PENGEMBANGAN APLIKASI PENENTUAN TEMA TUGAS AKHIR BERDASARKAN DATA ABSTRAK MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR

Ramadhan Rakhmat Sani[1], Junta Zeniarza[2], Ardytha Luthfiarta[3]
Teknik Informatika S1, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
Jalan Imam Bonjol No. 207, Semarang 50131, Jawa Tengah, Indonesia
Telp. (024) 3517261
E-mail: ramadhan_rs@dsn.dinus.ac.id, junta@dsn.dinus.ac.id,
ardytha.luthfiarta@dsn.dinus.ac.id

ABSTRAK

Semakin bertambahnya mahasiswa saat tahun ajaran baru pada suatu perguruan tinggi tentunya bertambah pula hasil tugas akhir yang dihasilkan dan terekam pada perpustakaan. Semua data tersebut ditempatkan dalam tempat tertentu yang disesuaikan dengan kategorinya dan data tersebut sudah terkomputerisasi dengan baik di dalam server database. Ketika harus mengidentifikasi dan mencari file laporan tugas akhir berdasarkan topik yang diangkat menjadi masalah pada penelitian ini. Penelitian ini bertujuan untuk mempermudah dalam melakukan pengelompokan tema atau topik yang nantinya digunakan sebagai acuan dalam penenpatan pada kategori yang disediakan. Metode yang digunakan dalam penelitian menggunakan Rapid Application Developmnet model dengan menggunakann algoritma klasifikasi K-Nearest Neighbor yang terbukti mencapai hasil akurasi yang baik dan sesuai dengan perhitungan yang diterapkan dalam sebuah aplikasi. Sehingga aplikasi ini dapat menjadi solusi untuk menyelesaikan masalah tersebut.
.

Kata Kunci: klasifikasi, text mining, aplikasi, k-nearest neighbor, prototype

1. PENDAHULUAN

1.1 Latar Belakang

Pada era informasi digital yang berkembang pesat telah berdampak pada meningkatnya volume informasi dalam bentuk teks. Dari banyaknya informasi digital, diperkirakan 80% dokumen digital merupakan data berpotret teks (Hamzah 2012). Salah satu instansi yang memanfaatkan pengolahan informasi digital adalah Perpustakaan. Bagian terpenting pada perpustakaan yang ada di Universitas yaitu dalam menyediakan buku-buku referensi untuk topik tugas akhir. Dimana sering terjadi kesulitan ketika perpustakaan harus mengenali buku-buku referensi tersebut sesuai dengan topik tugas akhir dikarenakan melimpahnya informasi yang tersimpan (Hamzah 2012).

Semakin berkembangnya informasi teks yang tidak terstruktur dalam analisis teks merupakan salah satu faktor berkembangnya *text mining*. Cara kerja *text mining* ialah informasi yang digali dari suatu teks yang tidak terstruktur diproses menjadi matriks *term* dokumen untuk dicari polanya (Hamzah 2012). Secara umum topik laporan tugas akhir dapat dijadikan panduan umum untuk mengerti isi suatu buku, namun isi dari laporan tugas akhir tesebut juga dapat di *retrieve* untuk menjelaskan atau mendapatkan informasi atau pengetahuan lain.

Salah satu penelitian yang mendiskusikan permasalahan tersebut adalah pendekatan *supervised learning* dengan klasifikasi atau kategorisasi teks yang saat ini mempunyai banyak

cara pendekatannya seperti berbasis numeris, misalnya pendekatan probabilistic, KNN, *Artificial Neural Network, Support Vector Machine*, dan juga berbasis non numeris salah satunya *Decision Tree*. Dalam pendekatan KNN yang berbasis numeris ada beberapa kelebihan diantranya, cepat, berakurasi tinggi dan sederhana (Aggarwal & Zhai 2012). Dalam algoritma KNN jarak atau perbedaan atribut kata yang hadir pada suatu dokumen menjadi dasar pengklasifikasian dokumen teks. Kinerja KNN sebagai algoritma klasifikasi cukup bagus ditunjukkan oleh beberapa penelitian yang menggunakannya. Berdasarkan penjelasan di atas penelitian pengklasifikasian topik laporan tugas akhir berdasarkan dokumen teks pada abstraknya, dilakukan dengan menggunakan metode KNN.

2. KAJIAN PUSTAKA

2.1 Penelitian Terdahulu

Beberapa penelitian terkait dengan penelitian ini pernah dilakukan (Darujati et al. 2012) mengembangkan sebuah aplikasi untuk klasifikasi teks bahasa Indonesia dengan menerapkan metode *naïve bayes classifier*.Hasil akurasi terbaik dari data uji yang bersumber dari situs web dengan data latih yang besar menghasilkan akurasi lebih dari 87% dan berjalan baik untuk data latih lebih dari 150 dokumen. Dengan metode yang sama (Hamzah 2012) mengelompokkan teks berita dan abstrak akademis. Akurasi yang dicapi untuk dokumen berita 91% sedangkan untuk dokumen akademik 82% dengan 450 dokumen abstrak akademik dan 1000 dokumen berita. Kelemahan dari metode ini dalam pengasumsian yang sulit dipenuhi, yaitu independensi fitur kata.

Februariyanti (2012) membangun aplikasi untuk mengimplementasikan klasifikasi dokumen berupa penggunaan ontologi dalam berita teks bahasa Indonesia dengan objek artikel berita berbahasa Indonesia dari *internet*. Luaran yang dihasilkan berupa halaman web yang mengandung kata kunci yang tersimpan pada file.

Selain itu (Zainal & Novan, 2012.) mengklasifikasikan dokumen berita berbahasa Indonesia menggunakan algoritma *single pass clustring* dengan menggunakan sampel berita dari media massa berbasis web. Hasil yang didapat dari pengujian dengan pemilihan nilai *threshold* yang tepat akan meningkatkan kualitas *information retrieval* dengan tingkat recall 79% dan precision 88%.

Adapun (Krisandi et al. 2013) Mengadopsikan algoritma *K-Neares Neighbor* untuk mengklasifikasi data hasil kelapa sawit. Dihasilkan dalam 6 cluster berdasarkan kesamaan hasil produksi dari kelompok tani yang ada sehingga dapat memperkirakan hasil produksi dimasa mendatang.

2.2 Rekayasa Perangkat Lunak

Menurut (Pressman, 2014) perangkat lunak merupakan instruksi-instruksi dalam bentuk program komputer yang memiliki fitur yang diinginkan, fungsi dan kinerja bila dieksekusi. Berisikan struktur data yang memungkinkan untuk memanipulasi informasi. Bisa dikatakan Rekayasa perangkat lunak merupakan disiplin ilmu yang mempelajari tentang prinsip pembuatan perangkat lunak yang ekonomis (Pressman, 2014) dari proses pembuatan hingga memasuki tahapan penggunaan (Sommerville, 2015).

2.3 *Unifid Modelling Language* (UML)

Merupakan sekumpulan konvensi yang digunakan untuk menentukan atau menggambarkan sebuah sistem perangkat lunak unutk memodelkan perangkat lunak berbasis objek.. Beberapa diagram yang dipakai dalam UML adalah *Use Case Diagram, Activity Diagram, Sequence*

ISBN: 978-979—3649-96-2 *Unisbank Semarang, 28 Juli 2016*

Diagram, State Chart Diagram, Class Diagram, Collaboration Diagram, Component Diagram dan Deployment Diagram.

2.4 *Information Retrieval*

Merupakan sebuah konsep untuk melakukan proses mencari dan mendapatkan apa yang dicari beserta prosedur-prosedur dan metode-metode untuk mendapatkan informasi yang tersimpan yang relevan. Seperti dalam panggilan (*searching*), indeks (*indexing*), pemanggilan data kembali (*recalling*) (Zainal & Novan, 2012.). Jenis pencarian data tersebut yang dapat ditemukan diantaranya teks, tabel, gambar (image), video, audio dengan cara atau menguragi dokumen pencarian yang tidak relevean atau meretrieve dokumen yang relevan yang bertujuan untuk memenuhi informasi pengguna..

2.5 *K-Nearest Neighbor* (KNN)

Merupakan salah satu metode berbasis NN yang paling tua dan popular di dalam melakukan pengkategorian teks (Toker & Kırmemiş n.d.; Yan et al. 2013; Yao & Vocational 2014). Nilai K yang digunakan disini menyatakan jumlah tetangga terdekat yang dilibatkan dalam penentuan prediksi label kelas pada data uji (Arifin et al. n.d. 2012). Dari K tetangga terdekat yang terpilih kemudian dilakukan voting kelas dari K tetangga terdekat tersebut. Kelas dengan jumlah suara tetangga terbanyaklah yang diberikan sebagai label kelas hasil prediksi pada data uji (Toker & Kırmemiş n.d.). Dengan data latih yang berjarak paling dekat dengan objek untuk melakukan klasifikasi dianggap sebagai metode terbaik dalam preses tersebut (Santoso et al. 2014). Adapun cara kerja dari KNN perlu adanya penetuan inputan berupa data latih, data uji dan nilai k. Kemudian menghitung jarak data yang diuji dengan data latih dengan mengurutkan data latih berdasarkan kedekatan jaraknya. Setelah itu Pengambilan k data latih teratas untuk menentukan kelas klasifikasi untuk kelas yang dominan dari k data latih yang diambil. Dekat atau jauhnya tetangga biasanya dihitung dari *Euclidean Distance* yang direpresentasikan dengan rumus sebagai berikut :

$$D(a,b) = \sqrt{\sum_{k=1}^{d}(a_k - b_k)^2} \quad (1)$$

3. METODE

3.1 Metode Pengumplulan data

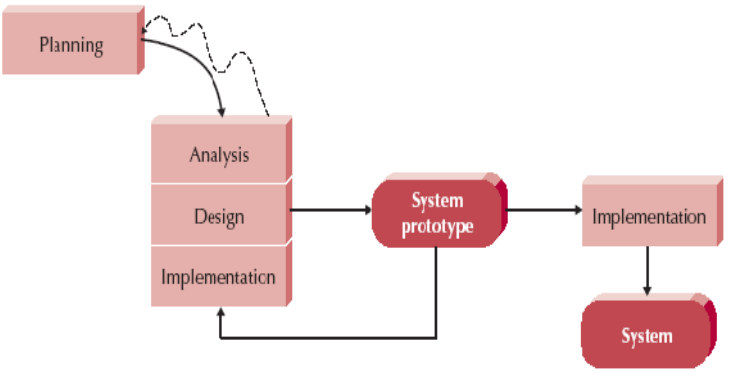
Data yang digunakan pada penelitian ini berupa data abstrak tugas akhir mahasiswa yang didapat dari beberapa sumber. Dokumen abstrak tersebut berjumlah 9 dokumen dan dibagi menjadi data training dan data testing. Terdapat tiga kategori class yaitu *Image*, IR (*Information Retrieval)*, dan Jarkom (Jaringan Komputer). Dari data dokumen tersebut 9 data abstrak dijadikan sebagai data training dan 1 data abstrak dijadikan sebagai data testing.

3.2 Metode Perancangan Sistem

Penelitian ini menggunakan metodologi pengembangan SDLC (*System Development Life Cycle)* dengan metode (*Rapid Application Development) prototyping* pada gambar 1 yang tahapannya diawali dengan *Planning* (Perencanaan) dilanjutkan dengan *Analysis* (Analisa), *Design* (Perancangan), *Implementation* (Implementasi) yang dilakukan untuk menghasilkan sistem *prototype* (purwa rupa). Perbaikan *prototype* dilakukan secara berulang-ulang dalam siklus (analisa-perancangan-implementasi), pengulangan perbaikan *prototype* berhenti ketika

ISBN: 978-979—3649-96-2 *Unisbank Semarang, 28 Juli 2016*

prototype merupakan sistem kerja yang lengkap untuk diimplementasikan pada tahap akhir menjadi sistem yang utuh.



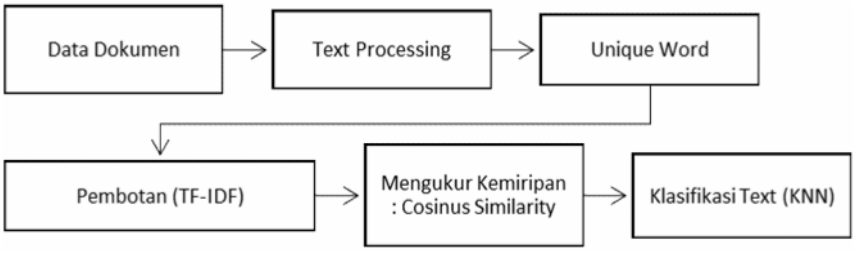
Gambar 1 RAD Prototyping

a. Tahap Perencanaan

Kegiatan dalam perencanaan dimulai dari Identifikasi dari pengguna dengan penganalisi untuk tujuan dar dibangunnya aplikasi penentuan tema atau topik pada abstrak tugas akhir mahasiswa. Dengan memanfaatkan perangkat komputer yang sudah ada pada perpustakaan.

b. Tahap Analisis

Analisis dilakukan dengan menganalisis data dan informasi yang diperoleh dari data dokumen yang terdiri dari dokumen *training* yang sudah memiliki kategori dan dokumen *testing* yang belum memiliki kategori merupakan dua jenis inputan dari sistem ini. Keduanya melalui proses *text proscesing* seperti pada gambar 2 dengan *tokenizing*, yaitu pemecahan dokumen menjadi frase atau *term* (kata), sesuai dengan dokumen masing - masing. Setelah dilakukan *tokenizing*, tahapan selanjutnya adalah dilakukan *stopword removal* atau penghilangan kata yang dirasa tidak diperlukan. Kemudian tahapan yang terakhir adalah *stemming*, yaitu penghilangan imbuhan sehingga menjadi kata dasar. Pada fase *feature weighting*, term-term dari dokumen *training* dan dokumen *testing* dihitung bobotnya. *Cosine similarity* digunakan untuk proses menghitung kemiripan antar dokumen. Kelas atau pengelompokan dari dokumen testing merupakan hasil dari proses classifier. Kemuadian membangun classifier dari bobot dokumen training dan dokumen testing pada proses klasifikasi. Algoritma KNN digunakan untuk membentuk *Classifier*.



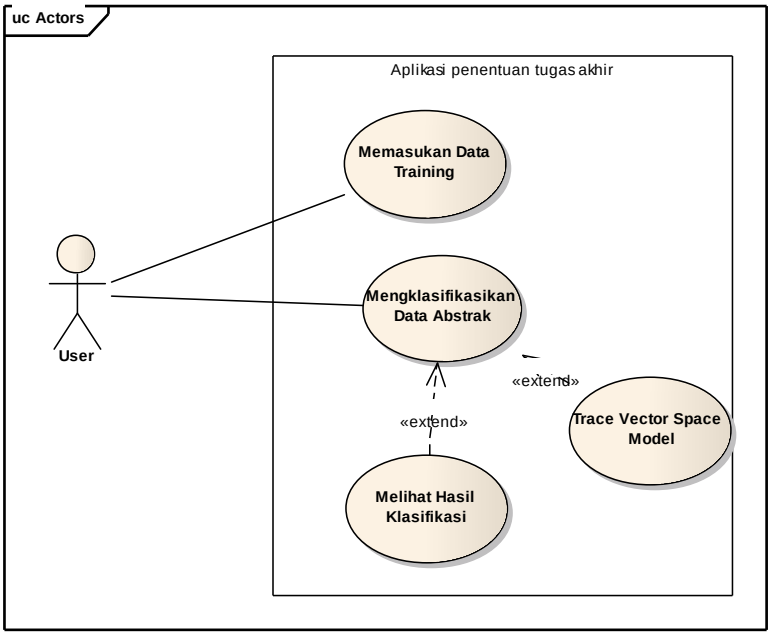
Gambar 2. Proses *text prosessing*

c. Tahap Desain

Perancangan dilakukan untuk mendapatkan deskripsi arsitektural perangkat lunak, deskripsi antarmuka, deskripsi data yang telah dilakukan pada tahap analisis, membuat sketsa antarmuka

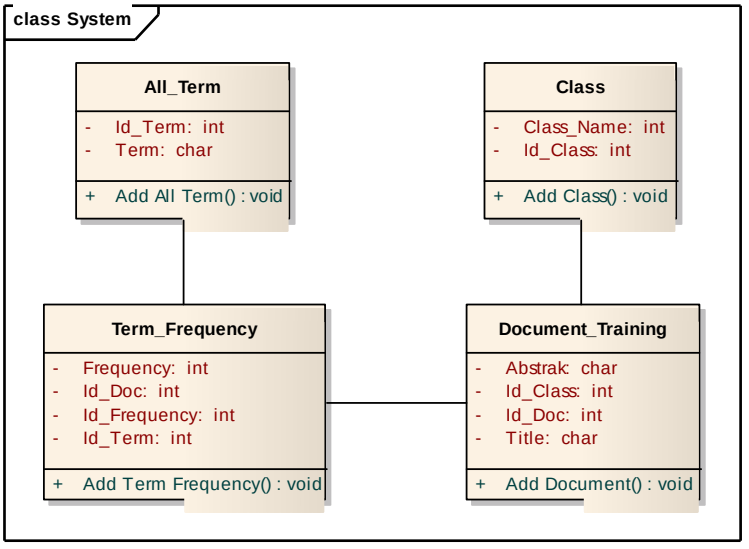
ISBN: 978-979—3649-96-2 *Unisbank Semarang, 28 Juli 2016*

dari aplikasi, memetakan obyektif pengguna ke dalam antarmuka yang spesifik. Pada tahapan ini digunakan bahasa pemodelan UML yang memberikan standar penulisan tersendiri pada sebuah sistem *blue print*, yang mencakup konsep proses bisnis, penulisan kelas-kelas pada bahasa program yang spesifik, skema database dan komponen-komponen yang dibutuhkan dalam sistem piranti lunak. Adapun gambaran dari *use case diagram* pada aplikasi penentuan tugas akhir ditampilkan pada gambar 3.



Gambar 3 *Use case diagram*

Pada gambar 4 menujukan class diagram sebagai perancangan basis data yang digunakan untuk menampilkan kelas-kelas maupun paket-paket yang ada pada suatu sistem yang digunakaan. Terdapat 4 kelas beserta operatornya yaitu All_Term, Class, Term_Frequency dan Document_Training. Jadi diagram ini dapat memberikan sebuah gambaran mengenai sistem maupun relasi-relasi yang terdapat pada sistem tersebut



ISBN: 978-979—3649-96-2 *Unisbank Semarang, 28 Juli 2016*

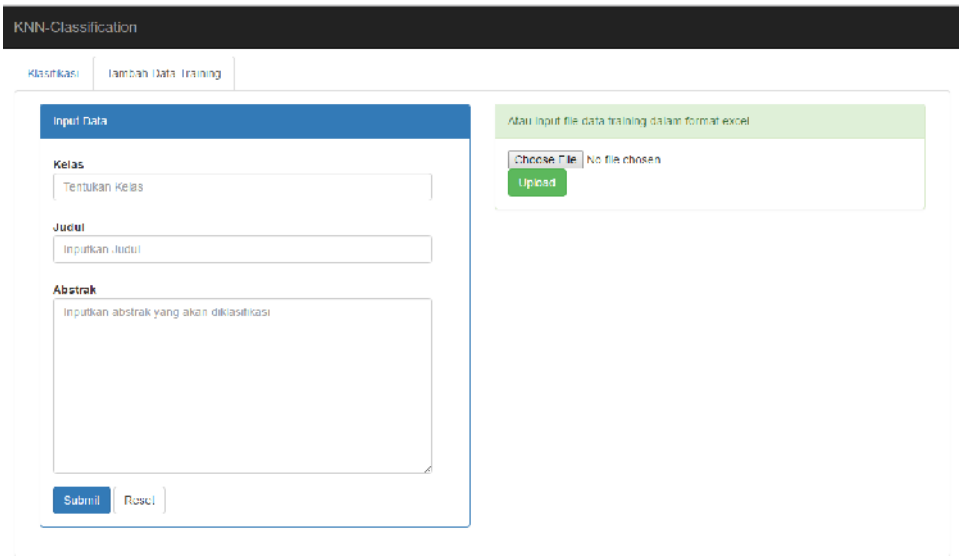
Gambar 4 *Class diagram*

d. Tahap Implementasi

Implementasi dilakukan dengan mengaplikasikan halaman aplikasi ke dalam bahasa pemrograman PHP menggunakan database MySQL. Dan proses pengujian dilakukan untuk mengetahui kemungkinan ternyadinya kesalahan pada skrip.

4. HASIL DAN PEMBAHASAN

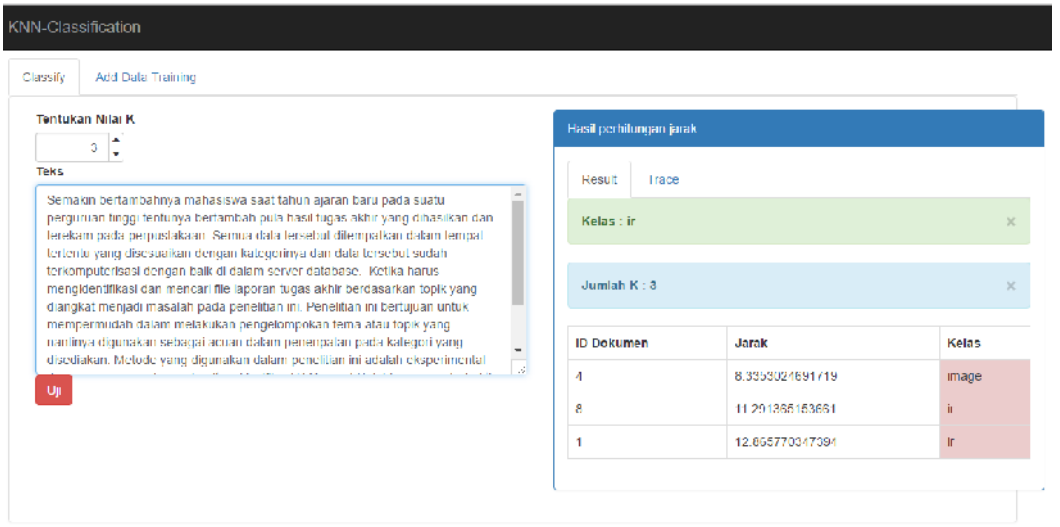
Realisasi yang dihasilkan dari penelitian ini adalah adanya penambahan didalam sistem pada gambar 3 menampilkan halaman untuk menambahkan data trainaing yang digunakan sebagai data acauan pada saat melakukan testing. Masukkan dapat berupa teks dan juga mampu untuk mengambil dari file berformat .xls.



Gambar 3 Halaman inputan data training

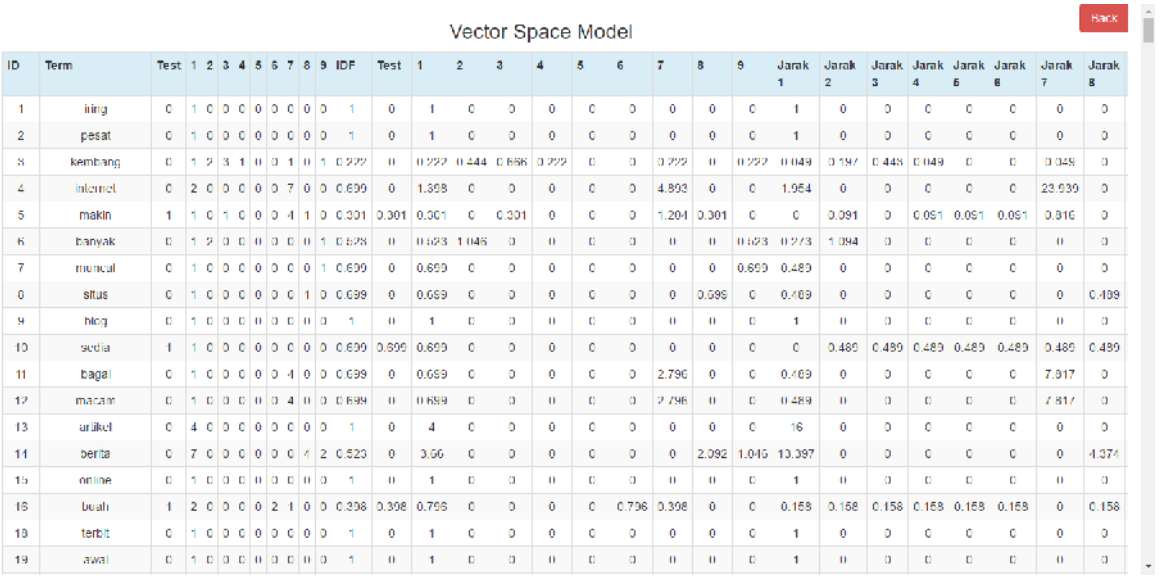
Hasil pengujian ditampilkan pada gambar 4 dengan memasukan data abstraksi sebagai data uji. Dari data tersebut dapat diketahui data abstraksi tersebut masuk pada kelas IR dengan menentukan k=3 berdasarkan dari pengurutaan dari jarak terdekat menggunakan algoritma *K-Nearest Neighbor*.

ISBN: 978-979—3649-96-2 *Unisbank Semarang, 28 Juli 2016*



Gambar 4. Hasil pengujian

Selain menampilkan hasil, dapat pula mengetahui proses vektor space model yang terjadi pada saat pemrosesan atau perhitungan TF-IDF *(Term Frequency – Inverse Document Frequency)* hingga menghitung kemiripan vektor dokumen *testing* dengan dokumen yang telah terklasifikasi



ID	Term	Test	1	2	3	4	5	6	7	8	9	IDF	Test	1	2	3	4	5	6	7	8	9	Jarak 1	Jarak 2	Jarak 3	Jarak 4	Jarak 5	Jarak 6	Jarak 7	Jarak 8
1	iring	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
2	pesat	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
3	kembang	0	1	2	3	1	0	0	1	0	1	0.222	0	0.222	0.444	0.666	0.222	0	0	0.222	0	0.222	0.049	0.197	0.443	0.049	0	0	0.049	0
4	internet	0	2	0	0	0	0	0	7	0	0	0.699	0	1.398	0	0	0	0	0	4.893	0	0	1.954	0	0	0	0	0	23.939	0
5	makin	1	1	0	1	0	0	0	4	1	0	0.301	0.301	0.301	0	0.301	0	0	0	1.204	0.301	0	0	0.091	0	0.091	0.091	0.091	0.916	0
6	banyak	0	1	2	0	0	0	0	0	0	1	0.523	0	0.523	1.046	0	0	0	0	0	0	0.523	0.273	1.094	0	0	0	0	0	0
7	musical	0	1	0	0	0	0	0	0	0	1	0.699	0	0.699	0	0	0	0	0	0	0	0.699	0.489	0	0	0	0	0	0	0
8	situs	0	1	0	0	0	0	0	0	1	0	0.699	0	0.699	0	0	0	0	0	0	0.699	0	0.489	0	0	0	0	0	0	0.489
9	blog	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
10	sedia	1	1	0	0	0	0	0	0	0	0	0.699	0.699	0.699	0	0	0	0	0	0	0	0	0	0.489	0.489	0.489	0.489	0.489	0.489	0.489
11	bagai	0	1	0	0	0	0	0	4	0	0	0.699	0	0.699	0	0	0	0	0	2.796	0	0	0.489	0	0	0	0	0	7.817	0
12	macam	0	1	0	0	0	0	0	4	0	0	0.699	0	0.699	0	0	0	0	0	2.796	0	0	0.489	0	0	0	0	0	7.817	0
13	artikel	0	4	0	0	0	0	0	0	0	0	1	0	4	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0
14	berita	0	7	0	0	0	0	0	0	4	2	0.523	0	3.66	0	0	0	0	0	0	2.092	1.046	13.397	0	0	0	0	0	0	4.374
15	online	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
16	buah	1	2	0	0	0	0	2	1	0	0	0.398	0.398	0.796	0	0	0	0	0.796	0.398	0	0	0.158	0.158	0.158	0.158	0.158	0.158	0	0.158
17	terbit	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
18	awal	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0

Gambar 5. *Trace vektor space model*

ISBN: 978-979—3649-96-2

5. KESIMPULAN

Aplikasi ini sudah mampu melakukan proses klasifikasi data abstrak tugas akhir untuk menentukan kelasnya dengan baik.Akan tetapi proses klasifikasi semakin akurat jika data training atau latihnya yang digunakan dalam pembelajaran berjumlah semakin banyak. Aplikasi yang dikembangkan untuk penentuan tema tugas akhir pada perpustakaan dapat menjadi solusi bagi operator pada perpustakaan dalam pendataan.

DAFTAR PUSTAKA

Aggarwal, C. & Zhai, C., (2012). Mining Text Data - Aggarwal-Zhai.pdf. Mining Text Data, 4(2(63)), pp.889–903. Available at: http://www.springerlink.com/index/10.1007/978-1-4614-3223-4\nhttp://books.google.com/books?hl=en&lr=&id=vFHOx8wfSU0C&oi=fnd&pg=PR5&dq=Mining+Text+Data&ots=oaadUEiCTx&sig=WWPHKZpR-2ehG_cnB72GdOPemVk\nhttp://books.google.com/books?hl=en&lr=&id=vFHOx8wfSU0C&.

Arifin, A.D., Arieshanti, I. & Arifin, A.Z., (2012). Implementasi Algoritma K-Nearest Neighbor Yang Berdasarkan One Pass Clustering Untuk Kategorisasi Teks. , pp.1–7.

Darujati, C. et al., (2012). Pemanfaatan Teknik Supervised Untuk. Jurnal Link, 16(1), pp.1–8.

Februariyanti, H., (2012). Klasifikasi Dokumen Berita Teks Bahasa Indonesia menggunakan Ontologi. Jurnal Teknologi Informasi Dinamik, 17(1), pp.14–23.

Hamzah, A., (2012). Klasifikasi teks dengan naïve bayes classifier (nbc) untuk pengelompokan teks berita dan abstract akademis. Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Periode III, (2011), pp.269–277.

Krisandi, N., Helmi & Prihandono, B., (2013). ALGORITMA k-NEAREST NEIGHBOR DALAM KLASIFIKASI DATA HASIL PRODUKSI KELAPA SAWIT PADA PT. MINAMAS KECAMATAN PARINDU. Buletin Ilmiah Math. Stat. dan Terapannya (Bimaster), 02(1), pp.33–38.

Pressman, R. S. (2014). Software Engineering a Practitioner's Approach. McGraw-Hill Education.

Santoso, D., Ratnawati, D.E. & Indriati, (2014). Perbandingan Kinerja Metode Naive Bayes, K-Nearest Neighbor, dan Metode Gabungan K-Means dan LVQ dalam Pengkategorian Buku Komputer Berbahasa Indonesia berdasarkan Judul dan Sinopsis. Repositori Jurnal Mahasiswa PTIIK UB, 4(9).

Toker, G. & Kırmemiş, Ö., TEXT CATEGORIZATION USING k-NEAREST NEIGHBOR CLASSIFICATION.

Sommerville, I. (2015). Sofeware Engineering (10th ed.). Addison-Wesley.

Yan, X. et al., (2013). Weighted K-Nearest Neighbor Classification Algorithm Based on Genetic Algorithm. TELKOMNIKA, 11(10), pp.6173–6178.

Yao, M. & Vocational, B., (2014). Research on Learning Evidence Improvement for k NN B ased Classification Algorithm. International Jurnal of Database Theory and Application, 7(1), pp.103–110.

Zainal, A. & Novan, A.,(2012). Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia dengan Algoritma Single Pass Clustering. Online.

ISBN: 978-979—3649-96-2 *Unisbank Semarang, 28 Juli 2016*