Rancang Bangun Aplikasi *Text Mining* dalam Mengelompokkan Judul Penelitian Dosen Menggunakan Metode *Shared Nearest Neighbor* dan *Euclidean Similarity*

Lisna Zahrotun, Mushlihudin
Program Studi Teknik Informatika, Fakultas Teknologi Industri, Universitas Ahmad Dahlan
Kampus3, Jln. Prof. Dr. Supomo, Janturan, Yogyakarta 55164
e-mail: lisna.zahrotun@tif.uad.ac.id

Abstract

Data mining is the process of extracting hidden information into a know ledge. Some types of data in data mining are web mining, text mining, sequence mining, graph mining, temporal data mining, spatial data mining, distributed data mining and multimedia mining. Document grouping is one of the techniques of text mining. The purpose of this research is to build the application of lecturer research title classification using Shared nearest Neighbor method. The method used in the research is one of the methods of grouping in text mining that is Shared Nearest Neighbor (SNN) with Euclidean Similarity. Testing is done using black box test. The result of this research is text mining application that able to clustering the title of research lecturer.

Keywords: *text mining; shared nearest neighbor; euclidean similarity*

Abstrak

Data mining adalah proses untuk mengekstrak informasi tersembunyi menjadi sebuah pengetahuan. Beberapa jenis data dalam *data mining* adalah *web mining, text mining, sequence mining, graph mining, temporal data mining, mining spatial data, Mining data* terdistribusi *dan* multimedia *mining.* Pengelompokan dokumen merupakan salah satu teknik dari *text mining*. Tujuan penelitian ini adalah untuk membangun aplikasi pengelompokkan judul penelitian dosen menggunakan metode *shared nearest neighbor*. Metode yang digunakan dalam penelitian merupakan salah satu metode pengelompokkan dalam *text mining* yaitu *shared nearest neighbor* (SNN) dengan *euclidean similarity*. Pengujian dilakukan menggunakan *black box test*. Hasil dari penelitian ini adalah aplikasi *text mining* yang mampu mengelompokkan judul penelitian dosen.

Kata kunci*:* *text mining; shared nearest neighbor; euclidean similarity*

1. Pendahuluan

Data mining adalah proses untuk mengekstrak informasi tersembunyi menjadi sebuah pengetahuan. Beberapa jenis data dalam *data mining* adalah *web mining, text mining, sequence mining, graph mining, temporal data mining, mining spatial data, Mining data* terdistribusi dan multimedia *mining* [1]. *Text mining* merupakan proses penggalian data tersembunyi dengan data berbentuk *text*. Salah satu teknik dalam *text mining* adalah pengelompokkan. Pengelompokkan merupakan teknik yan digunakan untuk membentuk kelompok-kelompok yang memiliki kemiripan atau kesamaan dalam data dalam setiap kelompoknya [2]. Ini berarti satu objek memiliki kemiripan dalam setiap kelompok dan ketidak miripan terhadap objek pada kelompok lain. Metode *shared nearest neighbor* (SNN) merupakan metode pengelompokkan yang baik, dimana dalam metode ini pengelompokkan ditentukan dengan mencari ketetanggaan terdekat dari semua titik yang telah ditentukan sebelumnya [3]. Beberapa penelitian tentang *text mining* diantaranya penelitian yag membandingkan *cosine similarity* dan *jaccard similarity* pada metode SNN[4]. Perancangan aplikasi pengelompokkan

judul penelitian dosen meggunakan metode SNN juga pernah dilakukan, dimana dalam penelitian ini dihasilkan desaign aplikasi *text mining* [5]. Penggunaan *metode K-Nearest Neigbour* (KNN) untuk kategorisasi teks [6], pengelompokan teks data dengan *fuzzy c-means* [7], pengelompokkan dengan AHC *single linked* dan *K-Means* [8], dan *text mining* pengelompokkan judul kerja praktek dengan menggunakan metode *K-Means* yang dikolaborisakan dengan metode AHC untuk menentukan titik pusat awalnya [9].

Dengan banyaknya penelitian dan skema-skema dari penelitian maka tentunya penelitian yang dihasilkan juga akan sangat beragam dan bervariasi. Salah satu kendala dosen dalam penelitian adalah mencari pasangan yang tepat yang sesuai dengan bidang keilmuan. Padahal bidang keilmuan ini dapat dilakukan dengan dosen antar program studi. Jika para dosen mengetahui bidang minat dan riwayat dari penelitian-penelitian sebelumnya dosen lain tentu ini akan memudahkan dosen dalam berkolaborasi dengan dosen lain untuk melakukan penelitian.

Dengan demikian, dengan melihat dari penelitian-penelitian sebelumnya dan agar judul-judul penelitian dosen dari UAD dapat bermanfaat maka dilakukan penelitian dengan judul "Rancang Bangun Aplikasi *Text Mining* dalam Mengelompokkan Judul Penelitian Dosen Menggunakan Metode *Shared Nearest Neighbor* dan *Euclidean Similarity*". Dalam penelitian ini dilakukan pengembangan dari penelitian sebelumnya yaitu perancangan *text mining* pengelompokkan judul penelitian dosen menjadi sebuah aplikasi berbasis *web* yang mampu mengelompokkan judul penelitian dosen.

2. Metode Penelitian

Metode penelitian yangg dilakukan adalah pengumpulan data, analisis sistem, perancangan, implementasi dan pengujian. Pengumpulan data dilakukan dengan studi literatur dari berbagai macam buku, artikel, publikasi ilmiah untuk mempelajari mekanisme *teks Mining* dalam mengelompokkan data, mekanisme pengelompokan data menggunakan metode SNN dan algoritma SNN dan *euclidean similarity*. Selain itu juga dilakukan wawancara dengan cara bertanya langsung terhadap nara sumber yaitu kepada karyawan yang menangani tentang pendataan penelitian dosen UAD. Metode observasi juga dilakukan dengan cara melihat proses pendataan judul penelitian dosen melalui *website* lpp.uad.ac.id dan juga pendataan judul penelitian dosen menggunakan *form hardcopy*.

Analisis sistem dilakukan dengan beberpa tahap dalam yang mengcu pada prose *text mining* yaitu *cleaning*, *tokenizing, filtering*. Kemudian penghitungan *similarity* menggunakan *Euclidean similarity* dan pemahaman dari algoritma *shared nearest neighbor* (SNN). Algoritma *shared nearest neighbor* (SNN) merupakan proses pengelompokkan pada data yang memiliki dimensi tinggi yang telah dikembangkan [3]. Algoritma SNN memerlukan 3 masukan parameter yaitu *k* yang merupakan jumlah tetangga terdekat, *e* yang merupakan nilai ambang ketetanggaan yang dimiliki secara bersama dan *mint* yang merupakan jumlah minimal data untuk setiap kelompok.

Langkah-langkah Algoritma *shared nearest neighbor* (SNN)

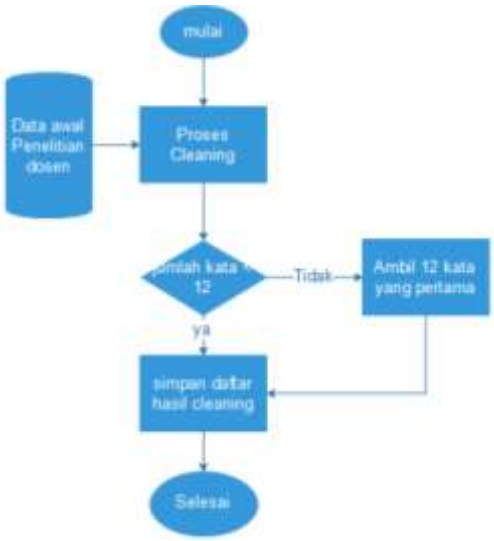
a) Menghitung nilai kesamaan dari data yang ada
b) Membentuk daftar k-tetangga terdekat masing-masing titik data untuk *k* data
c) Membentuk graph ketetanggaan dari daftar *k* tetangga terdekat
d) Mencari kepadatan untuk setiap data
e) Menemukan titik-titik representatif
f) Membentuk kelompok dari titik-titik representatif tersebut

Sedangkan untuk menghitung jarak kemiripan antar judul digunakan *euclidean similarity. Euclidean similarity* merupakan penentuan akar perbedaan persegi antara koordinat sepasang objek. Untuk jarak vektor x dan y (x, y) ditunjukkan dalam persamaan 1 [10]

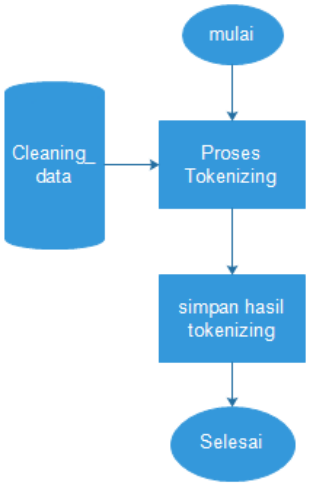
$$\mathrm{Sim}\,(\mathrm{x},\mathrm{y}) = d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

Dimana x dan y adalah vektor n-dimensi

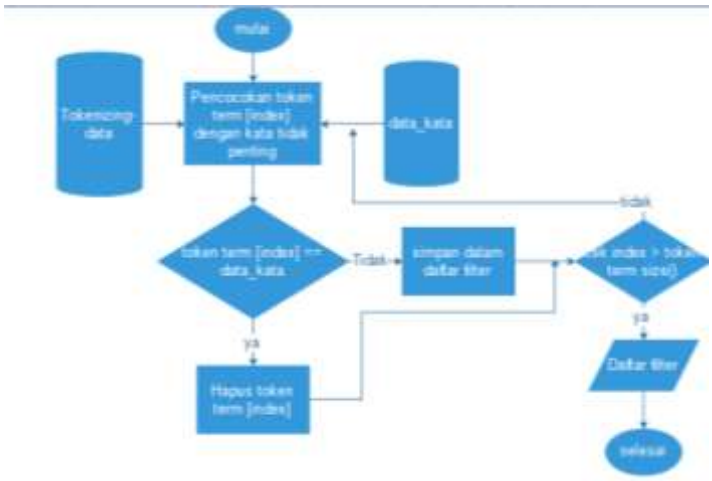
Setelah melalui tahap analisis maka dilakukan prose peracangan. Tahap perancangan dilakukan untuk menurunkan hasil analisis ke tahap yang lebih detail. Beberapa tahap perancangan yang dilakukan adalah perancangan *flowchart*, perancangan basis data dan perancangan *user interface*. Perancangan *flowchart* meliputi tahapan *text mining* yang meliputi proses *cleaning, tokenizing* dan *filtering*. Dalam proses *cleaning* dilakukan pemotongan kata pada judul kerja praktek yang melebihi 12 kata. Sehingga dari data awal dalam *data base jika* ditemukan judul yang memiliki lebih dari 12 kata, maka kata ke 13 sampai terakhir dihilangkan. Tampilan *flowchart* proses *cleaning* ditunjukkan dalam Gambar 1. Proses *tokenizing* dalam penelitian ini di mulai dengan pengambilan data judul kerja praktek dalam *data base,* dari data judul kerja praktek tersebut kemudian dilakukan proses *tokenizing*. Hasil dari proses *tokenizing* ini disimpan dalam *data base* kembali. *Flowchart* proses *tokenizing* ditunjukkan dalam Gambar 2. Dalam penelitian ini proses *filtering* dilakukan dengan menggunakan *modle stoplist* atau membuang kata-kata yang tidak penting. Pertama kata-kata yang dianggap tidak penting di simpan dalam *data base* yaitu di, ke, dari, dan, untuk, pada, atau. Setelah di simpan dalam *data base* maka kaa yang tidak penting tersebut akan di panggil untuk dicocokkan terhadap kata pada setiap judul. Jika ditemukan salah satu dari daftar *stoplist* di dalam judul kerja praktek maka kata tersebut akan dihapus oleh *system*. Hasil proses *filtering* ini kemudian disimpan dalam *data base. Flowchart* proses *filering* ditampilkan dalam Gambar 3.



Gambar 1. *Flowchart* proses *cleaning*

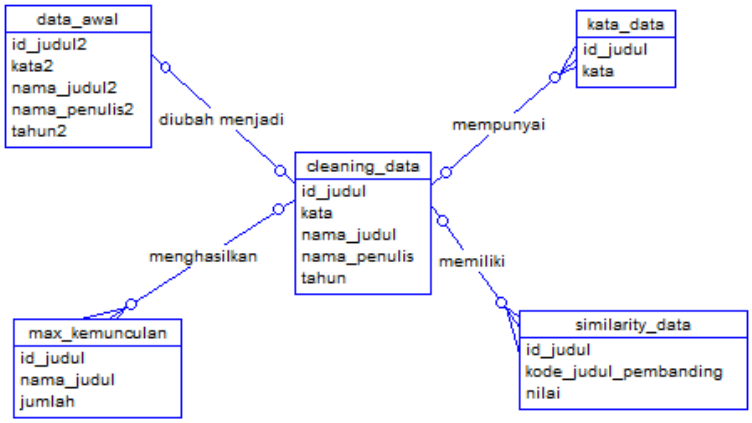


Gambar 2. *Flowchart* proses *tokenizing*



Gambar 3. *Flowchart* proses *filtering*

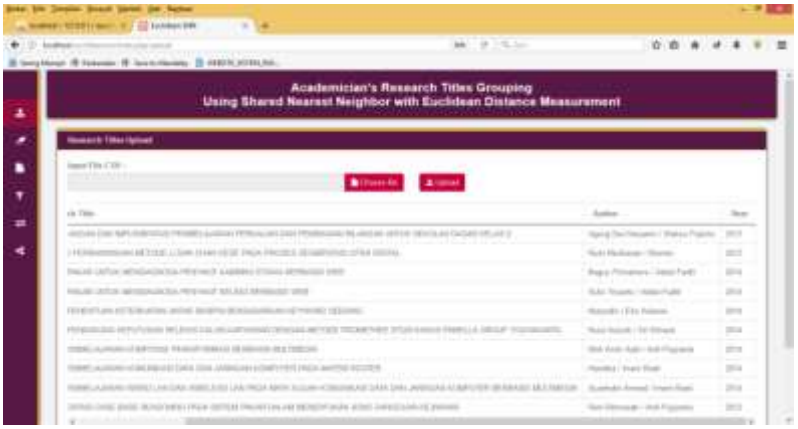
Perancangan basis data meliputi tabel data kata untuk menyimpan setiap kata hasil dari *tokenizing* dan *filtering* untuk setiap judul, tabel data frekuensi untuk menyimpan frekuensi kemunculan setiap kata, tabel data nilai, untuk menyimpan nilai kemiripan setiap judul kerja praktek dengan judul kerja praktek yang lain dan tabel data *cluster*, untuk menyimpan judul yang yang telah dikelompokkan. Rancangan basis data digambarkan dalam bentuk ERD (*entity relationship diagram*) yang digunakan untuk pemodelan bisnis data *relational*. Gambar 4 menggambarkan tenatang rancangan basis data sistem dalam bentuk ERD. Tahap yang terakhir adalah impelentasi program dan pengujian



Gambar 4. *Entity relationship diagram (ERD)*

3. Hasil dan Pembahasan

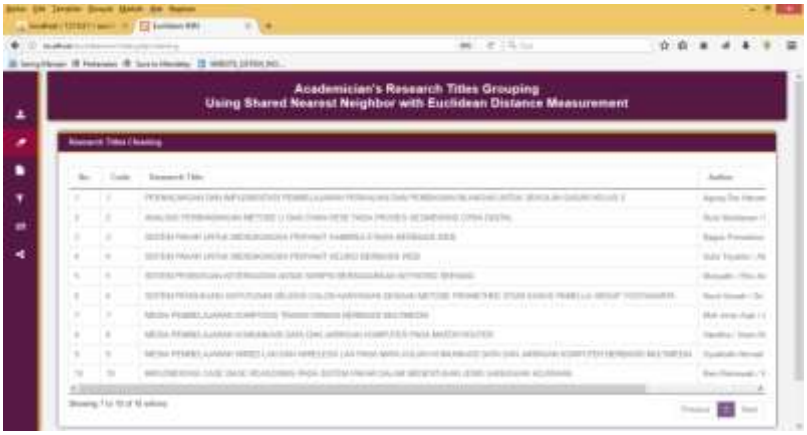
Hasil dari penelitian ini adalah sebuah aplikasi pengelompokkan judul penelitian dosen, dengan sampel data yag digunakan adalah sejumlah 13 data. Aplikasi yang dibuat Dalam proses ini sistem mampu mengambil data dari excel. Tipe data yang dapat diakses oleh sistem ini adalah xls. Tampilan hasil proses *load* data ditunjukkan dalam Gambar 5.



Gambar 5. Tampilan hasil proses *load* data

3.1 Proses *cleaning*

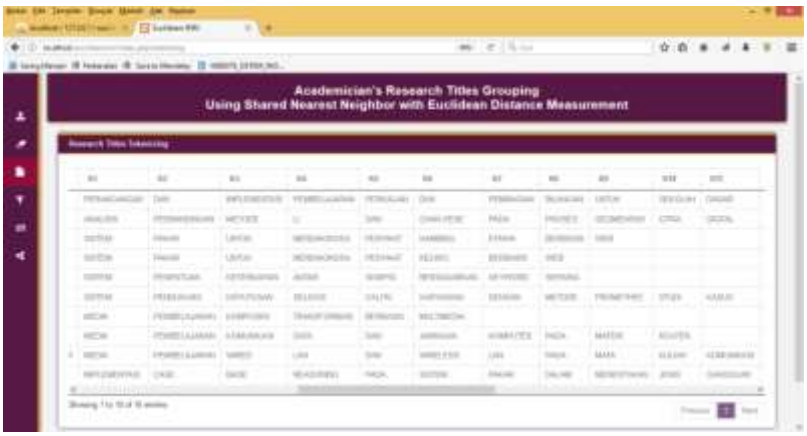
Proses *cleaning* atau pembersihan data dalam aplikasi *text mining* ditunjukkan dalam Gambar 6.



Gambar 6. Tampilan proses hasil *cleaning*

3.2 Proses *tokenizing*

Proses *tokenizing* data dalam aplikasi *text mining* ditunjukkan dalam Gambar 7.



Gambar 7. Tampilan hasil proses *tokenizing*

3.3 Tampilan *filtering*

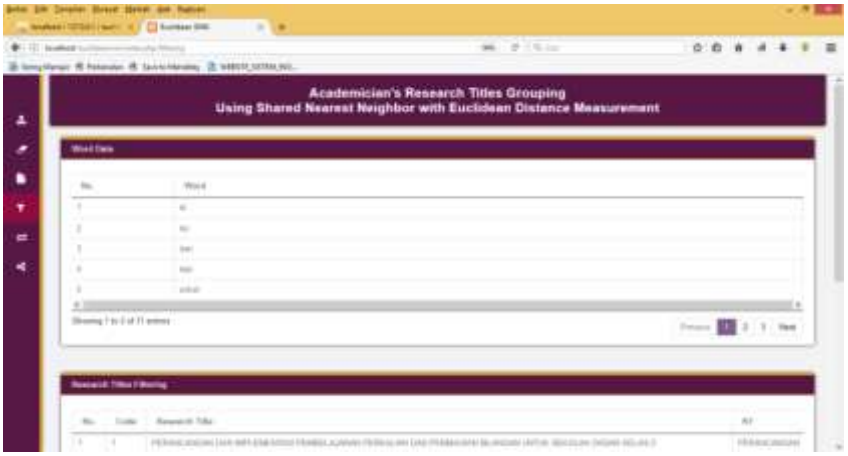
Tampilan daftar kata dalam aplikasi *text mining* ditunjukkan dalam Gambar 8.

3.4 Proses penghitungan *similarity*

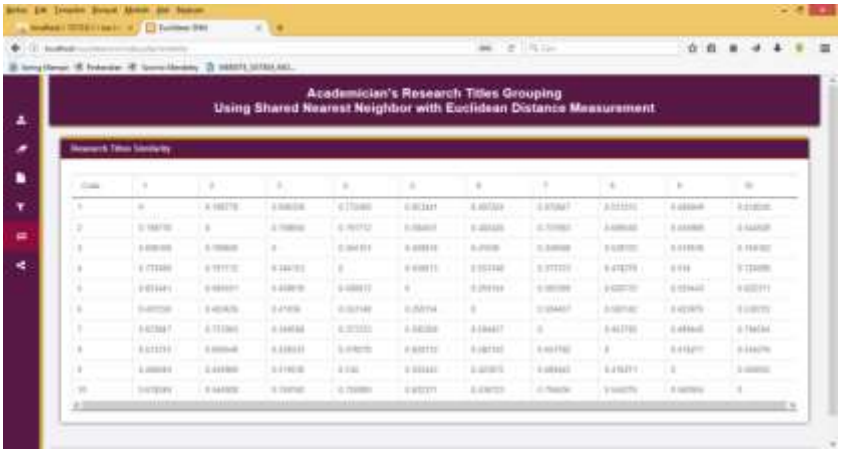
Proses penghitungan *similarity* data dalam aplikasi *text mining* ditunjukkan dalam Gambar 9

3.5 Proses pengelompokkan menggunakan metode SNN

Proses pengelompokkan menggunakan metode SNN dalam aplikasi *text mining*. Tampilan hasil pengelompokkan direpresentasikan dalam dua bentuk yaitu daftar judul setiap kelompok yang ditunjukkan dalam Gambar 10 dan hasil pengelompokkan dalam bentuk grafik yang ditunjukkan dalam Gambar 11.

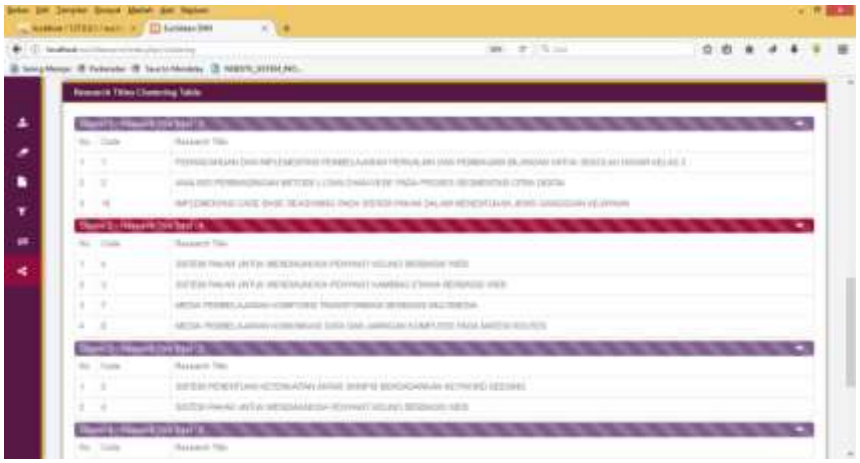


Gambar 8. Tampilan *filtering*



Gambar 9. Tampilan hasil penghitungan *similarity*

Gambar 10. Tampilan hasil pengelompokkan dalam bentuk grafik



Gambar 11. Tampilan hasil pengelompokkan dalam bentuk *list* judul

3.6 Pengujian

Hasil dari pengujian *black box test* ini dilakukan terhadap ahli *data mining*. Dimana hasil dari *testing blackbox test*.

Tabel 1 .*Form blackbox test*

Test ID	*Function Name / Process Name*	*Description*	*Expected Results*	*Actual Results*
1.	Menu *data set*	Klik menu *Load* Data	Menampilkan daftar data judul penelitian dosen	√
		Klik *button Choose File*	Menampilkan *pop up windows* untuk memilih *file* yang akan di *upload*	√
		Klik *button Upload*		
		Syarat : *file* yang di *upload* berekstensi .csv	Sistem akan mengunggah *file* yang dipilih ke dalam *data base* kemudian menampilkan daftar data judul penelitian dosen	√

		Syarat : *File* yang di *upload* tidak berekstensi .csv	Menampilkan pesan *error*, file yang di *upload* harus berekstensi .csv	√
2.	Menu *Cleaning*	Klik Menu *Cleaning*	Menampilkan daftar data judul penelitian dosen yang sudah melalui proses *cleaning* data (*cleaning:* menghapus data yang kosong)	√
3.	Menu *tokenizing*	Klik menu *Selection*	Menampilkan daftar data konsumen hasil seleksi,	√
4.	Menu *filtering*	Klik menu *Filtering*	Menampilkan daftar kata Menampilkan daftar hasil *filtering*	√
7.	Menu *similarity*	Klik menu *similarity*	Menampilkan halaman hasil penghitungan *similarity*	√
8.	Menu *Clustering*	Klik menu *clustering*	Menampilkan hasil pengelompokkan judul penelitian dosen dalam bentuk data dan grafik	√
		Syarat : *form input* tidak diisi semua	Muncul peringatan form validasi yang belum diisi	√
		Syarat : *form input* diisi semua	Menampilkan halaman *fold-growth* dengan hasil pola assosiasi	√

Dari hasil penilaian dan pengujian *blackbox test*, dapat disimpulkan bahwa aplikasi *data mining segmentasi* konsumen sudah mampu melakukan proses *clustering* dengan baik dan berfungsi sebagaimana mestinya.

3.7 Representasi Pengetahuan dan Evaluasi Pola

Dari hasil pengujian *Black box test* telah dihasilakan sebuah aplikasi *data mining* berbasis *web* yang mampu melakukan pengelompokkan judul penelitian dosen. Dari hasil aplikasi yang ada maka dilakukan analisis terhadap hasil pengelompokkan.

4. Kesimpulan

Dari penelitian yang telah dilakukan dapat disimpulkan bahwa aplikasi *text mining* yang dibangun dapat berjalan dengan baik. Hal ini berdasarkan hasil pengujian dari *black box test.* Selain menampilkan judul penelitian dosen aplikasi yang dibangun dapat menampilkan proses *cleaning, tokenizing*, dan *filtering* dan proses pengelompokkan judul penelitian dosen menggunakan metode SNN dan *eucliean similarity*.

5. Ucapan Terima Kasih

Penelitian ini telah didukung oleh hibah penelitian RISTEK DIKTI dengan skema Penelitian Dosen Pemula (PDP) tahu anggaran 2017

Referensi

[1] R. Janani and S. Vijayarani, “Text Mining Research : A Survey,” *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 4, no. 4, pp. 6564–6571, 2016.
[2] B. Santosa, *D*ata Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis. Yogyakarta: Graha Ilmu, 2007.
[3] R. F. Zainal and A. Djunaidy, “Algoritma Shared Nearest Neighbor Berbasis Data Shrinking,” *JUTI*, vol. 7, pp. 1–8, 2008.
[4] L. Zahrotun, “Comparison Jaccard similarity , Cosine Similarity and Combined Both of the Data Clustering With Shared Nearest Neighbor Method,” vol. 5, no. 1, pp. 11–18, 2016.

[5] Mushlihudin and L. Zahrotun, "Perancangan Text mining Pengelompokkan Penelitian Dosen Menggunakan Metode Shared Nearest Neihbor dengan Euclidean Similarity," in *Seminar Nasional Teknologi dan Informatika (SNATIF)*, 2016, pp. 849–855.
[6] S. Jiang, G. Pang, W. Meiling, and K. Limin, "An Improved K-Nearest-Neighbor Algoritm for Text Categorization," *Expert Syst. with Appl.*, vol. 39.1, pp. 1503–1509, 2012.
[7] C. Li and L. Nan, "A Novel Text Clustering Algorithm," *Energy Procedia*, vol. 13, pp. 3583–3588, 2011.
[8] R. Handoyo, S. M. Nasution, P. Studi, S. Komputer, S. Linkage, and S. Coefficient, "Perbandingan Metode Clustering Mengggunakan metode Single Linkage dan K-Means Pada Pengelompokkan Dokumen," *JSM STMIK Mikroskil*, vol. 15, no. 2, pp. 73–82, 2014.
[9] T. Alfina and B. Santosa, "Analisa Perbandingan Metode Hierarchical Clustering, K-Means dan Gabugan Keduanya dalam Membentuk Cluster Data (Studi Kasus : Problem Kerja Praktek Jurusan Teknik Industri ITS)," *J. Tek. POMITS*, vol. 1, no. 1, pp. 1–5, 2012.
[10] A. K. Patidar, J. Agrawal, and N. Mishra, "Analysis of Different Similarity Measure Functions and their Impacts on Shared Nearest Neighbor Clustering Approach," *Int. J. Comput. Appl.*, vol. 40, no. 16, pp. 1–5, 2012.