ISSN. 2720-9180

RANCANG BANGUN SISTEM KLASTERISASI DOKUMEN MENGGUNAKAN METODE K-MEANS UNTUK IDENTIFIKASI TOPIK DOKUMEN TUGAS AKHIR PROGRAM STUDI TEKNIK INFORMATIKA UNIVERSITAS ISLAM SULTAN AGUNG

Rahmah widya astuti[1], Badieah, S.T., M.Kom [2], Bagus Satrio.W.P [3]
1. Teknik Informatika, Universitas Islam Sultan Agung
2. Dosen Teknik Informatika, Universitas Islam Sultan Agung
3. Dosen Teknik Informatika, Universitas Islam Sultan Agung

Correspondence Author : rahmahwidyaastuti@std.unissula.ac.id

Abstract **-** Dalam ruang baca Fakultas Teknologi Industri Unissula belum terdapat sistem untuk mengetahui topik dari judul tugas akhir tersebut, Dan mahasiswa juga harus membaca keseluruhan atau isi dari laporan tugas akhir tersebut. Mahasiswa pada program studi ini juga belum memiliki akses digital terhadap tugas akhir yang pernah dilakukan serta mahasiswa harus mencari tugas akhir secara fisik pada Ruang baca fakultas. Hal ini tentu membuat kesulitan mahasiswa dalam pencarian sumber pustaka serta sulit untuk mengetahui topik dari topik tugas akhir tersebut yang tepat untuk dilakukannya karena harus membaca keseluruhan isi dokumen tugas akhir. mengimplementasikan metode K-Means dalam pengelompokan dokumen.

Data dokumen yang digunakan adalah dokumen pada ruang baca Fti Unissula. Dari data tersebut akan ditentukan jumlah *cluster* yang akan dibentuk. Kemudian menentukan titik pusat *centroid* secara random dan mengitung jarak terdekat setiap data kepusat kelompok dengan menggunakan rumus *Eucludian Distance.* Hasil dari perhitungan jarak tersebut akan dikelompokan berdasarkan jarak *eucludian*nya jika masih ada data yang berubah maka prosesnya akan masuk ke iterasi berikutnya, namun jika data *cluster*nya tetap maka proses akan dihentikan. Berdasarkan implementasi sistem. Dengan kemampuan *clustering* teks tersebut, algoritma K-Means *clustering* dapat menjadi solusi untuk identifikasi topik dokumen tugas akhir Program Studi Teknik Informatika Unissula.

Kata Kunci : Metode K-Means, Dokumen TA, *Clustering*

1. PENDAHULUAN

Tugas akhir adalah karya tulis yang disusun oleh mahasiswa yang telah menyelesaikan kurang lebih 130 sks dengan dibimbing oleh dosen pembimbing guna mendapatkan gelar pendidikan

Dalam ruang baca Fakultas Teknologi Industri Unissula belum terdapat sistem untuk mengetahui topik dari judul tugas akhir tersebut, Dan mahasiswa juga harus membaca keseluruhan atau isi dari laporan tugas akhir tersebut. Mahasiswa pada program studi ini juga belum memiliki akses digital terhadap tugas akhir yang pernah dilakukan serta mahasiswa harus mencari tugas akhir secara fisik pada Ruang baca fakultas. Hal ini tentu membuat kesulitan mahasiswa dalam pencarian sumber pustaka serta sulit untuk mengetahui topik dari topik tugas akhir tersebut yang tepat untuk dilakukannya karena harus membaca keseluruhan isi dokumen tugas akhir. Dokumen tugas akhir dapat dikelompokan menjadi beberapa *cluster* yang mempresentasikan topik-topik tugas akhir maka, diperlukan proses *clustering.* identifikasi topik tugas akhir, hasil pengelompokan juga dapat memperlihatkan topik yang banyak diambil mahasiswa dan yang sering diambil mahasiswa dan yang jarang diambil mahasiswa pada waktu tertentu. Pengelompokan data penelitian atau dokumen tugas akhir yang umumnya berbentuk text dan dapat dilakukan dengan text mining dengan metode K-Means *clustering*. dan dengan menggunakan algoritma ini agar dapat lebih cepat memecahkan permasalahan dari judul tugas akhir ini sehinga dapat menyelesaikannya dengan menggunakan algoritma K-Means *clustering*.

Algoritma K-Means merupakan sebuah metode, K dimaksudkan sebagai konstanta jumlah cluster yang diinginkan, Means dalam hal ini berarti nilai suatu rata-rata dari suatu grup data yang dalam hal ini didefinisikan sebagai *cluster*, sehingga K-Means *Clustering* adalah suatu metode penganalisaan data atau metode data mining yang melakukan proses pemodelan tanpa *supervisi* (*unsupervised*) dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi. Metode K-Means berusaha mengelompokkan data yang ada ke dalam beberapa kelompok, dimana data dalam satu kelompok mempunyai karakteristik yang sama satu sama lainnya dan mempunyai karakteristik yang berbeda dengan data yang ada didalam kelompok yang lain. Dengan kemampuan *clustering* teks tersebut, algoritma K-Means *clustering* dapat menjadi solusi untuk identifikasi topik dokumen tugas akhir Program Studi Teknik Informatika Unissula.

2. DASAR TEORI

2.1 Text Mining

Text mining adalah satu langkah dari analisis teks yang dilakukan secara otomatis oleh komputer untuk menggali informasi yang berkualitas dari suatu rangkaian teks yang terangkum dalam sebuah dokumen. Prosedur utama dalam metode ini terkait dengan menemukan kata-kata yang dapat mewakili isi dari dokumen untuk selanjutnya dilakukan analisis keterhubungan antar dokumen dengan menggunakan metode statistik tertentu seperti analisis kelompok, klasifikasi dan asosiasi. Tahapan dalam text mining secara umum adalah *tokenizing*, *filtering*, *stemming*, *tagging*, dan *analyzing*. *Tokenizing* merupakan tahapan untuk memisah-misahkan setiap kata (token) pada dokumen input. *Filtering* merupakan proses seleksi terhadap kata-kata yang dihasilkan dari proses *tokenizing*, dapat dilakukan dengan algoritma *stop list* maupun *word list*. Algoritma *stop list* akan membuang kata-kata yang tidak penting seperti kata ganti, kata keterangan, kata sambung, kata depan dan kata sandang. Sebaliknya, algoritma *word list* akan menyimpan kata-kata yang penting. Proses *stemming* kemudian dilakukan untuk mencari kata dasar dari setiap kata yang telah lolos proses *filtering*. Terdapat 4 varian algoritma untuk proses stemming ini, yaitu: (1) *Table lookup*, seluruh kata dasar disimpan dalam memori untuk selanjutnya dijadikan acuan dalam pemeriksaan dokumen input. (Prilianti & Kunci, 2014)

2.2 Dokumen Tugas Akhir

Clustering dokumen Tugas Akhir dikerjakan untuk mengetahui topik pada tema tugas akhir tersebut yang diinputkan termasuk dalam dokumen tugas Akhir program studi teknik informatika unissula.

2.3 Pembobotan TF-IDF (Term Frequency-inverse Document Frequency)

Pembobotan TF-IDF merupakan suatu jenis pembobotan yang sering digunakan dalam IR dan *text mining*. Pembobotan merupakan pengukuran statistik untuk mengukur seberapa penting sebuah kata dalam kumpulan dokumen. Tingkat kepentingan sebuah kata meningkat ketika kata muncul beberapa kali.
(Amburika & Chrisnanto, 2016)

2.4 Metode K-Means *Clustering*

Data *Clustering* merupakan salah satu metode Data Mining yang bersifat tanpa arahan (*unsupervised*). Ada dua jenis data *clustering* yang sering dipergunakan dalam proses pengelompokan data yaitu *hierarchical* (hirarki) data *clustering* dan *non-hierarchical* (non hirarki) data *clustering*.
Secara umum metode K-Means Cluster menggunakan algoritma sebagai berikut:
1. Tentukan k sebagai jumlah *cluster* yang di bentuk. Untuk menentukan banyaknya *cluster* k dilakukan dengan beberapa pertimbangan seperti pertimbangan teoritis dan konseptual yang mungkin diusulkan untuk menentukan berapa banyak *cluster*.
2. Bangkitkan k *Centroid* (titik pusat cluster) awal secara random. Penentuan *centroid* awal dilakukan secara random/acak dari objek-objek yang tersedia sebanyak k *cluster*, kemudian untuk menghitung *centroid cluster* ke-i berikutnya, digunakan rumus sebagai berikut :

$$v = \frac{\sum_{i=1}^{n} x_1}{n} ; i = 1,2,3, \dots n \quad (1)$$

Dimana; v= *centroid* pada *cluster*
Xi= objek ke-i
n= banyaknya objek/jumlah objek yang menjadi anggota *cluster*

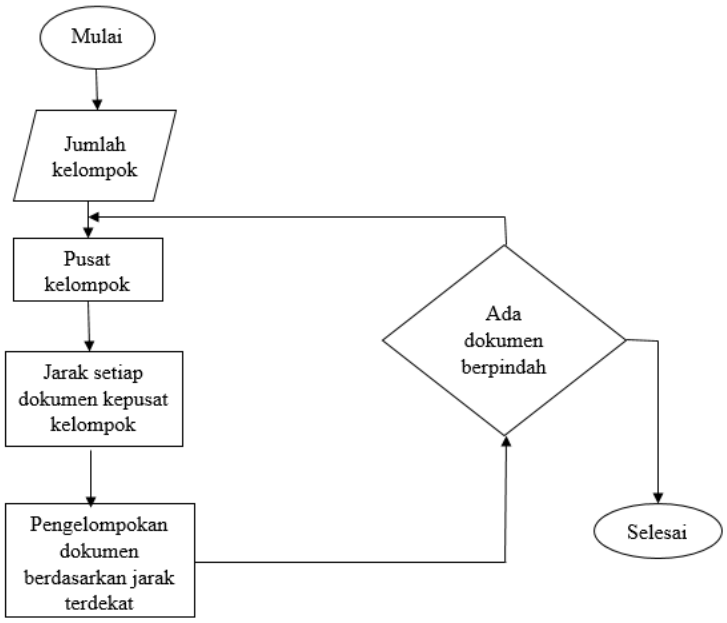
3. Hitung jarak setiap objek ke masing-masing *centroid* dari masing-masing *cluster*. Untuk menghitung jarak antara objek dengan *centroid* penulis menggunakan *Euclidian Distance*
Rumus;

$$d(x,y) = \|x - y\| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad ; i = 1,2,3, \dots, n \quad (2)$$

Dimana; xi= objek x ke-i
yi= objek y ke-i
n= banyaknya objek

ISSN. 2720-9180

4. Alokasikan masing-masing objek ke dalam *centroid* yang paling terdekat. Untuk melakukan pengalokasian objek kedalam masing-masing *cluster* pada saat iterasi secara umum dapat dilakukan dengan dua cara yaitu dengan *hard* k-means, dimana secara tegas setiap objek dinyatakan sebagai anggota *cluster* dengan mengukur jarak kedekatan sifatnya terhadap titik pusat *cluster* tersebut.
5. Lakukan *iterasi*, kemudian tentukan posisi *centroid* baru dengan menggunakan persamaan Ulangi langkah 3 jika posisi *centroid* baru tidak sama. Pengecekan konvergensi dilakukan dengan membandingkan matriks group *assignment* pada iterasi sebelumnya dengan matrik group *assignment* pada iterasi yang sedang berjalan. Jika hasilnya sama maka algoritma k-means *cluster analysis* sudah *konvergen*, tetapi jika berbeda maka belum *konvergen* sehingga perlu dilakukan iterasi berikutnya. (Lynda & Widya, 2014)



Gambar 2.1 Flowchart K-Means

2.5 Pengujian sistem *Black Box*

Pengujian sistem yang dilakukan pada sistem informasi identifikasi topik dokumen tugas akhir yaitu menggunakan pengujian *blackbox(blackbox testing)*. Pengujian *blackbox* dilakukan untuk mengetahui apakah fungsi-fungsi yang ada pada sistem berjalan sesuai harapan atau tidak dengan melakukan inputan dan akan dilihat pada hasil keluaran yang akan ditampilkan.

2.6. Elbow

Metode Elbow merupakan suatu metode yang digunakan untuk menghasilkan informasi dalam menentukan jumlah *cluster* terbaik dengan cara melihat persentase hasil perbandingan antara jumlah *cluster* yang akan Membentuk siku pada suatu titik.

Berikut merupakan tahapan metode Elbow dalam menentukan nilai k pada K-Means:

1. Menginisialisai awal nilai k
2. Menaikan nilai k
3. Menghitung hasil sum of square error dari tiap nilai k
4. Analisis hasil sum of sequare error dari nilai k yang mengalami penurunan secara drastis
5. Cari dan tetapkan nilai k yang berbentuk siku.
6. Pada metode Elbow nilai cluster terbaik yang akan diambil dari nilai Sum of Square Error (SSE) yang mengalami penurunan yang signifikan dan berbentuk siku.

Untuk menghitung SSE menggunakan rumus :

$$SSE = \sum_{K=1}^{K} \sum_{XissK} ||X_i - C_k||^2$$

Dimana :

K= jumlah *cluster*

Xi=data ke-i

Ck=centroid *cluster*

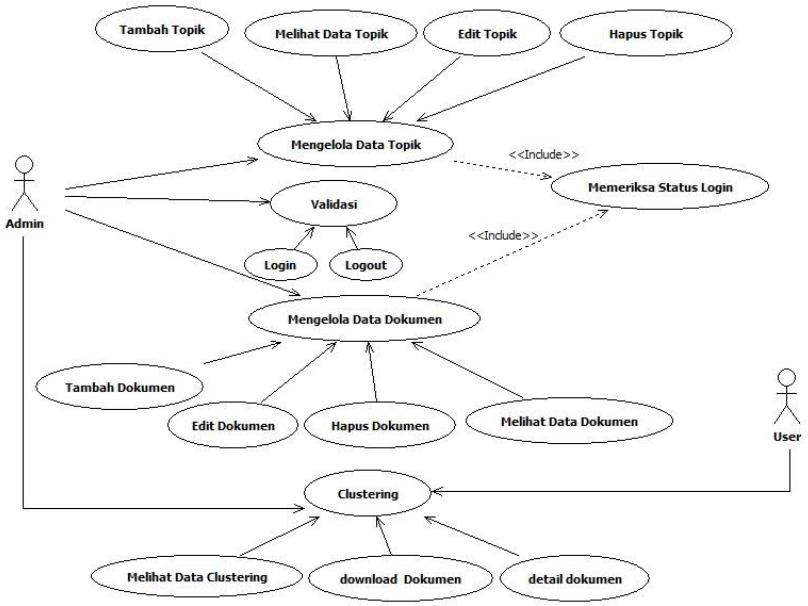
3. METODE PENELITIAN

3.1 Pengembangan Sistem

Berdasarkan tujuan peneliti agar mempermudah dalam penelitian sehingga metodelogi yang dapat digunakan dalam penelitian yaitu:

a) Data-data yang hanya diperlukan atau mendukung masalah berdasarkan pada rumusan masalah untuk fokus penelitian. Studi *literature* adalah cara yangdapat digunakan untuk menghimpun data-data atau sumber-sumber yang berhubungan dengan tujuan topik yang digunakan dalam penelitian. Studi *literature* didapatkan dari berbagai sumber, seperti jurnal,buku,laporan tugas akhir, internet dan pustaka
b) menganalisa data
   Menganalisa dan menetapkan data yang akan menjadi parameter. Seperti penganalisaan pada data tema/judul tugas akhir berdasarkan kebutuhan penelitian pada rumusan masalah.
c) Penentuan metode
   Metode yang digunakan dalam penentuan topik tugas akhir ini yang banyak telah digunakan yaitu dapat menggunakan metode K-Means *Clustering*.

3.2 *usecase*

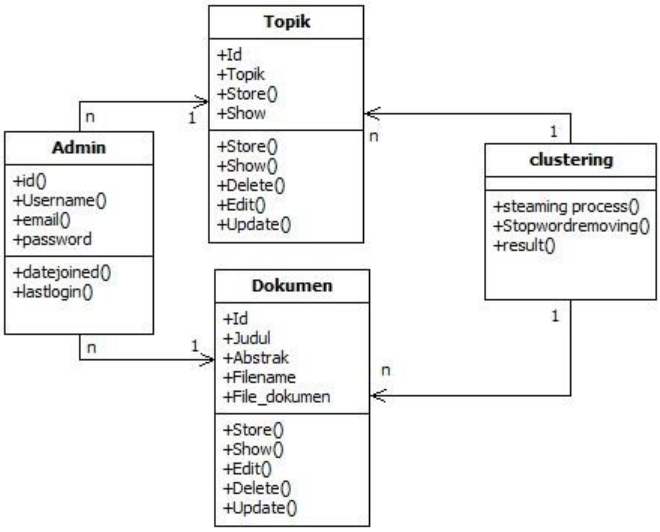


Gambar 3.2 *Use Case* Diagram

Pada gambar 2 merupakan *use case* diagram dari sistem identifikasi topik pada dokumen tugas akhir program Studi Teknik Informatika Unissula.

3.3 *Class* diagram

ISSN. 2720-9180



Gambar 3.3*Class* Diagram

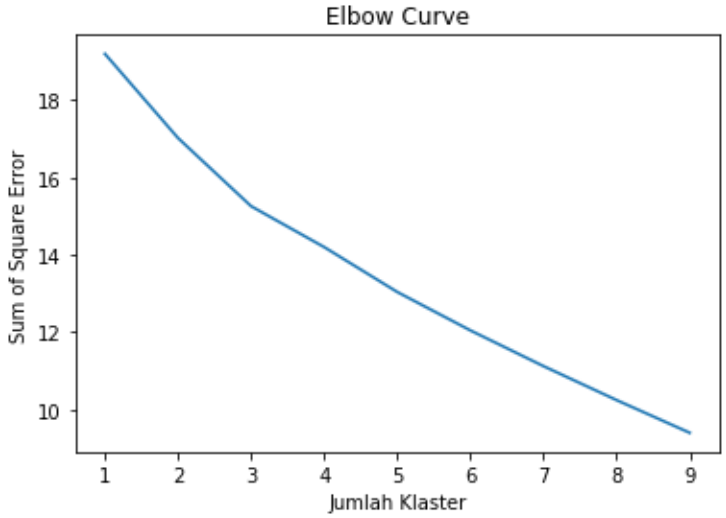
Perancangan pada gambar 3.10 adalah *Class* diagram memperlihatkan struktur sistem dari segi pendefinisian kelas-kelas yang akan dibuat untuk membangun sebuah system, yaitu terdapat kelas admin, topik, dokumen, dan clustering.

4. HASIL DAN ANALISI PENELITIAN

4.1 Pengujian Algoritma menggunakan Elbo

Tabel 4.1 Hasil *Sum Of Square*

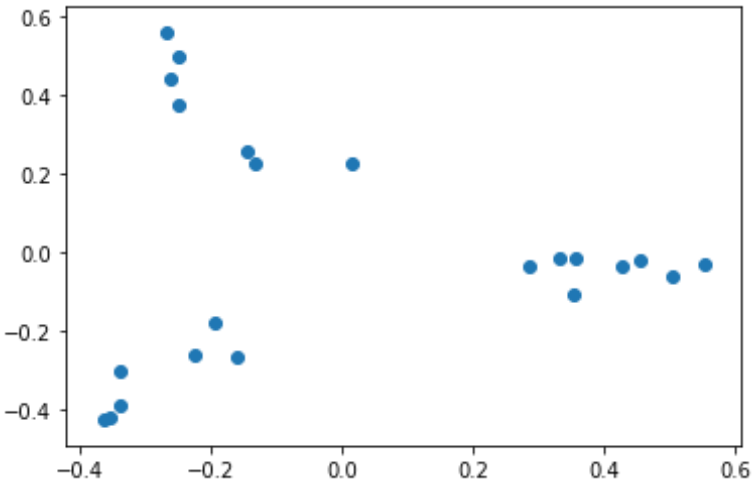
JUMLAH CLUSTER	SUM OF SQUARE ERROR
1	19.198907078445405
2	17.020838863463
3	15.25782078603296
4	14.200176230328404
5	13.038710203473572
6	12.045949185111498
7	11.121498285902637
8	10.243305048063071
9	9.398058547154838



Gambar 4.16 Grafik Elbow

Keterangan:

Gambar 4.16 yaitu kurva atau grafik Elbow yang dapat menjelaskan representasi dari tabel sebelumnya. Garis bawah kurva menunjukan jumlah Cluster yang diuji, dari 1 sampai dengan 9. Dan hasil garis biru menunjukan pergerakan SSE pada jumlah cluster ke 1 sampai dengan 3 berkurang terus menerus sebanyak 2, sedangkan dari hasil cluster 4 sampai 9 berkurangnya hanya 1. Jumlah cluster 3 menjadi titik Elbow (jumlah cluster terbaik)karena pada cluster 3 mengubah atau mempengaruhi penurunan hasil SSE awalnya berkurang 2 terus menerus menjadi berkurang 1 Kurva menunjukan bentuk siku pada cluster ke 3.

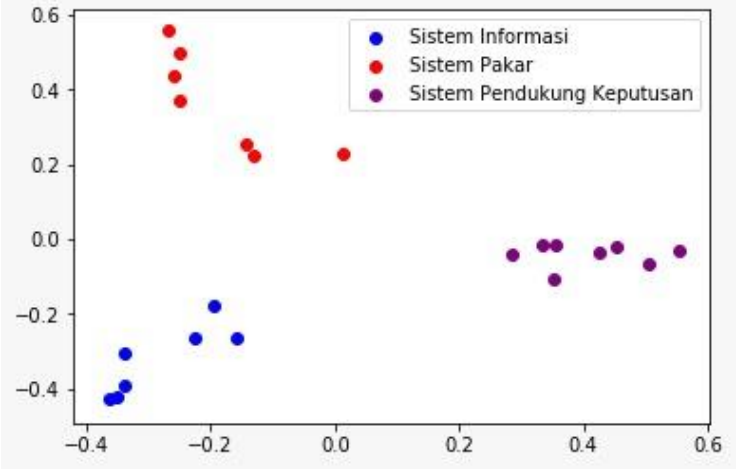


Gambar 4.17 Hasil plot Tf-Idf

Keterangan:

Gambar 3.17 merupakan hasil plot atau visualisasi Tf-Idf yang menjelaskan bahwa titik-titik biru pada plot menunjukan sebuah dokumen, atau jarak antara titik-titik biru pada plot diatas sama dengan jarak antara dokumen pada data.

ISSN. 2720-9180



Gambar 4.18 Plot hasil Clustering

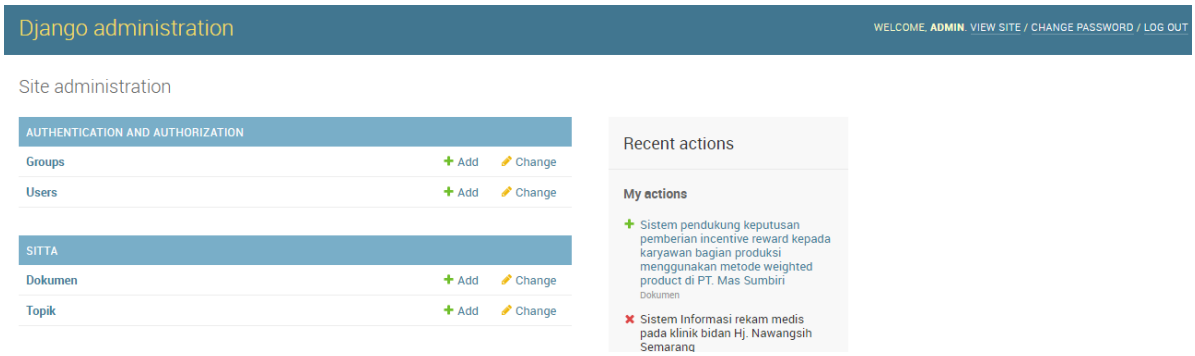
Keterangan:

Gambar 4.18 merupakan plot atau visualisasi hasil clustering yang menjelaskan bahwa pada plot diatas, tiap titik menunjukan sebuah dokumen, tiap dokumen memiliki warna pada titik dan menunjukan titik tersebut ada pada cluster bagian mana, dan titik x merupakan centroid dari setiap cluster. Pada titik biru merupakan cluster Sistem informasi, merah sistem pakar dan warna ungu merupakan sistem pendukung keputusan dapat dilihat pada plot bahwa dokumen masuk kedalam cluster yang benar tidak ada yang berpindah cluster.

4.2 Hasil sistem

Gambar 4.1 Halaman awal sistem

Pada halaman awal sistem pengguna tidak harus login,pengguna dapat langsung memilih topik yang diinginkan dapat ,memilih detail ataupun download dokumen dalam kelompok topik.



Gambar 4.2 halaman dasboard admin

Pada halaman hasil dasboard admin sistem, admin dapat melihat ,menambah, mengedit,menghapus dokumen dan topik.

5 KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian dan implementasi Sistem penentuan Topik TA menggunakan algoritma K-Means, maka dapat disimpulkan bahwa:

1. Hasil penelitian menunjukan bahwa algoritma yang diterapkan yaitu algoritma K-Means untuk *identifikasi* topik pada dokumen Tugas Akhir informatika Unissula dapat mengelompokan dokumen pada topik yang sesuai, dan untuk mengetahui jumlah k menggunakan rumus SSE (*Sum of Square Error*) dihasilkan 3 cluster yang memiliki nilai maksimal atau terbaik. Dengan jumlah cluster 3 menjadi titik Elbow jumlah cluster terbaik karena mengubah penurunan hasil SSE yang awalnya berkurang 2 terus menerus menjadi berkurang 1 sehingga terbentuk *cluster* 1,2 dan 3.
2. Sistem identifikasi topik pada judul/ dokumen TA dapat memberikan gambaran kepada mahasiswa apa saja judul yang masuk dalam kelompok topik tertentu sehingga sistem ini dapat membantu mahasiswa menentukan topik yang akan diambil.
3. Sistem identifikasi topik pada judul/dokumen TA dapat membantu proses pengelompokan dokumen dalam ruang baca Fakultas Teknologi Industri Unissula.

5.2. Saran

Adapun saran dari peneliti tentang penelitian yang telah dilakukan adalah:

1. Penelitian selanjutnya dapat memperbanyak topik karena topik yang diambil pada penelitiaan ini masih belum mencukupi atau belum memenuhi topik-topik yang terdapat didokumen tugas akhir informatika Fti Unissula.
2. Penelitian selanjutnya dapat membuat sistem dimana Topik dapat dibuat secara otomatis dan tidak ditambahkan oleh admin.

.

DAFTAR PUSTAKA

Agarwal, S. (2014). Data mining: Data mining concepts and techniques. In *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013*. https://doi.org/10.1109/ICMIRA.2013.45

Amburika, & Chrisnanto, D. (2016). Teknik Vector Space Model (VSM) dalam Penentuan Penanganan Dampak Game Online Pada Anak. *Prosiding SNST Ke-7 Tahun 2016*, 10–27. https://doi.org/10.1103/PhysRevC.6.1023

Ediyanto, Mara, & Dkk. (2013). Pengklasifikasian Karakteristik Dengan Metod K-Means Cluster Analysis. *Buletin Ilmiah*, *02*(2), 133–136.

ISSN. 2720-9180

Handoyo, & Rumani, D. (2014). Perbandingan Metode Clustering Menggunakan Metode Single Linkage Dan K-Means Pada Pengelompokan Dokumen. *JSM STMIK Mikroskil*, *15*(2), 73–82.

Lynda, & Widya, D. (2014). *ANALISA CLUSTERING MENGGUNAKAN METODE K-MEANS DAN HIERARCHICAL CLUSTERING (STUDI KASUS : DOKUMEN SKRIPSI JURUSAN KIMIA , FMIPA , 2 . 3 Term Weighting dengan Term Frequency*. *Volume 3*

N. munigsih & kiswati. (2015). 570-1255-1-Pb. *Jurnal Bianglala Informatika*, *3*(1).

Prabowo, & Fauzi, D. (2017). Peringkasan Teks Ekstraktif Kepustakaan Ilmu Komputer Bahasa Indonesia Menggunakan Metode Normalized Google Distance dan K-means | Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, *1*(December), 1697–1707.

Prilianti, & Kunci. (2014). Aplikasi Text Mining untuk Automasi Penentuan Tren Topik Skripsi dengan Metode K-Means Clustering. *Cybermatika*, *2*(1), 1–6. Retrieved from http://www.mendeley.com/research/aplikasi-text-mining-untuk-automasi-penentuan-tren-topik-skripsi-dengan-metode-kmeans-clustering

Putu, N., & Merliana Dkk. (n.d.). *Analisa Penentuan Jumlah Cluster Terbaik pada Metode K-Means*. 978–979.

Rahman, A. T. (2017). Coal Trade Data Clustering Using K-Means (Case Study Pt. Global Bangkit Utama). *ITSMART: Jurnal Teknologi Dan Informasi*, *6*(1), 24–31. https://doi.org/10.20961/ITS.V6I1.11296

Somantri, & Wiyono, D. (2016). Metode K-Means untuk Optimasi Klasifikasi Tema Tugas Akhir Mahasiswa Menggunakan Support Vector Machine (SVM). *Scientific Journal of Informatics*, *3*(1), 34–45. https://doi.org/10.15294/sji.v3i1.5845