TEXT MINING DALAM PENENTUAN KLASIFIKASI DOKUMEN SKRIPSI DI PRODI TEKNIK INFORMATIKA FAKULTAS ILMU KOMPUTER BERBASIS WEB

Teuku Muhammad Johan dan Riyadhul Fajri
Program Studi Teknik Informatika Fakultas Ilmu Komputer Universitas Almuslim

ABSTRAK

Plagiarisme dalam penulisan karya ilmiah adalah perilaku curang yang sangat merugikan mahasiswa di sebuah perguruan tinggi. Skripsi adalah sebuah karya ilmiah yang sering menjadi objek plagiat dari kalangan mahasiswa. Banyaknya kasus plagiat dikalangan mahasiswa sudah menjadi rahasia umum sehingga untuk menghindari hal tersebut maka perlunya dilakukan identifikasi kemiripan naskah dokumen skripsi. Dibutuhkan sebuah system yang dapat mendeteksi tingkat kemiripan judul skripsi. Algoritma K-Nearest Neighbor yang digunakan dalam klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek data training yang telah dilatih untuk menghasilkan nilai kemiripan naskah dokumen skripsi. Algoritma text mining dapat digunakan dalam melakukan pendeteksian naskah dokumen skripsi yaitu dengan mencari nilai cosine similarity. Melalui sistem ini, diharapkan mahasiswa prodi Teknik Informatika dan univesitas Almuslim dapat melihat tingkat persentase kesamaan naskah document skripsi dengan document yang telah ada sehingga dapat menghindari plagiasi. Sistem ini diharapkan mampu mengidentifikasi dan menglasifikasikan kemiripan naskah skripsi yang satu dengan yang lainnya dan disertai nilai kemiripan judul berdasarkan bobot serta akan memberikan informasi mengenai daftar judul skripsi yang telah ada. Dalam penerapan text mining dalam mengklasifikasi dokumen naskah skripsi sesuai dengan tingkat kemiripan judul dan studi kasus. Hal ini dilakukan sebelum naskah skripsi tersebut dipublikasikan atau disidangkan sehingga dapat meminimalisir tingkat kecurangan mahasiswa dalam menulis karya ilmiah. Hal ini dengan dilihat dari tingkat presentase kesamaan judul antara satu mahasiswa dengan mahasiswa yang lain. Tujuan penelitian ini adalah agar memudahkan pihak prodi, fakultas dan universitas dalam melihat kesamaan tingkat document skripsi berbasis web dan dari pihak mahasiswa dapat melihat presesntase nilai kemiripan dengan naskah documen skripsi yang telah ada. Sehingga semua dokumen skripsi jurusan informatika unimal dan skripsi yang ada di universitas almuslim terhindar dari tindak plagiarisme.

Kata kunci: Skripsi, Identifikasi, Cosine Similarity, K-Nearest Neighbor

PENDAHULUAN

Karya ilmiah yaitu karya tulis yang telah diakui dalam bidang ilmu pengetahuan, teknologi atau seni. Pada dasarnya karya ilmiah merupakan perwujudan kegiatan ilmiah yang dikomunikasikan lewat bahasa tulisan. Penulisan karya ilmiah juga menjadi salah satu kegiatan pokok di perguruan tinggi. Karya ilmiah ditulis sesuai dengan tata cara ilmiah dan mengikuti pedoman atau konvensi ilmiah yang telah disepakati atau ditetapkan oleh suatu lembaga pendidikan tinggi. Skripsi merupakan salah satu syarat kelulusan bagi mahasiswa/i strata satu. Plagiat dan tindak kecurangan dalam pengajuan judul skripsi masih kerap terjadi dan menjadi fenomena umum di dalam dunia pendidikan. Hal ini terjadi karena tidak adanya suatu sistem yang menyediakan informasi mengenai daftar judul-judul yang telah diambil serta memberikan persentase kemiripan terhadap judul yang telah ada.

Berbagai metode telah diteliti untuk mencari cara terbaik, waktu tercepat maupun tingkat kemiripan yang paling tepat pada dokumen teks seperti karya ilmiah. Maka dari permasalahan tersebut penulis akan meneliti salah satu metode yang dapat diterapkan dalam menghitung tingkat kemiripan untuk mengetahui kemiripan dokumen karya ilmiah denganmetode *K-Nearest Neighbor* yaitu sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut.

Untuk mengatasi masalah tersebut maka dibutuhkan suatu sistem dalam bidang *text mining* berupa sistem yang mampu mendeteksi persentase kemiripan judul sehingga dapat

menyediakan informasi tersebut kepada seluruh mahasiswa. Salah satu metode yang dapat diterapkan dalam melakukan klasifikasi teks untuk mengetahui kemiripan suatu dokumen teks adalah menggunakan *cosine simirarity*, yaitu sebuah metode *K-Nearest Neighbor* untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut.

Penelitian ini memberikan alternatif bagi para mahasiswa dalam mengidentifikasi kemiripan naskah documen skripsi dan membandingkan dengan naskah yang sudah ada yang telah ada melalui beberapa tahapan dalam teks mining yaitu *tokenizing, filtering(stoplist dan wordlist).* Melalui sistem ini, pihak prodi dan universitas hanya perlu memasukkan naskah document skripsi yang akan diajukan ke formulir yang telah disediakan, kemudian sistem akan mengecek secara otomatis dan menampilkan hasilnya. Hasil tersebut bisa dijadikan sebagai pertimbangan dalam menentukan diterima atau ditolak judul tersebut.

METODE PENELITIAN

Tahapan-Tahapan Penelitian

Metode yang digunakan pada penelitian ini menggunakan Algoritma Text mining. langkah pertama adalah user sebagai mahasiswa melakukan input skripsi yang diajukan, kemudian diproses dalam case folding dan tekinizing yaitu Pada tahap tokenizing semua huruf dalam naskah skripsi akan diubah menjadi huruf kecil dan hanya “a” sampai huruf “z” yang diterima. Untuk menyaring kata-kata penting tersebut akan menggunakan daftar kata stoplist yang telah disimpan di tabel tb_stoplist, Pada penelitian ini untuk melakukan stemming menggunakan algoritma stemming porter dan daftar kata dasar yang telah disimpan pada tabel tb_katadasar. Selanjutnya mengindeks kata-kata tiap dokumen yang kemudian disimpan dalam kata hasil indeks. Pembobotan document menggunakan algoritma *nearest neighbor* proses perhitungan bobot hanya dilakukan ketika proses pengujian dokumen dijalankan. Namun nilai tf (*term frequency*) akan disimpan terlebih dahulu di tabel *tb_indeks* pada proses *training* dokumen sampel.

Lokasi Penelitian

Penelitian tentang Penerapan Text Mining Penentuan Klasifikasi naskah document skripsi di Universitas Almuslim, Fakultan Teknik Program Studi Teknik Informatika.

Peubah yang diamati/diukur

Berdasarkan peubah yang diamati/diukur adalah:

1. kata dalam judul yang telah diinput ke sistem akan melewati rangkaian *preprocessing* yaitu *tokenizing*, *filtering* dan *stemming* yang bertujuan untuk mendapatkan kata dasar dari setiap kata yang sebelumnya terdapat beberapa imbuhan.
2. Pembobotan judul pada sistem klasifikasi ini, proses perhitungan bobot hanya dilakukan ketika proses pengujian judul dijalankan. Untuk menghitung nilai tf (*term frequency*) setiap judul pada kode program Fungsi tersebut akan otomatis menghitung jumlah kata yang sama di dalam *array* dan membuatnya menjadi bentuk *array asosiasi* yang berisi kata dan jumlah frekuensinya.
3. Proses *training* document skripsi sampel yaitu melakukan *preprocessing* terhadap banyaknya judul sampel skripsi yang telah diambil dari Prodi Informatika Universitas Almuslim. Hasil dari *training* judul sampel ini berupa nilai tf (*term frequency*) atau frekuensi setiap kata dari masing-masing judulyang diinput dan disimpan pada tabel *tb_indeks*.

Model yang digunakan

Model yang digunakan pada penelitian ini Text Mining dengan menggunakan *cosine similarity* sedangkan dalam Penentuan Klasifikasi Dokumen Skripsi untuk pengujian dengan menggunakan model Algoritma *K-nearest neighbor* (k-NN atau KNN).

Rancangan Penelitian

Rancangan penelitian yang digunakan pada pengujian ini akan dilakukan pada naskah skripsi diluar sampel dan membandingkannya dengan banyaknya judul sampel yang telah di *preprocessing* terlebih dahulu sebelumnya ketika proses *training* judul sampel. Tujuan dari pengujian tersebut adalah agar dapat mengetahui judul skripsi yang diuji pada sistem tersebut dapat terklasifikasi pada kategori mana yang sesuai dengan data yang telah dilatih(*training*) sebelumnya.

Pemrosesan data dilakukan setelah sistem mendapatkan data-data masukan dari pengguna. Data-data tersebut diproses untuk mendapatkan hasil berupa persentase dan daftar judul-judul skripsi. Data tersebut yang akan digunakan dalam proses penentuan persentase kemiripan judul.

Teknik Pengumpulan Data

Studi Kepustakaan

Sebelum memulai penelitian yang dilakukan terlebih dahulu adalah studi kepustakaan mengenai referensi tentang algoritma *K-Nearest Neighbor*, *Similiarity*, Algoritma *tf/idf* dan teori pendukung lainnya. Setelah memperoleh referensi tersebut, kemudian merancang sistem untuk mengidentifikasi kemiripan karya ilmiah dengan menerapkan beberapa metode berdasarkan dari studi kepustakaan yang dilakukan tersebut.

Pengumpulan Data

Adapun sebelum membangun aplikasi untuk mengidentifikasi kemiripan karya ilmiah tersebut maka diperlukan beberapa dokumen karya ilmiah baik sampelataupun yang diuji. Pengumpulan data berupa dokumen tersebut dilakukan untuk menyiapkan juga menguji kemampuan aplikasi yang akan dibangun.

Analisa Data

1. Menganalisa masalah yang ditemukan pada proses penentuan judul skripsi dan mempelajari sistem, memahami permalsalahn yang ada. Sebelum mengambil tindakan akhir dalam pembuatan sistem.
2. Merancang aplikasi sistem pendeteksian kemiripan judul skripsi.
3. Perancangan dan implementasi dengan menggunakan alat bantu *Data Flow Diagram* dengan menggambarkan proses-proses yang ada pada sistem sehingga akan mempermudah dalam menyelesaikan program.
4. Analisa pembuatan program berbasis web menggunakan bahasa pemrograman PHP dan database MySQL.
5. Pengujian terhadap aplikasi pengujian terhadap program yang telah dibuat dengan melakukan beberapa tes terutama pada penerapan algoritma yang digunakan dan menganalisa keluaran yang dihasilkan untuk mendapatkan kesalahan sehingga kesalahan tersebut bisa diperbaiki.

HASIL DAN PEMBAHASAN

Analisa Sistem

Analisa sistem bertujuan untuk mengidentifikasi permasalahan yang ada pada sistem, dimana aplikasi yang dibangun meliputi lingkungan operasi, *user* dan elemen-elemen yang terkait. Analisa terhadap sistem diperlukan sebagai dasar untuk tahapan perancangan sistem, yaitu meliputi desain sistem, perancangan dan implementasi sistem.

Penelitian ini dirancang untuk dapat mencari dokumen yang sesuai dengan dokument yang dimasukkan oleh pengguna, selanjutnya sistem akan mendapat hasil presentasi dari nilai kemiripan berdasarkan bobot dengan nilai tertinggi hingga terendah. Dalam lingkungan uji coba telah disiapkan beberapa file dokumen dalam berbagai ukuran dan ekstensi yang berbeda untuk mengetahui apakah sistem dapat berjalan dengan baik atau tidak.

Dalam dari pengujian sistem yang telah disiapkan beberapa file dokumen dalam berbagai ukuran dan ekstensi yang berbeda untuk mengetahui apakah sistem dapat berjalan dengan baik atau tidak. File dokumen yang mempunyai ekstensi *.txt, *.pdf dan *.doc/*.docx dalam berbagai ukuran dengan menggunakan model sistem informasi dalam pendeteksian dengan sistem dan hasil secara umum.

Perancangan Sistem

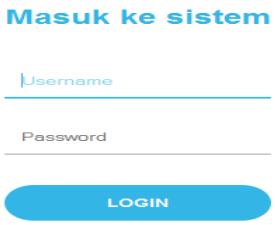
Perancangan Sistem (Desain Sistem) merupakan gambaran atau sketsa dari alur proses sistem pengolahan data. Rancangan suatu sistem dapat menggunakan Diagram Arus Data (DAD) atau *Data Flow Diagram* (DFD).

Diagram aliran data hanya memuat satu proses dan menunjukkan sistem secara keseluruhan. Adapun bentuk diagram konteks dari Text Mining dalam Penentuan Plagiarisme Klasifikasi Dokumen Skripsi di Prodi Teknik Informatika Fakultas Ilmu Komputer Berbasis Web

Implementasi Program Tahap Awal

Form Login

Adapun menu tampilan form login ini sebagai tampilan menu awal program yang berisi pendaftar, login dan keluar. Adapun tampilan form login dapat dilihat pada gambar berikut:



Gambar 1. Form Login

Form Daftar Dokumen Skripsi

Adapun tampilan form menu utama terdiri dari data home, daftar dokumen, klasifikasi dan bantuan. Berikut tampilan gambar sebagai berikut:

Gambar 2. Form Utama

Form Input Dokumen Skripsi

Form menu Penentuan Klasifikasi Dokumen Skripsi data yang akan dimasukkan ke dalam sistem berikut ini tampilan gambar nya:

Gambar 3. Penentuan Klasifikasi Dokumen Skripsi

Form Cari Daftar Dokumen Skripsi

Form menu Penentuan Klasifikasi Dokumen Skripsi data yang akan dimasukkan ke dalam sistem berikut ini tampilan gambar nya:

Gambar 4. Penentuan Klasifikasi Dokumen Skripsi

Form Pengujian Klasifikasi Dokumen Skripsi

Form menu Penentuan Klasifikasi Dokumen Skripsi data yang akan dimasukkan ke dalam sistem berikut ini tampilan gambar nya:

Gambar 5. Penentuan Klasifikasi Dokumen Skripsi

Form Pengujian Hasil Klasifikasi Dokumen Skripsi

Form menu Penentuan Klasifikasi Dokumen Skripsi data yang akan dimasukkan ke dalam sistem berikut ini tampilan gambar nya:

Gambar 6. Tampilan Hasil

PENUTUP

Simpulan

Dari hasil dan pembahasan Implementasi Text Mining dalam Penentuan Klasifikasi Dokumen Skripsi di Prodi Teknik Informatika Fakultas Ilmu Komputer Berbasis Web dapat mengambil kesimpulan adalah sebagai berikut:

1. Algoritma *k-nearest neighbor*yang diterapkan pada sistem identifikasi ini terbukti mampu mengidentifikasi dengan baik kemiripan dokumen karya ilmiah yang diuji dengan membandingkannya pada kumpulan dokumen sampel yang diinput dan di *training* terlebih dahulu.
2. Dengan adanya aplikasi ini proses identifikasi persentase kemiripan naskah dokumen skripsi menjadi lebih cepat dan akurat karena menggunakan text mining dengan moetode *Cosine Similarity* dan klasifikasinya menggunakan Algoritma *K-nearest neighbor* (k-NN atau KNN)
3. Dapat membantu pihak prodi, jurusan, perpustakaan dalam melihat kesamaan naskah dokumen skripsi di prodi Teknik Informatika fakultas ilmu komputer dan universitas Almuslim

Saran

Berdasarkan penelitian yang saat ini sedang berjalan, berikut adalah saran yang dapat disampaikan:

1. Implementasi Text Mining dalam Penentuan Klasifikasi Dokumen Skripsi di Prodi Teknik Informatika Fakultas Ilmu Komputer Berbasis Web, akan lebih baik sistem ini dicoba dengan menggunakan metode yang lain sehingga dapat diketahui kekurangan dan kelebihan dari masing-masing metode.

2. Perancangan berikutnya diharapkan dapat menyempurnakan bagian desain agar tampak lebih menarik.

DAFTAR PUSTAKA

Berry, M.W. & Kogan, J. 2010. *Text Mining Aplication and theory*. WILEY: United Kingdom.

Feldman, R & Sanger, J. 2007. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press: New York.

Dragut, E., Fang, F., Sistla, P., Yu, S. & Meng, W. 2009. *Stop Word and Related Problems in Web Interface Integration*. http://www.vldb.org/pvldb/2/vldb09-384.pdf. Diakses tanggal 10 Januari 2016.

T. Winarto Yunita, *Karya Tulis Ilmiah Sosial: Menyiapkan, Menulis, dan Mencermatinya*, Yayasan Obor Indonesia, Jakarta, 2004.

Hermawati, Fajar Astuti, *Data Mining*, Edisi I, Penerbit: Andi, Yogyakarta, 2013.

Kusrini, Andi Koniyo, 2007. *Tuntunan Praktis Membangun Sistem Informasi Akuntansi dengan Visual Basic dan Microsoft SQL Server*.Yogyakarta: ANDI.

Nugroho A, 2011, *Perancangan dan Implementasi Sistem Basis Data*, Yogyakarta: Penerbit Andi

Sarno, R., dkk, 2012, "*Semantic Search*", Yogyakarta. Andi.

Susanto, B., 2011, *Text Mining*, Teknik Informatika UKDW Yogyakarta.

Silberschatz, 2002, *Database System Concept*, Fourth Edition, McGraw-Hill Inc., New York

Tala, F.Z., 2003, "*A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*", Master of Logic Project, Institute for Logic, Language and Computation, Universiteit van Amsterdam.