Task 4 – Parallel Programming

Description

Given a set of text files in your favorite programming language (C#, Java, C, C++, Python, Javascipt etc.) do the following tasks.

- Every text should be cleared before further processing. (Remove any numbers, special characters, brackets and keywords etc.).
- For each word compute its frequency $x(w)$ (this is the number of occurrences of a given word divided by the number of all words in the cleared text).
- For each file find a set s, consisting of words which are **frequent** in the cleared file. A word is considered **frequent** if its frequency is at least $K\%$ of words in the cleared text.
- Using previously computed values, compute the set S and frequency $X(w)$ consisting of words **frequent** in all texts. (Results should be identical as if you computed s and $x(w)$ for concatenation of all input text files – but this way is inefficient and it is forbidden to do it this way.)
- Print N most frequent words with theirs frequencies $X(w)$ and list of files satisfying $|s \cap S|/|s| > 0.5$.

Technical aspects

1. Parameters N, K and directory with input files on *hdfs://* are configurable.
2. To complete the task you can use any library operating on a distributed file system (e.g. Spark).