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 Penguji 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.	(.....)
Penguji 1	:	Yusuf Lestanto, S.T., M.Sc.	(.....)
Penguji 2	:	Berkah I. Santoso, S.T., M.T.I.	(.....)

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 membatu 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 dalambentuk 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 presentase stabil, sedangkan pada *dataset* abstrak Bahasa Indonesia hanya terdapat dua *minimum support* yang memiliki presentase 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

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR RUMUS

DAFTAR KODE SUMBER

DAFTAR SINGKATAN

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

DAFTAR LAMPIRAN