	tf_idf
1	Alices_Adventures_in_Wonderland.txt	mock turtle	mock	turtle	54	0.026785714	1.7917595	0.047993557
2	Through_the_Looking-Glass.txt	red queen	red	queen	56	0.021815349	1.7917595	0.039087858
3	Through_the_Looking-Glass.txt	humpty dumpty	humpty	dumpty	53	0.020646669	1.7917595	0.036993865
4	Great_Expectations.txt	miss havisham	miss	havisham	236	0.017538644	1.7917595	0.031425032
5	Alices_Adventures_in_Wonderland.txt	march hare	march	hare	31	0.015376984	1.7917595	0.027551857
6	Through_the_Looking-Glass.txt	white queen	white	queen	33	0.012855473	1.7917595	0.023033916
7	A_Tale_of_Two_Cities.txt	miss pross	miss	pross	144	0.011921517	1.7917595	0.021360490
8	Alices_Adventures_in_Wonderland.txt	white rabbit	white	rabbit	22	0.010912698	1.7917595	0.019552931
9	A_Tale_of_Two_Cities.txt	madame defarge	madame	defarge	113	0.009355079	1.7917595	0.016762051
10	A_Tale_of_Two_Cities.txt	doctor manette	doctor	manette	75	0.006209123	1.7917595	0.011125255

Table 1: Top 10 bigrams overall

Table 1 shows the top 10 bigrams of all the books combined which is sorted by the tf_idf values.

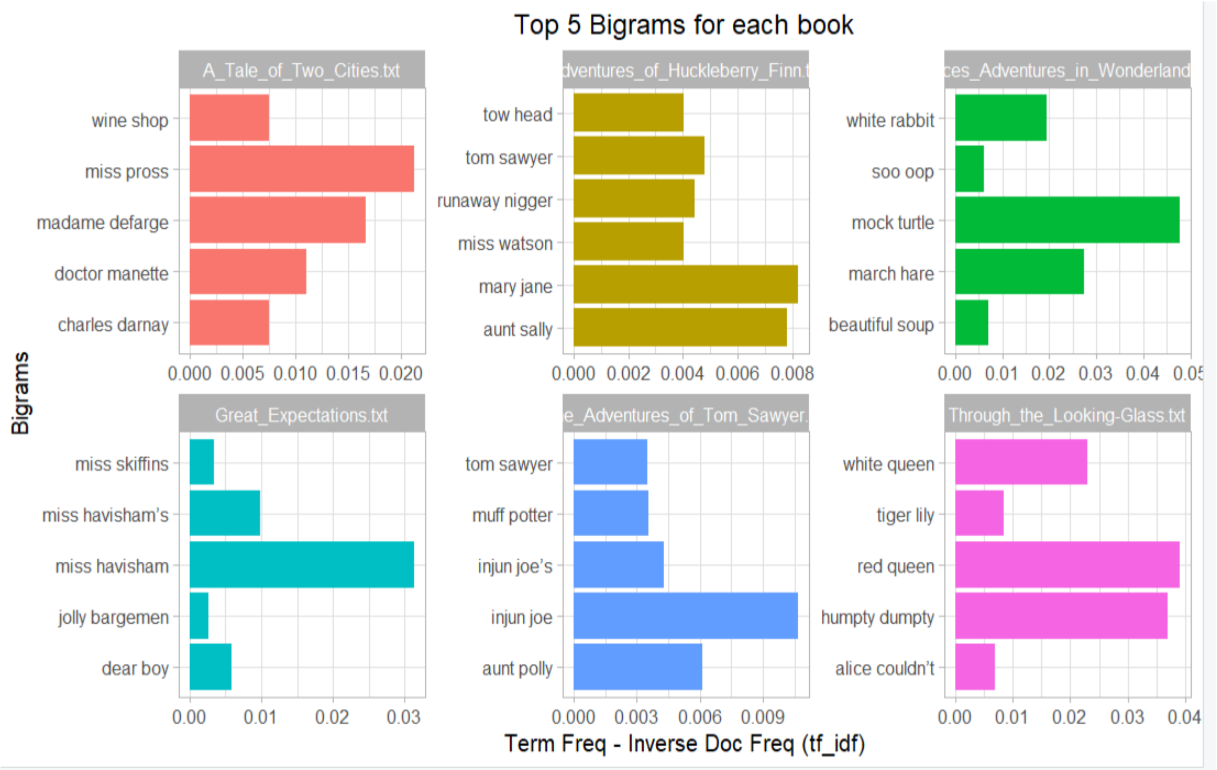


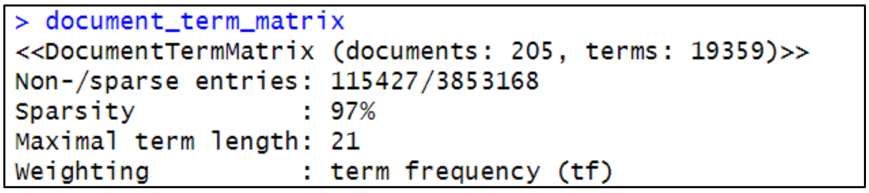
Fig 1: Top 5 Bigrams Plot

The above Fig 1 shows top 5 bigrams of the individual books and their term frequency. It gives us the top consecutive terms and its frequency of occurrence.
In the given books the top bigrams are mostly referring to the characters in the stories like Miss Pross, Tom Sawyer, Injun Joe, Aunt Polly, Mary Jane, Madame Defarge etc. which suggests that there is a lot going around these characters in the stories.

Some bigrams like Alice couldn't in the Through the Looking Glass book tells us about something that the character probably could not manage to do.

Overall, the bigrams with maximum values for term frequencies of are referring to characters implying the involvement of characters in the stories.

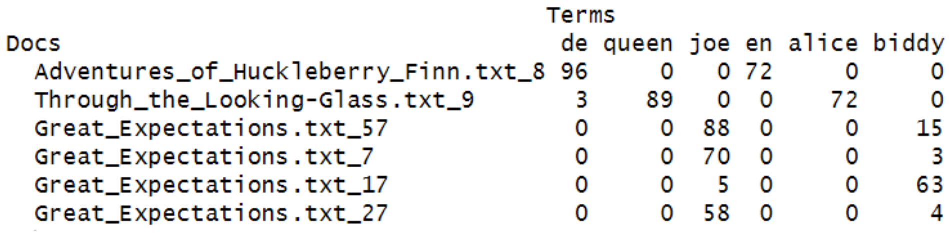
2. Document term matrix



```
> document_term_matrix
<<DocumentTermMatrix (documents: 205, terms: 19359)>>
Non-/sparse entries: 115427/3853168
Sparsity           : 97%
Maximal term length: 21
Weighting          : term frequency (tf)
```

Table 2: DTM details

The DTM consists of 205 documents and 19359 terms. The Non-/sparse entries shows that out of the total entries in the DTM, 115427 are non-zero and 3853168 entries are zeros which gives us a sparsity of 97%. This implies that most of the words do not appear in any of the document. The longest word in the document is of 21 characters.



```
                                      Terms
Docs                                   de queen joe en alice biddy
  Adventures_of_Huckleberry_Finn.txt_8 96     0   0 72     0     0
  Through_the_Looking-Glass.txt_9       3    89   0  0    72     0
  Great_Expectations.txt_57             0     0  88  0     0    15
  Great_Expectations.txt_7              0     0  70  0     0     3
  Great_Expectations.txt_17             0     0   5  0     0    63
  Great_Expectations.txt_27             0     0  58  0     0     4
```

The above output shows the subset of DTM with top 6 rows and columns of the top 6 terms and documents.

3. LDA model

Considering the DTM, a three topic LDA model is implemented on the results which will help in examining the top words in each topic. Based on the associated words within the topic an analysis can be made on how the words in each topic are coherent. Below visualisations are created to analyse more on LDA.

i. Beta Plot

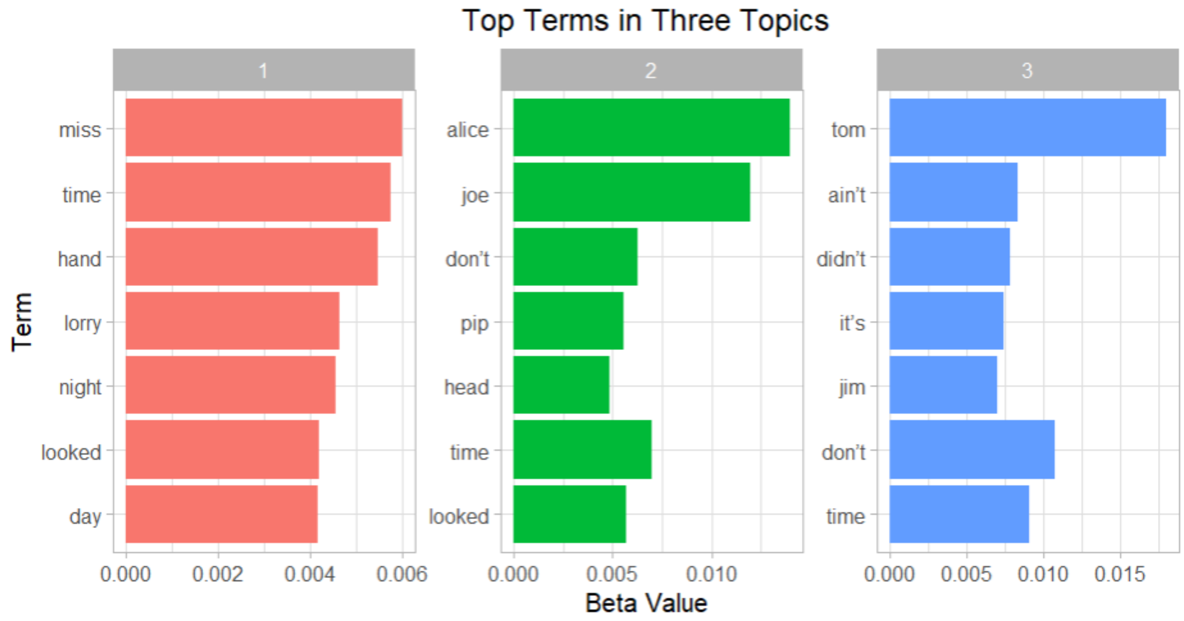


Fig 2: Beta plot – Top terms in three topics

Topic 1 (in red) with words "miss", "time", "hand", "lorry", "night", "looked" and "day" relates to a daily life of someone or a routine. It indicates a more of a day and night story with some descriptions of the actions because of the terms looked and hand.

In Topic 2 (in green), the words Alice and Joe representing characters along with words like "pip", "looked" which are verbs and "time, "head" and "don't" which suggest a third person narration or a dialogue-based story.

The third topic (in blue) the terms "tom", "ain't", "didn't", "its", "Jim", "don't" and "time" also portray a focus on a character named Tom and Jim. The other terms indicate a spoken language or a conversation happening in the document with more contraction words like don't, didn't and ain't.

Apart from individual analysis of the topics the term "time" which seems to appear in all the topics suggest time related actions happening in the stories. However, topic wise its probability varies.

The term "looked" appearing in two different topics tell us that there is shared narrative element different chapters of stories in the books.

After evaluating the topics, the next step is to find out the frequency of the topic per document which is done by plotting gamma plot.

ii. Gamma Plot

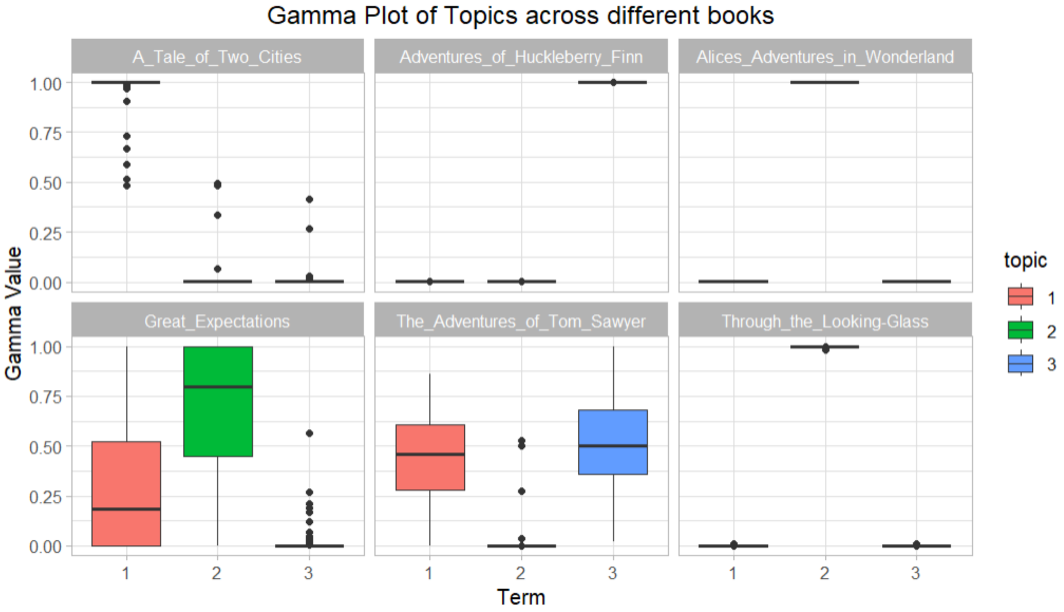


Fig 3. Gamma Plot

The above plot gives us a distribution of the above three topics across six different books with respect to their gamma values.

"A Tale of Two Cities": Topic 1 has high median gamma value with probability 1, indicating that the terms "miss", "time", "hand" appears in most chapters on the book 'A Tale of Two Cities'. Whereas Topics 2 and 3 also show zero probability but with some outliers, suggesting there might be some chapters which has the associated terms.

"Adventures of Huckleberry Finn": Topic 3 terms (Tom, don't, ain't) have the most probability i.e. median value in the book Adventures of Huckleberry Finn which suggests a conversation/dialogue with respect to Tom. While Topic 1 and 2 has zero probability in the book with no outliers.

"Alice's Adventures in Wonderland": Topic 2 terms (Alice, Joe, looked) have the most probability in the book Alice's Adventures in Wonderland which makes sense as the book is about story of Alice. While Topic 1 and 3 has zero probability in the book with no outliers.

"Great Expectations": The clear box-plot with a high median value shows significant presence of Topic 2 terms in several chapters of the book. Topic 1 terms also appear in the book with a probability ranging from 0 to 50% with a median value of 25% while Topic 3 terms appear least in this book.

"The Adventures of Tom Sawyer": There is a balance between Topic 1 and 3 terms and they both have a probability of occurrence of 50%. Whereas Topic 2 shows a low median gamma value, implying it doesn't appear in the book though there are few outliers which might suggesting there might be some terms that may appear.

"Through the Looking-Glass": Similar to Alices Adventure in Wonderland Topic 2 is predominant in this book which suggests the story of Alice in this book as well. On the other hand, Topic 1 and 3 has the least probability of occurrence.