**Topic Modeling and the Writing Process**

Kevin Russell

**Introduction**

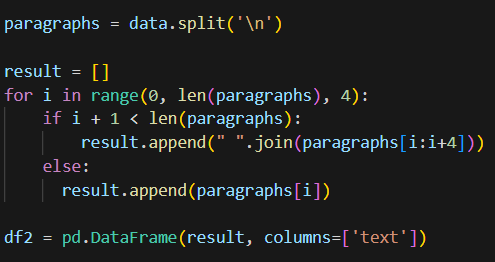
Topic modeling is a process that may seem a bit like magic to those unfamiliar with it. Through a mechanism called Latent Dirichlet Allocation (LDA), it can extract topics from documents that comprise a corpus– that is, it pulls out groups of related words from chunks of text. It is what is known as an unsupervised learning method, meaning that there are no predefined outcome variables or metrics that we are aiming to predict. Essentially, topic modeling pulls topics out of thin air– though, depending on the construction of the model, those topics may or may not be meaningful.

This process has a number of uses in many different contexts. Topic modeling can give great exploratory analysis for the study of text data, lending itself to possible organization of big ideas, summary of corpora through various lenses, and eventual classification of documents into different groups. Generally, topic modeling is performed early in text study, and it is done to obtain some surface-level understanding of what is discussed and how it is discussed in often lengthy corpora. This broad list of applications means that linguists, statisticians, computer scientists, and historians alike may make use of this tool.

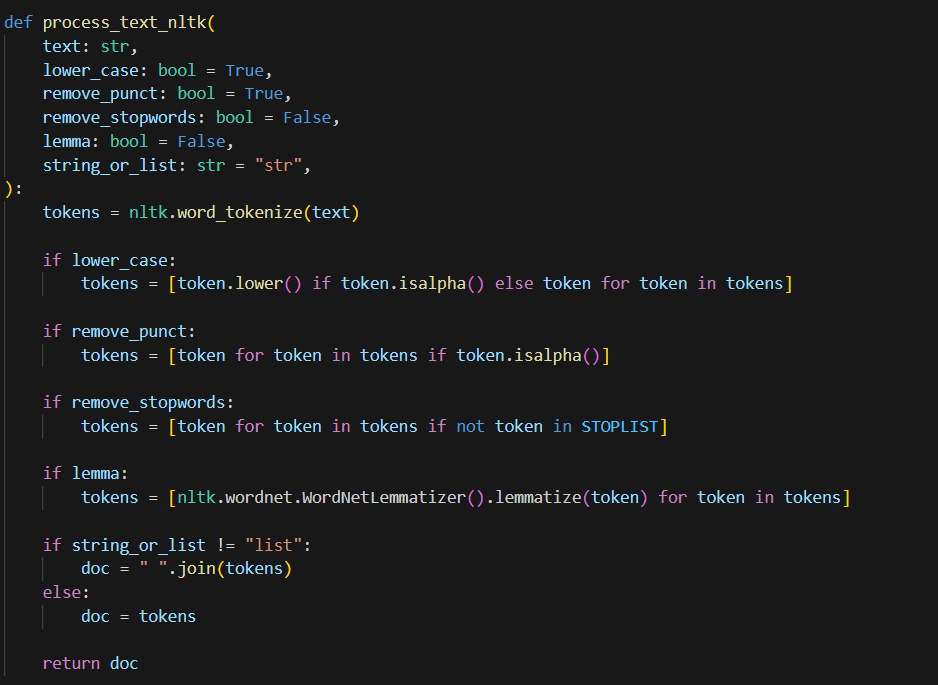
LDA works under the assumption that words in the studied corpus are generated independently from some prior distribution. This bag-of-words understanding throws out the word association that comes with sentence structure or parts of speech or frequent embeddings– each word may well be thrown in a bag, shaken up, and taken in any order. LDA also assumes that each document in the corpus is generated by one or many topics– groups of related words. In this way, LDA works backwards– it probabilistically determines what set of topics is most likely to have generated the documents in the corpus. The output of this process will be a number of topics that essentially amount to lists of words with probabilities attached to them– words with higher probability ratings in the list are more reflective of the overall topic, and therefore are more useful in determining how the model grouped the topic together. If “bus”, “car”, and “drive” were the top three words by probability in a topic generated by LDA, we may have reason to believe that topic was about transportation.

Our example data source to demonstrate the power of topic modeling as it relates to the writing process comes from two drafts of novels that I wrote over the course of the past two years. Draft 1 is around 80,000 words long, and draft 2 is around 50,000 words long. Draft 2 is a follow-up to the first, so most of the characters are the same in both drafts. We will attempt to extract insights into the writing process and style of these drafts using topic modeling and LDA.

We are not yet ready to create our model, however. First, we’ll have to preprocess the text. One of the most important decisions is document length. Documents comprise the corpus, and to build the model, we’ll have to decide just how many documents we want to have. I decided to split documents into four paragraph chunks, but I could have gone with single paragraphs or chapters, as two other options. More documents might give better topics, but they would give less information about the topics of specific portions of the text.

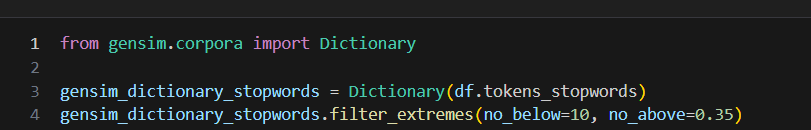


Next, we’ll remove punctuation and stop words. We’ll also convert all words to lowercase. Punctuation won’t be important in pulling raw words out of our documents to make topics, and stop words (like and, the, or a) may muddy the topics the model derives. We’ll use a very versatile function to accomplish this.



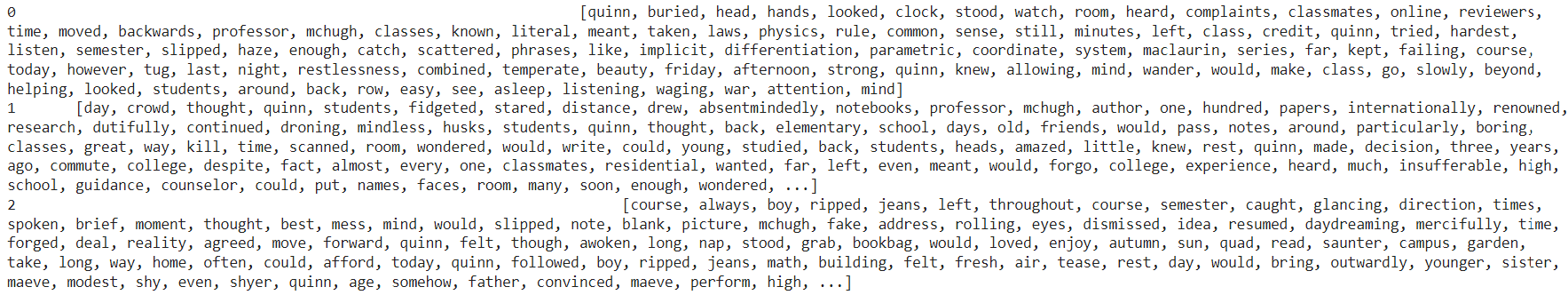
This process\_text() function can do other preprocessing steps as well, including lemmatizing words, or turning the words into either a string or a list. Essentially, these steps make the text data easier to work with. The all lowercase words parameter ensures that words that start a sentence, for instance, aren’t counted separately from the same words that are in the middle of a sentence. However, if words may be shouted, or if inflection matters in analysis, it may be wise not to lowercase every word. Lemmatizing treats a word only as its root– helpful if the analyst wants the word “runs” and “running” to be treated the same way.

Finally, we’ll remove words that appear in more than 35% of the documents we have created, or less than 10 total documents. We’ll do this because we don’t want very common non-stopwords to dominate the topics either, and we don’t want the model to pick up on very rare words that may be a fluke relating to a specific passage.



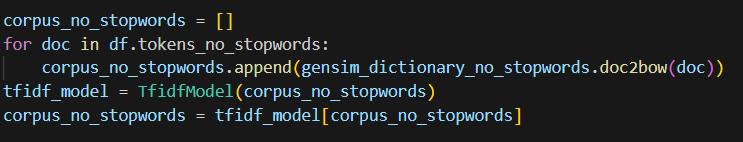
We could have also included bi- or tri-grams, which combine words that frequently appear together, but those preprocessing steps were tested and not found to improve the quality of the final model.

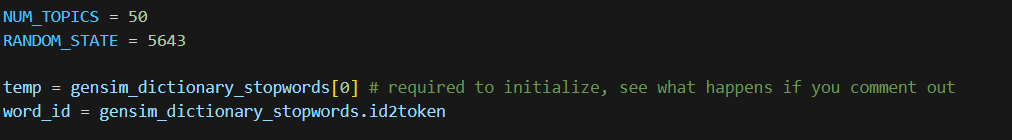
Here’s how the data looks now:



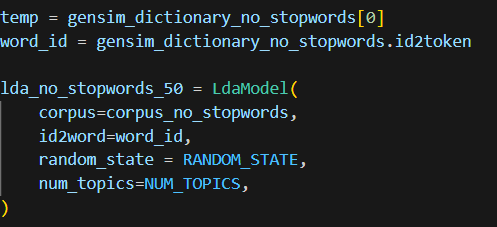
Each row is a long list of words: all lowercase, without stopwords, and filtered so that only the relatively rare ones appear.

Now, we’re ready to begin the modeling process. The first decision we can make here is how to represent the words. We could just use raw counts of words in each document to create the dictionary we’ll use, but we could also use TF/IDF weighting, which assigns larger values to words that are “important” to specific documents. For this exercise, we’ll use TF/IDF.

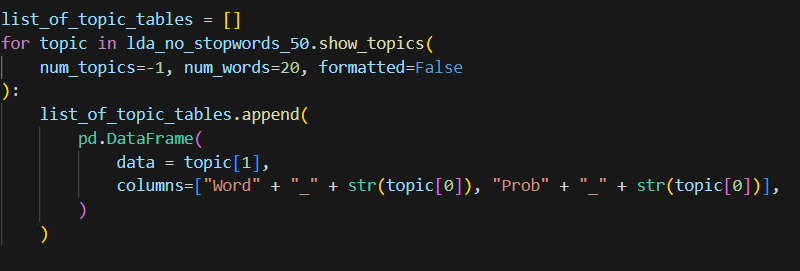


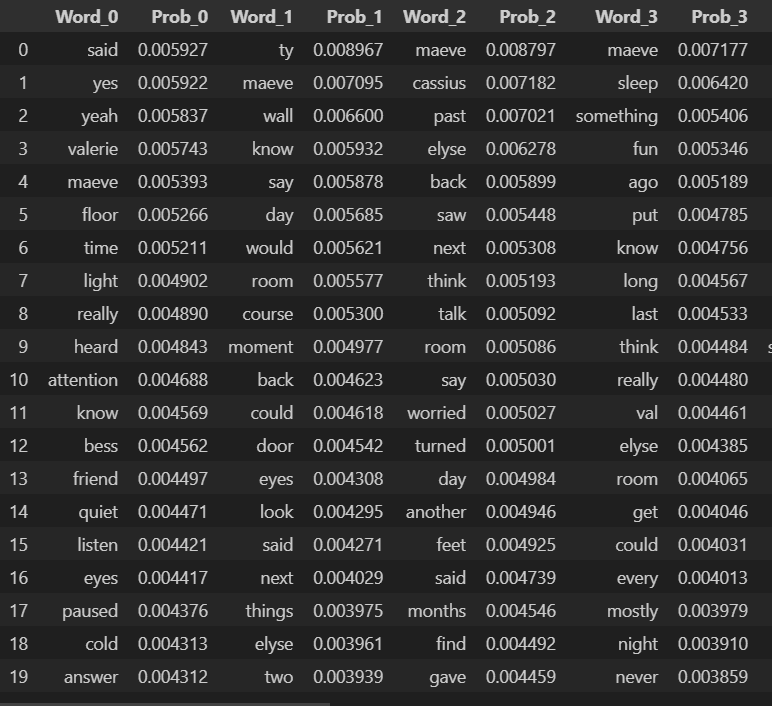
Next, we’ll decide upon how many topics we want. This involves a lot of refining throughout the modeling process– theoretically, less topics might be more focused and better reflective of the corpus, but less might also mean less insight into the nuances of the corpus. Much of topic number selection comes down to what the researcher is looking for from their analysis, and what the topics they are getting look like. I used fifty in the construction of my model. 

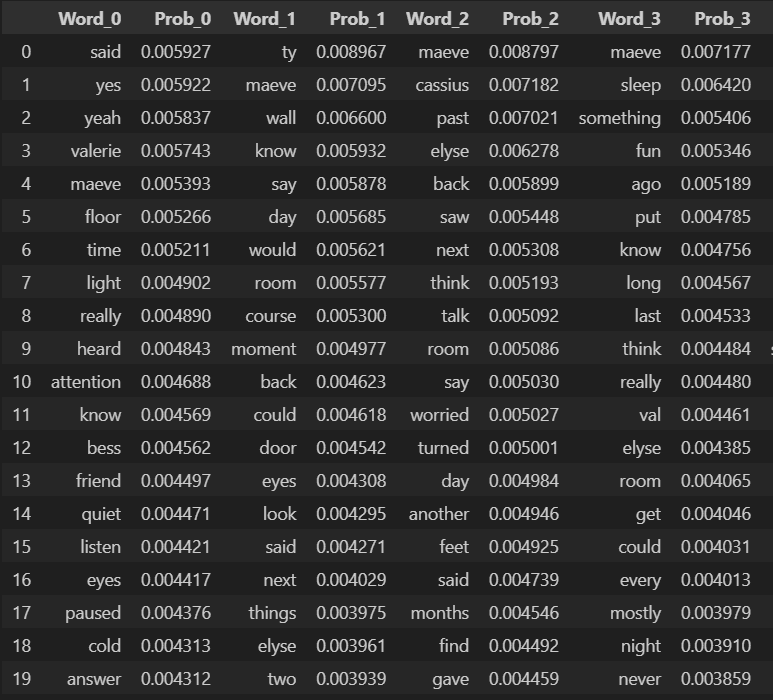
With all of these preprocessing and modeling decisions in mind, we are finally ready to create the model. Here’s how it looks:



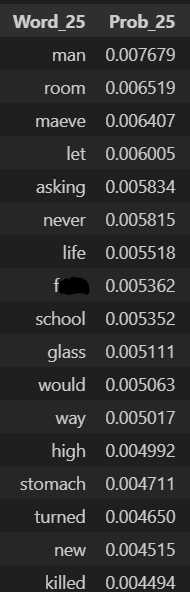
What has this model given us? We can check the topics in a dataframe.

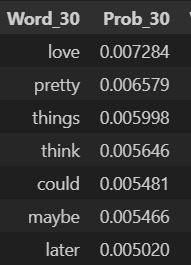


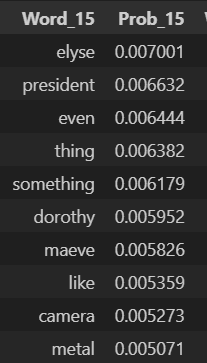
This code outputs the following (for the first few of the fifty topics):



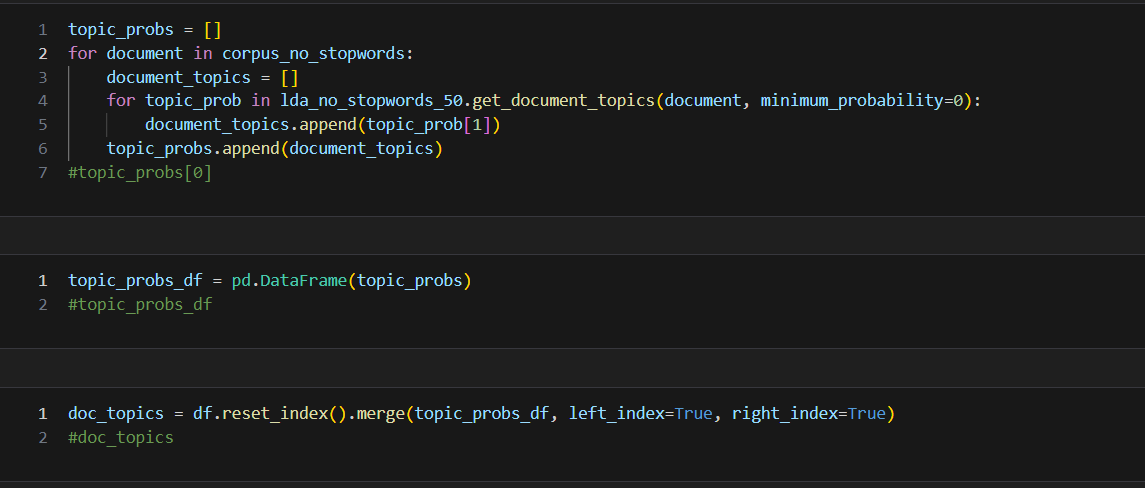
We can glance at these topics and see if they make any sense. Topic 0 seems to have words that indicate conversation, attention, and agreement– said, yes, yeah, heard, attention, know, listen, and answer all seem to fall under that category. There are a lot of characters in these topics, as well– as an author, I know that Topic 2 starting with Maeve and Cassius likely has to do with their relationship. There are some other interesting topics that the model included, such as:

Topic 25: Significant Conflict: words like never, f—, killed, stomach, and turned, indicate that this topic captures areas of high stakes and immediate conflicts. Additionally, because significant events happen in the school one of the characters attends and in Room 137 (the number was filtered out from the preprocessing steps we took) involving glass, I can get a feel for what this topic is about. It’s important to remember that we’re using topic modeling in the context of writing, so authors should use their own knowledge of the story they’ve crafted to understand the topics the model produces. Documents that are given high probability scores in this topic should be analyzed for succinct writing, appropriate emotion from characters involved, and proper establishment of drama. This topic is a great way to see if I, as the author, am writing my big, climactic scenes with enough gravity and effect. 

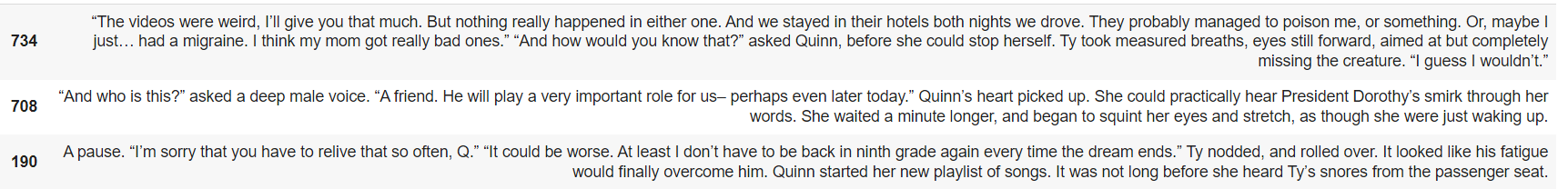
Topic 30: Hedging Words: this is an interesting topic for sure. The model decided to group together words like pretty, think, could, and maybe, along with informal descriptors like love, things, and later. Generally, authors should refrain from using these kinds of words in their prose, as they indicate a lack of precision and conviction in storytelling. They are frequently used in speech, however, so the appearances of these words in dialogue-heavy documents may not be a bad thing. At any rate, it is important for an author, upon seeing this topic, to ensure that these hedging words and informal descriptors are being used appropriately. 



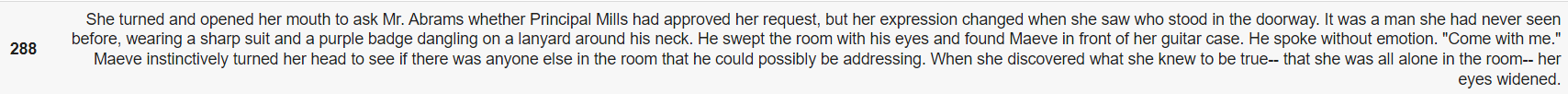
Topic 15: Maeve in Captivity: There are a lot of characters (Dorothy, Elyse, Maeve) and objects (camera, metal) that have to do with what happens to one of the characters later in the book. This topic theoretically should unite some topics about what is happening to Maeve during one section of the book. As an author, this topic is quite interpretable, and it should lend itself to focused document review. I should make sure the passages that are generated by this topic according to the model are consistent– the character attitudes, the setting, and the style should reflect what is going on thematically.

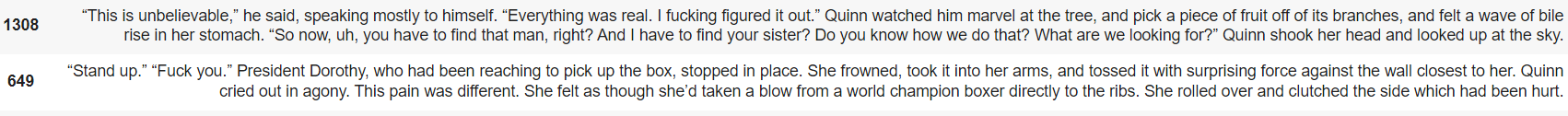
It is very interesting to see the results from our model– it grouped words together in interesting and (in regards to the hedging words topic) unexpected ways. But how do we know these topics are valid? And how can I, the author, use these topics in a way that actually helps my writing? To accomplish both of these feats, we need to dig into the documents that are assigned the highest probabilities of having been generated by these topics. We can say the model is validated if we can generate documents from specific topics that are related in a clear, meaningful way to the top words in each of those specific topics. 

Above is the code to append the topic probability to each document. Each probability will be between zero and one, and the sum of each of the fifty topic probabilities for each document will add up to one. We can then sort by the topic we want to find the documents that should be most closely related, stylistically or semantically, to the topics the model parsed. Let’s look at the top three for topic 30: the hedging words.

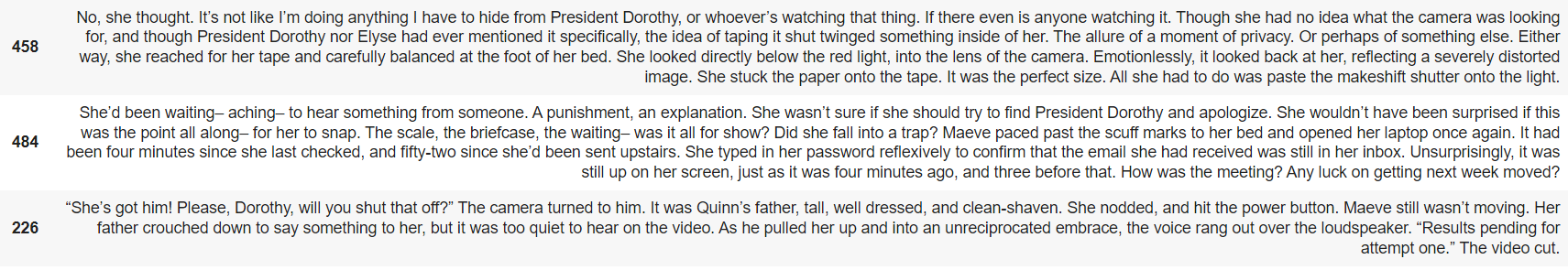


We can see that there is, indeed, a lot of dialogue in these documents– really, probably, maybe, think, guess, perhaps, and could all appear in dialogue as examples of informal speech terms or hedging words. There are also some that do not occur in the course of dialogue, but in the course of the third person narration. These instances should be reviewed– they aren’t always bad or wrong, especially when the narration is from Quinn’s perspective, but they should be reviewed nonetheless. These sentences do seem to validate the topic, and give a fresh lens for an author to view their work. Let’s look at another topic– 25, involving significant conflict.





The top sentences from this topic are, indeed, during very pivotal moments in the book, and involve hatred or pain. The writing here should be sharp and to-the-point, which was not always the case when I examined the top documents from this topic. In this way, we find the model validated and more evidence that the topic model is doing a good job aggregating documents together in meaningful ways.



Finally, we have Topic 15. These top three documents all do, in fact, have to do with the characters and scenario I thought they would– but they come from various perspectives and various points in the course of the two drafts. I would not have grouped these documents together this way myself, but I can see how the model did what it did.

Ultimately, we can see that there is reason to believe that the topic model we’ve generated is worth something to an author– it was able to come up with some meaningful topics, and upon closer inspection, the documents with the highest probabilities of having been generated from those topics made sense. That is not to say, however, that this topic modeling process was without any fault. A lot of the topics were hard to interpret– common words, like said or know or would, that were not stop words made it into several topics. This made many of them difficult to interpret using the top words alone. Oftentimes, there would be some interesting relationship between top documents for a topic, but it would be tough to predict without going into the documents blindly and then looking dozens of words down for the link that made some sense. Also, more preprocessing steps could have been attempted– I tried 20, 50, and 100 topics, but I could have tried more, or I could have limited less common words from the analysis.

Finally, there is the biggest limitation– subjectivity. Topic modeling is very subjective, and tweaking the smallest model attribute or preprocessing step can give wildly varying results. It would have been nice to establish another perspective for topic analysis– another human with deep knowledge of the source material– but as an author in the drafting phase, that is often not possible. Results should be taken with a metaphorical grain of salt, but I find that if they have some meaning to the author, they can still be valuable in a meaningful way.