

Programming Assignments: Frequency Analysis

Jan Baumann

December 2024

1 Description

1.1 Problem

Your task is to write a function, which takes a text as input and counts how many times each individual character occurs.

1. Only letters from a to z are relevant.
2. The program should ignore case-sensitivity, e.g. "E" and "e" are the same.
3. The function should return the result in an adequate data structure.

1.2 Example

Take this example text:

Frequency analysis is based on the fact that, in any given stretch of written language, certain letters and combinations of letters occur with varying frequencies. Moreover, there is a characteristic distribution of letters that is roughly the same for almost all samples of that language. For instance, given a section of English language, E, T, A and O are the most common, while Z, Q, X and J are rare. Likewise, TH, ER, ON, and AN are the most common pairs of letters (termed bigrams or digraphs), and SS, EE, TT, and FF are the most common repeats. The nonsense phrase "ETAOIN SHRDLU" represents the 12 most frequent letters in typical English language text. In some ciphers, such properties of the natural language plaintext are preserved in the ciphertext, and these patterns have the potential to be exploited in a ciphertext-only attack.

Source: https://en.wikipedia.org/wiki/Frequency_analysis

Here, the letter distribution is as follows: 'f': 15, 'r': 49, 'e': 94, 'q': 4, 'u': 14, 'n': 51, 'c': 22, 'y': 7, 'a': 61, 'l': 27, 's': 48, 'i': 42, 'b': 5, 'd': 14, 'o': 40, 't': 75, 'h': 31, 'g': 18, 'v': 6, 'w': 4, 'm': 18, 'p': 17, 'z': 1, 'x': 6, 'j': 1, 'k': 2

2 Tasks

As with every problem you encounter, sketch the problem first. A picture might give you a better idea on how to tackle the problem.

Continue by writing your function. If you have knowledge in unit testing, this is a good practice example. Think about the **valid cases**, **edge cases** and **invalid cases**.