System Design and Programming

Assignment Part 2

The problem of *authorship attribution* has a long history and in recent decades computer programs have played a significant role in the many new techniques developed to address it. One such technique, known as n-gram analysis, has had some success.

n-gram analysis calculates the frequency of occurrence of n-length sequences of characters or words. The use of the most frequently occurring n-grams can then be analysed both within and between the collective works of authors. A paper describing this (and other) approaches is attached.

The Task.

Write a Python program that will permit the generation of the *f* most frequently occurring *n*-grams (the program should support character and word n-grams) in a text sample. Run the program across the sample scripts of each of the authors (provided as adata.zip on the VLE) and store the author occurrence averages and standard deviations of the 20 most frequently occurring n-grams for each author.

A further data sample (unknown.txt) contains text written by one of the authors. Analyse the text and then use the data that you have constructed to speculate as to the likely author (a statistical analysis is neither required nor expected).