PENGUKURAN KEMIRIPAN DOKUMEN BERBAHASA INDONESIA MENGGUNAKAN METODE RABIN-KARP DAN LEVENSHTEIN DISTANCE

Diajukan Sebagai Syarat untuk Menyelesaikan

Pendidikan Program Strata-1

di Jurusan Teknik Informatika

Oleh :

DARA D KARNINDO

09121002053

JURUSAN TEKNIK INFORMATIKA

FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA

2019

LEMBAR PENGESAHAN TUGAS AKHIR

PENGUKURAN KEMIRIPAN DOKUMEN BERBAHASA INDONESIA MENGGUNAKAN ALGORITMA RABIN-KARP DAN LEVENSHTEIN DISTANCE

Oleh :

Dara D Karnindo
09121002053

Palembang, Juli 2019

Mengetahui,
Ketua Jurusan Teknik Informatika

Rifkie Primartha, M.T.
NIP. 197706012009121004

Pembimbing,

Novi Yusliani, M.T.
NIP. 198211082012122001

TANDA LULUS UJIAN SIDANG TUGAS AKHIR

Pada hari Selasa tanggal 30 Juli 2019 telah dilaksanakan ujian sidang komprehensif oleh Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Nama : Dara D Karnindo

NIM : 09121002038

Judul : Pengukuran Kemiripan Dokumen Bernahasa Indonesia Menggunakan Algoritm Rabin-Karp dan Levenshtein Distance

1. Ketua Penguji

Novi Yusliani, M.T.
NIP. 198211082012122001

.....

2. Penguji I

Alvi Syahrini Utami, M.Kom., Ph.D.
NIP. 197812222006042003

.....

3. Penguji II

Hardini Novianti, M.T.
NIP. 197911012014042002

.....

Mengetahui,
Ketua Jurusan Teknik Informatika

Rifkie Primartha, M.T.
NIP. 197706012009121004

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini :

Nama	:	Dara D Karnindo
NIM	:	09121002053
Program Studi	:	Teknik Informatika
Judul Skripsi	:	Pengukuran Kemiripan Dokumen Berbahasa Indonesia Menggunakan Algoritma Rabin-Karp dan Levenshtein Distance
Hasil Pengecekan Software *Turnitin*	:	19 %

Menyatakan bahwa Laporan Projek saya merupakan hasil karya sendiri dan bukan hasil penjiplakan/plagiat. Apabila ditemukan unsur penjiplakan/plagiat dalam laporan projek ini, maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya sesuai dengan ketentuan yang berlaku.

Demikian, pernyataan ini saya buat dengan sebenarnya dan tidak ada paksaan oleh siapapun.

Palembang, Juli 2019

Dara D Karnindo
NIM. 09121002053

Motto :

"Keep Fighting! Because if you fight, many doors open "

"To deal with big problem: Faith, Fight, Family, Friends and Fun"

Kupersembahkan karya tulis ini
Kepada :

- Kedua orang tua tercinta
- Kakak dan Adik Saya tercinta
- Mute dan Parjim Squad
- Almamaterku
- IF Reg 2012
- Teman-teman dan Sahabat

PENGUKURAN KEMIRIPAN DOKUMEN BERBAHASA INDONESIA MENGGUNAKAN RABIN-KARP DAN LEVENSHTEIN DISTANCE

Oleh:

Dara D Karnindo

09121002053

ABSTRAK

Pengukuran kemiripan menjadi sangat penting untuk menghindari perilaku plagiat. Terdapat beberapa metode yang digunakan untuk mendeteksi plagiarisme. Beberapa diantaranya adalah algoritma *Rabin-Karp* dan *Levenshtein Distance*. Pada penelitian ini, penulis mencoba menggabungkan algoritma *Rabin-Karp* dan *Levenshthein Distance* untuk pengukuran kemiripan dokumen berbahasa Indonesia. Algoritma *Rabin-Karp* merupakan algoritma pencarian pola pada teks menggunakan teknik hashing. Pencocokan string akan dilakukan dengan membandingkan nilai hash antara kedua dokumen menggunakan *Levenshtein Distance*. Pengujian dilakukan dengan membandingkan dokumen uji dan dokumen asli. Kombinasi nilai n-gram, base dan modulo juga diterapkan pada pengujian ini. Hasil pengujian menunjukan, penggabungan kedua metode menghasilkan persentase kemiripan yang cukup baik. Penerapan berbagai kombinasi nilai n-gram base dan modulo menghasilkan persentase kemiripan yang berbeda-beda.

Kata kunci: mengukur kemiripan, pencocokan string, bahasa Indonesia, *rabin-karp, levenshtein distance*,

Palembang, Juli 2019

Mengetahui,
Ketua Jurusan Teknik Informatika

Rifkie Primartha, M.T.
NIP. 197706012009121004

Pembimbing,

Novi Yusliani, M.T.
NIP. 198211082012122001

SIMILARITY MEASURE FOR INDONESIAN DOCUMENTS USING RABIN KARP ALGORITHM AND LEVENSHTEIN DISTANCE

Oleh:

Dara D Karnindo

09121002053

ABSTRACT

Similarity measure is very important to avoid plagiarism behavior. There are several methods and appproaches used to detect plagiarism. Some of them are Rabin-Karp algorithm and Levenshtein Distance. In this study, the author tried to combine the Rabin-Karp algorithm and Levenshthein Distance to measure the similarity of Indonesian documents. Rabin-Karp algorithm is a search algorithm that searhes pattern in text using hashing techniques. String matching will be done by comparing the hash value between the two documents using Levenshtein Distance. The trial software is done by comparing the modifief document from the original document. A combination of n-gram values, base and modulo is also applied to this test. From the test result, it is known that combining the two methods produces a pretty good percentage of similarity. The application of various combinations of base, n-gram and modulo also produces different percentages of similarity.

Keyword : similarity measure, string matching, Indonesian, rabin-karp, levenshtein distance

Palembang, Juli 2019

Mengetahui,
Ketua Jurusan Teknik Informatika

Rifkie Primartha, M.T.
NIP. 197706012009121004

Pembimbing,

Novi Yusliani, M.T.
NIP. 198211082012122001

KATA PENGANTAR

Alhamdulillah, puji dan syukur kehadirat Allah SWT atas segala nikmat, rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan Tugas Akhir yang berjudul "Pengukuran Kemripan Dokumen Berbahasa Indonesia Menggunakan Algoritma Rabin–Karp dan Levenshtein Distance". Tugas akhir ini disusun untuk memenuhi salah satu persyaratan kelulusan tingkat sarjana pada Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Penulis menyadari bahwa dalam penulisan Tugas Akhir ini tidak terlepas dari dukungan, doa, bantuan, pengarahan maupun bimbingan dari berbagai pihak. Untuk itu penulisa mengucapkan terima kasih setulus-tulusnya kepada :

1. Kedua orang tua saya dan kedua saudara saya yang selalu memberikan semangat dalam penulisan Tugas Akhir ini;
2. Pemerintah dan Universitas Sriwijaya yang telah memberikan saya kesempatan dan berbagai fasilitas dalam perkuliahan sehingga saya dapat menyelesaikan Tugas Akhir ini;
3. Bapak Jaidan, S.Pd., M.T. selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya;
4. Bapak Rifkie Primartha, M.T. selaku Pembimbing Akademik dan Ketua Jurusan Teknik Informatika.
5. Ibu Novi Yusliani, M.T.selaku Pembimbing Tugas Akhir atas bimbingan dan masukannya penulis dapat menyelesaikan Tugas Akhir ini dengan baik;

6. Ibu Alvi Syahrini Utami, M.Kom., Ph.D. selaku dosen penguji yang telah memberikan koreksi dan masukan untuk Tugas Akhir ini;
7. Ibu Hardini Novianti, M.T. selaku dosen penguji yang telah memberikan koreksi dan masukan untuk Tugas Akhir ini;
8. Segenap staf pengajar di Fakultas Ilmu Komputer Universitas Sriwijaya yang telah mengajar, membimbing dan memberikan pemahaman tentang ilmu computer;
9. Segenap karyawan Fakultas Ilmu Komputer Universitas Sriwijaya, terutama Kak Ricy Firnando selaku Admin Jurusan Teknik Informatika atas bantuannya selama ini;
10. Mute ,Wenty Octaviani, Putri Septria, Lisa Desta Sari dan Riza Gamal Fuad yang telah memberi semangat, membantu dan menemani saya dalam penulisan Tugas Akhir ini;
11. Teman – teman Parjim Squad: Dwi Erviana, Elbananda Permana, Putri Septria, Rahmi Fadhillah Busyra dan Wenty Octaviani yang selalu menunggu, mendukung dan mendoakan saya selama penulisan Tugas Akhir ini
12. Semua anggota Teknik Informatika Regular 2012 yang telah memberi semangat kepada saya dalam menyelesaikan Tugas Akhir ini;
13. Semua teman-teman dan sahabat yang telah memberikan semangat kepada saya dalam penulisan Tugas Akhir ini.

Akhir kata, penulis menyadari bahwa laporan tugas akhir ini masih jauh dari kesempurnaan karena keterbatasan ilmu yang dimiliki penulis. Oleh karena itu, penulis mengharapkan kritik dan saran yang membangun untuk kesempurnaan tugas akhir ini. Semoga tugas akhir yang sederhana ini dapat memberikan manfaat bagi yang membutuhkan.

Palembang, Juli 2019

Penulis

DAFTAR ISI

Halaman

DAFTAR TABEL

Halaman

DAFTAR GAMBAR

Halaman

BAB I

PENDAHULUAN

1.1 Pendahuluan

Pada bab ini membahas latar belakang masalah, rumusan masalah, tujuan, dan manfaat penelitian, serta batasan masalah. Bab ini akan memberikan penjelasan umum mengenai keseluruhan penelitian.

Pendahuluan dimulai dengan penjelasan mengenai tantangan dan tujuan proses menemukan pengetahuan baru pada deteksi kemiripan teks serta penelitian yang berkaitan dengan menerapkan algoritma *rabin-karp* dan *levenshtein distance* yang menjadi latar belakang dari penelitian ini.

1.2 Latar Belakang

Dalam dunia akademik, terutama dalam proses penulisan karya ilmiah atau skripsi, ditemukan banyak terjadi peniruan atau penjiplakan isi tulisan, baik dalam bentuk sejumlah kalimat, paragraf, bab, dan bahkan penjiplakan total tanpa ada yang diubah sama sekali, dan tanpa memberikan sumber identitas sang penulis asli. Peristiwa ini disebut sebagai plagiarisme. Kata plagiarisme berakar dari dua kata Latin, *plagiarius* yang berarti penculik, dan *plagiare* yang berarti mencuri. Plagiarisme dapat dikategorikan sebagai tindakan kriminal pelanggaran hak cipta karena dilakukan dengan cara menerbitkan karya orang lain dan menjadikannya seolah-olah karya milik sendiri (Joy & Luck, 1999).

Bull, et. al. (2001) pada penelitiannya mengkhawatirkan pengembangan yang pesat dalam penggunaan internet sebagai sarana belajar dan penelitian, akan meningkatkan praktik plagiat di berbagai bidang akibat dari kemudahan mengakses materi di internet. Masalah plagiarisme ini tumbuh menjadi sangat serius di kalangan akademis terutama universitas-universitas di Indonesia. Ini dibuktikan dengan banyaknya ditemukan plagiarisme pada *paper* atau proyek tugas akhir mahasiswa (Krisnawati & Schulz, 2014)

Untuk mengatasi hal tersebut perlu dilakukan pendeteksian terhadap dokumen-dokumen yang dicurigai sabagai dokumen plagiat. Deteksi plagiarisme dapat dilakukan secara manual dan otomatis. Deteksi secara manual dilakukan oleh manusia, namun memiliki kelemahan dari segi keefektifan jika dilakukan pada dokumen dalam jumlah yang sangat banyak (Clough, 2003). Oleh karena itu diperlukan suatu sistem yang dapat mendeteksi plagiarisme secara otomatis.

Beberapa metode pencocokkan string seperti *Levenshtein Distance* (LD) dan *Smith-Waterman* (SW) (Su et al., 2008), *Conceptual Similarity* (CS) dan *Graph-Based* (GB) (Osman, Salim, Binwahlan, Hentably, & Ali, 2011), *Rabin-Karp* (RK) (Firdaus, 2003) ((Salmuasih & Sunyoto, 2013) telah diusulkan untuk mendeteksi plagiarisme pada dokumen. Metode LD mempunyai kelebihan akurasi yang baik pada proses pencocokan string, yaitu dengan cara membandingkan rangkaian string secara keseluruhan, namun algoritma ini lemah pada efisiensi waktu jika diterapkan pada aplikasi deteksi yang berskala besar (berbasis web) (Su et al, 2008). Metode RK mempunyai kelebihan sangat baik untuk pencarian *string* dengan pola banyak dan

ukuran yang besar (Firdaus, 2003), dan tetap baik jika digunakan dalam aplikasi berbasis web (Mutiara & Agustina, 2008). Namun, r*abin-karp* hanya menghitung berdasarkan jumlah *hash* yang memiliki nilai yang sama pada kedua dokumen yang dibandingkan. Jumlah *identical hash* tentu saja akan berpengaruh pada persentasi kemiripan kedua dokumen. Karena itu, penulis mencoba mengkombinasikan *levenshtein distance* dan *rabin-karp* pada penelitian ini.

Levenshtein distance akan digunakan untuk menggantikan perhitungan kemiripan *hash* pada *rabin-karp*. Algoritma ini bekerja dengan menghitung jarak antara *string* satu dengan lainnya sehingga derajat kemirpan kedua dokumen dapat ditentukan berdasarkan bobot jaraknya. Dengan menggunakan *levenshtein distance* hasil perhitungan jarak *hash* antara kedua dokumen diharapkan mampu mempengaruhi dan menghasilkan tingkat kemiripan yang lebih baik.

1.3 Perumusan Masalah

Berdasarkan latar belakang masalah yang telah dijelaskan, rumusan masalah pada penelitian ini adalah:

1. Bagaimana mengukur kemiripan dari dua dokumen menggunakan algoritma *rabin-karp* dan *levenshtein distance*?
2. Bagaimana persentase kemiripan dokumen berbahasa indonesia yang dihasilkan menggunakan algoritma r*abin-karp dan levenshtein distance*?
3. Bagaimana pengaruh nilai n-gram, base dan modulo terhadap persentase kemiripan dokumen yang dihasilkan?

1.4 Tujuan Penelitian

Tujuan penelitian ini adalah sebagai berikut :

1. Menerapkan algoritma *rabin-karp* dan *levenshtein distance* ke dalam perangkat lunak untuk pengukuran kemiripan dokumen berbahasa Indonesia.
2. Mengetahui persentase kemiripan dokumen menggunakan algoritma r*abin-karp* dan l*evenshtein distance*.
3. Mengetahui pengaruh n-gram, base dan modulo terhadap persentase kemiripan yang dihasilkan .

1.5 Manfaat Penelitian

Manfaat penelitian ini adalah sebagai berikut :

1. Hasil penelitian ini dapat digunakan sebagai rujukan dalam penelitian yang terkait dengan deteksi plagiarisme;
2. Memahami Algoritma *Rabin Karp* dan *Levenshtein Distance* sebagai metode pada pendeteksian kemiripan dokumen berbahasa Indonesia

1.6 Batasan Masalah

Batasan masalah pada penelitian ini adalah sebagai berikut:

1. Data set yang digunakan adalah dokumen bahasa Indonesia.
2. Penelitian ini hanya mengukur persentase kemiripan dari dua dokumen.

3. Perangkat lunak hanya menerima dokumen berektensi*.txt
4. Data yang diuji hanya berupa teks.
5. Sistem tidak memperhatikan kesalahan ejaan ataupun penulisan pada dokumen.

1.7 Sistematika Penulisan

Sistematika penulisan skripsi ini adalah sebagai berikut:

BAB I. PENDAHULUAN

Bab I berisi uraian mengenai latar belakang, perumusan masalah, tujuan dan manfaat penelitian, batasan masalah, metodologi penelitian, model proses pengembangan perangkat lunak dan sistematika penulisan.

BAB II. KAJIAN LITERATUR

Bab II berisi landasan dasar teori yang digunakan dalam penelitian ini, seperti analisis *preprocessing*, algoritma *Rabin-Karp*, *Levenshtein Distance*, dan penelitian-penelitian lain yang relevan dengan penelitian ini.

BAB III. METODOLOGI PENELITIAN

Pada bab III ini akan membahas mengenai tahapan yang akan dilaksanakan pada penelitian ini. Pada tahapan penelitian ini akan dibahas dengan rinci dengan mengacu pada suatu kerangka kerja. Serta pada bagian akhir bab ini akan dijabarkan perancangan manajemen proyek perangkat lunak untuk pelaksanaan penelitian ini.

1.8 Kesimpulan

Pada peneltian ini metode yang digunakan untuk mengukur kemiripan teks bahasa Indonesia yaitu algoritma *rabin-karp* dan *levenshtein distance*. Pengujian akan dilakukan dengan melihat persentase hasil kemiripan teks dari sistem.Kemudian hasil pengujian akan dibandingkan dengan hasil kemiripan yang telah ditentukan secara manual.

DAFTAR PUSTAKA

Arifin, A. Z., Mahendra, I. P. A. K., & Ciptaningtyas, H. T. (2009). ENHANCED CONFIX STRIPPING STEMMER AND ANTS ALGORITHM FOR CLASSIFYING NEWS DOCUMENT IN INDONESIAN LANGUAGE. *The 5th International Conference on Information & Communication Technology and Systems*.

Baruah, D., & Kakoti Mahanta, A. (2013). A New Similarity Measure with Length Factor for Plagiarism Detection. *International Journal of Computer Applications*, *72*(14), 14–17. https://doi.org/10.5120/12561-8671

Clough, P. (2003). Old and new challenges in automatic plagiarism detection. *National Plagiarism Advisory Service*.

Firdaus, H. B. (2003). Deteksi Plagiat Dokumen Menggunakan Algoritma Rabin-Karp. *Jurnal Ilmu Komputer Dan Teknologi Informasi*.

Gusmita, R. H., Durachman, Y., Harun, S., Firmansyah, A. F., Sukmana, H. T., & Suhaimi, A. (2014). A rule-based question answering system on relevant documents of Indonesian Quran Translation. *2014 International Conference on Cyber and IT Service Management, CITSM 2014*. https://doi.org/10.1109/CITSM.2014.7042185

Haryanto, E. V. (2011). Rancang Bangun Prototype Mesin Pencari String Menggunakan Metode Fuzzy String Matching. *Konfrensi Nasional Sistem Dan Infromatika*, *1*(Pencarian Infromasi), 1–26.

Joy, M., & Luck, M. (1999). Plagiarism in programming assignments. *IEEE Transactions on Education*. https://doi.org/10.1109/13.762946

Krisnawati, L. D., & Schulz, K. U. (2014). *Plagiarism Detection for Indonesian Texts*. https://doi.org/10.1145/2539150.2539213

Martin, B. (1994). Plagiarism: A Misplaced Emphasis, by Brian Martin. *Information Ethics*.

Retrieved from https://www.uow.edu.au/~bmartin/pubs/94jie.html

Mutiara, A. B., & Agustina, S. (2008). *Anti Plagiarism Application with Algorithm Karp-Rabin at Thesis in Gunadarma University*. 3–10. Retrieved from http://arxiv.org/abs/0811.4349

Nugroho, E. (2011). Perancangan Sistem Deteksi Plagiarisme Dokumen Teks Dengan Menggunakan Algoritma Rabin-Karp. In *Journal of Strategic Studies*. https://doi.org/10.1080/01402390.2011.569130

Osman, A. H., Salim, N., Binwahlan, M. S., Hentably, H., & Ali, A. M. (2011). Conceptual similarity and graph-based method for plagiarism detection. *Journal of Theoretical and Applied Information Technology*.

Pressman, R. S., & Maxim, B. R. (2015). Software Engineering : A Practitioner's Approach, Eighth Edition. In *ACM SIGSOFT Software Engineering Notes*. https://doi.org/10.1145/1226816.1226822

Salmuasih, & Sunyoto, A. (2013). Implementasi Algoritma Rabin Karp untuk Pedeteksi Plagiat Dokumen Teks Menggunakan Konsep Similarity. *Seminar Nasional Aplikasi Teknologi Informasi (SNATi)*.

Stein, B., & zu Eissen, S. M. (2006). *Near Similarity Search and Plagiarism Analysis*. https://doi.org/10.1007/3-540-31314-1_52

Su, Z., Ahn, B. R., Eom, K. Y., Kang, M. K., Kim, J. P., & Kim, M. K. (2008). Plagiarism detection using the Levenshtein distance and Smith-Waterman algorithm. *3rd International Conference on Innovative Computing Information and Control, ICICIC'08*. https://doi.org/10.1109/ICICIC.2008.422