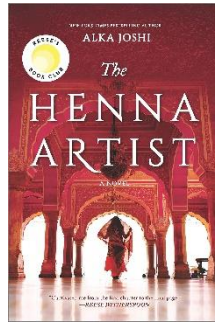Tulsi Shrivastava
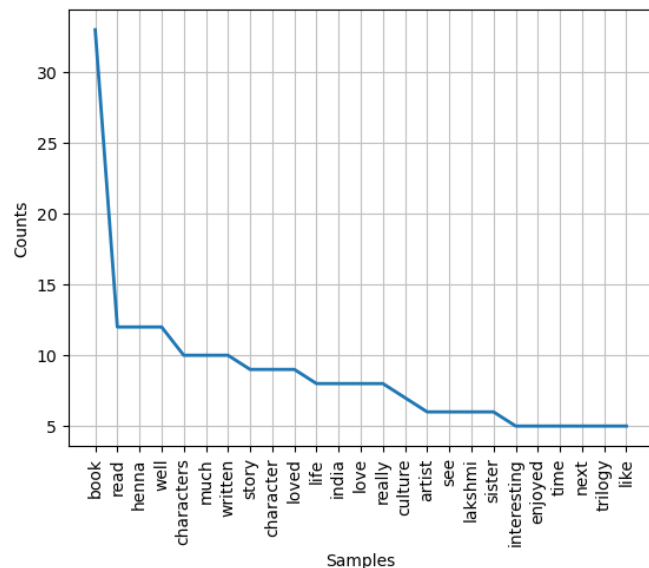
Sentiment Analysis with NLP Python and NLTK

In this project, I analyzed Amazon reviews for the 2020 fiction novel "The Henna Artist", written by author Alka Joshi. First, I collected 25 revies at random from the 22,528 listed on the product site and combined them into one text file. Then, I imported the file into Anaconda Jupyter notebook for analysis using Python and NLTK Library. The product description and reviews can be found at https://www.amazon.com/Henna-Artist-Novel-Alka-Joshi/dp/0778331474/ref=sr_1_1?keywords=the+henna+artist&qid=1680727318&sprefix=the+henna+art%2Caps%2C76&sr=8-1.

Next, I generated the **word frequency counts** and a **word frequency plot** for the reviews, excluding stop words that do not aid in understand sentiment or meaning.

Top 25 Most Frequent Words Used in the Reviews

	token	count
0	book	33
1	read	12
2	henna	12
3	well	12
4	characters	10
5	much	10
6	written	10
7	story	9
8	character	9
9	loved	9
10	life	8
11	india	8
12	love	8
13	really	8
14	culture	7
15	artist	6
16	see	6
17	lakshmi	6
18	sister	6
19	interesting	5
20	enjoyed	5
21	time	5
22	next	5
23	trilogy	5
24	like	5

The list shows the top 25 most frequently appearing words in the reviews. From these, we can make a few conjectures about the book itself:

1. It is expected for the term "book" to appear at the top of the list with 33 appearances, as it is likely being used as a descriptor in the reviews. If we did not know what kind of product we were analyzing, this key term would be a helpful identifier.

2. Similarly, words 2 and 15 echo the title of the book "The Henna Artist" and are likely used as descriptor, explaining what it will be about.

3. The words India", "Lakshmi", "culture" and "trilogy" indicate that the book is centered around the culture of India, has a character named Lakshmi, and is part of a three-part series written by the author.

4. Lastly, even though they are included in the list, words such as "book", "read", "next", and "see" offer little to the overall understanding of the novel or the reviews. Even after removing stop words, it is important to continue refining the list.

Finally, we can conduct a basic sentiment analysis of the review to better understand what readers perceive of the novel. This information can be helpful feedback for publishers, sellers, other readers looking to read the book, and the authors themselves. Knowing your audience is vital in ensuring the success of a novel.

1. Terms such as "love/loved", "interesting", and "enjoyed" carry more weight as they indicate emotions. Readers likely experienced joy after reading the novel. Without looking at the rating score, we can conclude that it is likely highly rated on Amazon.

2. While emphasis words such as "really" and "much" offer little explanation alone, they provide context on the degree to which the sentiment was seen. For example, if combined with the word "enjoyed", "really enjoyed" indicates a higher level of satisfaction than "enjoyed" alone. On a similar note, if we combine the words "well" and "written", we can assume that readers enjoy the author's style of writing. In this scenario, it would be helpful to employ a bi-gram language model to identify terms that go together and contribute to the overall review sentiment.

To conclude, employing NLP methods can help make important conclusions and decisions around the purchase and sale of novels on Amazon. In this scenario, we can suggest that "The Henna Artist" is well-favored among readers from the 25 randomly collected reviews.