AFP 2015 – Course work

Your task is to write a program to compare similarities of texts, as explained below. The calculations are based on comparison of sentences, separated by ".", ";", "!", and "?". If there is a comma "," or a colon ":" in the file it should be ignored in the comparisons (we consider they do not separate sentences). Uppercase and lowercase characters are considered equal.

We say that the distance between sentences S and Q is the smallest number of words we have to delete from the sentences S and Q to make the word sets of the sentences equal. For instance, the distance between "Hello, one world!" and "the world is a round one" is 5. The word sets will be the same after deleting "Hello", "the", "is", "a", and "round". Deleting fewer words is not enough.

We consider several methods to compute the distance between files F1 and F2.

In the first method, the distance between files F1 and F2 is the sum of distances that we get, when all sentences in F1 are compared with all sentences in F2.

In the second method, the distance between files F1 and F2 is the minimum sum of distances, when each sentence of F1 is compared to exactly one of the sentences in F2 (notice that a sentence in F2 may be compared to many sentences, or may not be compared at all).

In the third method, each sentence of F1 can be compared with at most one sentence of F2 and vice versa. Let's suppose that a sentence S1 in F1 is compared with sentence S2 in F2. Then,

- a sentence before S1 in F1 can only be compared with a sentence before S2 in F2,
- a sentence before S2 in F2 can only be compared with a sentence before S1 in F1,
- a sentence after S1 in F1 can only be compared with a sentence after S2 in F2,
- a sentence after S2 in F2 can only be compared with a sentence after S1 in F1,
- it is not necessary that all sentences participate in the comparison,
- given a selection E of which sentences are compared with which sentences, the E-distance obtained is the sum of the distances of the comparisons plus the sum of the lengths of all sentences not participating in the comparisons, and
- the distance between F1 and F2 is the minimum E-distance that can be obtained (making an optimal selection, of course).

Your program will be given three arguments in the command line: first, the comparison method (one of **first** , **second** or **third**) and then two filenames: the names of the files to be compared.

Your program is to produce the following output. If you have not implemented that comparison method, then output **Not implemented** and else output as follows:

- In the first line, output the distance (an integer) and nothing else.
- For the first comparison method, no other output is required.

For the second and third comparison method, output the pairs of sentences for which the distance was calculated and included in the result. Each pair is output on *one line* as follows: first output a sentence from file F1 (you may or may not output periods, commas and such), then output the character "&" (which will not appear in the input) and then output a sentence from F2. The order of the pairs in your output is not important.