

***Spelling Checker Menggunakan Algoritma Damerau Levenshtein
Distance dan Cosine Similarity***

*Diajukan Sebagai Syarat Untuk Menyelesaikan
Pendidikan Program Strata-1 Pada
Jurusan Teknik Informatika*



Oleh:

Nur Hamidah
09021181520128

**JURUSAN TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA
2019**

LEMBAR PENGESAHAN SKRIPSI

SPELLING CHECKER MENGGUNAKAN ALGORITMA DAMERAU LEVENSHTTEIN DISTANCE DAN COSINE SIMILARITY

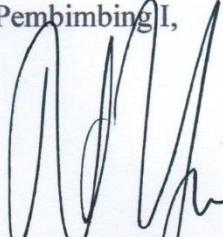
Oleh :

NUR HAMIDAH

NIIM : 09021181520128

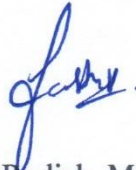
Palembang, Juli 2019

Pembimbing I,



Novi Yusliani, M.T.
NIP. 198211082012122001

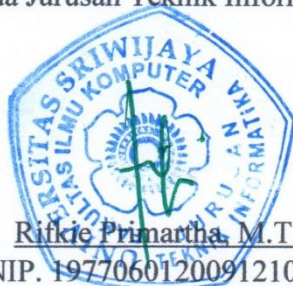
Pembimbing II,



Desty Rodiah, M.T.
NIK. 1671016112890005

Mengetahui,

Ketua Jurusan Teknik Informatika



Rifkie Primartha, M.T.
NIP. 197706012009121004

TANDA LULUS UJIAN SIDANG SKRIPSI

Pada hari Selasa tanggal 23 Juli 2019 telah dilaksanakan ujian sidang skripsi oleh
Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya

Nama : Nur Hamidah
NIM : 09021181520128
Judul : *Spelling Checker Menggunakan Algoritma Damerau Levenshtein
Distance dan Cosine Similarity*

1. Pembimbing I

Novi Yusliani, M.T.
NIP. 198211082012122001



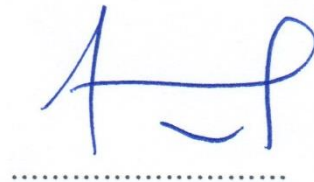
2. Pembimbing II

Desty Rodiah, M.T.
NIK. 1671016112890005



3. Penguji I

M. Fachrurrozi, M.T.
NIP. 198005222008121002



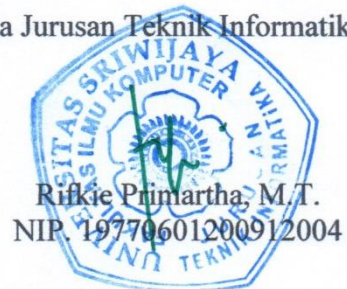
4. Penguji II

Rizki Kurniati, M.T.
NIP. 199107122019032016



Mengetahui,

Ketua Jurusan Teknik Informatika



HALAMAN PERNYATAAN

Yang bertanda tangan dibawah ini :

Nama : Nur Hamidah
NIM : 09021181520128
Program Studi : Teknik Informatika
Judul Skripsi : *Spelling Checker Menggunakan Algoritma Damerau Levenshtein Distance dan Cosine Similarity*

Hasil Pengecekan Software *iThenticate/Turnitin* : 11%

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



Nur Hamidah
NIM. 09021181520128

MOTTO DAN PERSEMBAHAN

Jangan berfikir apa yang saya dapat, tapi berfikirilah apa yang saya beri
~ J.F. Kennedy ~

You will never fall if you are afraid to climb, but there is no joy in
living your entire life on the ground

~ Anonim ~

“Sesungguhnya bersama kesulitan ada kemudahan. Maka apabila
engkau telah selesai (dari sesuatu urusan), tetaplah bekerja keras
(untuk urusan yang lain). Dan hanya kepada Tuhanmulah engkau
berharap”

~ QS : Al-Insyirah, ayat 6-8 ~

**Tak ada sidang yang lebih berat dan menakutkan melainkan sidang
dihadapan Allah Subhanahuwa ta'ala**

~ Nur Hamidah ~

Kupersembahkan karya tulis ini kepada :

- ♥ Allah Subhanahuwa ta'ala
- ♥ Kedua orang tua tercinta
- ♥ Kedua adik tercinta
- ♥ Keluarga besarku
- ♥ Sahabat dan teman seperjuangan
- ♥ Almamaterku

KATA PENGANTAR

Puji dan syukur atas kehadiran Allah ta'ala atas segala nikmat, rahmat, dan karuaniaNya sehingga penulis dapat menyelesaikan tugas akhir ini. Tugas akhir yang berjudul "*Spelling Checker Menggunakan Algoritma Damerau Levenshtein Distance dan Cosine Similarity*" disusun untuk memenuhi salah satu persyaratan kelulusan tingkat sarjana (S1) pada Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Pada kesempatan ini, penulis ingin mengucapkan terima kasih kepada pihak-pihak yang telah memberikan dukungan, bimbingan, kerjasama, dan motivasi kepada penulis dalam penyelesaian tugas akhir ini, diantaranya :

1. Bapak Sofyan dan Ibu Tina Harnita selaku orang tua penulis, Muhammad Agil Febriansyah dan Muhammad Aris Wijaya selaku saudara penulis serta Keluarga besar yang telah memberikan dukungan baik moril maupun materil kepada penulis
2. Bapak Jaidan Jauhari, M.T. selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya
3. Bapak Rifkie Primartha, M.T. selaku Ketua Jurusan Teknik Informatika Universitas Sriwijaya
4. Ibu Rifka Widyastuti S.Kom, M.Ti, M.Im. selaku pembimbing akademik
5. Ibu Novi Yusliani, M.T selaku pembimbing I Tugas Akhir dan ibu Desty Rodiah, M.T. selaku pembimbing II Tugas Akhir yang telah memberikan

pengarahan, bimbingan, bantuan, serta masukan kepada penulis sehingga Tugas Akhir ini dapat diselesaikan dengan baik

6. Bapak M. Fachrurrozi, M.T. selaku penguji I Tugas Akhir dan Ibu Rizki Kurniati, M.T. selaku penguji II Tugas Akhir yang telah memberikan banyak saran dan kritik serta arahan selama penulisan Tugas Akhir
7. Seluruh Bapak dan Ibu Dosen Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya yang telah memberikan ilmu pengetahuan kepada penulis selama perkuliahan
8. Seluruh karyawan dan karyawanati Fakultas Ilmu Komputer Universitas Sriwijaya yang telah membantu penulis dalam urusan administrasi selama kegiatan akademik
9. Sahabat Asrama Palembang Titi, Tika, Amel, Yulik, Okta, Inda, Vivi, Iga, dan Marina yang selalu memberikan asupan semangat dan doa kepada penulis
10. Mbak Fit dan Lupina, sahabat penulis dari awal perkuliahan yang selalu ada disampingku sebagai tempatku bersandar serta tempat berbagi cerita dan selalu memberi semangat, doa, dan dukungan kepada penulis
11. Teman satu bimbingan Tugas Akhir, dalam berbagi informasi dan pengetahuan kepada penulis
12. Teman seperjuangan penulis seluruh mahasiswa/mahasiswi Teknik Informatika Reguler Angkatan 2015
13. Kakak dan adik tingkat serta keluarga Mahasiswa Fasilkom Unsri
14. Seluruh pihak yang telah membantu dalam penyusunan dan penyempurnaan tugas akhir ini yang tidak dapat disebutkan satu persatu

Penulis menyadari bahwa tugas akhir ini jauh dari kata sempurna. Untuk itu, penulis mengharapkan kritik dan saran yang membangun untuk kesempurnaan tugas akhir ini di masa mendatang. Semoga tugas akhir ini dapat bermanfaat.

Palembang, Juli 2019

Penulis

DAFTAR ISI

	Halaman
HALAMAN JUDUL.....	i
LEMBAR PENGESAHAN SKRIPSI.....	ii
TANDA LULUS UJIAN SIDANG SKRIPSI.....	iii
HALAMAN PERNYATAAN.....	iv
MOTTO DAN PERSEMBAHAN.....	v
ABSTRACT.....	vi
ABSTRAK.....	vii
KATA PENGANTAR	viii
DAFTAR ISI.....	x
DAFTAR TABEL.....	xv
DAFTAR GAMBAR	xvii
DAFTAR LAMPIRAN.....	xx
BAB I PENDAHULUAN	
1.1 Pendahuluan	I-1
1.2 Latar Belakang.....	I-1
1.3 Rumusan Masalah.....	I-4
1.4 Tujuan Penelitian.....	I-4
1.5 Manfaat Penelitian.....	I-5
1.6 Batasan Masalah.....	I-5
1.7 Sistematika Penulisan.....	I-6
1.8 Kesimpulan.....	I-8

BAB II KAJIAN LITERATUR

2.1 Pendahuluan	II-1
2.2 Landasan Teori	
2.2.1 <i>Spelling Checker</i>	II-1
2.2.2 <i>Typographical Error</i>	II-2
2.2.3 <i>Preprocessing</i>	II-5
2.2.4 <i>Damerau Levenshtein Distance</i>	II-6
2.2.5 <i>N-gram</i>	II-7
2.2.6 <i>Term Frequency-Inverse Document Frequency (tf-idf)</i>	II-8
2.2.7 <i>Cosine Similarity</i>	II-9
2.2.8 <i>Mean Reciprocal Rank (MRR)</i>	II-10
2.2.9 <i>Rational Unified Process (RUP)</i>	II-11
2.3 Penelitian Lain yang Relevan.....	II-12
2.4 Kesimpulan.....	II-14

BAB III METODOLOGI PENELITIAN

3.1 Pendahuluan.....	III-1
3.2 Metode Pengumpulan Data.....	III-1
3.3 Tahapan Penelitian	
3.3.1 Menetapkan Kerangka Kerja/ <i>Framework</i>	III-2
3.3.2 Menetapkan Kriteria Pengujian.....	III-5
3.3.3 Menetapkan Format Data Pengujian.....	III-5
3.3.4 Menentukan Alat yang Digunakan dalam Pelaksanaan Penelitian	III-6
3.3.5 Melakukan Pengujian Penelitian.....	III-7
3.3.6 Analisis Hasil Pengujian dan Membuat Kesimpulan Penelitian.....	III-7

3.4 Metode Pengembangan Perangkat Lunak.....	III-8
3.4.1 Fase Insepsi.....	III-8
3.4.2 Fase Elaborasi.....	III-9
3.4.3 Fase Konstruksi.....	III-9
3.4.4 Fase Transisi.....	III-10
3.5 Manajemen Proyek Penelitian.....	III-10

BAB IV PENGEMBANGAN PERANGKAT LUNAK

4.1 Pendahuluan.....	IV-1
4.2 Fase Insepsi.....	IV-1
4.2.1 Pemodelan Bisnis.....	IV-1
4.2.2 Kebutuhan Sistem.....	IV-3
4.2.3 Analisis dan Desain.....	IV-3
4.2.3.1 Analisis Kebutuhan Perangkat Lunak.....	IV-4
4.2.3.2 Analisis Data.....	IV-5
4.2.3.3 Analisis Preprocessing.....	IV-6
4.2.3.4 Analisis Identifikasi Kesalahan.....	IV-7
4.2.3.5 Analisis Damerau Levenshtein Distance.....	IV-8
4.2.3.6 Analisis Bigram.....	IV-10
4.2.3.7 Analisis TF/IDF.....	IV-11
4.2.3.8 Analisis Cosine Similarity.....	IV-12
4.2.3.9 Analisis Mean Reciprocal Rank (MRR).....	IV-15
4.2.3.10 Desain Perangkat Lunak.....	IV-16
4.3 Fase Elaborasi.....	IV-22
4.3.1 Pemodelan Bisnis.....	IV-22
4.3.2 Perancangan Data.....	IV-23

4.3.3 Perancangan Antarmuka.....	IV-23
4.3.4 Kebutuhan Sistem.....	IV-24
4.3.5 Diagram Aktivitas (<i>Activity Diagram</i>).....	IV-24
4.3.6 Diagram Alur (<i>Sequence Diagram</i>).....	IV-27
4.4 Fase Konstruksi.....	IV-30
4.4.1 Kebutuhan Sistem.....	IV-30
4.4.2 Diagram Kelas (<i>Class Diagram</i>).....	IV-30
4.4.3 Implementasi.....	IV-32
4.4.3.1 Implementasi Kelas.....	IV-32
4.4.3.2 Implementasi Antarmuka.....	IV-35
4.5 Fase Transisi.....	IV-36
4.5.1 Pemodelan Bisnis.....	IV-36
4.5.2 Kebutuhan Sistem.....	IV-36
4.5.3 Rencana Pengujian.....	IV-37
4.5.3.1 Rencana Pengujian <i>Use Case</i> Melakukan Muat Data.....	IV-37
4.5.3.2 Rencana Pengujian <i>Use Case</i> Melakukan Pengecekan.....	IV-37
4.5.3.3 Rencana Pengujian <i>Use Case</i> Memilih Kandidat Kata.....	IV-38
4.5.3.4 Rencana Pengujian <i>Use Case</i> Melakukan Hitung MRR.....	IV-38
4.5.4 Implementasi.....	IV-39
4.5.4.1 Pengujian <i>Use Case</i> Melakukan Muat Data	IV-40
4.5.4.2 Pengujian <i>Use Case</i> Melakukan Pengecekan.....	IV-41
4.5.4.3 Pengujian <i>Use Case</i> Memilih Kandidat Kata	IV-42
4.5.4.4 Pengujian <i>Use Case</i> Melakukan Hitung MRR.....	IV-43

4.6 Kesimpulan.....	IV-44
---------------------	-------

BAB V HASIL DAN ANALISIS PENELITIAN

5.1 Pendahuluan.....	V-1
5.2 Data Hasil Penelitian.....	V-1
5.2.1 Konfigurasi Percobaan.....	V-1
5.2.2 Data Hasil Konfigurasi.....	V-2
5.3 Analisis Hasil Penelitian.....	V-14
5.4 Kesimpulan.....	V-16

BAB VI KESIMPULAN DAN SARAN

6.1 Kesimpulan.....	VI-1
6.2 Saran	VI-2

DAFTAR PUSTAKA

DAFTAR TABEL

	Halaman
Tabel II-1. Contoh dari Bentuk <i>Typographical Error</i>	II-3
Tabel II-2. Contoh Proses Pembentukan <i>Bigram</i>	II-8
Tabel III-1. Rancangan Tabel Hasil Pengujian.....	III-5
Tabel III-2. Rancangan Tabel Keseluruhan Hasil Pengujian Sistem.....	III-6
Tabel III-3. Penjadwalan Penelitian Dalam Bentuk <i>Work Breakdown Structure (WBS)</i>	III-11
Tabel IV-1. Kebutuhan Fungsional.....	IV-3
Tabel IV-2. Kebutuhan Non-Fungsional.....	IV-3
Tabel IV-3. Hasil Pemecahan <i>Bigram</i>	IV-10
Tabel IV-4. Hasil Perhitungan TF/IDF.....	IV-11
Tabel IV-5. Perhitungan Pembobotan <i>Term</i>	IV-12
Tabel IV-6. Hasil Perhitungan Panjang Vektor.....	IV-13
Tabel IV-7. Hasil Perangkingan Kandidat Kata.....	IV-15
Tabel IV-8. Contoh Perhitungan MRR.....	IV-15
Tabel IV-9. Definisi Aktor.....	IV-18
Tabel IV-10. Definisi <i>Use Case</i>	IV-18
Tabel IV-11. Skenario <i>Use Case</i> Muat Data.....	IV-19
Tabel IV-12. Skenario <i>Use Case</i> Pengecekan.....	IV-20
Tabel IV-13. Skenario <i>Use Case</i> Memilih Kandidat.....	IV-21

Tabel IV-14. Skenario <i>Use Case</i> Menghitung Nilai MRR.....	IV-21
Tabel IV-15. Rancangan Data.....	IV-23
Tabel IV-16. Implementasi Kelas dalam Bahasa Java.....	IV-32
Tabel IV-17. Rencana Pengujian <i>Use Case</i> Melakukan Muat Data.....	IV-33
Tabel IV-18. Rencana Pengujian <i>Use Case</i> Melakukan Pengecekan.....	IV-34
Tabel IV-19. Rencana Pengujian <i>Use Case</i> Memilih Kandidat Kata.....	IV-38
Tabel IV-20. Rencana Pengujian <i>Use Case</i> Melakukan Hitung MRR.....	IV-38
Tabel IV-21. Pengujian <i>Use Case</i> Melakukan Muat Data.....	IV-40
Tabel IV-22. Pengujian <i>Use Case</i> Melakukan Pengecekan.....	IV-41
Tabel IV-23. Pengujian <i>Use Case</i> Memilih Kandidat.....	IV-42
Tabel IV-24. Pengujian <i>Use Case</i> Hitung MRR.....	IV-43
Tabel V-1. Hasil Algoritma <i>Damerau Levenshtein Distance</i> (DLD).....	V-3
Tabel V-2. Hasil Pengurutan Kandidat Kata Berdasarkan Perhitungan <i>Cosine Similarity</i> Jenis Kesalahan <i>Deletion</i>	V-9
Tabel V-3. Hasil Pengurutan Kandidat Kata Berdasarkan Perhitungan <i>Cosine Similarity</i> Jenis Kesalahan <i>Insertion</i>	V-10
Tabel V-4. Hasil Pengurutan Kandidat Kata Berdasarkan Perhitungan <i>Cosine Similarity</i> Jenis Kesalahan <i>Substitution</i>	V-12
Tabel V-5. Hasil Pengurutan Kandidat Kata Berdasarkan Perhitungan <i>Cosine Similarity</i> Jenis Kesalahan <i>Transposition</i>	V-13
Tabel V-6. Hasil <i>Mean Reciprocal Rank</i> (MRR).....	V-15

DAFTAR GAMBAR

	Halaman
Gambar II-1. Ilustrasi Proses Penghapusan Huruf (<i>Deletion Error</i>).....	II-3
Gambar II-2. Ilustrasi Proses Penambahan Huruf (<i>Insertion Error</i>).....	II-4
Gambar II-3. Ilustrasi Proses Penggantian Huruf (<i>Substitution Error</i>).....	II-4
Gambar II-4. Ilustrasi Proses Penukaran Posisi Huruf (<i>Transposititon Error</i>)....	II-4
Gambar II-5. Ilustrasi Proses <i>Case Folding</i>	II-5
Gambar II-6. Ilustrasi Proses <i>Tokenizing</i>	II-5
Gambar II-7. <i>Pseudocode</i> Algoritma <i>Damerau Levenshtein Distance</i>	II-7
Gambar II-8. Arsitektur <i>Rational Unified Process</i> (RUP).....	II-11
Gambar III-1. Ilustrasi Kerangka Kerja Sistem.....	III-3
Gambar III-2. Penjadwalan Penelitian Tahap Menentukan Ruang Lingkup dan Unit Penelitian.....	III-15
Gambar III-3. Penjadwalan Penelitian Tahap Menentukan Dasar Teori yang Berkaitan dengan Penelitian.....	III-16
Gambar III-4. Penjadwalan Penelitian Tahap Menentukan Kriteria Pengujian.....	III-16
Gambar III-5. Penjadwalan Penelitian Tahap Menentukan Alat yang Digunakan untuk Pelaksanaan Penelitian Pada Fase Insepsi.....	III-17
Gambar III-6. Penjadwalan Penelitian Tahap Menentukan Alat yang	

	Digunakan untuk Pelaksanaan Penelitian Fase Elaborasi.....	III-18
Gambar III-7. Penjadwalan Penelitian Tahap Menentukan Alat yang		
	Digunakan untuk Pelaksanaan Penelitian Fase Kontruksi.....	III-19
Gambar III-8. Penjadwalan Penelitian Tahap Menentukan Alat yang		
	Digunakan untuk Pelaksanaan Penelitian Fase Transisi.....	III-20
Gambar III-9. Penjadwalan Penelitian Tahap Melakukan Pengujian		
	Penelitian.....	III-21
Gambar III-10. Penjadwalan Penelitian Tahap Melakukan Analisa Hasil		
	Pengujian dan Membuat Kesimpulan.....	III-21
Gambar IV-1. Diagram Sistem <i>Preprocessing</i>		IV-6
Gambar IV-2. Contoh Dokumen.....		IV-6
Gambar IV-3. Hasil <i>Case Folding</i>		IV-6
Gambar IV-4. Hasil <i>Tokenizing</i>		IV-7
Gambar IV-5. Diagram Sistem Identifikasi Kesalahan.....		IV-7
Gambar IV-6. Hasil Identifikasi Kesalahan.....		IV-7
Gambar IV-7. Diagram Sistem <i>Damerau Levenshtein Distance</i>		IV-8
Gambar IV-8. Hasil Perhitungan <i>Edit Distance</i>		IV-8
Gambar IV-9. Hasil Algoritma <i>Damerau Levenshtein Distance</i>		IV-9
Gambar IV-10. Diagram Sistem <i>Bigram</i>		IV-10
Gambar IV-11. Diagram Sistem TF/IDF.....		IV-11
Gambar IV-12. Diagram <i>Use Case</i>		IV-17

Gambar IV-13. Rancangan Antarmuka Perangkat Lunak.....	IV-24
Gambar IV-14. Diagram Aktivitas Muat Data.....	IV-25
Gambar IV-15. Diagram Aktivitas Pengecekan.....	IV-26
Gambar IV-16. Diagram Aktivitas Memilih Kandidat.....	IV-27
Gambar IV-17. Diagram Aktivitas Hitung MRR.....	IV-27
Gambar IV-18. <i>Sequence Diagram</i> Muat Data.....	IV-28
Gambar IV-19. <i>Sequence Diagram</i> Hitung MRR.....	IV-28
Gambar IV-20. <i>Sequence Diagram</i> Pengecekan.....	IV-29
Gambar IV-21. Diagram Kelas Perangkat Lunak.....	IV-31
Gambar IV-22. Antarmuka Perangkat Lunak.....	IV-36

DAFTAR LAMPIRAN

1. Dokumen yang Digunakan
2. Nilai Cosine Similarity Masing-Masing Kandidat Kata
3. Kode Program

SPELLING CHECKER USING DAMERAU LEVENSHTTEIN DISTANCE AND COSINE SIMILARITY ALGORITHM

By :
Nur Hamidah
09021181520128

ABSTRACT

Writing is an embodiment of the author's ideas that are to be conveyed to others. A writer often experiences typos in typing the script, so that it can influence the meaning of the text. Therefore, a system is needed to detect word errors. In this study, checking is done by using the Dictionary Lookup method and giving the candidate words using the Damerau Levenshtein Distance algorithm. Candidates will then determine the ranking by breaking the word into Bigram form and calculating the similarity value using the Cosine Similarity algorithm. The test results based on the data used yield different Mean Reciprocal Rank (MRR) values for each type of error. The type of error deletion produces an MRR value of 88.89%, the type of insertion error produces an MRR value of 97.78%, the type of substitution error produces an MRR value of 88.89%, the type of transposition error produces an MRR value of 89%.

Keywords : Dictionary Lookup, Damerau Levenshtein Distance, Spelling Checker, Cosine Similarity, Mean Reciprocal Rank

Palembang, Juli 2019

Pembimbing I,


Novi Yustiani, M.T.

NIP. 198211082012122001

Pembimbing II,


Desty Rodiah, M.T.

NIK. 1671016112890005

Mengetahui,

Ketua Jurusan Teknik Informatika



Rikie Primartha, M.T.

NIP. 197706012009121004

**SPELLING CHECKER MENGGUNAKAN ALGORITMA DAMERAU
LEVENSHTEIN DISTANCE DAN COSINE SIMILARITY**

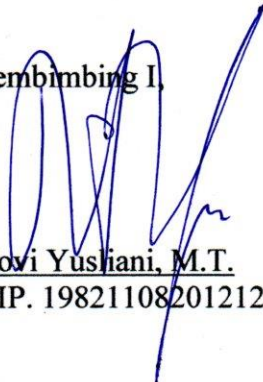
**Oleh :
Nur Hamidah
09021181520128**

ABSTRAK

Tulisan adalah sebuah perwujudan gagasan penulis yang hendak disampaikan kepada orang lain. Seorang penulis seringkali mengalami kesalahan pengetikan dalam pengetikan naskahnya, sehingga dapat mempengaruhi arti dari tulisan tersebut. Oleh karena itu, diperlukan suatu sistem untuk mendeteksi kesalahan kata. Pada penelitian ini, pengecekan dilakukan dengan menggunakan metode *Dictionary Lookup* dan memberikan kandidat kata menggunakan algoritma *Damerau Levenshtein Distance*. Kandidat kata kemudian akan dilakukan penentuan ranking dengan memecah kata menjadi bentuk *Bigram* dan menghitung nilai kemiripan dengan menggunakan algoritma *Cosine Similarity*. Hasil pengujian berdasarkan data yang digunakan menghasilkan nilai *Mean Reciprocal Rank* (MRR) yang berbeda pada masing-masing jenis kesalahan. Jenis kesalahan *deletion* menghasilkan nilai MRR sebesar 88.89%, jenis kesalahan *insertion* menghasilkan nilai MRR sebesar 97.78%, jenis kesalahan *substitution* menghasilkan nilai MRR sebesar 88.89%, jenis kesalahan *transposition* menghasilkan nilai MRR sebesar 89%.


Kata Kunci : *Dictionary Lookup, Damerau Levenshtein Distance, Spelling Checker, Cosine Similarity, Mean Reciprocal Rank*

Pembimbing I,



Novi Yushiani, M.T.
NIP. 198211082012122001

Palembang, Juli 2019

Pembimbing II,


Desty Rodiah, M.T.
NIK. 1671016112890005

Mengetahui,
Ketua Jurusan Teknik Informatika


Rifkie Primartha, M.T.
NIP. 197706012009121004



BAB I

PENDAHULUAN

1.1 Pendahuluan

Pada bab ini berisi tentang latar belakang masalah, rumusan masalah, tujuan dan manfaat penelitian, batasan masalah serta sistematika penulisan.

Bab ini juga berisi penjelasan mengenai tantangan dan tujuan proses menemukan pengetahuan baru pada sistem pengecekan kata salah ejaan. Pendahuluan ini berkaitan dengan menerapkan algoritma *Damerau Levenshtein Distance* dan algoritma *Cosine Similarity* pada sistem pengecekan kata salah ejaan.

1.2 Latar Belakang Masalah

Sebuah tulisan adalah sebuah perwujudan gagasan penulis yang hendak disampaikan kepada orang lain. Keterampilan menulis merupakan keterampilan yang penting dalam kehidupan, baik dalam kehidupan pendidikan maupun masyarakat. Aktivitas menulis merupakan salah satu manifestasi kemampuan dan keterampilan berbahasa paling akhir yang dikuasai pengguna bahasa setelah mendengarkan, membaca, dan berbicara (Luqman, 2009).

Seorang penulis seringkali mengalami kesalahan pengetikan dalam penulisan naskahnya, sehingga dapat mempengaruhi arti dari tulisan tersebut. Pengoreksian secara manual dapat memakan waktu yang lama karena dilakukan secara berulang-ulang untuk mendapatkan hasil yang benar-benar bebas dari kesalahan pengetikan (*typographical error*).

Terdapat beberapa jenis *typographical error* yaitu *insertion*, *deletion*, *substitution*, dan *transposition*. Dalam mendeteksi *typographical error* pada teks dibutuhkan suatu aplikasi yang disebut *spelling checker*. *Spelling checker* melakukan proses pengecekan terhadap pengejaan kata-kata yang mengalami kesalahan ejaan dan memberikan *suggestion* berupa kandidat kata (Soleh and Purwarianti, 2011).

Sistem perbaikan ejaan kata dapat dibagi menjadi dua yaitu pencocokan *string* kira-kira (*inexact string matching*) dan pencocokan *string* secara tepat (*exact string mathcing*). Pencocokan *string* kira-kira (*inexact string matching*) terbagi menjadi dua yaitu *approximate string matching* yang merupakan pencocokan *string* berdasarkan kemiripan penulisannya (jumlah karakter dan susunan karakter) dan *phonetic string matching* yang merupakan pencocokan *string* berdasarkan kemiripan ucapan (Rochmawati dan Kusumaningrum, 2015).

Penelitian mengenai *approximate string matching* telah banyak diteliti sebelumnya. Penelitian (Viny, Rudy, and Naga, 2018) melakukan perbandingan antara algoritma *Damerau Levenshtein Distance* dan algoritma *Levenshtein Distance* yang dikombinasikan dengan metode *Bigram* dan metode *Trigram*,

menyimpulkan bahwa kombinasi algoritma *Damerau Levenshtein Distance* dengan metode *Bigram* memiliki akurasi sebesar 84,62%.

Algoritma *Damerau Levenshtein Distance* adalah algoritma pencocokan string *word to word* sebagai pengembangan dari algoritma *Levenshtein Distance*. Algoritma *Damerau Levenshtein Distance* menentukan nilai operasi minimum yang dibutuhkan untuk mengubah satu string menjadi string lain, dimana operasi yang digunakan yaitu *insertion*, *deletion*, *substitution*, dan *transposition* (Damerau, Corporation, and Heights, 1964).

Penelitian (Maghfira, Cholissodin, and Widodo, 2017) menggunakan algoritma *Damerau Levenshtein Distance* mengatakan bahwa algoritma *Damerau Levenshtein Distance* menghasilkan kandidat kata dengan nilai jarak edit operasi terkecil. Namun kandidat kata hasil pengecekan tidak diurutkan.

Kandidat kata hasil pengecekan kata dapat diurutkan berdasarkan nilai probabilitas kemiripan dengan menggunakan algoritma *Cosine Similarity* (Abdillah dan Muktyas, 2013). Algoritma *Cosine Similarity* merupakan perhitungan derajat kesamaan dengan menentukan sudut antara dua vektor (Fahma, Cholissodin, dan Perdana, 2018).

Berdasarkan uraian sebelumnya, maka akan dilakukan penelitian dengan menggunakan algoritma *Damerau Levenshtein Distance* dan *Cosine Similarity* untuk mengecek kata salah ejaan dan memberikan saran kata berdasarkan kata salah ejaan.

1.3 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan diatas, rumusan masalah dalam penelitian ini adalah apakah algoritma *Damerau Levenshtein Distance* dan algoritma *Cosine Similarity* dapat memberikan saran kata berdasarkan kata salah ejaan. Untuk menyelesaikan permasalahan diatas, maka penelitian ini dibagi dalam beberapa pertanyaan penelitian yaitu :

1. Bagaimana mengembangkan sistem pengecekan kata salah ejaan menggunakan algoritma *Damerau Levenshtein Distance* ?
2. Bagaimana nilai *Mean Reciprocal Rank* (MRR) algoritma *Damerau Levenshtein Distance* dan algoritma *Cosine Similarity* dalam memberikan saran kata berdasarkan kata salah ejaan ?

1.4 Tujuan Penelitian

Tujuan penelitian ini adalah :

1. Mengembangkan sistem pengecekan kata salah ejaan menggunakan algoritma *Damerau Levenshtein Distance*
2. Mengetahui nilai *Mean Reciprocal Rank* (MRR) algoritma *Damerau Levenshtein Distance* dan algoritma *Cosine Similarity* dalam memberikan saran kata berdasarkan kata salah ejaan

1.5 Manfaat Penelitian

Manfaat dalam penelitian ini adalah :

1. Sistem yang dihasilkan dapat memberikan saran kata berdasarkan kata salah ejaan
2. Hasil dari penelitian ini dapat digunakan sebagai referensi penelitian selanjutnya dalam mengembangkan sistem pengecekan kata salah ejaan

1.6 Batasan Masalah

Batasan masalah pada penelitian ini adalah :

1. Sistem hanya dapat mengecek kata berbahasa Indonesia
2. Sistem hanya berfokus mengecek kata salah ejaan dan memberikan kandidat kata, tidak sampai perbaikan
3. Keluaran kata yang disarankan sebanyak 5 kandidat kata dengan probabilitas tertinggi
4. Kamus yang digunakan yaitu Kamus Besar Bahasa Indonesia (KBBI)
5. Sistem hanya dapat mengecek kesalahan tulisan berupa kata, bukan singkatan, lambang, nama orang ataupun kata khusus lainnya

1.7 Sistematika Penulisan

Untuk memahami lebih jelas penelitian ini, pemaparan materi dikelompokkan menjadi beberapa bab dengan sistematika penulisan sebagai berikut :

BAB I PENDAHULUAN

Pada bab ini diuraikan mengenai latar belakang, perumusan masalah, tujuan dan manfaat penelitian, batasan masalah/ ruang lingkup, metodologi penelitian, dan sistematika penulisan.

BAB II KAJIAN LITERATUR

Pada bab ini akan dibahas dasar-dasar teori yang digunakan dalam penelitian, seperti *text preprocessing*, *typography error*, dan metode pencocokan *string*.

BAB III METODOLOGI PENELITIAN .

Pada bab ini akan dibahas mengenai tahapan yang akan dilaksanakan pada penelitian ini. Masing-masing rencana tahapan penelitian dideskripsikan dengan rinci dengan mengacu pada suatu kerangka kerja. Di akhir bab ini berisi perancangan manajemen proyek pada pelaksanaan penelitian.

BAB IV PENGEMBANGAN PERANGKAT LUNAK

Pada bab ini membahas mengenai proses pengembangan perangkat lunak yang digunakan sebagai alat penelitian, penulis mengembangkan perangkat lunak dengan metode pemrograman berorientasi obyek berdasarkan panduan *Rational Unified Process*. Di dalam *Rational Unified Process* terdapat empat fase pengembangan perangkat lunak yaitu fase inepsi, elaborasi, konstruksi, dan transisi. Setiap fase memiliki pemodelan bisnis, kebutuhan, analisis dan desain, implementasi, dan pengujian.

BAB V HASIL DAN ANALISIS PENELITIAN

Pada bab ini akan menyajikan kata salah ejaan beserta kandidatnya dan evaluasi perhitungan nilai MRR menggunakan perangkat lunak.

BAB VI KESIMPULAN DAN SARAN

Pada bab ini akan membahas kesimpulan dan saran dari penelitian ini.

1.8 Kesimpulan

Berdasarkan uraian diatas, penelitian ini akan menggunakan algoritma *Damerau Levenshtein Distance* dan algoritma *Cosine Similarity* untuk mengecek kata salah ejaan, dengan tujuan agar menghasilkan kandidat kata berdasarkan kata salah ejaan. Penjelasan mengenai algoritma *Damerau Levenshtein Distance* dan algoritma *Cosine Similarity* serta penelitian yang relevan akan dijelaskan di bab 2.

DAFTAR PUSTAKA

- Abdillah, A. A., & Muktyas, I. B. (2013). Implementasi vector space model untuk pencarian dokumen, (May 2015), 0–7.
- Damerau, F. J., Corporation, I. B. M., & Heights, Y. (1964). A Technique for Computer Detection and Correction of Spelling Errors *, (3), 171–176.
- Fahma, A. I., Cholissodin, I., & Perdana, R. S. (2018). Identifikasi Kesalahan Penulisan Kata (Typographical Error) pada Dokumen Berbahasa Indonesia Menggunakan Metode N-gram dan Levenshtein Distance, 2(1), 53–62.
- Gan, J. Q. (2015). ORIGINAL RESEARCH A query suggestion method combining TF-IDF and Jaccard Coefficient for interactive web search, 4(2), 119–125. <https://doi.org/10.5430/air.v4n2p119>
- Herwijayanti, B., Ratnawati, D. E., & Muflikhah, L. (2018). Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan Cosine Similarity, 2(1), 306–312.
- Jupin, J., Shi, J. Y., & Obradovic, Z. (2012). Understanding Cloud Data Using Approximate String Matching and Edit Understanding Cloud Data Using Approximate String Matching and Edit Distance, (May 2016). <https://doi.org/10.1109/SC.Companion.2012.149>
- Kornain, A., & Yansen, F. (n.d.). PENERAPAN ALGORITMA JARO-WINKLER DISTANCE UNTUK SISTEM PENDETEKSI PLAGIARISME PADA DOKUMEN TEKS BERBAHASA INDONESIA Hal-1 Kesamaan Hal-2, 1–10.
- Lahitani, A. R., Permanasari, A. E., & Setiawan, N. A. (2016). Cosine Similarity to Determine Similarity Measure : Study Case in Online Essay Assessment.
- Luqman. (2009). Program Aplikasi Pengoreksian Ejaan Bahasa Indonesia Berbasis Web, 13–19.
- Maghfira, T. N., Cholissodin, I., & Widodo, A. W. (2017). Deteksi Kesalahan Ejaan dan Penentuan Rekomendasi Koreksi Kata yang Tepat Pada Dokumen Jurnal JTIK Menggunakan Dictionary Lookup dan Damerau-Levenshtein Distance, 1(6), 498–506.
- Rochmawati, Y., & Kusumaningrum, R. (2015). Studi Perbandingan Algoritma Pencarian String dalam Metode Approximate String Matching untuk Identifikasi Kesalahan Pengetikan Teks, 125–134.
- Soleh, M. Y., & Purwarianti, A. (2011). A Non Word Error Spell Checker for Indonesian using Morphologically Analyzer and HMM.

Sugianto, S. A., Liliana, & Rostianingsih, S. (n.d.). PEMBUATAN APLIKASI PREDICTIVE TEXT MENGGUNAKAN METODE N-GRAM-BASED.

Viny, C. M., Rudy, & Naga, D. S. (2018). Fast and Accurate Spelling Correction Using Trie and Damerau-levenshtein Distance Bigram, *16*(2), 827–833.
<https://doi.org/10.12928/TELKOMNIKA.v16i2.6890>