Assignment I

Q1: Find the collocations in text5

Q2: Define a variable my_sent to be a list of words. Convert my_sent into string and then split it as list of words.

Q3: Find the index of the word *sunset in text9.*

Q4: ompute the vocabulary of the sentences sent1 ... sent8

Q5: What is the difference between the following two lines:

>>> sorted(set([w.lower() for w in text1]))

>>> sorted([w.lower() for w in set(text1)])

Q6: Write the slice expression that extracts the last two words of text2

Q7: Find all the four-letter words in the Chat Corpus (text5). With the help of a frequency distribution (FreqDist), show these words in decreasing order of frequency

Q8: Use a combination of for and if statements to loop over the words of the movie script for *Monty Python and the Holy Grail* (text6) and print all the uppercase words

Q9: Write expressions for finding all words in text6 that meet the following conditions.

a. Ending in *ize*

b. Containing the letter *z*

c. Containing the sequence of letters *pt*

d. All lowercase letters except for an initial capital (i.e., titlecase)

Q10: Define sent to be the list of words ['she', 'sells', 'sea', 'shells', 'by', 'the', 'sea', 'shore']. Now write code to perform the following tasks:

a. Print all words beginning with *sh*.

b. Print all words longer than four characters

Q11: What does the following Python code do? sum([len(w) for w in text1]) Can you use it to work out the average word length of a text?

Q12: Define a function called vocab_size(text) that has a single parameter for the text, and which returns the vocabulary size of the text.

Q13: Define a function percent(word, text) that calculates how often a given word occurs in a text and expresses the result as a percentage.