

Klasterisasi Abstraksi Karya Ilmiah Menggunakan Metode K-Means

Muhammad Nur Ikhwani
110411100030

Pembimbing 1 : Mula'ab, S.Si, M.Kom
Pembimbing 2 : Iwan Santosa, S.T., M.T.

ABSTRAK

Banyaknya jumlah karya ilmiah yang tersedia menyebabkan dokumen sangat banyak dan beragam sehingga pengelompokan menjadi sulit. Oleh Karena itu, dibutuhkan suatu metode yang dapat mengelompokkan dokumen sesuai kategori. Klasterisasi merupakan metode *data mining* untuk mengelompokkan data sesuai dengan karakteristik, sehingga dapat membantu pengguna dalam proses pencarian dokumen dengan cepat. Metode yang digunakan dalam penelitian ini adalah metode *k-means* yang merupakan salah satu metode klaster yang bersifat *unsupervised learning*. Tujuan dari penelitian ini adalah untuk mengetahui tingkat akurasi metode *k-means* dalam klasterisasi karya ilmiah. Pada proses ini pengelompokan abstrak dibagi tiga kategori yaitu hukum, pertanian dan teknik. Pengujian hasil *clustering* menggunakan *k-means* menghasilkan nilai rata-rata *precision* dan *recall* masing-masing 55.2 %, 51.1% serta akurasi sebesar 50.3%.

Kata kunci: karya ilmiah, data mining, klasterisasi, k-means

[Halaman ini sengaja dikosongkan]

DAFTAR ISI

LEMBAR PENGESAHAN	v
HALAMAN PERNYATAAN ORISINALITAS	vii
KATA PENGANTAR.....	ix
LEMBAR PERSEMBAHAN	xi
ABSTRAK	xiii
DAFTAR ISI	xv
DAFTAR GAMBAR	xvii
DAFTAR TABEL	xix
 BAB I PENDAHULUAN.....	 1
1.1. Latar Belakang	1
1.2. Perumusan Masalah.....	2
1.3. Tujuan Penelitian.....	2
1.4. Manfaat Penelitian.....	2
1.5. Batasan Masalah.....	2
1.6. Metodologi Penelitian.....	3
1.7. Sistematika Penulisan	4
 BAB II STUDI PUSTAKA.....	 6
2.1. Landasan Teori.....	6
2.2.1. <i>Text Mining</i>	6
2.2.2. <i>Preprocessing</i>	7
2.2.3. <i>Clustering</i>	9
2.2.4. <i>K-Means</i>	11
2.2. Penelitian Sebelumnya.....	18
 BAB III ANALISIS DAN PERANCANGAN SISTEM.....	 23
3.1. Metode Pengembangan Sistem SDLC Waterfall	23
3.2. Analisis Sistem	25
3.2.1. Analisis Permasalahan.....	
3.2.2. Analisis Kebutuhan Fungsional	25
3.2.3. Analisis Kebutuhan Non Fungsional.....	26
3.3. Perancangan Sistem	26
3.3.1. <i>Flowchart</i>	27
3.3.1.1. <i>Flowchart</i> Rancangan Sistem	27
3.3.1.2. <i>Flowchart Preprocessing</i>	28
3.3.1.3. <i>Flowchart</i> proses perhitungan <i>tf-idf</i> ..	36

3.3.1.4. <i>Flowchart k-means</i>	37
3.3.2. <i>Use Case Diagram</i>	38
3.3.3. <i>Activity Diagram</i>	39
3.3.4. Perancang Antarmuka	40
3.3.4.1. Antarmuka Menu Utama	41
3.3.4.2. Antarmuka Menu Input	41
3.3.4.3. Antarmuka Menu <i>K-Means</i>	42
3.3.5. Perancangan <i>Database</i>	43
3.3.5.1. CDM (<i>Conceptual Data Model</i>).....	43
3.3.5.2. PDM (<i>Physical Data Model</i>)	44
3.3.6. Desain Tabel <i>Database</i>	44
BAB IV IMPLEMENTASI SISTEM	47
4.1. Lingkungan Implementasi	47
4.1.1. Spesifikasi Perangkat Keras.....	47
4.1.2. Spesifikasi Perangkat Lunak.....	47
4.2. Implementasi Kode Program	48
4.2.1. Case Folding.....	48
4.2.2. Tokenizing.....	49
4.2.3. Term Frequency	49
4.2.4. Menghitung IDF.....	49
4.2.5. Menghitung TF-IDF dan Panjang Vektor.....	50
4.2.6. K-Means	50
4.3. Pengujian Sistem.....	51
4.4. Implementasi Program	52
4.4.1. Proses Input Data	53
4.4.2. Proses Stemming.....	54
4.4.3. Proses Preprocessing.....	54
4.4.4. Proses tf-idf	56
4.4.4. Proses <i>k-means</i>	57
4.5. Implementasi Uji Coba.....	59
4.6. Analisis Hasil Uji Coba	61
BAB V PENUTUP.....	63
5.1. Kesimpulan.....	63
5.2. Saran	63
DAFTAR PUSTAKA.....	65
LAMPIRAN.....	67

DAFTAR GAMBAR

Gambar 1.1. Rancangan Penelitian.....	3
Gambar 2.1. Proses <i>Text Mining</i>	7
Gambar 2.2. Proses <i>case folding</i>	8
Gambar 2.3. Proses <i>tokenizing</i>	9
Gambar 2.4. Proses <i>filtering</i>	9
Gambar 2.5. Proses <i>stemming</i>	10
Gambar 2.6. Ilustrasi <i>cluster</i>	12
Gambar 2.7. Ilustrasi Proses <i>k-means</i>	13
Gambar 3.1. Model Pengembangan Sistem Waterfall	24
Gambar 3.2. Flowchart Perancangan sisitem	28
Gambar 3.3. Flowchart Preprocessing.....	29
Gambar 3.4. Flowchart case folding.....	30
Gambar 3.5. Flowchart Tokenizing.....	31
Gambar 3.6. Flowchart Stemming.....	33
Gambar 3.7. Flowchart proses tf idf.....	37
Gambar 3.8. Flowchart k-means	38
Gambar 3.9. <i>Use Case Diagram</i> Sistem	39
Gambar 3.10 Activity Diagram Sistem.....	40
Gambar 3.11 Antarmuka menu utama	41
Gambar 3.12 Antarmuka menu input	42
Gambar 3.13 Antarmuka menu <i>k-means</i>	42
Gambar 3.14 Rancangan CDM Sistem.....	43
Gambar 3.15 Rancangan PDM Sistem	44
Gambar 4.1. Proses Input Data.....	53
Gambar 4.2. Proses Setelah Input.....	53
Gambar 4.3. Proses <i>stemming</i>	54
Gambar 4.4. Perbandingan kalimat sebelum dan sesudah di <i>processing</i>	55
Gambar 4.5. Hasil potongan <i>term</i> setelah <i>Preprocessing</i>	55
Gambar 4.6. Hasil <i>document frequency</i> setiap <i>term</i>	56
Gambar 4.7. Hasil <i>tf-idf</i>	57
Gambar 4.8. Hasil Perhitungan K-Means	58
Gambar 4.9. Hasil Pengelompokkan Teks.....	58

[Halaman ini sengaja dikosongkan]

DAFTAR TABEL

Tabel 2.1. Data Mentah	16
Tabel 2.2. Titik Pusat Awal <i>cluster</i>	17
Tabel 2.3. Hasil perhitungan jarak setiap data	17
Tabel 2.4. Posisi <i>cluster</i> pada iterasi pertama	18
Tabel 2.5. Penelitian Sebelumnya	21
Tabel 3.1. Kombinasi awalan dan akhiran yang tidak diizinkan	34
Tabel 3.2. Pemenggalan Awalan <i>Stemmer</i> Nazief Adriani.....	35
Tabel 3.3. Tabel Kalimat	45
Tabel 3.4. Tabel Term	45
Tabel 3.5. Tabel term_detail	46
Tabel 4.1. Skenario Pengujian sistem.....	51
Tabel 4.2. Pengujian fungsional sistem	52
Tabel 4.3. Hasil Pengujian <i>clustering</i> skenario pertama.....	59
Tabel 4.4. Hasil Pengujian <i>clustering</i> skenario kedua	60
Tabel 4.5. Hasil Pengujian <i>clustering</i> skenario ketiga.....	61
Tabel 4.6. Rata-rata nilai <i>precision</i> , <i>recall</i> dan <i>accuracy</i>	62