Dokumentasi

Implementasi TF-IDF

BIG DATA A

1. Deskripsi Dataset

Nama dataset : dataset.tsv

Ukuran file : 39kb

Jumlah Kolom : 3

Details Kolom :

* ID
* Tanggal Berita
* Judul + Isi Berita

Jumlah Baris : 21  
Sumber dataset : viva.co.id

1. Tahapan proses datamining

* Buka dataset
* Lakukan lowercase pada dataset agar pencarian dapat dilakukan dengan mudah
* Lalu pisahkan tiap kolom yang ditandai dengan tab
* Lalu pisahkan tiap kata dari kolom ke 3 yang ditandai dengan “ “
* Lalu simpan id yang berada di kolom ke 1 sebagai key
* Buat hash table dan tf table dari kata kata yang telah di split tadi
* Lalu buat idf, dimana ditentukan minimal document fequensinya 1
* Lalu kalkulasi dari hasil tf/idf setelah mendapat tf dan idf pada proses sebelumnya
* Setelah itu, value yang akan dicari katanya dibuat lowercase agar bisa menemukan kata yang sama tanpa mempedulikan case
* Lalu setelah itu value yang dicari tadi dihitung hashnya
* Lalu kata tadi dicari dengan kata yang sama berdasarkan hashnya tadi
* Setelah ketemu, lalu ditampilkan id yang mana nilai tf/idf nya paling besar