Zadanie 4 – Programowanie Równoległe

Opis zadania

Dany jest zbiór plików tekstowych z tekstami w ulubionym języku programowania (C#, Java, C, C++, Python, Javascript etc.).

- Każdy tekst przed obróbką ma być przetworzony do postaci oczyszczonej. (Usuwamy liczby, znaki przestankowe, nawiasy, słówka kluczowe danego języka etc.).
- Każdemu słowu w przypisujemy częstość $x(w)$ (liczba wystąpień podzielona przez liczbę słów w tekście) występowania w tym zbiorze.
- W każdym pliku wyszukujemy zbiór s, słów które **często** się pojawiają w tekstach. Słowo uznaje się za **częste** w tekście, jeśli stanowi co najmniej $K\%$ słów w oczyszczonym tekście.
- Na podstawie wyliczonych wcześniej wartości tworzymy zbiór słów S oraz częstość $X(w)$ występowania słów **częstych** we wszystkich tekstach. (Tak jakby wyliczyć zbiór słów **częstych** na konkatenacji wszystkich plików wejściowych – ale takie rozwiązanie jest niewydajne i tak zrobić nie wolno, trzeba sprytniej).
- Wypisujemy N najczęstszych słów wraz z częstościami $X(w)$ oraz listę plików, w których $|s \cap S|/|s| > 0.5$.

Aspekty techniczne

1. Parametry N, K oraz folder z plikami na *hdfs://* są konfigurowalne.
2. Do wykonania zadania można użyć dowolnej biblioteki operującej na rozproszonym systemie plików (np. Spark).