KOMPARASI ALGORITMA WINNOWING DAN JARO - WINGKLER MENDETEKSI KEMIRIPAN TUGAS MAHASISWA

Ida Bagus Ketut Surya Arnawa
Program Studi Sistem Informasi
Institut Teknologi dan Bisnis STIKOM Bali
arnawa@stikom-bali.ac.id

ABSTRACT

Education is a conscious effort made by humans in order to develop their potential through the learning process. In the learning process, especially in tertiary institutions, students get assignments from lecturers, which is something that usually happens. Giving assignments aims to train responsibility and build student initiative and creativity as well as measure the level of understanding of the material presented. When correcting lecturers often find plagiarism actions on student assignments, especially assignments in essay form which require students to answer according to the student's own opinion. Plagiarism acts This is a bad act and should be avoided. In an effort to overcome this act of plagiarism, especially in correcting student assignments, a lecturer finds many obstacles, namely the difficulty in measuring the level of plagiarism of one student's work with one another considering the number of student assignments that must be corrected manually. In Text Mining there are techniques that allow for matching. between documents, including Jaro-Winkler, Winnowing and so on. Each algorithm has advantages and disadvantages. From the problems that have been described, the authors compare the performance of the Jaro-Winkler and Winnowing algorithms. The result of this research is that the winnoing algorithm has better performance than the Jaro-Winkler algorithm.

Keywords*: comparison, jaro-winkler , winnowing, manber.*

ABSTRAK

Pendidikan merupakan usaha sadar yang dilakukan oleh manusia agar dapat mengembangkan potensi dirinya melalui proses pembelajaran. Dalam proses pembelajaran khususnya di perguruan tinggi mahasiswa mendapatkan tugas dari dosen merupakan sesuatu yang lazim terjadi. Pemberian tugas bertujuan untuk melatih tanggung jawab dan membangun inisiatif serta kreatifitas mahasiswa erta mengukur tingkat pemahaman terhadap materi yang disampaikan.Saat mengkoreksi dosen sering menemukan tindakan plagiarisme pada tugas mahasiswa terutama tugas dalam bentuk essay yang mewajibkan mahasiswa untuk menjawab sesuai pendapat mahasiswa itu sendiri.Tindakan plagiarisme merupakan tindakan yang tidak baik dan patut untuk dihindari diatasi. Dalam upaya mengatasi tindakan plagiarisme ini khususnya dalam mengkoreksi tugas mahasiswa seorang dosen banyak menemukan kendala yaitu kesulitan dalam mengukur tingkat plagiarisme tugas mahasiswa yang satu dengan yang lainnya mengingat banyaknya tugas mahasiswa yang harus dikoreksi dengan cara manual.Dalam Text Mining terdapat teknik yang memungkinkan untuk melakukan pencocokan antar dokumen yaitu diantaranya Jaro-Winkler, Winnowing dan lain sebagainya. Setiap algoritma memiliki kekurangan dan kelebihan. Dari permasalah yang telah dijabarkan penulis membandingkan unjuk kerja algoritma Jaro-Winkler dan Winnowing. Hasil dari penelitian ini yaitu algoritma winnoing memiliki unjuk kerja yang lebih baik dari pada algoritma Jaro-Winkler.

Kata Kunci : komparasi, jaro-winkler, winnowing.

PENDAHULUAN

Pendidikan merupakan salah satu aspek yang penting untuk membangun pendidikan di Indonesia. Hakekat pendidikan yaitu memanusiakan manusia. Seluruh warga negara Indonesia berhak untuk mendapatkan pendidikan yang layak sesuai yang diamanatkan dalam UUD 1945 BAB XIII, Pasal 31 ayat (1).Pendidikan merupakan usaha sadar yang dilakukan oleh manusia agar dapat mengembangkan potensi dirinya melalui proses pembelajaran. Untuk dapat mengembangkan potensi diri yang lebih maksimal dapat ditempuh dengan melanjutkan pendidikan tinggi sampai perguruan tinggi. Dalam proses belajar mengajar khususnya di perguruan tinggi mahasiswa mendapatkan tugas dari dosen merupakan sesuatu yang lazim terjadi. Pemberian tugas bertujuan untuk melatih tanggung jawab dan membangun inisiatif serta kreatifitas mahasiswa. Disamping itu juga pemberian tugas juga dapat dijadikan salah satu parameter dalam mengukur tingkat pemahaman mahasiswa dalam menerima materi pembelajaran.

Dalam kasus mengkoreksi tugas mahasiswa sering kali seorang dosen menemukan ada kemiripan antar tugas mahasiswa satu dengan mahasiswa yang lainnya terutama tugas dalam bentuk essay yang mewajibkan mahasiswa untuk menjawab sesuai pendapat mahasiswa itu sendiri. Kemiripan tugas mahasiswa dapat terjadi baik secara sengaja maupun tidak sengaja. Namun terjadinya kemiripan tugas mahasiswa baik sengaja maupun tidak sengaja sudah termasuk tindakan plagiarisme. Tindakan plagiarisme merupakan tindakan yang tidak baik dan patut untuk dihindari. Plagiarisme merupakan pengambilan ide atau gagasan atau karya orang lain tanpa sepengetahuan atau seijin orang yang memiliki ide atau gagasan atau karya tersebut. Banyak faktor yang mendorong terjadinya tindakan plagiarisme ini khusus di lingkungan perguruan tinggi yaitu tingginya sikap arogansi dan kurang bertanggungjawabnya mahasiswa dalam mengerjakan tugas. Tindakan plagiarisme ini harus dicegah supaya tidak menjadi kebiasaan mahasiswa. Dalam upaya mengatasi tindakan plagiarisme ini khususnya dalam mengkoreksi tugas mahasiswa seorang dosen banyak menemukan kendala yaitu kesulitan dalam mengukur tingkat plagiarisme tugas mahasiswa yang satu dengan yang lainnya mengingat banyaknya tugas mahasiswa yang harus dikoreksi dengan cara manual.

Salah satu upaya yang dapat dilakukan untuk mengatasi serta meminimalisir tindakan plagiarisme khususnya dilingkungan Perguruan Tinggi yaitu dengan cara mencocokan satu dukumen tugas mahasiswa dengan mahasiswa lainnya dengan memanfaatkan teknik pencocokan string. Dengan teknik pencocokan string ini kita bisa mengetahui tingkat plagiarisme antar dokumen. Dalam Text Mining terdapat teknik yang memungkinkan untuk melakukan pencocokan antar dokumen. Banyak algoritma Text Mining yang dapat dimanfaatkan dalam mengukur tingkat plagiarisme suatu dokumen yaitu diantaranya Jaro-Winkler, Winnowing dan lain sebagainya. Algoritma yang terdapat dalam Text Mining masing-masing memiliki kekurangan dan kelebihan dalam melakukan pengukuran tingkat plagiarisme suatu dokumen. Merujuk dari permasalahan yang sudah dijabarkan penulis tertarik untuk melakukan analisis unjuk kerja algoritma Jaro-Winkler dengan Winnowing untuk mengetahui algoritma yang memiliki unjuk kerja yang lebih baik dalam mengukur tingkat plagiarisme suatu dokumen tugas mahasiswa [3].

TINJAUAN PUSTAKA

Algoritma Winnowing

Algoritma Winnowing merupakan salah satu algoritma dalam text mining yang dapat dimanfaatkan untuk menghitung tingkat kemiripan suatu dokumen dalam bentuk text. Dalam menentukan nilai hash, algoritma winnowing menggunakan Rolling hash, dimna nilai hash adalah nilai numerik yang terbentuk dari perhitungan ASCII. Adapun cara kerja dari algoritma winnowing yaitu sebagai berikut [1][5]:

1. Langkah Pertama

Menghilangkan atau menghapus karakter yang tidak relevan yaitu berupa spasi, tanda baca serta spesial karakter seperti !,@,#,$,%,^,&,*,(,),_,-?

Contoh : Latihan pemrograman c++

Akan dirubah menjadi

latihanpemrogramanc

2. Langkah Kedua

Penentuan rangkaian n-gram yaitu dengan cara membentuk rangkaian karakter sepanjang n dari hasil pembuangan karakter yang tidak relevan pada langkah pertama. Dari text diatas telah dibersihkan dengan ukuran k=5

latih atiha tihan ihanp hanpe anpem npemr
pemro emrog mrogr rogra ogram grama
raman amanc

3. Langkah Ketiga

Menghitung fungsi hash untuk setiap n-gram menggunakan rolling hash untuk menghitung nilai hash dalam algoritma winnowing. Rolling hash merupakan suatu teknik untuk mentransformasikan sebuah string menjadi nilai yang unik dengan panjang tertentu yang berfungsi sebagai penanda string tersebut. Fungsi hash H(c1...ck) didefinisikan sebagai berikut :

H(ck) = c1*b(k-1) + c2*b(k-2) + ... + ck*b(k-k)

Keterangan :

c = nilai ascii karakter
b = basis (bilangan prima)
k = banyak karakter

hasil rolling hash dari kalimat diatas yaitu

l725630 1588519 1851872 1688788
l666414 1581359 1773116 1788777
l638938 1762281 1830629 1777214
l672790 1812547 1578302

4. Langkah Keempat

Pembentukan window dari nilai hash dari window dengan ukuran 3 yaitu sebagai berikut :

{ 1725630 1588519 1851872 }
{ 1588519 1851872 1688788 }
{ 1851872 1688788 1666414 }
{ 1688788 1666414 1581359 }
{ 1666414 1581359 1773116 }
{ 1581359 1773116 1788777 }
{ 1773116 1788777 1638938 }
{ 1788777 1638938 1762281 }
{ 1638938 1762281 1830629 }
{ 1762281 1830629 1777214 }
{ 1830629 1777214 1672790 }
{ 1777214 1672790 1812547 }
{ 1672790 1812547 1578302 }

5. Langkah Kelima

Langkah terakhir yaitu memilih nilai terkecil dari setiap window untuk dijadikan fingerprint, hasil dari nilai fingerprintnya sebagai berikut:

[1588519,1] [1666414,4] [1581359,5] [1638938,8] [1762281,9] [1672790,12] [1578302,14]

Nilai fingerprint yang dibentuk dari algoritma winnowing digunakan untuk mengukur

prosentase kemiripan teks pada persamaan Jaccard Coeficient. Persamaan Jaccard Coefficient digunakan untuk menghitung kemiripan (similarity) dari kumpulan kata-kata yang telah dihitung nilai hash nya. Berikut ini rumus persamaan Jaccard Coefficient.

$$\text{Similarity} = \frac{\text{Jumlah\_fingerprint\_sama}}{\text{Total\_seluruh\_fingerprint}} \text{ X } 100$$

Algoritma Jaro-Winkler

Algoritma Jaro-Winkler merupakan salah satu algoritma dalam text mining yang mengukur kesamaan antara dua string dan termasuk varian dari Jaro Distance. Dua buah string dikatakan mirip jika nilai Jaro-Winkler semakin tinggi. Dalam Jaro-Winkler memiliki nilai normal yaitu 0 dan 1 dimana 0 artinya tidak ada kesamaan sedangkan 1 artinya ada kesamaan. Jaro Winkler memiliki tiga bagian dasar yaitu [2][4] :

1. Menghitung panjang string.
2. Menemukan jumlah karakter yang sama di dalam dua string.
3. Menemukan jumlah transposisi.

Adapun rumus yang digunakan dalam menghitung jarak (dj) antara dua string (s1,s2) pada algoritma Jaro Winkler yaitu

$$d_j = \frac{1}{3} + \left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m\text{-}t}{m} \right)$$

Dimana :

M = jumlah karakter yang sama persis

|s1| = panjang string1

|s2| = panjang string 2

t= jumlah transposisi

Jarak teoritis dua buah karakter yang disamakan dapat dibenarkan jika tidak melebihi :

$$\left(\frac{max\ (|s_1|,|s_2|)}{2} \right) \text{-} 1$$

Algoritma Jaro-Winkler memakai prefix scale (p) dan prefix length (l), dimana prefix scale (p) memberikan tingkat pengukuran yang lebih sedangkan prefix length (l) merupakan panjang karakter yang sama sampai ditemukan ketidakmiripan. Jika dua buah string dibandingkan s1 dan s2, maka Jaro Winkler distance (dw) menggunakan rumus berikut :

$$d_w = d_j + (lp(1\text{-}d_w))$$

Dimana :

dj = Jaro distance untuk string s1 dan s2

l = panjang prefiks umum di awal string

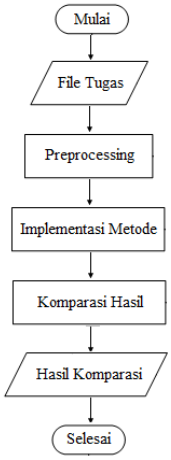
p = konstanta scaling factor.

Nilai standar untuk konstanta ini menurut Winkler adalah p = 0.1.

METODE PENELITIAN

Arsitektur Sistem

Arsitektur sistem yang dirancang untuk mengukur tingkat kemiripan dokumen tugas mahasiswa. Inputan yang diperlukan yaitu dokumen tugas mahasiswa yang berupa text. Sebelum diproses lebih lanjut dilakukan tahapan preprosessing yaitu case folding, tokenizing, filtering dan stemming. Setelah melalui proses preprocessing dilanjutkan dengan pengukuran kemiripan antar dokumen dengan algoritma winnowing dan jaro-winkler. Tahap selanjutkan dilakukan komparasi unjuk kerja kedua algoritma untuk mengetahui algoritma yang memiliki unjuk kerja yang lebih baik. Gambar 1 merupakan gambaran arsitektur sistem.



Gambar 1. Arsitektur Sistem

HASIL DAN PEMBAHASAN

Implementasi Perbandingan Algoritma

Implementasi dari algoritma *Winnowing* dan *Jaro-Winkler* dalam mendeteksi kemiripan tugas mahasiswa. Sample tugas mahasiswa yang dijadikan percobaan yaitu sebagai berikut :

Tabel 1. Sample Pengujian

Percobaan 1. Tugas Mahasiswa dengan tingkat kemiripan 100 %	
Dokumen Asli	**Dokumen Uji**
Bahasa pemrograman, atau sering diistilahkan juga dengan bahasa komputer atau bahasa pemrograman komputer, adalah instruksi standar untuk memerintah komputer. Bahasa pemrograman ini merupakan suatu himpunan dari aturan sintaks dan semantik yang dipakai untuk mendefinisikan program komputer.	Bahasa pemrograman, atau sering diistilahkan juga dengan bahasa komputer atau bahasa pemrograman komputer, adalah instruksi standar untuk memerintah komputer. Bahasa pemrograman ini merupakan suatu himpunan dari aturan sintaks dan semantik yang dipakai untuk mendefinisikan program komputer.
Percobaan 2. Tugas Mahasiswa dengan tingkat kemiripan 50 %	
Dokumen Asli	**Dokumen Uji**
Bahasa pemrograman, atau sering diistilahkan juga dengan bahasa komputer atau bahasa pemrograman komputer, adalah instruksi standar untuk memerintah komputer. Beberapa contoh Bahasa pemrograman yaitu C++, php, javascript, java, visual basic. Setiap Bahasa pemrograman memiliki aturan yang berbeda-beda dan memiliki kelebihan dan kekurangan masing-masing.	Bahasa pemrograman, atau sering diistilahkan juga dengan bahasa komputer atau bahasa pemrograman komputer, adalah instruksi standar untuk memerintah komputer. Untuk menjadi seorang programmer diwajibkan menguasai dasar-dasar pemrograman diantaranya struktur penulisan dan struktur logika. Selain itu juga paham dengan variabel, konstanta, perulangan, percabangan.
Percobaan 3. Tugas Mahasiswa dengan tingkat kemiripan 30 %	
Dokumen Asli	**Dokumen Uji**
Bahasa pemrograman, atau sering diistilahkan juga dengan bahasa komputer atau bahasa pemrograman komputer, adalah instruksi standar untuk memerintah komputer. Beberapa contoh Bahasa pemrograman yaitu C++, php, javascript, java, visual basic. Setiap Bahasa pemrograman memiliki aturan yang berbeda-beda dan memiliki kelebihan dan kekurangan masing-masing.	Bahasa pemrograman, atau sering diistilahkan juga dengan bahasa komputer. Untuk menjadi seorang programmer diwajibkan menguasai dasar-dasar pemrograman diantaranya struktur penulisan dan struktur logika. Selain itu juga paham dengan variabel, konstanta, perulangan, percabangan. Jika sudah menguasai hal tersebut untuk belajar Bahasa yang lain akan lebih mudah

1. Hasil window dan kemiripan terbaik algoritma winnowing

Tabel 2. Percobaan 1. Tugas Mahasiswa dengan tingkat kemiripan 100 %

uji coba	n-gram (n)	window (w)	kemiripan (%)
1	2	3	100
2	3	3	100
3	4	3	100
4	5	3	100
5	6	3	100
6	7	5	100
7	8	5	100
8	9	5	100
9	10	5	100

Tabel 3. Percobaan 2. Tugas Mahasiswa dengan tingkat kemiripan 50 %

uji coba	n-gram (n)	window (w)	kemiripan (%)
1	2	3	48,0
2	3	3	32,2
3	4	3	26,5
4	5	3	26,1
5	6	3	26,8
6	7	5	27,2
7	8	5	27,0
8	9	5	27,6
9	10	5	26,4

Tabel 4. Percobaan 3. Tugas Mahasiswa dengan tingkat kemiripan 30 %

uji coba	n-gram (n)	window (w)	kemiripan (%)
1	2	3	45,4
2	3	3	28,2
3	4	3	15,2
4	5	3	12,8
5	6	3	12,0
6	7	5	10,0
7	8	5	9,90
8	9	5	8,69
9	10	5	7,69

2. Hasil kemiripan algoritma Jaro-Winkler

Tabel 5. Percobaan 1. Algoritma Jaro-Winkler

uji coba	Kemiripan(%)	Hasil (%)
1	100	100
2	50	54,0
3	30	42,0

3. Hasil Perbandingan kemiripan algoritma Winnowing dengan Jaro-Winkler

Tabel 6. Perbandingan Winnowing dan Jaro-Winkler

Tingkat kemiripan (%)	Hasil Winnowing (%)	Hasil JaroWinkler (%)
100	100	100
50	48,0	54,0
30	28,2	42,0

SIMPULAN

Berdasarkan hasil uji coba yang sudah dilakukan dalam mengukur kemiripan dokumen tugas mahasiswa menggunakan algoritma winnowing dan jaro-winkler, maka dapat disimpulkan sebagai berikut :

1. Algoritma winnowing memiliki unjuk kerja yang lebih baik dibandingkan dengan algoritma jaro-winkler.
2. Terdapat perbedaan mendasar dari kedua algoritma yaitu pada algoritma winnowing memiliki window dan k-gram sedangkan algoritma jaro-winkler tidak memiliki.

DAFTAR PUSTAKA

[1] Astutik, Sariyanti, et al. " Sistem Penilaian Esai Otomatis Pada E-learning Dengan Algoritma Winnowing." *Jurnal Informatika University Petra Kristian*, vol. 12, no. 2, Nov. 2014, pp. 47-52.

[2] Frando, J., Ruslianto, I., & Hidayati, R. (2019). Penerapan Jaro Winkler Distance dalam Aplikasi Pengoreksi Kesalahan Penulisan Bahasa Indonesia Berbasis Web. Coding Jurnal Komputer dan Aplikasi, 7(03).

[3] Rahardjo, Andi Bekto. "Penerapan Data Mining Untuk Mengklasifikasi Penerima dan Bukan Penerima Kartu Identitas Miskin (KIM) Kelurahan Sumurrejo Gunungpati dengan Metode Naive Bayes

Classifier." (2015): 1-8.
[4] Hakim, L. (2019). Penggunaan N-Gram dan Jaro Winkler Distance pada Aplikasi Kelas Daring untuk Deteksi Plagiat. Prosiding Semnastek.
[5] Setiawan, A. (2017). IMPLEMENTASI ALGORITMA WINNOWING UNTUK DETEKSI KEMIRIPAN JUDUL SKRIPSI STUDI KASUS STMIK BUDIDARMA. Majalah Ilmiah INTI (Informasi dan Teknologi Ilmiah), 12(1).