**How classy is your character?**

Text Mining Project by Audrey Lewis

**Project Overview:**

Did you know? Not all of Shakespeare is in iambic pentameter – only “high class” characters of people in formal situations spoke that way in his plays. For this project, I got the complete works of Shakespeare from Gutenberg. Then, I created a dictionary of characters for each play in the complete works, and for each of those characters counted how many lines they had, and how many were in pentameter.  
  
**Implementation:**

First, I fetched and pickled the complete works of Shakespeare. To format the document, I split it into a (very long!) list of words, with all the whitespace removed, and the ends of lines denoted by adding in the word ENDLINE. I also removed an irritating copyright statement that was arbitrarily sprinkled throughout.

I found the beginning of the plays by looking for the keyword Dramatis, which begins the character list, and used those locations to find the title and the end of each character list.

I went through each character list and found all the words that could possibly refer to a character in the play. This was by far the hardest part, because you have lines in the character list like “RICHARD, DUKE OF GLOUCESTER, afterwards KING RICHARD III”, and the code should add the words RICHARD, DUKE, KING and GLOUCESTER to the dictionary, but not 'afterwards'. (fun fact: the script ends up prefacing this particular character's lines with GLOUCESTER, but how are we supposed to know that?).

Then I took each play's dictionary and its text and used the character dictionary to figure out who said each line. These lines were then passed to a pentameter checker, which utilized a syllable counter I got off GitHub. Each time, this function incremented the total times a character spoke, and if it passed the pentameter checker, the total times in pentameter.

**Results*:***

I successfully identified the characters in all the plays, and counted their lines, but I'm seriously doubting most of the pentameter data. I think the syllable counter program I'm using might just not be that good. I'm hoping that's the case, because it means *my* implementation works well. However accurate the data is, it exists, and I'm proud to say that I can enter any character name in any Shakespeare play and get how many lines they have, and how many are in pentameter. Here are two examples from the terminal:

audrey@MercyofKalr:~/TextMining$ python process\_shakespeare.py

Which character would you like to see?PUCK

This character is in the play A MIDSUMMER NIGHT'S DREAM and has

46 total lines, 5 of which are in iambic pentameter.

audrey@MercyofKalr:~/TextMining$ python process\_shakespeare.py

Which character would you like to see?MARIA

This character is in the play LOVE'S LABOUR'S LOST and has

26 total lines, 1 of which are in iambic pentameter.

This character is in the play TWELFTH NIGHT; OR, WHAT YOU WILL and has

80 total lines, 3 of which are in iambic pentameter.

**Reflection*:***

I'm glad that in this, the first project where we started without predefined functions, I think I was able to split up my code into appropriate functions and document well what each function did. I think my code's pretty readable. I didn't use doctests at all, doing unit testing just at the end of the document instead, which probably wasn't ideal. As far as functionality goes, I could definitely improve this by pickling the dictionaries so I don't have to generate them every time, and there are a couple other lines here and there where I realize a faster way to do them, now that I'm done. If part of the point of this project was to get comfortable with all the different sequence types, I think it probably succeeded – one of the things I output is a list of dictionaries of tuples of strings. :)