

ABSTRAK

Pada beberapa kasus, kesamaan konten dua buah *website* perlu ditentukan. Sebagai contoh dalam memprediksi konsistensi akses *website* berdasarkan *website* dan *software behavior log*, sebagian fitur hanya dapat dihitung dengan adanya kesamaan konten antar dua *website*. Pada tugas akhir ini, akan dibangun sistem deteksi kesamaan konten antar dua buah *website*. Nilai kesamaan ditentukan dari metode *clustering* dengan *k-means clustering* dan *information retrieval* dengan similaritas konten. Berdasarkan pengujian, mekanisme *clustering* berbasis *website* dan *software* menghasilkan kata-kata yang representatif. Selain itu algoritma perhitungan *cosine similarity* yang dihasilkan sesuai dengan perkiraan.

Kata kunci: *cosine similarity*, *k-means clustering*, similaritas konten, *website*.



ABSTRACT

In some cases, the similarity between two websites are needed to be calculated. For example, on predicting web access consistency based on website and software behavior log, some features can only be calculated based on the similarity content between two websites. In this thesis, a system for detecting the silimarity content will be proposed. Similarity value is determined from k-means clustering and cosine similarity. Based on evaluation, resulted website and software are considered as effective since their most frequent terms are representative words. Beside that, the algorithm for calculating cosine similarity is produce according to predictions.

Keywords: content similarity, cosine similarity, k-means clustering, website.

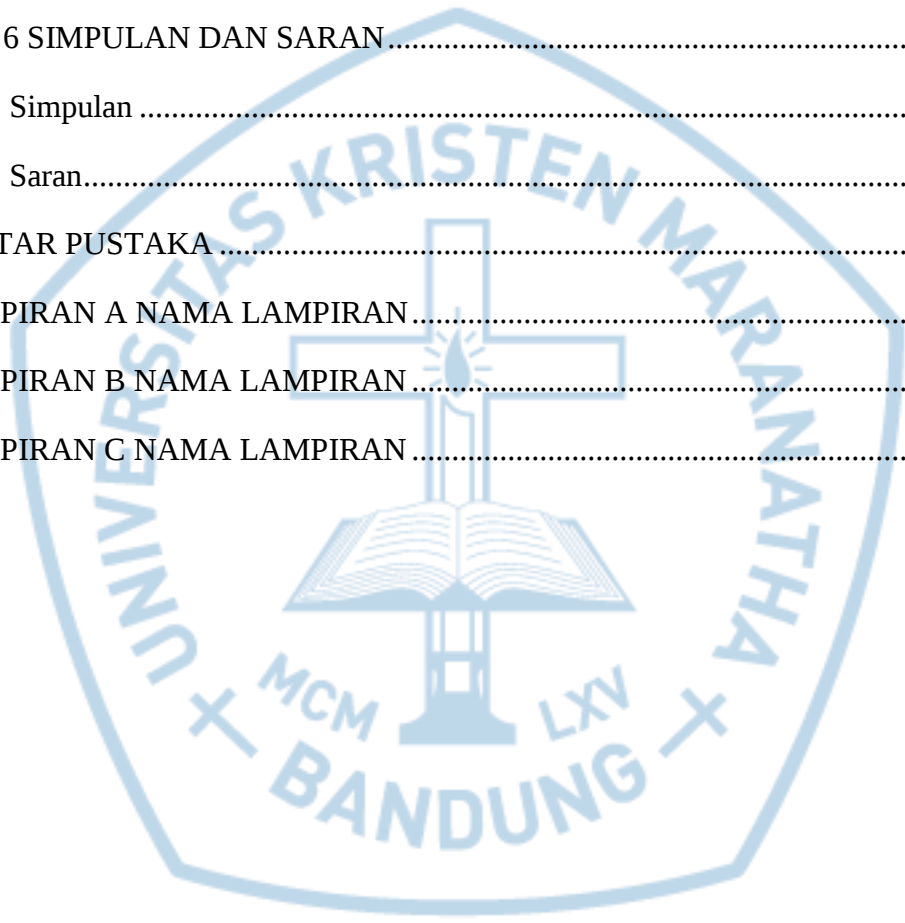


DAFTAR ISI

LEMBAR PENGESAHAN	ii
PERNYATAAN ORISINALITAS LAPORAN PENELITIAN.....	iii
PERNYATAAN PUBLIKASI LAPORAN PENELITIAN	iv
PRAKATA.....	v
ABSTRAK	vii
ABSTRACT.....	viii
DAFTAR ISI.....	ix
DAFTAR GAMBAR	xii
DAFTAR TABEL.....	xv
DAFTAR NOTASI/ LAMBANG.....	xviii
DAFTAR SINGKATAN	xix
BAB 1 PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	1
1.3 Tujuan Pembahasan	1
1.4 Ruang Lingkup.....	2
1.5 Sumber Data.....	2
1.6 Sistematika Penyajian	2
BAB 2 KAJIAN TEORI	4
2.1 <i>Machine Learning</i>	4
2.1.1 <i>Classification</i>	4
2.1.2 <i>Clustering</i>	4
2.2 <i>K-Means Clustering</i>	4
2.3 <i>Cosine Similarity</i>	5

2.4 Weka	5
2.5 Jsoup	5
2.6 Riset Terkait	6
BAB 3 ANALISIS DAN RANCANGAN SISTEM.....	7
3.1 Data Set	7
3.2 Metodologi Deteksi Kesamaan Konten Website dengan Website	8
3.3 Metodologi Deteksi Kesamaan Konten Software dengan Website	10
3.4 Use Case Diagram.....	13
3.5 Activity Diagram	14
3.5.1 Pengambilan Data Software	14
3.5.2 Pengambilan Data Website	15
3.5.3 Pengelompokan Data Software	16
3.5.4 Pengelompokan Data Website	17
BAB 4 IMPLEMENTASI.....	19
4.1 Implementasi Deteksi Kesamaan Konten Website dengan Website.....	19
4.1.1 Proses Pengambilan Data Web	19
4.1.2 Proses Konversi Data	20
4.1.3 Proses Pengelompokan.....	22
4.2 Implementasi Deteksi Kesamaan Konten Website dengan Software	23
4.2.1 Proses Pengambilan Data Software	23
4.2.2 Proses Konversi Data	24
4.2.3 Proses Pengelompokan.....	27
4.3 Implementasi Pengecekan <i>Stop Words</i>	28
BAB 5 PENGUJIAN	29
5.1 Pengujian Efektifitas Clustering Berbasis Website	29
5.1.1 Pengujian Efektifitas Cluster Untuk Kata – Kata Berbahasa Inggris ..	29

5.1.2 Pengujian Efektifitas Cluster Untuk Kata – Kata Berbahasa Mandarin	39
5.2 Pengujian Efektifitas Clustering Berbasis Software	49
5.3 Pengujian Efektifitas <i>Cosine Similarity</i>	50
5.4 Perkiraan Topik <i>Clustering</i> Berbasis <i>Website</i>	55
5.5 Pengujian Efektifitas Cluster Untuk Kata – Kata Berbahasa Inggris Tanpa Stop Words.....	56
BAB 6 SIMPULAN DAN SARAN.....	68
6.1 Simpulan	68
6.2 Saran.....	68
DAFTAR PUSTAKA	69
LAMPIRAN A NAMA LAMPIRAN.....	A-1
LAMPIRAN B NAMA LAMPIRAN.....	B-1
LAMPIRAN C NAMA LAMPIRAN.....	C-1



DAFTAR GAMBAR

Gambar 2.1 Tampilan User Interface Weka	5
Gambar 2.2 Contoh <i>code Jsoup</i>	6
Gambar 3.1 Tahap-tahap metode deteksi kesamaan konten website dengan website.....	8
Gambar 3.2 <i>Pseudocode</i> pengambilan data pada log.....	8
Gambar 3.3 <i>Pseudocode</i> validasi <i>url</i>	9
Gambar 3.4 <i>Pseudocode</i> pengambilan data pada <i>url</i>	9
Gambar 3.5 <i>Pseudocode</i> pengambilan data pada file teks.....	9
Gambar 3.6 <i>Pseudocode</i> menyimpan hasil kemunculan kata.....	9
Gambar 3.7 <i>Pseudocode</i> mencari indeks dari kata yang dipisah	9
Gambar 3.8 <i>Pseudocode</i> pengisian kata yang belum ada	9
Gambar 3.9 <i>Pseudocode</i> penambahan jumlah kata yang sudah ada.....	9
Gambar 3.10 <i>Pseudocode</i> menyimpan hasil konversi ke dalam file teks	10
Gambar 3.11 <i>Pseudocode</i> merubah format file teks	10
Gambar 3.12 <i>Pseudocode</i> proses pengelompokan.....	10
Gambar 3.13 <i>Pseudocode</i> menyimpan hasil pengelompokan ke dalam file teks .	10
Gambar 3.14 Tahap-tahap metode deteksi kesamaan konten software dengan website.....	11
Gambar 3.15 <i>Pseudocode</i> pengambilan data pada log.....	11
Gambar 3.16 <i>Pseudocode</i> pengambilan data pada <i>url Wikipedia</i>	11
Gambar 3.17 <i>Pseudocode</i> menyimpan data kedalam list.....	12
Gambar 3.18 <i>Pseudocode</i> pengecekan file	12
Gambar 3.19 <i>Pseudocode</i> pengambilan data pada file teks.....	12
Gambar 3.20 <i>Pseudocode</i> menyimpan hasil kemunculan kata.....	12
Gambar 3.21 <i>Pseudocode</i> mencari indeks dari kata yang dipisah	12
Gambar 3.22 <i>Pseudocode</i> pengisian kata yang belum ada	12
Gambar 3.23 <i>Pseudocode</i> penambahan jumlah kata yang sudah ada.....	12
Gambar 3.24 <i>Pseudocode</i> menyimpan hasil konversi ke dalam file teks	13
Gambar 3.25 <i>Pseudocode</i> merubah format file teks	13
Gambar 3.26 <i>Pseudocode</i> proses pengelompokan.....	13

Gambar 3.27 <i>Pseudocode</i> menyimpan hasil pengelompokan ke dalam file teks .	13
Gambar 3.28 <i>Use case</i> web software clusterer	14
Gambar 3.29 <i>Activity</i> pengambilan data software.....	15
Gambar 3.30 <i>Activity</i> pengambilan data website	16
Gambar 3.31 <i>Activity</i> pengelompokan data <i>software</i>	17
Gambar 3.32 <i>Activity</i> pengelompokan data website	18
Gambar 4.1 Kode program pengambilan data pada log.....	19
Gambar 4.2 Kode program validasi <i>url</i>	19
Gambar 4.3 Kode program mengakses url untuk pengambilan data	20
Gambar 4.4 Kode program pengambilan data pada file teks	20
Gambar 4.5 Kode program menyimpan kemunculan data.....	20
Gambar 4.6 Kode program pengecekan indeks kata pada array.....	21
Gambar 4.7 Kode program pengisian kata baru pada list	21
Gambar 4.8 Kode program penambahan jumlah kemunculan kata pada list.....	21
Gambar 4.9 Kode program menyimpan hasil tabel kedalam file teks	22
Gambar 4.10 Kode program merubah format file.....	22
Gambar 4.11 Kode program menentukan <i>cluster</i>	23
Gambar 4.12 Kode program menyimpan hasil <i>cluster</i>	23
Gambar 4.13 Kode program proses pengambilan data	23
Gambar 4.14 Kode program mengambil data	24
Gambar 4.15 Kode program menggabungkan data software.....	24
Gambar 4.16 Kode program pengambilan data pada file teks	25
Gambar 4.17 Kode program menyimpan kemunculan data.....	25
Gambar 4.18 Kode program pengecekan indeks kata pada array.....	25
Gambar 4.19 Kode program pengisian kata baru pada list	26
Gambar 4.20 Kode program penambahan jumlah kemunculan kata pada list.....	26
Gambar 4.21 Kode program menyimpan hasil tabel kedalam file teks	26
Gambar 4.22 Kode program merubah format file.....	27
Gambar 4.23 Kode program menentukan <i>cluster</i>	27
Gambar 4.24 Kode program menyimpan hasil <i>cluster</i>	27
Gambar 4.25 Kode program <i>stop words Inggris</i>	28
Gambar 4.26 Kode program <i>stop words Mandarin</i>	28

Gambar 4.27 Kode program pengecekan stop words	28
Gambar 5.1 Relevansi <i>cluster</i> berbahasa inggris	38
Gambar 5.2 Relevansi cluster berbahasa mandarin	49
Gambar 5.3 Relevansi cluster berbasis software.....	50
Gambar 5.4 Penujian Efektifitas <i>Cosine Similarity</i>	55
Gambar 5.5 Relevansi cluster berbahasa inggris tanpa stop words	66



DAFTAR TABEL

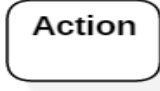
Tabel 5.1 Tabel hasil kemunculan kata terbanyak grup 1 sampai 10	29
Tabel 5.2 Tabel relevansi grup 1-10	29
Tabel 5.3 Tabel hasil kemunculan kata terbanyak grup 11 sampai 20	30
Tabel 5.4 Tabel relevansi grup 11-20	30
Tabel 5.5 Tabel hasil kemunculan kata terbanyak grup 21 sampai 30	30
Tabel 5.6 Tabel relevansi grup 21-30	31
Tabel 5.7 Tabel hasil kemunculan kata terbanyak grup 31 sampai 40	31
Tabel 5.8 Tabel relevansi grup 31-40	31
Tabel 5.9 Tabel hasil kemunculan kata terbanyak grup 41 sampai 50	32
Tabel 5.10 Tabel relevansi grup 41-50	32
Tabel 5.11 Tabel hasil kemunculan kata terbanyak grup 51 sampai 60	33
Tabel 5.12 Tabel relevansi grup 51-60	33
Tabel 5.13 Tabel hasil kemunculan kata terbanyak grup 61 sampai 70	33
Tabel 5.14 Tabel relevansi grup 61-70	34
Tabel 5.15 Tabel hasil kemunculan kata terbanyak grup 71 sampai 80	34
Tabel 5.16 Tabel relevansi grup 71-80	34
Tabel 5.17 Tabel hasil kemunculan kata terbanyak grup 81 sampai 90	35
Tabel 5.18 Tabel relevansi grup 81-90	35
Tabel 5.19 Tabel hasil kemunculan kata terbanyak grup 91 sampai 100	36
Tabel 5.20 Tabel relevansi grup 91-100	36
Tabel 5.21 Tabel hasil kemunculan kata terbanyak grup 101 sampai 110	36
Tabel 5.22 Tabel relevansi grup 101-110	37
Tabel 5.23 Tabel hasil kemunculan kata terbanyak grup 111 sampai 118	37
Tabel 5.24 Tabel relevansi grup 111-118	37
Tabel 5.25 Tabel hasil kemunculan kata terbanyak grup 1 sampai 10	39
Tabel 5.26 Tabel relevansi grup 1-10	39
Tabel 5.27 Tabel hasil kemunculan kata terbanyak grup 11 sampai 20	40
Tabel 5.28 Tabel relevansi grup 11-20	40
Tabel 5.29 Tabel hasil kemunculan kata terbanyak grup 21 sampai 30	41
Tabel 5.30 Tabel relevansi grup 21-30	41

Tabel 5.31 Tabel hasil kemunculan kata terbanyak grup 31 sampai 40	41
Tabel 5.32 Tabel relevansi grup 31-40	42
Tabel 5.33 Tabel hasil kemunculan kata terbanyak grup 41 sampai 50	42
Tabel 5.34 Tabel relevansi grup 41-50	43
Tabel 5.35 Tabel hasil kemunculan kata terbanyak grup 51 sampai 60	43
Tabel 5.36 Tabel relevansi grup 51-60	43
Tabel 5.37 Tabel hasil kemunculan kata terbanyak grup 61 sampai 70	44
Tabel 5.38 Tabel relevansi grup 61-70	44
Tabel 5.39 Tabel hasil kemunculan kata terbanyak grup 71 sampai 80	45
Tabel 5.40 Tabel relevansi grup 71-80	45
Tabel 5.41 Tabel hasil kemunculan kata terbanyak grup 81 sampai 90	45
Tabel 5.42 Tabel relevansi grup 81-90	46
Tabel 5.43 Tabel hasil kemunculan kata terbanyak grup 91 sampai 100	46
Tabel 5.44 Tabel relevansi grup 91-100	47
Tabel 5.45 Tabel hasil kemunculan kata terbanyak grup 101 sampai 110	47
Tabel 5.46 Tabel relevansi grup 101-110	47
Tabel 5.47 Tabel hasil kemunculan kata terbanyak grup 111 sampai 118	48
Tabel 5.48 Tabel relevansi grup 91-100	48
Tabel 5.49 Tabel hasil kemunculan kata berbasis software	49
Tabel 5.50 Tabel relevansi berbasis software	49
Tabel 5.51 Tabel Pengujian Cosine Similarity	50
Tabel 5.52 Tabel Perkiraan Topik <i>Clustering</i> Berbasis <i>Website</i>	55
Tabel 5.53 Tabel hasil kemunculan kata terbanyak grup 1 sampai 10	57
Tabel 5.54 Tabel relevansi grup 1-10	57
Tabel 5.55 Tabel hasil kemunculan kata terbanyak grup 11 sampai 20	58
Tabel 5.56 Tabel relevansi grup 11-20	58
Tabel 5.57 Tabel hasil kemunculan kata terbanyak grup 21 sampai 30	58
Tabel 5.58 Tabel relevansi grup 21-30	59
Tabel 5.59 Tabel hasil kemunculan kata terbanyak grup 31 sampai 40	59
Tabel 5.60 Tabel relevansi grup 31-40	60
Tabel 5.61 Tabel hasil kemunculan kata terbanyak grup 41 sampai 50	60
Tabel 5.62 Tabel relevansi grup 41-50	60

Tabel 5.63 Tabel hasil kemunculan kata terbanyak grup 51 sampai 60	61
Tabel 5.64 Tabel relevansi grup 51-60	61
Tabel 5.65 Tabel hasil kemunculan kata terbanyak grup 61 sampai 70	62
Tabel 5.66 Tabel relevansi grup 61-70	62
Tabel 5.67 Tabel hasil kemunculan kata terbanyak grup 71 sampai 80	62
Tabel 5.68 Tabel relevansi grup 71-80	63
Tabel 5.69 Tabel hasil kemunculan kata terbanyak grup 81 sampai 90	63
Tabel 5.70 Tabel relevansi grup 81-90	64
Tabel 5.71 Tabel hasil kemunculan kata terbanyak grup 91 sampai 100	64
Tabel 5.72 Tabel relevansi grup 91-100	64
Tabel 5.73 Tabel hasil kemunculan kata terbanyak grup 101 sampai 110	65
Tabel 5.74 Tabel relevansi grup 101-110	65
Tabel 5.75 Tabel hasil kemunculan kata terbanyak grup 111 sampai 118	66
Tabel 5.76 Tabel relevansi grup 111-118	66



DAFTAR NOTASI/ LAMBANG

Jenis	Notasi/ Lambang	Nama	Arti
UML (Use Case)		<i>Actor</i>	Pengguna istem
		<i>Use Case</i>	Entitas yang berhubungan dengan sistem dan pengguna.
		<i>Directed Association</i>	Hubungan satu arah antara dua objek
UML (Flowchart)		<i>Process</i>	Proses yang dilakukan secara komputerisasi.
		<i>Flow</i>	Menunjukkan arah proses.
		<i>Data</i>	Sebuah data.
UML (Activity Diagram)		<i>Action</i>	Aktivitas yang dilakukan pengguna.
		<i>Control Flow</i>	Menunjukkan arah proses.
		<i>Decision</i>	Sebuah pilihan yang dilakukan pengguna.
		<i>Initial</i>	Menandakan mulainya sebuah aktivitas.
		<i>Final</i>	Menandakan berakhirnya sebuah aktivitas.

Referensi:

Notasi/ Lambang UML dari *The Unified Modeling Language Reference Manual* [1].

DAFTAR SINGKATAN

UML	Unified Modelling Language
CSV	Comma Separated Value
WEKA	Waikato Environment for Knowledge Analysis
HTTP	Hypertext Markup Language

