Project Overview

I wanted to look at how similar in content and style the Jane Austen and Sir Arthur Conan Doyle work was because these authors wrote in roughly similar time periods but about vastly different subjects and in different formats (Austen focusing on books, and Doyle being more of a magazine writer).

I used term frequency, inverse document frequency (TF-IDF) to determine what were the most "important" words in each text and examining what they said about the style of the book. Most important was determined by finding words that are more frequent in one book than another (so words such as 'the', 'and', and 'or' could be excluded).

Implementation

My initial approach was just to compute the frequency of all the words in a document and print the most frequent, but when I tried that approach, I found that (as could be predicted) the most frequent words were 'the', 'and', 'or', etc. Because of this, I decided to alter my approach by finding the TF-IDF score, therefore finding the words that are uniquely important to a text (ie. are more frequent in one book, less frequent in another. TF-IDF takes into account the frequency of a word in a book and also the number of other documents in the analyzed set that word also appears in.

To implement this approach, I created a dictionary for each book with the words as keys and the frequency of the word as the value. Each book was processed by being transformed into a list of all the words in the book before processed into a dictionary. To find the term frequency, I simply divided the frequency by the total number of words in a book. Next, I found the inverse document frequency with the formula idf = log(number of documents analyzed/number of documents word appears in). The TF-IDF score is just the product of the term frequency and the inverse document frequency. I then sorted the dictionaries by decreasing TF-IDF and printed the top 5 words for each book. My results can be found below.

Results:

<u>Jane Austen</u>

Pride and Prejudice:

Word: Darcy, TF-IDF: 0.00318
Word: Elizabeth, TF-IDF: 0.00277
Word: Bennet, TF-IDF: 0.00248
Word: Bingley, TF-IDF: 0.0022
Word: Collins, TF-IDF: 0.00148

Sense and Sensibility:

Word: Elinor, TF-IDF: 0.00516
Word: Marianne, TF-IDF: 0.00373
Word: Elinor,, TF-IDF: 0.00291

Word: Marianne,, TF-IDF: 0.00244
Word: Dashwood, TF-IDF: 0.00182

Emma:
Word: Emma, TF-IDF: 0.00366
Word: Weston, TF-IDF: 0.00251
Word: Knightley, TF-IDF: 0.00212
Word: Elton, TF-IDF: 0.00201
Word: Emma,, TF-IDF: 0.00192

Persuasion:
Word: Mrs, TF-IDF: 0.00608
Word: Mr, TF-IDF: 0.00517
Word: Elliot, TF-IDF: 0.00263
Word: Wentworth, TF-IDF: 0.00219
Word: Walter, TF-IDF: 0.00164

Sir Arthur Conan Doyle

Study in Scarlet:
Word: Drebber, TF-IDF: 0.00131
Word: [Footnote, TF-IDF: 0.00096
Word: Jefferson, TF-IDF: 0.00095
Word: Ferrier, TF-IDF: 0.0009
Word: “I, TF-IDF: 0.0009

Sign of Four:
Word: "I, TF-IDF: 0.00195
Word: Morstan, TF-IDF: 0.00127
Word: Thaddeus, TF-IDF: 0.001
Word: "It, TF-IDF: 0.00098
Word: Sholto, TF-IDF: 0.00098

The Adventures of Sherlock Holmes:
Word: “I, TF-IDF: 0.0013
Word: Holmes,, TF-IDF: 0.00093
Word: Holmes, TF-IDF: 0.00084
Word: Sherlock, TF-IDF: 0.00054
Word: “It, TF-IDF: 0.00046

The Memoirs of Sherlock Holmes:
Word: “I, TF-IDF: 0.00113
Word: Holmes, TF-IDF: 0.00097
Word: Holmes,, TF-IDF: 0.00063
Word: Straker, TF-IDF: 0.00044
Word: “You, TF-IDF: 0.00041

What’s significant about the Jane Austen novels is that all of the highest TF-IDF scores were names of characters, except for Mrs, and Mr, in *Persuasion*. I suspect this is because in this version of the book, those prefixes were formatted differently than in the other Austen novels

(and likely also the Doyle novels). In Sir Arthur Conan Doyle's works, character names also appear on the top of the list, but one word that always makes an appearance is '"I' (with a quote before the 'I.' This suggests that Doyle's works include more dialouge in the first person than Austen's. In both the *The Adventures* and *The Memiors of Sherlock Holmes*. 'Holmes,' and 'Holmes,,' both appear. I suspect that's becasue these two works are short stories where the only consistent characters are Watson and Holmes (and the books are written from Watson's perspective). In the other two works, important characters such as the murder victims, perpetrators, and suspects show up as more frequent and more important words than 'Holmes.'

To me, this comparison suggests that while both bodies of works are very character focused, Austen's works are more so and Doyle's works contain more first person dialogue. More specifically, Austen's works revolve around two or three most important characters (usually a main female and main male character) whereas while Doyle's works are all about Sherlock Holmes, they reference him less often than they reference the characters involved more directly in the crime.

Reflection

From this processes, I learned how to make my code run faster and make it less computationally expensive. Many of my functions had to be modified in order to take less than a minute to run. The TF-IDF analysis returned good results and provided both an interesting and appropriately scoped project. One problem with my code is I did not have a good way to unit test. I tested with return and print statements along the way but I was not able to find a good way to include doctests, and that is one thing I would like to know how to do going into this project (and in the future). I also wish I had found a way to strip punctuation from the word I analyzed, because although looking at punctuation yielded some interesting results in terms of dialogue, it confused the algorithm where commas were concerned.