Abstrak

Data mining adalah proses eksplorasi dan analisa data yang berjumlah besar untuk mendapatkan pola yang berguna. Data mining merupakan proses gabungan antar bidang-bidang terutama *machine learning*, analisis statistik, dan basis data

Salah satu *task* yang penting dalam data mining adalah clustering. Clustering adalah proses mempartisi sekumpulan objek ke dalam cluster-cluster. Objek-objek yang mirip akan ditempatkan dalam cluster yang sama dan cluster yang berbeda akan ditempatkan dalam cluster yang berbeda.

Tugas akhir ini berusaha untuk mengelompokkan dokumen dengan menggunakan algoritma top-k scoring. Dokumen yang digunakan adalah dokumen keselamatan kerja di PT.Pertamina UP IV Cilacap, sebab dokumen di perusahaan ini tersusun atas dokumen teks yang tidak terstruktur dan kompleks, sehingga membutuhkan usaha yang besar untuk pencarian terhadap dokumen-dokumen untuk menghadapi suatu permasalahan. Adapun kemiripan antar dokumen diukur dengan penjumlahan sederhana dari kemunculan kata-kata pada dokumen yang dibandingkan.

Setelah dilakukan pengujian dengan beberapa threshold, hasil uji menunjukkan bahwa algoritma top-k scoring dapat digunakan untuk mengelompokkan dokumen berbahasa Indonesia dengan tingkat akurasi sampai dengan 96.67%. Tingkat akurasi ini dihitung dengan cara membandingkan hasil clustering dengan hasil pengelompokan secara manual.

Kata kunci : data mining, clustering, top-k scoring, cluster, WIDF.