SNESTIK

Seminar Nasional Teknik Elektro, Sistem Informasi, dan Teknik Informatika

https://ejurnal.itats.ac.id/snestik dan https://snestik.itats.ac.id

Informasi Pelaksanaan :
SNESTIK II - Surabaya, 26 Maret 2022
Ruang Seminar Gedung A, Kampus Institut Teknologi Adhi Tama Surabaya

Informasi Artikel:
DOI : 10.31284/p.snestik.2022.2812

Prosiding ISSN 2775-5126

Fakultas Teknik Elektro dan Teknologi Informasi-Institut Teknologi Adhi Tama Surabaya
Gedung A-ITATS, Jl. Arief Rachman Hakim 100 Surabaya 60117 Telp. (031) 5945043
Email : snestik@itats.ac.id

PENERAPAN METODE *CLUSTERING* K-MEDOIDS UNTUK PENGELOMPOKAN ABSTRAK SKRIPSI BERBASIS WEB

Ruly Adi Permana[1], Dian Puspita Hapsari[2], Rotul Rotul Muhima[3]

[1,2]Jurusan Teknik Informatika, Fakultas Teknik Elektro dan Teknologi Informasi
Institut Teknologi Adhi tama Surabaya
email : *permanaruly@gmail.com*

ABSTRACT

literature review, the results of field research, or the results of the development of trials or experiments. Currently, the search for scientific papers in the Department of Informatics ITATS is only based on the title. This study aims to build a search application that is able to automatically classify thesis documents using the k-medoids clustering algorithm based on the abstracts contained in the document. The data group is in the form of 105 thesis abstract text data in the Department of Informatics ITATS with a span of three years. The results of the simulation of grouping thesis abstracts obtained a more optimal K value when it was at K=4. Testing the confidence of the shaped cluster using the Silhouette Coefficient method. The Silhouette Coefficient value is -0.028975.

Keywords: *Clustering, K-Medoids, Silhouette Coefficient*

ABSTRAK

Karya ilmiah yang ditulis mahasiswa program S1 atau yang disebut skripsi ditulis berdasarkan hasil kajian pustaka, hasil penelitian lapangan, ataupun hasil pengembangan uji coba atau eksperimen. Kegiatan pencarian karya ilmiah skripsi di Jurusan Teknik Informatika ITATS saat ini hanya berdasarkan judul. Penelitian ini bertujuan untuk membangun aplikasi pencarian yang mampu mengelompokkan dokumen skripsi secara otomatis menggunakan algoritma clustering k-medoids berdasarkan abstrak yang ada dalam dokumen tersebut. Kelompok data berupa 105 data text abstrak skripsi di Jurusan Teknik Informatika ITATS dengan rentang waktu tiga tahun. Dilakukan preproses data text kemudian dihitung total simpangan dengan menghitung nilai total *distance* baru dengan total *distance* lama. Hasil simulasi pengelompokan abstrak skripsi diperoleh nilai K yang lebih optimal pada saat berada di K=4. Pengujian kepercayaan dari

cluster yang berbentuk menggunakan metode Silhouette Coefficient. Diperoleh nilai Silhouette Coefficient sebesar -0,028975.

Kata Kunci : *Clustering, K-Medoids, Silhouette Coefficient*

PENDAHULUAN

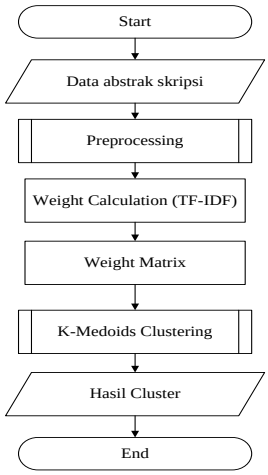
Sebuah karya ilmiah skripsi merupakan bentuk laporan hasil penelitian yang dilakukan oleh mahasiswa strata satu (S1). Saat ini perpustakaan ITATS mengelola banyak dokumen skripsi dari semua jurusan yang ada. Perpustakaan juga menjadi sumber referensi utama untuk mahasiswa dalam menyusun tinjauan pustaka yang menjadi bagian dari sebuah laporan skripsi. Kegiatan pencarian referensi menjadi kegiatan yang penting, sehingga membutuhkan sebuah aplikasi yang mampu mengelompokan tema skripsi berdasarkan abstrak. Saat ini belum ada aplikasi pencarian yang digunakan perpustakaan jurusan Teknik Informatika ITATS yang dilengkapi fitur pengelompokan tema secara otomatis berdasarkan abstrak skripsi. Pencarian informasi mengenai dokumen skripsi Jurusan Teknik Informatika pada sistem informasi perpustakaan ITATS saat ini hanya sebatas pencarian skripsi berdasarkan judul tanpa mengetahui isi dari penelitian yang tercantum pada abstrak. Hal ini membuat mahasiswa kesulitan dalam mencari referensi berupa informasi kategori terkait dengan topik penelitiannya yang akan dikerjakan.

Sehingga penlitian ini bertujuan untuk membangun aplikasi yang dilengkapi dengan fitur pencarian yang dapat mengelompokkan tema skripsi secara otomatis berdasarkan abstrak skripsi untuk memudahkan pencarian informasi yang dibutuhkan. Penelitian yang telah dilakukan menggunakan metode clustering dengan k-Medoids antara lain [1][3][5], menjelaskan kelebihan metode tersebut mampu menyelesaikan masalah K-means dan menghasilkan cluster kosong serta sensitif terhadap outlier atau noise. Untuk menilai tingkat kepercayaan dari hasil pengelompokan metode clustering k-medoids dalam penelitian ini menggunakan Silhouette Coefficient [2][4][6].

Berdasarkan latar belakang yang telah dijelaskan sebelumnya, maka dapat dirumuskan permasalahan. Yaitu bagaimana mengelompokkan data text dari abstrak skripsi mahasiswa jurusan teknik informatika ITATS dengan menggunakan metode clustering K-Medoids dan bagaimana menerapkan metode K-Medoids untuk mengelompokkan dokumen skripsi Jurusan Teknik Informatika berdasarkan abstraknya.Terdapat 105 data text abstrak skripsi yang akan digunakan dalam penelitian ini, Untuk mengelompokkan data text abstrak skripsi dengan menggunakan metode clustering K-Medoids dilakukan kegiatan preproses yang terdiri dari beberapa kegiatan antara lain tokenizier dan stopword. Penjelasan lebih lanjut dari penelitian ini akan dijelaskan pada sub bagian metode yang akan diulas selanjutnya. Manfaat yang diharapkan dapat memisah-misahkan dokumen skripsi Jurusan Teknik Informatika ITATS berdasarkan topik dan kesamaannya.

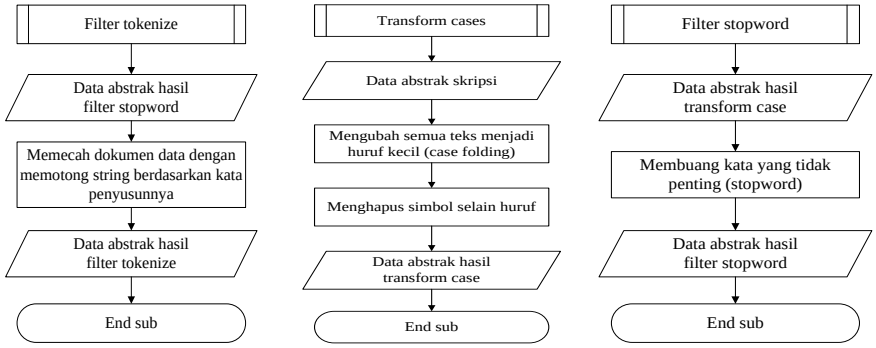
METODE

Perancangan Sistem



Gambar 1. *Flowchart Sistem Clustering* Abstrak Skripsi.

Flowchart sistem pada gambar 1 diatas menjelaskan alur proses sistem secara umum. Proses clustering abstrak skripsi dilakukan dengan cara, melakukan proses input data abstrak skripsi yang akan dicluster. Kedua, melakukan proses preprocessing data teks. Dimana preprocessing ini terdiri dari tiga langkah yaitu, transform cases, filter stopword dan filter tokenize. Langkah selanjutnya melakukan proses clustering menggunakan metode clustering K-Medoids. Terakhir, didapatkan hasil cluster untuk menentukan clustering data abstrak skripsi.



(a) (b) (c)

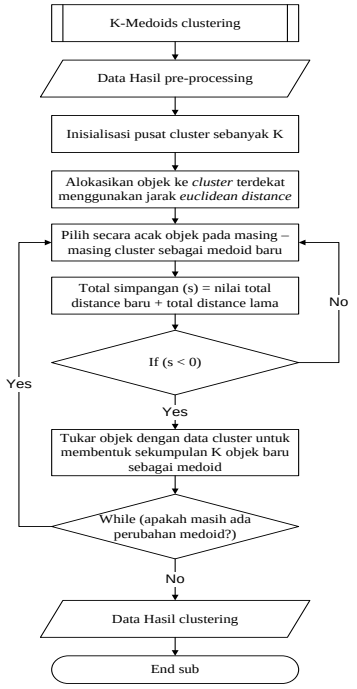
Gambar 2. a) *Flowchart Filter Tokenize*, b)) *Flowchart Transform Case*, c) Flowchart Filter Stopword

Pada gambar 2 merupakan diagram alir yang menjelaskan subproses dari *flowchart preprocessing* pada sistem *clustering* abstrak skripsi digambar 1. Proses pertama yang dilakukan adalah, melakukan proses input data abstrak skripsi hasi. Kedua, memecah dokumen data dengan memotong *string* berdasarkan kata penyusunnya. Terakhir, proses akan menampilkan hasil *output* dari subproses *filter tokenize*.

Dilanjutkan dengan *flowchart transform case* yang merupakan subproses dari *flowchart preprocessing* sistem *clustering* abstrak skripsi pada gambar1. Proses pertama yang dilakukan adalah, melakukan proses input data abstrak skripsi yang akan di*cluster*. Kedua, mengubah

semua teks menjadi huruf kecil (proses *case folding*). Kemudian, menghapus simbol selain huruf. Terakhir, proses akan menampilkan hasil *output* dari subproses *transform case*.

Kemudian masuk ke *flowchart filter stopword* yang merupakan subproses dari *flowchart preprocessing* sistem *clustering* abstrak skripsi pada gambar1. Proses pertama yang dilakukan adalah, melakukan proses input data abstrak skripsi hasil *transform case*. Kedua, membuang kata yang tidak penting (*stopword*). Terakhir, proses akan menampilkan hasil *output* dari subproses *filter stopword*.



Gambar 3. *Flowchart* Metode *K-Medoids Clustering*.

Pada gambar 3 adalah langkah-langkah metode k-medoids clustering yang akan dijalankan setelah subproses pada gambar 2 input data hasil preprocessing. Inisialisasi pusat cluster sebanyak K (jumlah cluster). Alokasikan setiap data (objek) ke cluster terdekat menggunakan persamaan ukuran jarak Euclidian Distance. Pilih secara acak objek pada masing-masing cluster sebagai kendidat medoid baru. Hitung jarak setiap objek yang berada pada masing-masing cluster dengan kandidat medoid baru. Hitung total simpangan (S) dengan menghitung nilai total distance baru total distance lama. Jika nilai S < 0. Maka tukar objek dengan data cluster untuk membentuk sekumpulan K objek baru sebagai medoid, jika tidak maka kembali pada langkah 4. While apakah masih ada perubahan pada medoid? Jika iya maka kembali pada langkah 4, jika tidak maka didapatkan data hasil clustering.

Pendekatan ekstraksi yang dilakukan dalam penelitian ini yaitu Term Frequency Inverse Document Frequensy (TF-IDF).

1. ***Document Frequency***
   Dokumen frekuensi (Df) adalah jumlah dokumen yang mengandung suatu *term* tertentu. Dokumen Frekuensi merupakan metode feature selection yang paling sederhana dengan waktu komputasi yang rendah.
2. ***Term Frequency***
   Term Frequency (Tf) merupakan salah satu metode untuk menghitung bobot tiap *term* dalam teks. Dalam metode ini tiap term diasumsikan memiliki nilai kepentingan yang sebanding dengan jumlah kemunculan *term* tersebut pada teks.

3. Inverse Document Frequency (IDF)

Merupakan Metode untuk menghitung kemunculan term dalam keseluruhan koleksi teks. Dalam hal ini, term yang jarang muncul pada koleksi keseluruhan term dinilai lebih berharga. Nilai kepentingan tiap term diasumsikan berbanding terbalik dengan jumlah teks yang mengandung term tersebut.

4. Term Frequency Inverse Document Frequency (TF-IDF)

Term Frequency Inverse Document Frequency (TF-IDF) merupakan pembobot yang dilakukan setelah ekstrasi artikel berita. Proses metode TF-IDF adalah menghitung bobot dengan cara integrasi antara term *frequency* (tf) dan *inverse document frequency* (idf). Langkah dalam TF-IDF adalah untuk menemukan jumlah kata yang kita ketahui (tf) setelah dikalikan dengan berapa banyak artikel berita dimana suatu kata itu muncul (idf).

HASIL DAN PEMBAHASAN

Pada klasifikasi dengan metode K-Medoids bisa dilihat dengan tingkat akurasi/kebenaran. Pengujian dilakukan dengan cara menguji data abstrak skripsi sebanyak 105 data.

Pengujian Metode Elbow Grafik

Pada pengujian aplikasi ini, pengujian dilakukan untuk dapat menentukan banyak K cluster yang paling optimal. Pengujian dilakukan dengan cara menguji 105 data abstrak skripsi yang kemudian diproses pada program clustering abstrak skripsi. Untuk cluster yang akan diujikan yaitu 2, 3, 4, dan 5

1. Cluster 2
   Cluster ke-1 mendapatkan 70 data
   Cluster ke-2 mendapatkan 35 data
   Rata – rata cluster 2 = $\frac{70+35}{2} = 52.5$
   Menghitung nilai SSE:
   $SSE = \sum_{K=1}^{K} \sum_{x_i e S_k} ||x_i - c_k||_2^2$
   $SSE(1) = |52.5 - 70|^2 + |52.5 - 35|^2$
   $= |17.5|^2 + |17.5|^2$
   $= 306.25 + 306.25 = 612.5$
2. Cluster 3
   Cluster ke-1 mendapatkan 31 data
   Cluster ke-2 mendapatkan 42 data
   Cluster ke-3 mendapatkan 32 data
   Rata – rata cluster 3 = $\frac{31+42+32}{3} = 35$
   Menghitung nilai SSE:

   $SSE = \sum_{K=1}^{K} \sum_{x_i e S_k} ||x_i - c_k||_2^2$
   $SSE(1) = |35 - 31|^2 + |35 - 42|^2 + |35 - 32|^2$
   $= |4|^2 + |7|^2 + |3|^2 = 16 + 49 + 9$
   $= 74$
3. Cluster 4
   Cluster ke-1 mendapatkan 34 data
   Cluster ke 2 mendapatkan 4 data
   Cluster ke 3 mendapatkan 27 data
   Cluster ke 4 mendapatkan 40 data

Rata – rata cluster 4 = $\frac{34+4+27+40}{4} = 26.25$

Menghitung nilai SSE:

$$SSE = \sum_{K=1}^{K} \sum_{x_i e S_k} ||x_i - c_k||_2^2$$

$$SSE(1) = |26.25 - 34|^2 + |26.25 - 4|^2 + |26.-27|^2 + |26.25 - 40|^2$$
$$= |7.25|^2 + |22.25|^2 + |1.25|^2 + |13.75|^2$$
$$= 52.56 + 495.06 + 1.56 + 189.06$$
$$= 738.24$$

4. Cluster 5

Cluster ke-1 mendapatkan 24 data
Cluster ke 2 mendapatkan 4 data
Cluster ke 3 mendapatkan 23 data
Cluster ke 4 mendapatkan 12 data
Cluster ke 5 mendapatkan 42 data

Rata–rata cluster 5 = $\frac{24+4+23+12+42}{5} = 21$

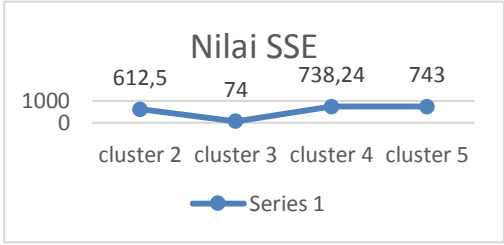
Menghitung nilai SSE:

$$SSE = \sum_{K=1}^{K} \sum_{x_i e S_k} ||x_i - c_k||_2^2$$

$$SSE(1) = |21 - 24|^2 + |21 - 4|^2 + |21 - 23|^2 + |21 - 12|^2 + |21 - 42|^2$$
$$= |3|^2 + |17|^2 + |2|^2 + |9|^2 + |21|^2$$
$$= 9 + 289 + 4 + 81 + 441$$
$$= 743$$

Berdasarkan perhitungan diatas maka didapatkan:
Nilai SSE Cluster 2 = 612.5
Nilai SSE cluster 3 = 74
Nilai SSE cluster 4 = 738.24

Nilai SSE cluster 5 = 743.



Gambar 4. Grafik *Elbow* Nilai *SSE*.

Pada gambar 4 merupakan visualisasi metode elbow dari hasil nilai SSE yang didapatkan. Dengan cluster 2 bernilai 612.5, cluster 3 bernilai 74, cluster 4 bernilai 738.24, cluster 5 bernilai 743. Berdasarkan grafik di atas maka dapat disimpulkan bahwa cluster yang membentuk pahatan berbentuk siku adalah cluster 3.

Pengujian Metode Silhouette

Pada pengujian aplikasi ini, pengujian dilakukan untuk mengetahui kualitas dari cluster yang terbentuk. Pengujian dilakukan dengan cara menguji 105 data abstrak skripsi yang kemudian diproses pada program clustering abstrak skripsi. Hasil dari metode Silhouette telah dihitung melalui program. Maka didapatkan nilai Silhouette dari tiap cluster, yaitu sebagai berikut:

Pada pengujian aplikasi ini, pengujian dilakukan untuk mengetahui kualitas dari *cluster* yang terbentuk. Pengujian dilakukan dengan cara menguji 105 data abstrak skripsi yang kemudian diproses pada program *clustering* abstrak skripsi. Hasil dari metode Silhouette telah dihitung melalui program. Maka didapatkan nilai Silhouette dari tiap cluster, yaitu sebagai berikut:

1. Hitung rata-rata jarak dari suatu dokumen dengan semua dokumen lain yang berada dalam satu cluster (a_i) :

   Tabel 1. Jarak Satu Dokumen Dengan Semua Dokumen Dalam Cluster Yang Sama

Cluster ke-	Jumlah a(i)	Rata-rata a(i)
2	1.13	0.010762
3	1.08	0.010286
4	1.05	0.01
5	1.06	0.010095
Rata-rata		0.01028575

   Dari tabel 1 didapatkan nilai rata-rata jarak dari satu dokumen dengan semua dokumen lain yang berada dalam satu cluster adalah 0,01028575.

2. Hitung rata-rata jarak dari dokmen dengan smeua dokumen di cluster lain:

   Tabel 2. Jarak Satu Dokumen Dengan Semua Dokumen

Cluster ke-	Jumlah d(i)	Rata-rata d(i)
2	1.13	0.010762
3	1.08	0.010286
4	1.07	0.01019
5	1.06	0.010095
	Rata-rata	0.01033325

   Dari tabel 2 didapatkan nilai rata-rata jarak dari satu dokumen dengan semua dokumen lain yang berada dalam satu cluster adalah 0,01033325.

3. Hitung nilai *Silhouette Coefficient*:

   Tabel 3. *Silhouette Coefficient*

Cluster ke-	Jumlah s(i)	Rata-rata s(i)
2	-1.4	-0.01333
3	-2.44	-0.02324
4	-3.35	-0.0319
5	-4.98	-0.04743
	Rata-rata	-0.028975

   Dari tabel 3 didapatkan hasil *Silhouette Coefficient* adalah $-0{,}028975$. Berdasarkan tabel 3 yang merupakan tabel kriteria pengukuran *silhouette coefficient* maka nilai S(C) sebesar $-0{,}028975$.

KESIMPULAN

Telah berhasil dibuat aplikasi pengelompokan data abstrak skripsi dengan metode *clustering* k-medoids. Berdasarkan hasil simulasi pengelompokan dengan metode *clustering*

dengan k-medoids pada data abstrak skripsi, memberikan luaran kelompok atau cluster yang paling optimal adalah 4. Untuk 105 data abstrak skripsi berhasil membentuk empat kelompok tema. Pengujian selanjutnya dilakukan untuk mengetahui tingkat kepercayaan dari cluster yang terbentuk menggunakan metode *Silhouette Coefficient*. Pengujian dilakukan pada 105 data abstrak skripsi dengan cluster yang diujikan adalah 2, 3, 4, dan 5. Diperoleh nilai *Silhouette Coefficient* sebesar $-0{,}028975$. Untuk penelitian yang akan datang dapat dikembangkan aplikasi pencarian yang dilengkapi dengan fitur klasifikasi data abstrak skripsi.

DAFTAR PUSTAKA

[1] Clara Anggita Ayu Setyaningrum. (2021). "Implementasi Algoritma K-Medoids Dalam Menentukan Tingkat Penyebaran Pandemi Covid-19 Di Indonesia", Prodi S1 Informatika. Fakultas Sains dan Teknologi. Universitas Sanata Dharma. Yogyakarta.

[2] Dewa Ayu Indah Cahya Dewi, Dewa Ayu Kadek Pramita. (2019). "Analisis Perbandingan Metode Elbow dan Sihouette pada Algoritma Clustering K-Medoids dalam Pengelompokkan Produksi Kerajinan Bali", JURNAL MATRIK, Vol. 9, No. 3.

[3] Pulungan, Nurliana. Suhada, Dedi Suhendra. (2019). "Penerapan Algoritma K-Medoids Untuk Mengelompokkan Penduduk 15 Tahun Keatas Menurut Lapangan Pekerjaan Utama", KOMIK (Konferensi Nasional Teknologi Informasi dan komputer). Volume 3, Nomor 1. DOI: 10.30865/komik.v3i1.1609. ISSN: 2597-4645 (Media Online). ISSN: 2597-4610 (Media Cetak).

[4] Rofiqi, ACH. Yasir. (2017). "Clustering Berita Olahraga Berbahasa Indonesia Menggunakan Metode K-Medoid Bersyarat", Jurnal SimanteC. Volume 6. Nomor 1. P-ISSN: 2088-2130. E-ISSN: 2502-4884.

[5] Silitonga, Sulastry. Eka Irawan, Saifullah, Muhammad Ridwan Lubis, Iin Parlina (2019). "Pengelompokkan Nilai Akademik Untuk Menentukan Kenaikkan Kelas Menggunakan Algoritma K-Medoids". Prosiding Seminar Nasional Riset Information Science (SENARIS). ISSN: 2686-0260.

[6] Sundari, Siti. Irfan Sudahri Damanik, Agus Perdana Windarto, Heru Satria Tambunan, Jalaluddin, Anjar Wanto. (2019). "Analisis K-Medoids Clustering Dalam Pengelompokkan data Imunisasi Campak Balita Di Indonesia", Prosiding Seminar Nasional Riset Information Science (SENARIS). ISSN: 2686-0260.