Laporan Tugas TF-IDF

**Deskripsi DataSet**

Data set berupa tab delimited text dengan nama file berita.txt

Data Set terdiri atas beberapa kolom yang dipisahkan dengan tab terdiri dari:

1. ID Berita
2. Judul Berita
3. Paragraf pertama berita (isi dari berita)
4. Link berita
5. Tanggal berita dimuat

Berita-berita yang diambil berasal dari beberapa sumber seperti detik.com, kompas.com dan sindonews.com.

**Proses Data Mining**

1. Lowering case untuk kolom paragraf pertama berita (agar case-insensitive)
2. Artikel dari berita akan di hashing dengan menggunakan modul HashingTF pada Spark
3. Melakukan penghitungan jumlah keyword dengan menggunakan TF (Text Frequency)
4. Menggunakan proses IDF untuk dibandingkan dengan setiap text yang ada pada setiap row.
5. User memasukkan kata yang akan dicari dengan menggunakan argument dan program akan mencari pada dokumen yang mana kata tersebut paling relevan.
6. Program mengeluarkan output berupa ID dokumen yang paling relevan dengan keyword yang dimasukkan user