E6998 – HW#1

Due 9/26/2017 at 11:59PM (23:59) - EST

1. Define a variable called ‘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. (5-points)

2. You have the following text “Wild brown trout are elusive“? Provide code that will append each word to a list (5-points).

3. Using the contents in classify.zip located on CourseWorks, for each topic (machinelearning,mathematics,fishing,hiking), create a frequency count for each *unique* word and output both the word and count to separate columns (word, freq) in separate CSV files; name the csv file the folder (topic) name. Please remove stopwords before you even process the frequency rankings; clean the text up by removing special characters and extra spaces. Aggregate all documents together for each specific category to perform the exercise (you don’t have to do this for each document in each topic). A sample CSV file would be named hiking.csv and look like below (these do NOT represent true counts): (90-points) (hint: look to the code class3.py for pointers)

word	freq
forest	100
leaves	96
animals	80

You can provide all code in one .py file, but please make sure you push your code to github and submit the location/path when you submit the exercise.