Petty Authors

Authors Stoker and Sir Arthur Conan Doyle want to sit with the classically cool Lady Shelly. To decide who she'll allow to sit with her, these authors most famous works, Dracula, Sherlock Holmes and Frankenstein respectively, were compared using an ngram model.

To achieve this, the books were downloaded from project gutenburg and cleaned of licenses and added info not written by the authors. The books were sentence tokenized to make setting start and stop probabilities easier for generation. Then each sentence was lowercased and word tokenized.

A standard measure of similarity in NLP is cosine similarity of tf-idf vectorized documents. Instead of simply using plain tokens, the term in 'tf-idf' is a bigram instead of a token. To generate counts of bigrams, a list of tuples of (sentence[i], sentence[i + 1]) was generated by looping over each sentence in each book. The count of each bigram in the list was then computed using nltk.FreqDist. The tf-idf for each book was then calculated to provide vectors that could be used in cosine similarity. This essentially calculated the euclidean distance between the two vectors (I looked up how to do this on billycambers.me tutorial on tf-idf and cosine similarity).

The similarities between the books were:

0.2323, 'SHolmes.txt', 'Dracula.txt'

0.1679, 'Frankenstein.txt', 'Dracula.txt'

0.1596, 'SHolmes.txt', 'Frankenstein.txt'

So the highest similarity Shelley (Frankenstein) has was with Stoker (Dracula). The loser, Doyle, would then have chosen Stoker as well.

To produce a possible conversation between Shelley and Stoker, a different architecture was needed. A 2d pandas dataframe of each books lexicon vs. lexicon was used to count bigrams df['a','be'] means the number of times 'a' was followed by 'be' in each book.

Originally `df.loc()` function was used to index the dataframe, but it proved to be the prohibitively slow. The lesser known alternative `df.at()` is much more optized for this use case. Swapping this access method out meant that a count table was possible for my machine to run, though very slow. Once counts were made, smoothing was done by adding 0.001 to all values. This is

essentially Laplace Smoothing, but skews the resulting probabilities far less than adding 1.0 to all values. Empirically the probability of “a” given start_token changed from 13.6% to 13.1%, a relatively small change.

With smoothing done, the probability for all combinations was calculated by summing each row and dividing the entries by the result. Then a function to choose a word given the probability table and a previous word was implemented by

1.) Taking the row of probabilities indexed by the previous word
2.) Choosing a random value between 0 and 1
3.) Walking through the row and subtracting each value from the random value until it was below 0
4.) Index the list of words to find the selected next word

This process was repeated until an end_token was produced to generate the conversation. Interestingly, but ultimately logically, using the smoothed probabilities led to generated sentences that made far less sense. The improved conversations are as follows:

```
Now that Shelly has chosen Stoker as the winner, here is
their conversation

Shelly:  how many scattered knots
Stoker:  a depraved wretch

Shelly:  i am sure wish to lord of the farther end had not i
have fled together
Stoker:  the lake of grief and wasted frame and in it was
replaced by these hopes i continued you may give no choice a
young man more mild a view of life should be an anxiety
returns upon the gallant and was absorbed in the snowy
mountains the chair my remissness by mountains i had obtained
my human goodness

Shelly:  the cracks between two years in america with ink
Stoker:  cease you could distinguish nothing contributes so
amiable and it was inspired me with perseverance and i am i
was called sister' or be no motion it with knowledge of life
```

will not comprised in various trains of voices that he dashed me to my fellow professor was consequently hired a sacrilege toward the endurance of my hands

Not quite sure how they became authors...

Poor Sir Aurthor Connon Doyle...
Here's his conversation with his best match: Stoker!
Take that Shelly!

Shelly: at the monotony of some two years ago
Stoker: on the ring which the crucifix on your slave and had been between the bloodstained voluptuous lips and i was ill quincey

Shelly: not deduce the official had put two of paul's wharf and i earn as far as follows to me through it was my own police court road is a woman's appearance at a false alarm took his former she no there's police
Stoker: he said laying the library

Shelly: the subject
Stoker: he said suddenly turn beginning of work friend from the shorthand and then he has never have wrecked the na vet of all the count asked him to be his wife the snow is not be necessary

Not quite sure how they became authors...
...Either