Our beyond topic is LDA, which is implemented to find common topics/themes of a book.

There are 7 books in total, and we want to find the topics in each book.

The example on the website has 4 main parts:

1. Specify the number of topics (k) we want to find. This must be chosen by us, there is no way to run an elbow plot to know how many topics would be optimal. Let’s say we want to find 3 themes.

2. Calculate the probability of a term being generated from a topic. This is the formula:

p(word w with topic t) = p(topic t | document d) \* p(word w | topic t)

from LDA notes. The goal is to find representative words for each topic (theme). From there, we can name the topics, such as “action, horror, adventure, etc.”.

3. Calculate the estimated proportion of words from a document that are generated from a topic. This is the formula:

p(topic t | document d)

Notice that this is the first part of the formula above. The goal of doing this is to know which topics are associated with each document (each chapter).

4. Assign each word in each document to a topic. The more words in a document are assigned to that topic, generally, the more weight will go on that document-topic classification.

IF YOU DON’T CARE ABOUT THE TECHNICALITIES, START READING FROM HERE!

For our purpose, we will only do step 1, 2, and 3 (doing 3 is optional). That is, if we only want to know what are the genres of each book. Note that, however, if we specify to find 3 topics, a document (chapter) will ALWAYS give back 3 topics, even if the words have a 0% probability of coming from topic 2 and 3.

Assuming that we still want to know the genres of EACH book, here’s our general work process:

- Take a book, divide into chapters

- Specify number of topics (genres) to find. Let’s work with 3

- Do step 2 to find the representative words from each topic. Then we can name the genres. Note that although every book (corpus) will have 3 topics, the specific names that we give to the topics might be different. We already have 3 topics, so we can stop here.

- Do step 3 to know how strongly each topic is associated with a book. That is, if we want to know the detailed distribution of genre (like 30% genre A, 40% B, and 30% C), we need this step. However, what we are having right now is how strongly each topic is associated with a chapter (the per-document classification code chunk). So, I propose we run the per-document classification as a per-book classification.

- Repeat the process for all 7 books.

- In the end, what we have will be something like this: Book 1’s 3 common genres are action, adventure, and romance. Among the 3 genres, 20% of the words in book 1 came from genre action, 79% came from adventure, and 1% from romance. So, although we asked the algorithm to explore 3 genres, there are only 2 prevalent genres here.