Deteksi Kemiripan Dokumen Menggunakan Metode *Document Clustering K-Means Clustering*

Diajukan Sebagai Syarat Untuk Menyelesaikan
Pendidikan Program Strata-1 Pada
Jurusan Teknik Informatika

Oleh **:**

Muhammad Ikhsan Kamil
NIM : 09021181823011

Jurusan Teknik Informatika
FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA
2023

LEMBAR PENGESAHAN SKRIPSI

Deteksi Kemiripan Dokumen Meggunakan Metode
Document Clustering K-Means Clustering

Oleh :

Muhammad Ikhsan Kamil
NIM : 09021181823011

Indralaya, 21 Desember 2023

Mengetahui Ketua Jurusan Teknik Informatika	Pembimbing
Alvi Syahrini Utami., M.Kom.	Dr. Abdiansah, S.Kom., M.Cs
NIP. 19781222200642003	NIP. 198411012009121005

TANDA LULUS UJIAN KOMPREHENSIF

Pada hari **Kamis** tanggal **21 Desember 2023** telah dilaksanakan ujian komprehensif skripsi oleh Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Nama : Muhammad Ikhsan Kamil
NIM : 09021181823011
Judul : Deteksi Kemiripan Dokumen Menggunakan *Document Clustering K-Means Clustering*.

dan dinyatakan **LULUS**.

1. Ketua Penguji

Rizki Kurniati, M.T.
NIP. 199107122019032016 .....

2. Penguji

Novi Yusliani, M.T.
NIP. 198211082012122001 .....

3. Pembimbing

Dr. Abdiansah, S.Kom., M.Cs
NIP. 198411012009121005 .....

Mengetahui,
Ketua Jurusan Teknik Informatika

Alvi Syahrini Utami, M.Kom.
NIP.197812222006042003

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini:

Nama	:	Muhammad Ikhsan Kamil
NIM	:	09021181823011
Program Studi	:	Teknik Informatika
Judul Skripsi	:	Deteksi Kemiripan Dokumen Menggunakan Metode *Document Clustering K-Means Clustering*

Hasil pengecekan Software *iThenticate/Turnitin* : 15%

Menyatakan bahwa Laporan Proyek saya merupakan hasil karya sendiri dan bukan hasil penjiplakan/plagiat. Apabila ditemukan unsur penjiplakan/plagiat dalam laporan proyek ini, maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya dan Ketentuan yang berlaku.

Demikian, pernyataan ini saya buat dengan sebenarnya dan tidak ada paksaan oleh siapapun.

Indralaya, 6 Januari 2024

Muhammad Ikhsan Kamil

NIM. 09021181823011

MOTTO DAN PERSEMBAHAN

"If you walk down the path that you believe is right, you cannot be wrong"

Ku persembahkan karya tulis ini kepada :

- Ayah, Ibu, dan Adik
- Teman-teman seperjuangan
- Dosen Pembimbing
- Fakultas Ilmu Komputer Universitas Sriwijaya

KATA PENGANTAR

Puji dan Syukur kehadirat Allah SWT atas segala nikmat, rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan Tugas Akhir yang berjudul "**Deteksi Kemiripan Dokumen Menggunakan Metode *Document Clustering K-Means Clustering***" . Tugas Akhir ini disusun untuk memenuhi salah satu persyaratan kelulusan tingkat sarjana pada jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Dalam menyelesaikan Tugas Akhir ini banyak pihak yang telah memberikan bantuan dan dukungan baik secara langsung maupun secara tidak langsung. Pada kesempatan ini, penulis ingin menyampaikan ucapan terimakasih kepada pihak-pihak yang telah membantu penulis dalam menyelesaikan Tugas Akhir ini, yaitu kepada:

1. Allah Subhanahu Wa Ta'ala yang telah memberikan hamba keimanan, kecerdasan, kemudahan dan kelancaran sehingga hamba dapat menyelesaikan tugas-tugas sebagai seorang mahasiswa.
2. Kedua Orang Tua penulis tercinta adik penulis yang selalu senantiasa mendukung dan percaya bahwa penulis dapat menyelesaikan Tugas Akhir ini.
3. Universitas Sriwijaya yang telah memberikan saya kesempatan dan berbagai fasilitas dalam perkuliahan.
4. Ibu Alvi Syahrini Utami, M.Kom Selaku Ketua Jurusan Teknik Informatika.
5. Bapak Dr. Abdiansah S.Kom., M.Cs sebagai pembimbing Tugas Akhir yang mengarahkan dan memberi masukan dalam proses pengerjaan sehingga penulis dapat menyelesaikan Tugas Akhir dengan baik.
6. Ibu Rizki Kurniati, M.T. dan ibu Novi Yusliani, M.T. selaku dosen penguji, yang telah memberikan masukan sehingga Tugas akhir ini menjadi lebih baik lagi.
7. Para teman-teman seperjuangan Ananda Meilizar Dwi Putra, Agung Sukrisna Jaya, Renaldi Budi Setiawan, Muhammad Reza Kurniawan Muhammad Sholeh, Muhammad Ariq Jagabaya, dan masih banyak yang lainnya telah membantu penulis saat kesulitan dalam mengerjakan Tugas Akhir, memberikan motivasi dan semangat.

8. Serta teman-teman seperjuangan Angkatan 2018 yang tidak tertuliskan dalam kata pengantar ini namun turut membantu dalam proses utnuk mencapai gelar sarjana ini.

Penulis menyadari bahwa laporan Tugas Akhir ini masih banyak kekurangan dan masih jauh dari kata sempurna karena keterbatasan ilmu yang dimiliki penulis. Oleh karena itu, penulis mengharapkan kritik dan saran yang membantu untuk kesempurnaan Tugas Akhir ini. Semoga Tugas Akhir ini dapat memberikan manfaat bagi orang banyak.

Inderalaya, 28 Desember 2023

Muhammad Ikhsan Kamil

ABSTRACT

Detection, is an action or process of identifying the presence of something that is concealed. This research aim to develop a software that can be used to detect the similarity between thesis using the K-Means Clustering method, which is one of the simplest and popular unsupervised machine learning algorithms. In this research the detection is done on 56 documents using the silhouette method to determine the optimal number of cluster and davies-bouldin index to evaluate the clustering result. The results of the research show that based on the documents studied, the optimal number of clusters was 35 clusters. In which there are 5 clusters that have a population of more than 2 documents.

Keywords : Detection, K-Means Clustering, Silhouette Score, Davies-Bouldin Index.

ABSTRAK

Deteksi, merupakan suatu tindakan atau proses mengidentifikasi keberadaan sesuatu yang disembunyikan. Penelitian ini bertujuan untuk mengembangkan perangkat lunak yang dapat digunakan untuk mendeteksi kemiripan antar skripsi dengan menggunakan metode *K-Means Clustering* yang merupakan salah satu algoritma unsupervised machine learning yang paling sederhana dan popular. Pada penelitian ini pendeteksian dilakukan terhadap 56 dokumen dengan menggunakan metode *Silhouette Score* untuk menentukan jumlah klister yang optimal dan *davies-bouldin index* untuk mengevaluasi hasil klasterisasi. Hasil penelitian menunjukkan bahwa berdasarkan dokumen yang diteliti, jumlah klaster yang optimal adalah 35 klaster. Dimana terdapat 5 klaster yang mempunyai populasi lebih dari 2 dokumen
.

Kata Kunci : Deteksi, *K-Means Clustering, Silhouette Score, Davies-Bouldin Index.*

DAFTAR ISI

Halaman

DAFTAR GAMBAR

Halaman

DAFTAR TABEL

Halaman

BAB I

PENDAHULUAN

1.1. Pendahuluan

Dalam bab ini dijelaskan mengenai pokok – pokok pikiran yang melandasi penelitian ini. Pokok – pokok pikiran dalam penelitian ini antara lain latar belakang, masalah penelitian, perumusan masalah/permasalahan penelitian, tujuan penelitian dan manfaat penelitian. Pokok – pokok pikiran dalam penelitian ini akan menjadi acuan dalam penelitian metode penelitian.

1.2. Latar Belakang

Perkembangan teknologi yang sangat pesat saat ini telah memberikan banyak manfaat serta kemudahan dalam kemajuan di berbagai aspek, salah satunya adalah aspek pendidikan. Seiring dengan kemajuannya, banyak terjadi pula pelaggaran di dalamnya. Salah satu dari pelanggaran ini adalah kemiripan suatu dokumen dengan dokumen lainnya.

Dalam penelitian yang dilakukan oleh Sraka dan Kaučič (2009) sebanyak 100 dari 138 mahasiswa yang mengikuti survey menyatakan bahwa mereka setidaknya pernah melakukan 1imana11sm setidaknya sekali selama masa studi mereka. Melalui penelitian ini dapat dilihat bahwa pelanggaran ini dapat dikatakan marak terjadi di kalangan mahasiswa. Untuk mengurangi perkembangan akademik yang buruk, penelitian mengenai deteksi kesamaan isi dari suatu dokumen sangat dibutuhkan.

Salah satu metode yang dapat digunakan untuk mendeteksi kesamaan isi dari suatu dokumen adalah dengan menggunakan klasterisasi dokumen. Metode klasterisasi yang dapat dipakai adalah *K – Means clustering*.

1.3. Rumusan Masalah

Rumusan Masalah dalam penelitian ini adalah sebagai berikut :

1. Bagaimana melakukan deteksi kemiripan dokumen menggunakan metode *K-Means Clustering* ?
2. Bagaimana kinerja metode *K-Means Clustering* dalam mendeteksi kemiripan dokumen?

1.4. Tujuan Penelitian

Tujuan Penelitian ini adalah sebagai berikut :

1. Membangun perangkat lunak unutk mendeteksi kemiripan dokumen menggunakan metode *K-Means Clustering*
2. Mengukur kinerja metode *K-Means Clustering* dalam mendeteksi kemiripan dokumen

1.5. Manfaat Penelitian

Manfaat dari penelitian ini adalah sebagai berikut :

1. Sistem yang dibuat dapat membantu pengguna untuk mendeteksi adanya kemiripan dokumen
2. Hasil penelitian dapat dijadikan sebagai rujukan untuk penelitian terkait di masa mendatang.

1.6. Batasan Masalah

Batasan Masalah dalam penelitian ini adalah sebagai berikut :

1. Data yang diperoleh merupakan tugas akhir mahasiswa di Jurusan Teknik Informatika, Fakultas Ilmu Komputer, Universitas Sriwijaya.
2. Metode yang digunkana adalah *K-Means Clustering*¸*Sillhoutte Coefficient,* dan *Davies-Bouldin Index*.

1.7. Sistematika Penulisan

Sistematika Penulisan dalam penelitian ini adalah sebagai berikut :

BAB I. PENDAHULUAN

Pada bab ini menjelaskan tentang latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian dan Batasan masalah. Hal ini akan menjadi dasar dan acuan dalam pengembangan penelitian pada bab selanjutnya.

BAB II. KAJIAN LITERATUR

Pada bab ini dibahas mengenai landasan teori yang digunakan di dalam penelitian termasuk di dalamnya mengenai *Document Clustering, K-Means Clustering* dan penelitian lainnya yang relevan.

BAB III. METODOLOGI PENELITIAN

Pada bab ini dibahas proses pengumpulan data dan tahapan – tahapan di dalam penelitian. Tahapan penelitian akan dibahas lebih rinci berdasarkan kerangka kerja tertentu. Dibagian akhir bab ini akan dimuat rancangan manajemen proyek penelitian.

BAB IV. PENGEMBANGAN PERANGKAT LUNAK

Pada bab ini akan dibahas mengenai perancangan dan lingkungan deteksi kemiripan dokumen menggunakan metode *K-Means Clustering*, hasil pelaksaan system, serta hasil pengujian.

BAB V. HASIL DAN ANALISIS PENELITIAN

Pada bab ini memaparkan hasil dari penerapan serta pengujian metode yang telah dirancang, yang akan disampaikan secara detail. Analisis yang disajikan akan menjadi dasar bagi kesimpulan yang akan diambil dalam penelitian

BAB VI. KESIMPULAN DAN SARAN

Pada bab ini akan dibahas mengenai kesimpulan dari penjabaran hasil dari penelitian. Selain itu, juga menyajikan saran atau rekomendasi yang diharapkan dapan memberikan manfaat dalam pengembangan sistem deteksi kemiripan dokumen.

1.8. Kesimpulan

Pada bab ini telah dibahas mengenai latar belakang penelitian serta acuan penting dalam penelitian seperti latar belakang, rumusan masalah, tujuan penelitian, Batasan masalah dan sistematika penulisan.

DAFTAR PUSTAKA

Aggarwal, C. C., & Zhai, C. X. (2013). Mining text data. In *Mining Text Data* (Vol. 9781461432, Issue August). https://doi.org/10.1007/978-1-4614-3223-4

Ahmadyfard, A., & Modares, H. (2008). Combining PSO and k-means to enhance data clustering. *2008 International Symposium on Telecommunications, IST 2008*, 688–691. https://doi.org/10.1109/ISTEL.2008.4651388

Anggodo, Y. P., Cahyaningrum, W., Fauziyah, A. N., Khoiriyah, I. L., Kartikasari, O., & Cholissodin, I. (2017). Hybrid K-Means Dan Particle Swarm Optimization Untuk. *Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK)*, *4*(2), 104–110.

Arief, M. rizal, Siahaan, D. O., Arie, I., & Shanti. (2010). KLASTERISASI TEKS MENGGUNAKAN METODE MAX-MAX ROUGHNESS (MMR) DENGAN PENGAYAAN SIMILARITAS KATA Mohammad Rizal Arief , Daniel O Siahaan , Isye Arieshanti. *Jurnal Ilmiah KURSOR*, *5*(4), 246–255.

Arifin, Z., Stefanus, S., & Soeleman, A. M. (2017). Klasterisasi Genre Cerpen Kompas Menggunakan Agglomerative Hierarchical Clustering- Single Linkage. *Jurnal Teknologi Informasi Cyberku*, *13*(2), 92–100.

Bisilisin, F. Y., Herdiyeni, Y., & Silalahi, B. P. (2017). Optimasi K-Means Clustering Menggunakan Particle Swarm Optimization pada Sistem Identifikasi Tumbuhan Obat Berbasis Citra. *Jurnal Ilmu Komputer Dan Agri-Informatika*, *3*(1), 37. https://doi.org/10.29244/jika.3.1.37-46

Cai, D., He, X., & Han, J. (2011). Locally consistent concept factorization for document clustering. *IEEE Transactions on Knowledge and Data Engineering*, *23*(6), 902–913. https://doi.org/10.1109/TKDE.2010.165

Cozzolino, I., & Ferraro, M. B. (2022). Document clustering. *Wiley Interdisciplinary Reviews: Computational Statistics*, *14*(6), 1–13. https://doi.org/10.1002/wics.1588

Cui, X., Potok, T. E., & Palathingal, P. (2005). Document clustering using particle swarm optimization. *Proceedings - 2005 IEEE Swarm Intelligence Symposium, SIS 2005*, *2005*(May 2014), 191–197. https://doi.org/10.1109/sis.2005.1501621

Desikan, K., & Huang, A. (n.d.). *Similarity Measures for Text Document Clustering Related papers ARABIC T EXT SUMMARIZAT ION BASED ON LAT ENT SEMANT IC ANALYSIS T O ENHANCE ARABI… Int ernat ional Journal of*

Dat a Mining & Knowledge Management Process (IJDKP), Abdelmon… Experiment al Es.

Fan, Y., Gongshen, L., Kui, M., & Zhaoying, S. (2018). Neural Feedback Text Clustering with BiLSTM-CNN-Kmeans. *IEEE Access*, *6*, 57460–57469. https://doi.org/10.1109/ACCESS.2018.2873327

G., K., & Wahi, A. (2011). Improving the Cluster Performance By Combining Pso and K-Means Algorithm. *ICTACT Journal on Soft Computing*, *01*(04), 206–208. https://doi.org/10.21917/ijsc.2011.0032

Gosno, E. B., Arieshanti, I., & Soelaiman, R. (2013). Implementasi KD-Tree K-Means Clustering. *Jurnal Teknik Pomits*, *2*(2), A432–A437.

Harliana, H., Herdian Bhakti, R. M., Saeful Bachri, O., & Sofian Efendi, F. (2021). Optimasi K-Means dengan Particle Swarm Optimization pada Pengelompokkan Daerah Stunting. *Jurnal Ilmiah Intech : Information Technology Journal of UMUS*, *3*(02), 95–101. https://doi.org/10.46772/intech.v3i02.457

Informasi, F. T. (2008). *PENGGALIAN FREQUENT TREES*.

Janani, R., & Vijayarani, S. (2019). Text document clustering using Spectral Clustering algorithm with Particle Swarm Optimization. *Expert Systems with Applications*, *134*, 192–200. https://doi.org/10.1016/j.eswa.2019.05.030

Kambey, G. E. I., & Dkk. (2020). Penerapan Clustering pada Aplikasi Pendeteksi Kemiripan Dokumen Teks Bahasa Indonesia. *Penerapan Clustering Pada Aplikasi Pendeteksi Kemiripan Dokumen Teks Bahasa Indonesia*, *15*(2), 75–82.

Kurnia Bakti, V., & Indriyatno, J. (2017). Klasterisasi Dokumen Tugas Akhir Menggunakan K-Means Clustering, Sebagai Analisa Penerapan Sistem Temu Kembali. *KOPERTIP : Jurnal Ilmiah Manajemen Informatika Dan Komputer*, *1*(1), 31–34. https://doi.org/10.32485/kopertip.v1i1.8

Michael Steinbach, George Karypis, and V. K. (2000). Technical Report : A Comparison of Document Clustering Techniques Michael. *University of Minnesota*.

Mr.Kaushi K Phukon MCA, P. H. K. B. (2013). Extension of the Fuzzy C Means Clustering Algorithm To Fit With the Composite Graph. *International Journal Of Cognitive Research In Science,Engineering and Education*, *1*(2).

Pednekar, A. M. (2019). *Optimal initialization of K-means using Particle Swarm Optimization*. http://arxiv.org/abs/1904.09098

Pradnyana, G., & Djunaidy, A. (2013). Metode Weighted Maximum Capturing Untuk Klasterisasi Dokumen Berbasis Frequent Itemsets. *Jurnal Ilmu Komputer*, *6*(2), 1–10.

Purnama, P. (2018). *Penggunaan Particle Swarm Optimization (PSO) pada K-Means untuk Pengelompokan Jenis Fluida Minyak Bumi*.

Riduwan, M., Fatichah, C., & Yuniarti, A. (2019). Klasterisasi Dokumen Menggunakan Weighted K-Means Berdasarkan Relevansi Topik. *JUTI: Jurnal Ilmiah Teknologi Informasi*, *17*(2), 146. https://doi.org/10.12962/j24068535.v17i2.a892

Shah, N., & Mahajan, S. (2012). Document Clustering: A Detailed Review. *International Journal of Applied Information Systems*, *4*(5), 30–38. https://doi.org/10.5120/ijais12-450691

Tagarelli, A. (2011). XML Document Clustering. *Handbook of Research on Innovations in Database Technologies and Applications*, 665–673. https://doi.org/10.4018/978-1-60566-242-8.ch071

Usino, W., Prabuwono, A. S., Allehaibi, K. H. S., Bramantoro, A., Hasniaty, A., & Amaldi, W. (2019). Document similarity detection using K-Means and cosine distance. *International Journal of Advanced Computer Science and Applications*, *10*(2), 165–170. https://doi.org/10.14569/ijacsa.2019.0100222

Xu, J., Wang, P., Tian, G., Xu, B., Zhao, J., Wang, F., & Hao, H. (2015). Short text clustering via convolutional neural networks. *1st Workshop on Vector Space Modeling for Natural Language Processing, VS 2015 at the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2015*, 62–69. https://doi.org/10.3115/v1/w15-1509