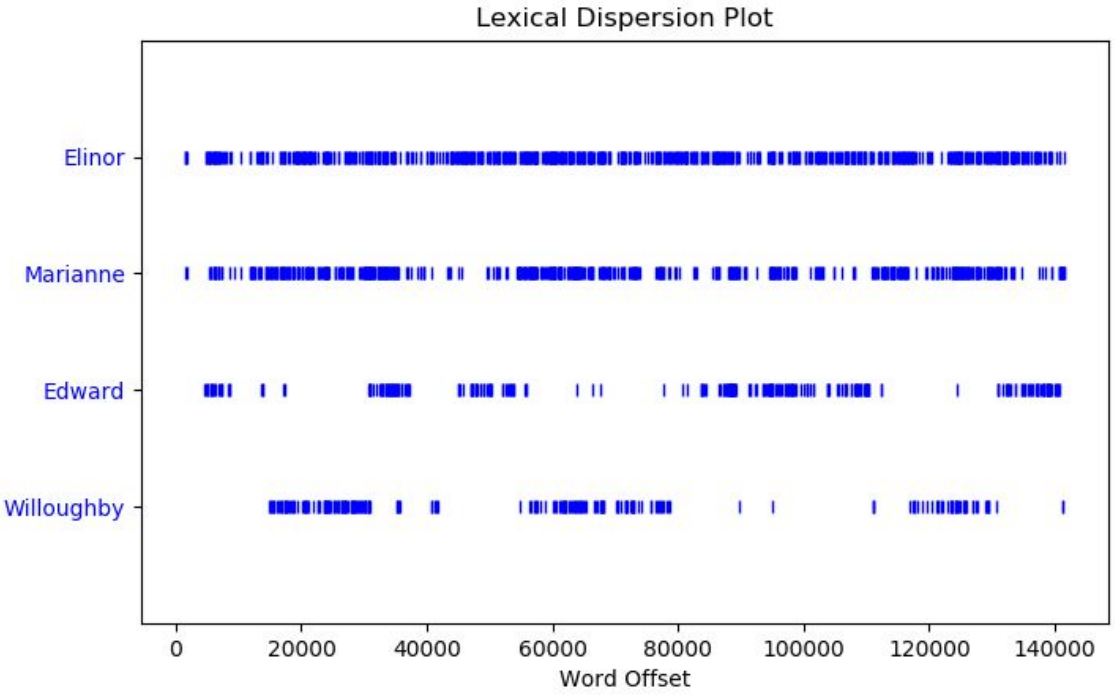
Name Jing Low
Section 27F3
Assignment 02
Due Date: January 25th, 2018

1. Compare the lexical diversity scores for humor and romance fiction in 1.1. Which genre is more lexically diverse?
    a. lexical_diversity(humor) = 0.231
    b. lexical_diversity(romance) = 0.121
    c. Because the lexical diversity is calculated by taking the length of the set of the text divided by the length of the original set, a higher number (closer to 1) means that the text is more lexically diverse. lexical_diversity(humor) > lexical_diversity(romance), therefore the genre humor is more lexically diverse.
2. Produce a dispersion plot of the four main protagonists in *Sense and Sensibility*: Elinor, Marianne, Edward, and Willoughby. What can you observe about the different roles played by the males and females in this novel? Can you identify the couples?
    a.



    b. According to the dispersion plot above, it can be inferred that Elinor is the main protagonist of the story as her name was present throughout almost the entire story; and it is possible that the story is narrated from Elinor's perspective in third person. Marianne's name comes second in terms of frequency, while the male characters appear rather sparsely in the novel. It is likely that Marianne serves as the second major protagonist along with Elinor as the occurrences and that

Marianne is a couple with Willoughby because Willoughby's appearances, though less frequent, parallel with Marianne's appearances. There is also a possibility that Willoughby and Edward do not know each other or are very unfamiliar with each other since their occurrences in the novel almost complement each other. Additionally, although it is hard to infer who Elinor's lover is due to her omnipresence in the book, Edward is most probably the man because of the given information that he is one of the four main protagonists of the book and because his occurrences do not parallel Marianne's at all.

3. What is the difference between the following two lines? Which one will give a larger value? Will this be the case for other texts?
    a. >>> sorted(set(w.lower() for w in text1))
        i. This will give a smaller size of output and will not consist duplicates of words.
    b. >>> sorted(w.lower() for w in set(text1))
        i. This will give a larger size of output and will consist duplicates of words.
    c. The main difference between these two lines of code is caused by the order of which the two functions, lower() and set(), are applied. In (a), all of the words in text1 were changed to lowercase, and then a set is taken on that modified text. In (b), a set is first taken for text1, treating uppercase and lowercase versions of the same words as different words; this results in a set larger than the one in (a) that would consist of same words with different capitalizations. And then, all of those words in the set are changed to lowercase, possibly causing there to be duplicates of the same words in the set.
    d. Nevertheless, there exists a situation where these two lines of code would output the same values. If the original text only consists of one version (either the uppercase or the lowercase) of each word, then the two outputs would be the same. However, this is not the case for text1.