

**IMPLEMENTASI ALGORITMA CLUSTERING BASED ON
FREQUENT WORD SEQUENCES (CFWS) UNTUK
CLUSTERING DOKUMEN ABSTRAK BAHASA INDONESIA
DAN INGGRIS**

TUGAS AKHIR



LILYANI BARRUNG

1132001004

PROGRAM STUDI INFORMATIKA

FAKULTAS TEKNIK DAN ILMU KOMPUTER

UNIVERSITAS BAKRIE

JAKARTA

2018

**IMPLEMENTASI ALGORITMA *CLUSTERING BASED ON
FREQUENT WORD SEQUENCES (CFWS)* UNTUK
CLUSTERING DOKUMEN ABSTRAK BAHASA INDONESIA
DAN INGGRIS**

TUGAS AKHIR

Diajukan sebagai salah satu syarat untuk memperoleh gelar
Sarjana Komputer



LILYANI BARRUNG

1132001004

PROGRAM STUDI INFORMATIKA

FAKULTAS TEKNIK DAN ILMU KOMPUTER

UNIVERSITAS BAKRIE

JAKARTA

2018

HALAMAN PERNYATAAN ORISINALITAS

Tugas Akhir ini adalah hasil karya saya sendiri,

Dan semua sumber baik yang dikutip maupun dirujuk

telah saya nyatakan dengan benar.

Nama : Lilyani Barrung

NIM : 1132001004

Tanda Tangan : 

Tanggal : 20 Agustus 2018

HALAMAN PENGESAHAN

Tuags Akhir ini diajukan oleh:

Nama : Lilyani Barrung
NIM : 1132001004
Program Studi : Informatika
Fakultas : Teknik dan Ilmu Komputer
Judul Tugas Akhir : Implementasi Algoritma *Clustering based on Frequent Word Sequences* (CFWS) untuk *Clustering* Dokumen Abstrak Bahasa Indonesia dan Inggris

Telah berhasil dipertahankan dihadapan Dewan Pengaji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjana Komputer pada Program Studi Informatika Fakultas Teknik dan Ilmu Komputer, Universitas Bakrie.

DEWAN PENGUJI

Pembimbing : Guson P. Kuntarto, S.T., M.Sc.



21/08/18

Pengaji 1 : Yusuf Lestanto, S.T., M.Sc.



24/08

Pengaji 2 : Berkah I. Santoso, S.T., M.T.I.



21/08

Ditetapkan di : Jakarta

Tanggal : 20 Agustus 2018

UNGKAPAN TERIMA KASIH

Puji syukur kepada Tuhan Yesus Kristus atas penyertaan, berkat dan kasih karunia-Nya lah, tugas akhir yang berjudul “Implementasi Algoritma Clustering Based on Frequent Word Sequence (CFWS) untuk Clustering Dokumen Abstrak Bahasa Indonesia dan Inggris” dapat terselesaikan dengan baik. Penulisan tugas akhir ini dilakukan guna memenuhi salah satu syarat dalam mencapai gelar Sarjana Komputer Program Studi Informatika pada Fakultas Teknologi dan Ilmu Komputer Universitas Bakrie.

Dalam proses penelitian dan penyusunan Tugas Akhir ini tentu saja tidak terlepas dari suka dan duka. Banyak pihak yang turut memberikan doa, nasihat, bantuan, motivasi, dan juga semangat selama penyusunan tugas akhir ini dilakukan. Oleh karena itu, dengan segala hormat Penulis menyampaikan rasa terima kasih kepada:

1. Kedua orang tua tercinta, Bapak Andarias Barrung, S.E., dan Ibu Yuliaty Pongarrang yang tidak pernah lelah untuk memberikan kasih sayang, dukungan, motivasi, dan doa.
2. Dosen pembimbing tugas akhir, Bapak Guson P. Kuntarto, S.T., M.Sc., yang senantiasa meluangkan waktu dan memberikan bimbingan, motivasi, nasihat, dukungan, semangat serta doa selama proses penyusunan tugas akhir ini.
3. Dosen pembahas dalam seminar proposal dan juga selaku penguji dalam sidang tugas akhir, Bapak Yusuf Lestanto, S.T, M.Sc., yang telah memberikan masukan dan perbaikan untuk penyusunan tugas akhir ini.
4. Dosen penguji kedua dalam sidang tugas akhir, Bapak Berkah I. Santoso, S.T., M.T.I., yang juga telah memberikan masukan dan perbaikan untuk penyusunan tugas akhir ini.
5. Ketua Program Studi, Bapak Prof. Dr. Hoga Saragih, S.T, MT, yang senantiasa memberikan motivasi dan semangat selama penulisan tugas akhir ini.
6. Dosen Pembimbing Akademik, Bapak Irwan Prasetya Gunawan, yang senantiasa memberikan bimbingan, masukan, motivasi, dan juga nasehat selama masa perkuliahan di Universitas Bakrie.

7. Seluruh keluarga besar Universitas Bakrie baik itu dosen maupun *staff*, khususnya untuk Program studi Informatika, yang senantiasa memberikan bantuan dan dukungan selama masa perkuliahan.
8. Kakak tercinta, Lindayani Barrung, yang selalu bersedia mendengarkan keluh kesah selama penyusunan tugas akhir ini, dan juga senantiasa mendukung serta memberikan semangat dan doa.
9. Teman-teman Informatika 2013, Muhammad Khalish Ramadhansyah, Febbie Ramadhini, Rizky Novriyedi Putra, Bagus Aryo Pamungkas, Fitriah Febriani, Millah Fatimah, Iman Nurmansyah, Salsa Ayu Kusumastuti, Jimmy, Ridho Gilang Fiesta, Fadillah Indra, Dede Muhammad Salim, Amelia Fahmi, Fildzah Adra Arifah, Gusti Maulana Arif, dan Yusuf Arwadi, atas kebersamaan, motivasi, semangat, bantuan, dukungan, dan suka duka selama masa perkuliahan di Universitas Bakrie.
10. Teman Kosan, Febbie Ramadhini dan Putri Amelia, yang selalu memberikan semangat dan dukungan, serta bersedia menjadi tempat curhat mengenai suka duka dalam penyusunan tugas akhir ini.
11. Senior maupun junior mahasiswa Informatika serta teman-teman seperjuangan angkatan 2013 Universitas Bakrie atas semangat, dukungan, dan juga doa.
12. Seluruh pihak yang terlibat, baik itu saudara atau pun teman yang telah membantu dan memberikan semangat serta doa dalam penyusunan tugas akhir ini.

Semoga Tuhan Yesus Kristus memberkati dan membalas kebaikan kepada kita semua. Tugas akhir ini tidak terlepas dari keterbatasan, untuk itu kritik dan saran sangat diharapkan sebagai masukan untuk perbaikan di masa mendatang. Semoga Tugas akhir ini dapat bermanfaat bagi semua kalangan pada bidang pendidikan, khususnya bidang Informatika.

Jakarta, 20 Agustus 2018

Lilyani Barrung

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI

Sebagai *civitas* akademik Universitas Bakrie, saya yang bertanda tangan di bawah ini:

Nama : Lilyani Barrung
NIM : 1132001004
Program Studi : Informatika
Fakultas : Teknik dan Ilmu Komputer
Jenis Tugas Akhir : Implementasi Algoritma

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Bakrie **Hak Bebas Royalti Noneksklusif (Non-exclusive Royalty-Free Right)** atas karya ilmiah saya yang berjudul:

Implementasi Algoritma *Clustering based on Frequent Word Sequences* (CFWS) untuk *Clustering* Dokumen Abstrak Bahasa Indonesia dan Inggris

beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Nonekslusif ini Universitas Bakrie berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mempublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta untuk kepentingan akademis.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Jakarta

Pada tanggal : 20 Agustus 2018

Yang menyatakan,



Lilyani Barrung

**IMPLEMENTASI ALGORITMA CLUSTERING BASED ON FREQUENT
WORD SEQUENCES (CFWS) UNTUK CLUSTERING DOKUMEN
ABSTRAK BAHASA INDONESIA DAN INGGRIS**

Lilyani Barrung

ABSTRAK

Klasterisasi merupakan salah satu metode *text mining* yang digunakan untuk mengelompokan dokumen. Terdapat banyak teknik klasterisasi yang dikembangkan, salah satunya adalah teknik yang merepresentasikan dokumen sebagai *bag-of-words* atau sekumpulan kata. Namun, teknik tersebut menyebabkan ukuran dimensi kata yang besar pada dokumen. Untuk mengatasi masalah tersebut digunakan sebuah teknik baru yaitu teknik *sequences-of-words* untuk klasterisasi dokumen. Penelitian ini berfokus untuk melakukan klasterisasi atau pengelompokan dokumen pada *dataset* yang digunakan dengan mengimplementasikan algoritma *Clustering based on Frequent Word Sequences* (CFWS) yang memanfaatkan teknik *sequences-of-words* serta membandingkan hasil klasterisasi dokumen dari kedua *dataset* dengan dua cara yaitu berdasarkan parameter *minimum support* dan berdasarkan stabilitas jumlah *cluster* yang dihasilkan setelah menggunakan *random dataset*. Adapun *dataset* yang digunakan adalah kumpulan dokumen bagian abstrak dari tugas akhir mahasiswa(i) sebanyak 300 dokumen (150 dokumen abstrak Bahasa Indonesia dan 150 dokumen abstrak Bahasa Inggris). Penelitian ini dimulai dengan pengolahan data menjadi *dataset*, kemudian melakukan implementasi CFWS untuk klasterisasi, serta melakukan validasi hasil dan komparasi. Klasterisasi dengan menggunakan algoritma CFWS dapat mengurangi dimensi kata pada dokumen. Dari komparasi yang telah dilakukan pada penelitian ini, kedua cara memiliki hasil yang serupa. Pada komparasi pertama, jumlah *cluster* yang dihasilkan oleh *dataset* abstrak Bahasa Inggris lebih unggul dari segi banyaknya jumlah *cluster* yang dihasilkan. Pada *minimum support* 5%, 20%, 25%, dan 30%, jumlah *cluster* dari *dataset* abstrak Bahasa Inggris cenderung lebih banyak dibandingkan dari *dataset* abstrak Bahasa Indonesia. Jumlah *cluster* terbanyak dihasilkan oleh *minimum support* 10% pada *dataset* abstrak Bahasa Indonesia dan *minimum support* 20% pada *dataset* abstrak Bahasa Inggris yaitu 7 *cluster*. Pada komparasi kedua, jumlah *cluster* yang dihasilkan *dataset* abstrak Bahasa Inggris juga lebih unggul dari segi stabilitas jumlah *cluster* yang dihasilkan karena cenderung lebih stabil. Pada *dataset* abstrak Bahasa Inggris terdapat tiga *minimum support* yang memiliki persentase stabil, sedangkan pada *dataset* abstrak Bahasa Indonesia hanya terdapat dua *minimum support* yang memiliki persentase stabil. Oleh karena itu, berdasarkan kedua hasil komparasi tersebut dapat disimpulkan bahwa *dataset* abstrak Bahasa Inggris cenderung lebih unggul daripada *dataset* abstrak Bahasa Indonesia.

Kata Kunci: Klasterisasi, algoritma CWFS, *cluster*, dokumen abstrak, stabilitas, *minimum support*.

IMPLEMENTATION OF *CLUSTERING BASED ON FREQUENT WORD SEQUENCES (CFWS) ALGORITHM FOR CLUSTERING OF INDONESIAN AND ENGLISH ABSTRACT DOCUMENT*

Lilyani Barrung

ABSTRACT

Clustering is a text mining method to group documents into a cluster. There are many clustering techniques has been developed, one of them is a technique that represents document as bag-of-words or a set of words. However, it causes large word dimensions in the document. One of the solutions to overcome this problem was used sequence-of-words technique for document clustering. This research focused on document clustering of two datasets (Indonesian and English dataset) by implemented the Clustering based on Frequent Word Sequences (CFWS) algorithm which used sequences-of-words technique. Then compared the results of clustering in two ways: based on minimum support parameters and based on stability of the number of clusters generated by random dataset. This research used 300 documents of abstract section in final thesis (150 documents in Indonesian and 150 documents in English). This research began with processing data into datasets, then implemented CFWS algorithm for clustering, and then validated results and comparisons. CFWS algorithm has been managed to reduce word dimension of document. There were two comparisons has been done in this research. The first comparison, the number of clusters generated by English datasets was superior in terms of the number of clusters. On minimum support 5%, 20%, 25%, and 30% the number of clusters from the English dataset more than Indonesian dataset. The largest number of clusters by Indonesian dataset was generated on minimum support 10% and on minimum support 20% by English dataset, it was 7 clusters. And the second comparison has same result as first comparison, the number of clusters generated by English dataset was superior in terms of stability of the number of clusters. There were three minimum support has stable percentages on English dataset. Meanwhile, there were two minimum support on Indonesian dataset. Thus, based on the result of two comparisons, it has been concluded that English dataset was superior than Indonesian dataset.

Keywords: *Clustering, CWFS algorithm, cluster, abstract documents, stability, minimum support.*

DAFTAR ISI

HALAMAN SAMPUL	
HALAMAN JUDUL.....	ii
HALAMAN PERNYATAAN ORISINALITAS	iii
HALAMAN PENGESAHAN.....	iv
HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI.....	v
ABSTRAK	viii
ABSTRACT.....	ix
DAFTAR ISI.....	x
DAFTAR GAMBAR	xiii
DAFTAR TABEL.....	xiv
DAFTAR RUMUS	xv
DAFTAR KODE SUMBER	xvi
DAFTAR SINGKATAN	xvii
DAFTAR LAMPIRAN.....	xviii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	4
1.3 Ruang Lingkup Penelitian	4
1.4 Tujuan Penelitian.....	5
1.5 Kontribusi Penelitian.....	5
1.6 Sistematika Penulisan Tugas Akhir.....	5
BAB II LANDASAN TEORI	7
2.1 Penelitian Terkait	7
2.2 <i>Text mining</i>	11
2.3 <i>Machine Learning</i>	13
2.3.1 <i>Supervised Learning</i>	13
2.3.2 <i>Unsupervised Learning</i>	14
2.3.3 <i>Semi-Supervised Learning</i>	15
2.4 <i>Clustering</i> Dokumen	15
2.5 Algoritma <i>Clustering based on Frequent Word Sequences</i> (CFWS)....	18

2.5.1	Deskripsi Algoritma CFWS	18
2.5.2	Tahapan Algoritma CFWS.....	18
2.6	Algoritma Apriori.....	21
2.7	<i>Generalized Suffix Tree</i>	22
2.8	Validasi Hasil <i>Clustering</i>	23
2.8.1	Validasi Eksternal	23
2.8.2	Validasi Internal	24
2.8.3	Validasi Relatif.....	24
2.9	<i>Assessing Stability</i>	25
BAB III	METODOLOGI PENELITIAN.....	27
3.1	Tahapan Penelitian	27
3.1.1	Studi Literatur	28
3.1.2	Pengumpulan Data	29
3.1.3	Pengolahan <i>Dataset</i>	30
3.1.4	Implementasi Algoritma CFWS.....	31
3.1.5	Validasi dan Komparasi Hasil Pengujian.....	33
3.1.6	Diskusi dan Pembahasan.....	33
3.1.7	Penyusunan Laporan Penelitian	33
3.2	Jenis Penelitian	34
BAB IV	IMPLEMENTASI DAN PEMBAHASAN	35
4.1	Pengumpulan Data	35
4.2	Pengolahan <i>Dataset</i>	38
4.2.1	<i>Case Folding</i>	40
4.2.2	<i>Filtering</i>	40
4.2.3	<i>Stopword Removal</i>	40
4.3	Implementasi Algoritma CFWS	42
4.3.1	Tahap <i>Finding Frequent Word Sequences</i>	42
4.3.2	Tahap <i>Collecting and Combining Cluster</i>	51
4.4	Komparasi Hasil <i>Clustering</i> Dokumen	65
4.4.1	Validasi Hasil <i>Clustering</i> Dokumen	66
4.4.2	Komparasi Hasil.....	70
4.5	Diskusi Hasil Penelitian	73

BAB V PENUTUP.....	77
5.1 Simpulan.....	77
5.2 Saran	78
DAFTAR PUSTAKA	79
LAMPIRAN	82

DAFTAR GAMBAR

Gambar 2.1 Arsitektur Sistem <i>Text Mining</i> (Feldman & Sanger, 2007)	11
Gambar 2.2 Taksonomi <i>Clustering</i> (Aggarwal & Zhai, 2012)	17
Gambar 2.3 Algoritma Apriori (Wu et al., 2008)	21
Gambar 2.4 Taksonomi Validasi <i>Clustering</i> (Brun et al., 2007)	23
Gambar 2.5 Tahap dalam Teknik <i>Resampling</i> untuk Menentukan Stabilitas (Levine & Domany, 2001)	26
Gambar 3.1 Tahapan Penelitian Implementasi Algoritma CFWS untuk <i>Clustering</i> Dokumen Abstrak Bahasa Indonesia dan Inggris	27
Gambar 3.2 Uraian Tahap Penelitian (Pengolahan <i>Dataset</i> , Implementasi Algoritma CFWS, serta Validasi dan Komparasi Hasil)	28
Gambar 3.3 Proses dalam Tahap <i>Finding Frequent Word Sequences</i>	31
Gambar 3.4 Proses dalam Tahap <i>Collecting and Combining Cluster</i>	32
Gambar 4.1 Tabel <i>Raw Document</i> Bahasa Indonesia pada <i>Database</i>	37
Gambar 4.2 Tabel <i>Raw Document</i> Bahasa Inggris pada <i>Database</i>	37
Gambar 4.3 Tabel <i>Preprocessing Document</i> Bahasa Indonesia pada <i>Database</i>	41
Gambar 4.4 Tabel <i>Preprocessing Document</i> Bahasa Inggris pada <i>Database</i>	41
Gambar 4.5 <i>Pseudocode For-loop Minimum Support</i> pada MainClass.java.....	43
Gambar 4.6 <i>Flowchart For-loop Minimum Support</i> pada MainClass.java	43
Gambar 4.7 Grafik Jumlah Kandidat <i>Cluster</i> dari <i>Dataset</i> Bahasa Indonesia.....	48
Gambar 4.8 Grafik Jumlah Kandidat <i>Cluster</i> dari <i>Dataset</i> Bahasa Inggris.....	50
Gambar 4.11 <i>Heatmap</i> Jumlah <i>Cluster</i> dari <i>Dataset</i> Bahasa Indonesia	67
Gambar 4.12 <i>Heatmap</i> Jumlah <i>Cluster</i> dari <i>Dataset</i> Bahasa Inggris	69
Gambar 4.13 Grafik Kecenderungan Jumlah <i>Cluster</i> dari <i>Dataset</i> Bahasa Indonesia dan Inggris	71

DAFTAR TABEL

Tabel 2.1 Rangkuman Penelitian Terkait.....	10
Tabel 3.1 Jumlah Dokumen Abstrak yang digunakan untuk <i>Dataset</i>	30
Tabel 4.1 Rincian Jumlah Dokumen Abstrak Bahasa Indonesia	36
Tabel 4.2 Rincian Jumlah Dokumen Abstrak Bahasa Inggris	36
Tabel 4.3 Jumlah Dokumen Abstrak yang digunakan untuk <i>Dataset</i>	37
Tabel 4.4 Contoh Dokumen Abstrak Bahasa Indonesia yang digunakan.....	38
Tabel 4.5 Contoh Dokumen Abstrak Bahasa Inggris yang digunakan	39
Tabel 4.6 Hasil <i>Preprocessing</i> Dokumen Abstrak Bahasa Indonesia	40
Tabel 4.7 Hasil <i>Preprocessing</i> Dokumen Abstrak Bahasa Inggris.....	41
Tabel 4.8 Besar Data dan Jumlah Kata pada <i>Raw Document, Preprocessing Document, Compact Document</i>	45
Tabel 4.9 Hasil F2WS, <i>Node</i> , FWS, dan Kandidat <i>Cluster</i> untuk <i>Dataset</i> Bahasa Indonesia	47
Tabel 4.10 Hasil F2WS, <i>Node</i> , FWS, dan Kandidat <i>Cluster</i> untuk <i>Dataset</i> Bahasa Inggris	49
Tabel 4.11 <i>Temporary Cluster</i> dari Hasil Proses <i>K-Mismatch</i>	52
Tabel 4.12 <i>Temporary Cluster</i> dari Hasil Proses <i>K-Mismatch</i> (Lanjutan)	53
Tabel 4.13 <i>Temporary Cluster</i> dari Hasil Proses <i>K-Mismatch</i> (Lanjutan)	54
Tabel 4.14 <i>Temporary Cluster</i> dari Hasil Proses <i>K-Mismatch</i> (Lanjutan)	55
Tabel 4.15 <i>Final Cluster</i> dari tiap Nilai <i>Threshold</i> (<i>Dataset</i> Bahasa Indonesia)..	57
Tabel 4.16 <i>Final Cluster</i> dari tiap Nilai <i>Threshold</i> (<i>Dataset</i> Bahasa Inggris).....	58
Tabel 4.17 <i>Final Cluster</i> dari Hasil Perhitungan Nilai <i>Overlapping</i>	63
Tabel 4.18 <i>Final Cluster</i> dari Hasil Perhitungan Nilai <i>Overlapping</i> (Lanjutan)..	64
Tabel 4.19 <i>Final Cluster</i> dari Hasil Perhitungan Nilai <i>Overlapping</i> (Lanjutan)..	65
Tabel 4.20 Jumlah <i>Cluster</i> dari <i>Random Dataset</i> Bahasa Indonesia	66
Tabel 4.21 Jumlah <i>Cluster</i> dari <i>Random Dataset</i> Bahasa Inggris	68
Tabel 4.22 Jumlah <i>Cluster</i> dari Dua <i>Dataset</i>	70
Tabel 4.23 Presentase Stabilitas Jumlah <i>Cluster</i> dari <i>Dataset</i> Bahasa Indonesia dan Inggris	72

DAFTAR RUMUS

Rumus (2.1) Nilai *Overlapping*

DAFTAR KODE SUMBER

Kode Sumber 4.1 Fungsi <i>Preprocessing</i> Dokumen.....	39
Kode Sumber 4.2 Metode <i>StartAprioriAlgorithm()</i> pada MainClass.java.....	44
Kode Sumber 4.3 Metode <i>StartGeneralizedSuffixTree()</i> pada MainClass.java....	46
Kode Sumber 4.4 Fungsi Proses <i>K-Mismatch</i>	51
Kode Sumber 4.5 Fungsi Menghitung Nilai <i>Overlapping Cluster</i>	56

DAFTAR SINGKATAN

CFWS	<i>Clustering based on Frequent Word Sequences</i>
F2WS	<i>Frequent 2-words Sequences</i>
FWS	<i>Frequent word sequences</i>
GST	<i>Generalized Suffix Tree</i>
VSM	<i>Vector Space Model</i>

DAFTAR LAMPIRAN

Lampiran 1: Dokumen Abstrak Bahasa Indonesia

Lampiran 2: Dokumen Abstrak Bahasa Inggris

Lampiran 3: *Dataset* Dokumen Abstrak Bahasa Indonesia

Lampiran 4: *Dataset* Dokumen Abstrak Bahasa Indonesia

Lampiran 5: *Compact Document* untuk *Dataset* Bahasa Indonesia

Lampiran 6: *Temporary Cluster* untuk *Dataset* Bahasa Indonesia

Lampiran 7: *Final Cluster* untuk *Dataset* Bahasa Indonesia

Lampiran 8: *Compact Document* untuk *Dataset* Bahasa Inggris

Lampiran 9: *Temporary Cluster* untuk *Dataset* Bahasa Inggris

Lampiran 10: *Final Cluster* untuk *Dataset* Bahasa Inggris