Computational Linguistics
Fall 2018

Practical 5

This will be a graded individual homework. You will submit your code and output files to GradeScope (detailed instructions to come).

1. Open the file ‘Lovers_on_Aran_messed.txt’. You’ll find that the poem is messed up (by the previous TA) with many unwanted space/tab characters, and “empty” lines which contain no words (but there might be some invisible space/tab/newline characters!!!). Write code to:

   (1) Read the file. Remove all superfluous space/tab characters and add a space character in front of any punctuation mark (e.g., `‘      To throw      wide arms  around a tide.'` –> `'To throw wide arms around a tide .'`). Remove all “empty” lines. Write the cleaned poem to a new file called ‘Lovers_on_Aran.txt’.

   (2) Read in the cleaned file you just created. Count the number of words (including punctuations) of the file. Try to find every word which either begins with a capital letter, or contains fewer than 3 letters. Print out the result (the first line is the count, followed by one target word each line). For example:

   (3) Read in the cleaned file you just created. Convert every word to lowercase. Use the code you wrote in the last practical session, compute the probability of each line of text (including title and the author name) under the **trigram** assumption.
   Write the result to a new file called ‘trigram_prob.txt’ in the following format: each line contains a line id (starting from 0) and the corresponding probability (to three decimal places), separated by a `:` and a space character. Sort the lines in a decreasing order by probability. For example:

```
3: 0.139
5: 0.103
0: 0.075
…
```