Assignment 2(a)

Q1: In the news genre of brown corpus, find the count of the words starting with wh, such as what, when, where, who and why?

Q2: Find the conditional frequency distribution of modals ['can', 'could', 'may', 'might', 'must', 'will'] in all the categories of brown corpus?

Q3: Find the year out of the filenames in the Inaugural Address Corpus?

Q4: Read in the texts of the State of the Union addresses, using the state_union corpus reader. Count occurrences of men , women , and people in each document. What has happened to the usage of these words over time?

Q5: Pick a pair of texts and study the differences between them, in terms of vocabulary, vocabulary richness, genre, etc.

Q6: Write a program to find all words that occur at least three times in the Brown Corpus.

Q7: Write a program to generate a table of lexical diversity scores (i.e., token/type ratios) for each genre. Include the full set of Brown Corpus genres (nltk.corpus.brown.categories()). Which genre has the lowest diversity

Q8: Write a function that finds the 50 most frequently occurring words of a text?