HLT

Homework 3: Morphology

Objective: Use Python, regex, and NLTK to process text.
Turn in: Your Python .py file.

Instructions:

1. Read in moby_dick.txt which you can download from Piazza.
2. Process the text:
    a. lower case
    b. use a string function to replace all occurrences of '--' with ' '
    c. use regex to remove all digits
    d. use regex to replace punctuation with a single space
3. Tokenize the text and print the number of tokens. Save the list of tokens for step 11.
4. Create a set of unique tokens and print the number of unique tokens.
5. Create a list of important words by removing stop words from the unique tokens list. Display the number of important words.
6. Using the list of important words, create a list of tuples of the word and stemmed word, like this:

    [('remarkably','remark'), ('prevented','prevent') …]

7. Create a dictionary where the key is the stem, and the value is a list of words with that stem, like this:

    {'achiev': ['achieved', 'achieve']
     'accident': ['accidentally', 'accidental'] … }

8. Print the number of dictionary entries.
9. For the 25 dictionary entries with the longest lists, print the stem and its list. One way to sort a dict by length of values:

    for k in sorted(stem_dict, key=lambda k: len(stem_dict[k]), reverse=True):

10. Using the dict from step 9, write a function to compute edit distance. Compute and print the edit distance between 'continue' and every word in the 'continu' list in the stem dict. See the 'iterative with full matrix' algorithm here: https://en.wikipedia.org/wiki/Levenshtein_distance
11. Perform POS tagging on the original text after step 3.
12. Create a dictionary of POS counts where the key is the POS, and the value is the number of words with that POS. Print the dictionary.

Element	Points
Python script runs without error	20
Appropriate comments and white space	10
Steps 1-9, 11-12	5*10 = 55
Step 10 Edit Distance function	15
Total	100