Assignment 7

due 21 November 2019

Place the code for the function described below a file named `a7.py`. Do not include any extraneous code: remove all diagnostic code (prints used during debugging etc.) prior to submission. Test your code thoroughly, but do not include any testing code in your `a7.py` file. See the sheet for Assignment 2 for how to seperate test code from your assignment code. Adhere to the programming style requirements as specified on earlier assignment sheets.

Write a Python function named `most_common(text, n)` that takes a string ‘‘`text`’’ representing a (possibly quite lengthy) block of English-language text and a positive integer ‘‘`n`’’ and that returns a list in alphabetical order of the ‘‘`n`’’ most frequently ocurring words within ‘‘`text`’’.

The frequency of occurence of a word is how many times that word appear within the text. We ignore capitalization (so “cat” and “Cat” are treated as the same word), but distinguish between different words sharing a common stem (e.g. the singular and plural forms of a noun, so “cat” and “cats” are treated as different).

A word is a maximal contiguous sequence of letters bounded before and after by a non-letter (space, newline, punctuation mark, beginning and end of string etc.). We assume that words within the text are separated from one another by one or more non-letters. “Words” according to this definition need not be valid English words as found in a dictionary. Internal punctuation (hyphens and apostrophes etc.) are non-letters and treated as word separators, so “high-performance” is treated as two seperate words. Similarly “cat’s” (singular possessive with apostrophe) is treated as two separate words (“cat” and “s”), while “cats’ ” (plural possessive with apostrophe) is treated as a single word (“cats”). All punctuation appearing just before or just after a word is not regarded as part of the word, so the words within “ ‘Meow,’ said the cat!” are “meow”, “said”, “the” and “cat”.

The Project Gutenberg website (`www.gutenberg.org`) contains a large number out-of-copyright literary works that might prove useful for target practice.