

KLASTERISASI BERITA OLAHRAHA BERBAHASA INDONESIA DENGAN METODE *K-MEANS* YANG DIMODIFIKASI

Wahit Abdulloh
(14.04.1.1.1.00070)

Pembimbing I : Sigit Susanto Putro, S.Kom, M.Kom

Pembimbing II : Yoga Dwitya Pramudita, S.Kom.,M.Cs

ABSTRAK

Berbagai cabang olahraga seperti sepak bola, tenis meja, bulu tangkis, dan lainnya mengharuskan setiap berita olahraga untuk dikelompokkan sesuai dengan cabangnya. Oleh karena itu diperlukan cara untuk mengelompokkan berita olahraga yang memiliki kesamaan. Pengelompokkan dapat dilakukan dengan banyak metode. Metode yang bisa menangani permasalahan ini salah satunya adalah *k-means clustering*. *K-means clustering* adalah metode pengelompokan yang bisa bekerja dengan dataset yang besar dengan efisien. Namun metode ini memiliki permasalahan disaat penentuan awal pusat *cluster*. Hasil akhir *cluster* dari *k-means* sangat bergantung pada penentuan awal ini. Metode *cosine similairty* dapat membantu *k-means clustering* mendapatkan pusat *cluster* yang baik. Pada penelitian ini pengelompokan berita olahraga berbahasa Indonesia dilakukan dengan metode *k-means clustering* yang telah dimodifikasi dengan *cosine similarity*. Hasil yang didapatkan dari percobaan adalah pusat *cluster* terbaik yang dihasilkan oleh metode *k-means* selalu terjadi perubahan, bergantung pada penentuan awal pusat *cluster*. Setelah dilakukan modifikasi pada penentuan awal, pusat *cluster* terbaik cenderung konstan, yaitu sebanyak 4 pusat *cluster* dengan nilai akurasi rata-rata 0.895.

Kata Kunci: berita olahraga, *text clustering*, *k-means clustering*, *modified k-means clustering*, *cosine similarity*.

DAFTAR ISI

HALAMAN PENGESAHAN.....	v
HALAMAN PERNYATAAN ORISINALITAS.....	vii
HALAMAN PERSEMBAHAN.....	ix
ABSTRAK.....	xi
KATA PENGANTAR.....	xii
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 dan Manfaat.....	2
1.3.1 Tujuan.....	2
1.3.2 Manfaat.....	3
1.4 Batasan Masalah.....	3
1.5 Sistematika Penulisan.....	3
BAB II TINJAUAN PUSTAKA.....	5
2.1 Representasi Dokumen.....	5
2.2 Preprocessing.....	6
2.2.1 <i>Case Folding</i>	6
2.2.2 Tokenization.....	6
2.2.3 Stopword Removal.....	7
2.2.4 Stemming.....	7
2.2.5 Term Weighting.....	8
2.3 Distance Space.....	8
2.4 K-means.....	9
2.5 Cosine Similarity.....	13
2.6 Perhitungan Akurasi.....	16
2.7 Penelitian Sebelumnya.....	16
BAB III PERANCANGAN SISTEM.....	19
3.1 Studi Literatur.....	19
3.2 Pengumpulan Data.....	20
3.3 Perancangan Sistem.....	20
3.3.1 Arsitektur Sistem.....	20
3.3.2 Diagram Alur Proses Sistem.....	21
3.3.2.1 Preprocessing.....	22



3.3.2.2 Term Weighting.....	23
3.3.2.3 Pengelompokan dengan <i>K-means</i>	24
3.4 Skenario Uji.....	26
3.4.1 Skenario Uji I.....	26
3.4.2 Skenario Uji II.....	27
3.4.3 Skenario Uji III.....	28
3.4.4 Skenario Uji IV.....	28
3.4.5 Skenario Uji V.....	28
3.4.6 Skenario Uji VI.....	28
BAB IV IMPLEMENTASI DAN PENGUJIAN.....	29
4.1 Lingkungan Uji Coba.....	29
4.2 Implementasi <i>User Interface</i>	29
4.2.1 Halaman Daftar Berita.....	29
4.2.2 Halaman Proses Pengelompokan.....	29
4.2.3 Halaman Hasil Pengelompokan.....	29
4.2.4 Halaman Detail Berita.....	30
4.3 Kode Program.....	31
4.3.1 Menentukan Pusat <i>Cluster</i>	32
4.3.2 Mengukur Jarak Pusat <i>Cluster</i> dengan Berita.....	34
4.3.3 Memilih Berita dengan Jarak Terdekat.....	34
4.3.4 Memperbarui Pusat <i>Cluster</i>	35
4.3.5 Mendapatkan Nilai Pengelompokkan yang Optimal.....	35
4.4 Pengujian.....	36
4.4.1 Hasil Skenario Uji I.....	36
4.4.2 Hasil Skenario Uji II.....	36
4.4.3 Hasil Skenario Uji III.....	36
4.4.4 Hasil Skenario Uji IV.....	37
4.4.5 Hasil Skenario Uji V.....	37
4.4.6 Hasil Skenario Uji VI.....	38
BAB V PENUTUP.....	47
5.1 Kesimpulan.....	47
5.2 Saran.....	47
DAFTAR PUSTAKA.....	49
LAMPIRAN.....	51
1 Lampiran A : Beberapa Dataset yang Digunakan.....	51
2 Lampiran B : Beberapa Hasil Klasterisasi Berita.....	88
3 Lampiran C : Kuisioner Uji Coba.....	92
BIOGRAFI PENULIS.....	93



DAFTAR GAMBAR

Gambar 2.1 Flowchart metode k-means [5].....	10
Gambar 2.2 Hasil terbaik dan terburuk dari k-means [12].....	11
Gambar 3.1 Arsitektur Sistem.....	20
Gambar 3.2 Diagram Alur Preprocessing.....	22
Gambar 3.3 Diagram Alur Term Weighting.....	23
Gambar 3.4 Diagram Alur Metode K-means.....	24
Gambar 3.5 Diagram Alur Metode <i>K-means</i> yang Dimodifikasi.....	25
Gambar 3.6 Skenario uji.....	27
Gambar 4.1 Halaman Daftar Berita.....	30
Gambar 4.2 Halaman Proses Pengelompokkan Berita.....	30
Gambar 4.3 Halaman Hasil Pengelompokkan Berita.....	31
Gambar 4.4 Halaman Detail Berita.....	31
Gambar 4.5 Grafik Perbandingan Hasil Akurasi I.....	44
Gambar 4.6 Grafik Perbandingan Hasil Akurasi II.....	44



DAFTAR TABEL

Tabel 2.1 Kumpulan kata hasil tokenization.....	5
Tabel 2.2 Kumpulan kata hasil stopword removal.....	6
Tabel 2.3 Kumpulan kata hasil stemming.....	7
Tabel 2.4 Kumpulan dokumen berita olahraga.....	8
Tabel 2.5 Perhitungan pemberian bobot.....	9
Tabel 2.6 Nilai pembobotan 5 berita olahraga.....	12
Tabel 2.7 Jarak berita dengan pusat cluster.....	13
Tabel 2.8 pusat cluster D1 yang baru.....	14
Tabel 2.9 pusat cluster D3 yang baru.....	15
Tabel 2.10 Penelitian-penelitian mengenai <i>k-means</i>	18
Tabel 4.1 Spesifikasi Perangkat.....	29
Tabel 4.2 Perbandingan Dokumen yang Relevan dengan 3 <i>Cluster</i>	38
Tabel 4.3 Perbandingan Akurasi dengan 3 Cluster.....	38
Tabel 4.4 Perbandingan Dokumen yang Relevan dengan 4 Cluster.....	39
Tabel 4.5 Perbandingan Akurasi dengan 4 Cluster.....	40
Tabel 4.6 Perbandingan Dokumen yang Relevan dengan 5 Cluster.....	41
Tabel 4.7 Perbandingan Akurasi dengan 5 Cluster.....	41
Tabel 4.8 Perbandingan Dokumen yang Relevan dengan 6 Cluster.....	42
Tabel 4.9 Perbandingan Hasil Akurasi I.....	42
Tabel 4.10 Perbandingan Akurasi dengan 6 Cluster.....	43
Tabel 4.11 Perbandingan Hasil <i>Silhouette Coefficient</i>	43
Tabel 4.12 Perbandingan Hasil Akurasi II.....	45

