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 I - Surabaya, 26 Juni 2021
Ruang Seminar Gedung A, Kampus Institut Teknologi Adhi Tama Surabaya

Informasi Artikel:
DOI : 10.31284/p.snestik.2021.1824

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 Algoritma K-Means Untuk Pengelompokan Topik Dokumen Studi Kasus:Dokumen Abstrak Skripsi Jurusan Teknik Informatika ITATS

Kurniana, I. R.[1], Muhima,R.R.[2], Wardana,S.[3], Hakimah,M.[4]
Jurusan Teknik Informatika ITATS
e-mail: rani.muhima@itats.ac.id

ABSTRACT

The inventory process in the ITATS Informatics Engineering library is still not optimal. This is indicated by the library material, especially student thesis reports in the ITATS e-library, which have not been grouped. Thesis report grouping in the library of the Department of Informatics will assist students in finding references or as a consideration in choosing a thesis topic. It is necessary to group the thesis reports in the departement based on the thesis topic. The method used in this thesis abstract document grouping is K-Means. This method is quite simple but can minimize the similarities between groups and maximize similarity in one group. The data used in this study were 182 thesis abstract document data in pdf format. Prior to the grouping process using the K-Means algorithm, data preprocessing was carried out. This process consists of case folding, tokenization, filtering and stemming. After the preprocessing stage the data was weighted using Tf-Idf. Evaluation of grouping results using the elbow method. The results showed that the optimal number of groups was at k = 3 with an SSE value of 41977.88

Keywords: *Document clustering, text mining, K-means*

ABSTRAK

Proses inventarisasi di perpustakaan jurusan Teknik Informatika ITATS masih belum optimal. Hal ini ditunjukkan dengan bahan pustaka khususnya laporan skripsi mahasiswa di *e-library* ITATS belum dikelompokkan. Pengelompokkan laporan skripsi di perpustakaan jurusan Teknik Informatika akan membantu mahasiswa dalam mencari referensi atau sebagai pertimbangan dalam memilih topik skripsi. Perlu dilakukan pengelompokkan laporan skripsi di Jurusan Teknik Informatika berdasarkan topik skripsi. Metode yang digunakan pengelompokan dokumen abstrak skripsi ini adalah K-Means. Metode ini cukup sederhana tetapi dapat meminimilasir kemiripan antar kelompok dan memaksimalkan kimiripan dalam satu kelompok. Data yang

digunakan pada penelitian adalah 182 data dokumen abstrak skripsi dengan format pdf. Sebelum proses pengelompokan dengan algoritma K-Means, dilakukan *preprocessing* data. Proses ini terdiri dari *case folding, tokenization, filtering* dan *stemming.* Setelah tahap *preprocessing* data kemudian dilakukan pembobotan dengan *Tf-Idf*. Evaluasi hasil pengelompokan menggunakan metode elbow. Hasil penelitian menunjukan jumlah kelompok optimal pada k=3 dengan nilai SSE sebesar 41977,88.

Kata kunci: pengelompokan dokumen, *text mining*, K-Means

PENDAHULUAN

Perpustakaan menurut Undang-Undang No.43 tahun 2007 pasal 3, memiliki peran sebagai wahana pendidikan, penelitian, pelestarian, informasi, serta sebagai wahana rekreasi untuk meningkatkan kecerdasan dan keberdayaan bangsa [1]. Pengolahan bahan pustaka adalah kegiatan utama yang dilakukan perpustakaan. Pengolahan bahan pustaka yang baik agar proses pencarian informasi dapat berjalan lancar. Pengolahan bahan pustaka meliputi kegiatan inventaris, klasifikasi, *input* data, pelabelan dan *shelving* [2]. Proses inventarisasi sendiri meliputi kegiatan pemeriksaan koleksi, pengelompokkan koleksi, pengecapan dan pencatatan.

Proses inventarisasi di perpustakaan jurusan Teknik Informatika ITATS masih belum optimal. Hal ini ditunjukkan dengan bahan pustaka khususnya laporan skripsi mahasiswa di *e-library* ITATS belum dikelompokkan berdasarkan topik. Pengelompokkan laporan skripsi di perpustakaan jurusan Teknik Informatika akan membantu mahasiswa dalam mencari referensi atau sebagai pertimbangan dalam memilih topik skripsi. Hal ini diharapkan agar yang sudah dikerjakan oleh mahasiswa sebelumnya tidak dikerjakan kembali untuk hal yang sama atau sebaliknya apa yang sudah dikerjakan dapat dikembangkan menjadi bahan skripsi yang baru. Berdasarkan hal tersebut pengelompokkan topik skripsi perlu dilakukan.

Abstrak suatu karya ilmiah merupakan tulisan singkat yang merupakan informasi dari sebuah karya ilmiah. Abstrak dapat mewakili isi tulisan dalam suatu karya ilmiah. Abstrak skripsi dapat memberikan informasi tentang isi skripsi. Sehingga pengelompokkan topik skripsi dapat didasarkan pada dokumen abstrak skripsi. Abstrak skripsi umumnya berbentuk teks. Pengelompokkan teks umumnya melibatkan data teks yang tidak terstruktur, maka solusi untuk menemukan pola yang diingikan untuk dijadikan kunci pengelompokkan dapat digunakan teknik *text mining* [3].

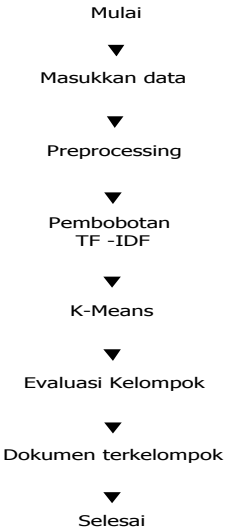
Kajian terkait *text mining* sudah banyak dilakukan. Penerapan *text mining* diimplementasikan pada laman blog di internet untuk menilai kinerja suatu organisasi dalam memberikan layanan publik. Penilain kinerja berdasarkan pendapat yang diberikan masyarakat sebagai umpan balik yang tertuang di internet [4]. Text mining juga diterapkan untuk mengetahui persepsi kecenderungan masyarakat terhadap implementasi UU ITE di Indonesia [5]. *Text mining* dengan metode clustering digunakan untuk mengelompokkan berita sebagai landasan untuk mengakategorikan berita informasi di provinsi Bali [3]. Pada penelitian tersebut algoritma yang digunakan adalah algoritma K-Means. Pengelompokan topik dokumen kedutaan besar Australia Jakarta juga menggunakan algortitma K-Means [6].

Algoritma K-Means salah satu algoritma pengelompokan yang cukup terkenal dengan algoritma cukup sederhana. Tujuan pengelompokan K-Means adalah meminimilasir kemiripan antar kelompok dan memaksimalkan kimiripan dalam satu kelompok [7]. Penelitian [3] berhasil mengelompokan data teks dengan purity sebesar 0,76. Hal ini menunjukkan bahwa dari data yang diolah, 76% data dikelompokkan oleh sistem dengan benar. Berdasarkan hal tersebut, pada pengelompokan topik skripsi jurusan Teknik Informatika berdasarkan abstrak skripsi berbasis *text mining* digunakan algoritma K-Means. Untuk mempermudah proses pengelompokan menggunakan metode K-Means, dilakukan *preprocessing* dokumen dahulu. Proses ini terdiri dari *case folding, tokenization, filtering* dan *stemming* [3].

METODE

Algoritma pengelompokan pada penelitian ini adalah algoritma K-Means. Data yang digunakan adalah data dokumen abstrak skripsi mahasiswa Jurusan teknik Informatika ITATS format

pdf dengan jumlah 182 data. Data dalam format pdf tersebut kemudian diubah dalam bentuk TXT. Sebelum data dikelompokan, dilakukan tahap *preprocessing* data kemudian dilakukan pembobotan term terlebih dahulu. Setelah proses pembobotan kemudian data dikelompokkan menggunakan metode K-means dengan evaluasi kelompok menggunakan *Sum Square Error* (SSE). Gambaran keseluruhan sistem dijelaskan dengan flowchart pada Gambar 1.



Gambar 1. *Flowchart* Sistem Pengelompokan Dokumen Abstrak Skripsi Menggunakan K-means

Preprocessing Data

Tahap *preprocessing* data ini terdiri dari beberapa tahap yaitu *case folding*, tokenisasi,*filtering* dan *stemming*. *Case folding* adalah tahap merubah semua huruf kapital menjadi huruf kecil dan menghilangkan karakter selain huruf. Kemudian tokenisasi merupakan tahap pemisahan kata. Tahap *filtering*, tahap penghilangan kata setelah proses tokenisasi berdasarkan *stopword removal*. Selanjutnya adalah tahap *stemming*. Tahap ini adalah tahap pengurangan kata menjadi bentuk dasar dari setiap kata. Pada penelitian ini, sastrawi digunakan untuk proses *stemming*.

Term Weighting (Pembobotan)

Term Weighting atau proses pembobotan tiap kata dilakukan setelah tahap *preprocessing* data. Tahap ini dilakukan untuk mengetahui bobot tiap kata pada setiap dokumen. Dari hal tersebut dapat diketahui ketersediaan dan kemiripan suatu kata pada suatu dokumen [8]. Metode yang digunakan dalam pembobotan ini adalah metode *Tf-Idf*.

Metode *Tf-Idf* merupakan penggabungan nilai *Tf* dan *Idf*. *Tf* atau *Term frequency* merupakan penentuan bobot dokumen berdasarkan kemunculan kata atau *term* pada dokumen tersebut. Semakin sering sebuah kata muncul maka semakin tinggi bobot dokumen tersebut atau sebaliknya [3]. *Idf* atau *Inverse document frequency* menunjukkan hubungan ketersediaan kata dalam seluruh dokumen. Nilai *Idf* yang semakin besar menunjukkan semakin sedikit jumlah dokumen yang mengandung kata tersebut [8]. Perhitungan *Tf-Idf* berdasarkan persamaan berikut:

$$Tf - Idf = Tf.log\frac{N}{df} \quad (1)$$

dimana:

$Tf - Idf$: bobot suatu dokumen

Tf : frekuensi term pada suatu dokumen
N : jumlah seluruh dokumen
df : jumlah dokumen yang mengandung term tersebut

K-Means

Algoritma K-Means yang digunakan untuk pengelompokan topik dokumen abstrak skripsi adalah sebagai berikut:

1. Tahap ini diawali dengan menentukan nilai k
2. Menentukan centroid awal secara acak.
3. Penghitungan jarak data ke centroid digunakan metode *Cosine Similarity*. Pengelompokan objek berdasarkan jarak terdekat objek dengan centroid.
4. Kembali Langkah 2 hingga centroid atau pusat cluster tidak mengalami perubahan.

Perhitungan jarak pada algoritma K-means ini dengan *Cosine Similarity. Cosine Similarity* merupakan metode pengukuran kemiripan kalimat dengan berdasarkan sudut dua vektor [3]. Kalimat disini dianggap vektor, dengan nilai cosinus sudut antara dua vektor tersebut sebagai parameter jarak. Parameter jarak sebagai parameter kemiripan dua vektor. Semakin dekat jarak antara dua vektor maka semakin mirip dua vektor tersebut. Persamaan pada metode *Cosine Similarity* adalah sebagai berikut:

$$Cosine(x,y) = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}} \quad (2)$$

dengan X dan Y merupakan vektor, X_i adalah bobot suatu kata i pada dokumen X, Y_i adalah bobot suatu kata i pada dokumen Y dan i adalah jumlah kata dalam suatu dokumen.

Evaluasi Cluster

Evaluasi *cluster* ini dilakukan untuk mengetahui nilai K yang optimal dalam pengelompokan menggunakan metode K-means. Metode yang digunakan pada evaluasi ini adalah metode Elbow. Metode ini didasarkan nilai SSE (*Sum Square Error*) setiap variasi K. Hasil pengelompokan yang optimal nantinya dijadikan dasar pelabelan tiap kelompok. Nilai k optimal diperoleh pada titik yang membentuk siku atau memiliki selisih nilai SSE terbesar [9].

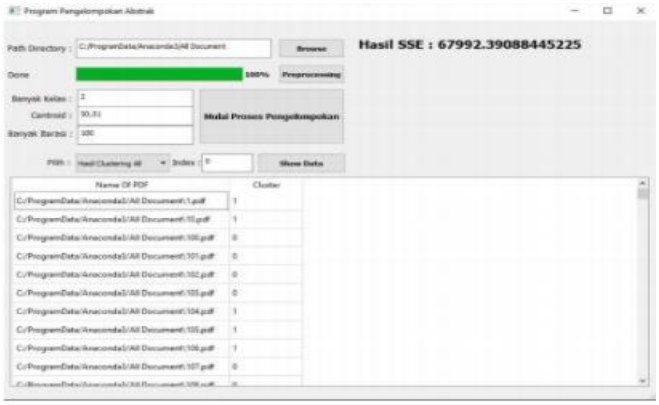
Algoritma metode elbow [10] adalah (1) inisialisasi nilai k, (2) menaikan nilai k, (3) menghitung nilai SSE dari setiap nikai k, (4) analisa nilai SSE dari nilai k. Jika terdapat penurunan drastis pada nilai SSE, maka nilai k tersebut merupakan nilai k yang benar. Perhitungan nilai SSE berdasarkan persamaan

$$SSE = \sum_{K=1}^{K} \sum_{x_i \in S_K} d^2 \quad (3)$$

dengan d merupakan jarak tiap term terhadap centroid dalam satu kelompok.

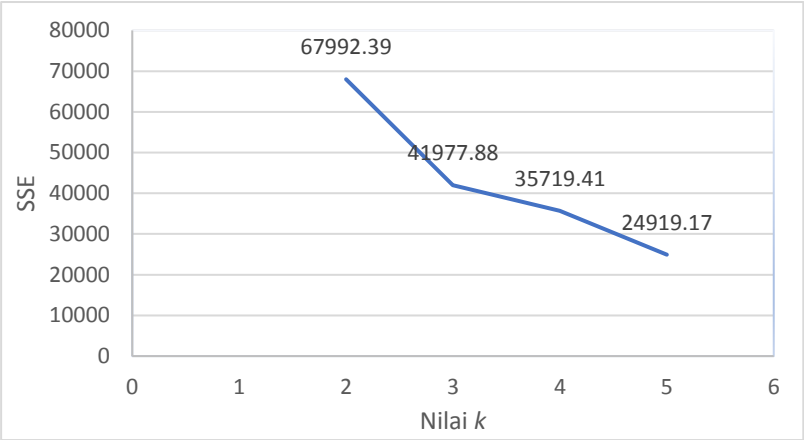
HASIL DAN PEMBAHASAN

Hasil antarmuka pengelompokan abstrak skripsi dapat dilihat pada gambar 2. Pada gambar dicontohkan hasil pengelompokan dengan nilai k=2. Pada penelitian ini nilai k yang digunakan dengan variasi k= 2, 3, 4, dan 5 serta data yang digunakan adalah 182 data dokumen abstrak skripsi. Pada k=2, centroid kelompok 1 adalah dokumen 51 dan kelompok 2 adalah dokumen 50. Untuk k=3, centroid kelompok 1 adalah dokumen 90, kelompok 2 adalah dokumen 1 dan kelompok 3 adalah dokumen 11. Sedangkan k=4, centroid kelompok 1 adalah dokumen 95, kelompok 2 adalah dokumen 37, kelompok 3 adalah 97 dan kelompok 4 adalah dokumen 75. Pada k=5, centroid kelompok 1 adalah dokumen 148, centroid kelompok 2 adalah dokumen 164, kelompok 3 adalah dokumen 150, kelompok 4 adalah dokumen 149 dan centroid kelompok 5 adalah dokumen 161.



Gambar 2. Antarmuka Sistem Pengelompokan Dokumen Abstrak Skripsi

Evaluasi *cluster* berdasarkan grafik elbow ditunjukkan pada gambar 3. Berdasarkan grafik tersebut, siku atau nilai penurunan terbesar pada k= 3. Hal ini menunjukkan bahwa pengelompokan dokumen abstrak skripsi optimal pada nilai k=3 dengan nilai SSE sebesar 41977,88.



Gambar 3. Grafik Elbow Pengelompokan dokumen abstrak skripsi dengan variasi nilai k=2, 3, 4 dan 5

KESIMPULAN

Berdasarkan hasil penelitian pengelompokan topik dokumen abstrak skripsi Jurusan Teknik Informatika ITATS, jumlah kelompok optimal pada k=3 dengan nilai SSE sebesar 41977,88. Hasil pengelompokan tersebut dapat dijadikan acuan untuk pelabelan kelompok topik. Pelabelan kelompok topik ini berdasarkan referensi dari pakar adalah topik sistem cerdas, rekayasa perangkat lunak dan jaringan komputer. Hasil pelabelan dapat dijadikan acuan dasar klasifikasi topik skripsi mahasiswa jurusan Teknik Informatika ITATS.

DAFTAR PUSTAKA

[1] "Undang Undang Republik Indonesia Nomor 43 Tahun 2007 Tentang Perpustakaan," 2007. [Online]. Available: https://www.perpusnas.go.id/law-detail.php?lang=id&id=170920114322Ir9g6HhRuc. [Accessed: 29-Apr-2021].

[2] Dinas Perpustakaan dan Kearsipan Kota Pekanbaru, "Proses Pengolahan Bahan Pustaka," 2018. [Online]. Available: https://dispusip.pekanbaru.go.id/proses-pengolahan-bahan-pustaka/. [Accessed: 30-Apr-2021].

[3] N. G. Yudiarta, M. Sudarma, and W. G. Ariastina, "Penerapan Metode Clustering Text Mining Untuk Pengelompokan Berita Pada Unstructured Textual Data," *Maj. Ilm. Teknol. Elektro*, vol. 17, no. 3, p. 339, 2018.

[4] F. Rahutomo, Z. Hanif Rachmat Adi, I. Fahrur Rozi, and P. Yoga Saputra, "Implementasi Text Mining Pada Website/Blog Di Internet Untuk Menilai Kinerja Suatu Organisasi," *INOVTEK Polbeng - Seri Inform.*, vol. 3, no. 2, p. 101, 2018.

[5] L. Hakim, T. F. Kusumasari, and M. Lubis, "Text Mining of UU-ITE Implementation in Indonesia," *J. Phys. Conf. Ser.*, vol. 1007, no. 1, 2018.

[6] W. Hardi, "Pengelompokan Topik Dokumen Berbasis Text Mining Dengan Algoritme K-Means : Studi Kasus Pada Dokumen Kedutaan Besar Australia Jakarta," vol. 21, no. 1, pp. 67–76, 2019.

[7] R. R. Muhima, M. Kurniawan, and O. T. Pambudi, "A LOF K - Means Clustering on Hotspot Data," *Int. J. Artif. Intell. Robot.*, vol. 2, no. 1, pp. 29–33, 2020.

[8] P. Yugianus, H. S. Dachlan, and N. Hasanah, "Pengembangan Sistem Penelusuran katalog Perpustakaan Dengan metode Rocchio Relevance Feedback," vol. 7, no. 1, pp. 47–52, 2013.

[9] R. A. Asroni, "Penerapan Metode K-Means Untuk Clustering Mahasiswa Berdasarkan Nilai Akademik Dengan Weka Interface Studi Kasus Pada Jurusan Teknik Informatika UMM Magelang," *Ilm. Semesta Tek.*, vol. 18, no. 1, pp. 76–82, 2015.

[10] R. Nainggolan, R. Perangin-Angin, E. Simarmata, and A. F. Tarigan, "Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method," *J. Phys. Conf. Ser.*, vol. 1361, no. 1, 2019.