Hadoop TFIDF

Implementacja algorytmu TFIDF (Term Frequency-Inverse Document Frequency) przy użyciu .

TFIDF

Algorytm TFIDF jest jednym z algorytmów obliczających statystyczne wagi termów (tokenów). Celem algorytmu jest klasyfikacja ważności danego słowa w odniesieniu do danego dokumentu oraz zbioru wszystkich dokumentów. Może być użyty np. w wyszukiwarkach.

Wynikiem algorytmu jest liczba

TF * IDF

Gdzie

TF = Nij * Cj
IDF = log(D/Di)

Nji - liczba wystąpień tokenu **i** w dokumencie **j**.
Cj - liczba wszystkich wystąpień termów w dokumencie **j**.
D - liczba wszystkich dokumentów
Di - liczba dokumentów w których występuje token **i**.

Implementacja MapReduce

Algorytm składa się z trzech kroków. Działanie algorytmu przedstawione jest na poniższym diagramie: