

PENERAPAN PENGHAPUSAN POLA KATA DALAM REKONSTRUKSI *STEMMING* BAHASA INDONESIA

Ana Tsalitsatun Ni'mah
09.04.1.1.1.00132

Pembimbing I : Firdaus Solihin, S.Kom., M.Kom.
Pembimbing II : Iwan Santosa, S.T., M.T.

ABSTRAK

Stemming adalah pemisahan kata dasar dari imbuhan. *Stemming* dalam bahasa Indonesia umumnya melakukan penghapusan berkali-kali dikarenakan imbuhan dalam bahasa Indonesia sangatlah beragam. Sehingga saat kata mendapatkan imbuhan awalan dan akhiran, maka yang terhapus lebih dahulu adalah akhiran kemudian setelah itu awalan. Hal ini tentu membuat proses *stemming* memerlukan waktu yang lebih lama.

Penelitian kali ini mencoba merekonstruksi algoritma *stemming* bahasa Indonesia dengan perbaikan pada proses yang berulang pada penghapusan imbuhan, penghapusan yang berulang tersebut direkonstruksi menjadi penghapusan serempak atau sekaligus. Penelitian ini menggunakan teori *regular expression* pada proses penghapusannya. Algoritma *stemming* dalam penelitian ini dinamakan *Indonesian Stemming Reconstruction (ISR)*.

ISR diuji coba pada 300 dokumen dari data Fadhila Tala dan 50 konten dari data hasil Crawler RISE. Penelitian ini memberikan hasil yang cukup baik yaitu perhitungan kebenaran kata mencapai 72% dan kecepatan 0,0243 detik pada 300 dokumen Fadhila Tala. Sedangkan pada 50 konten RISE didapatkan kecepatan sebesar 0,0052 detik dengan persentase kebenaran kata sebesar 50%.

Kata kunci : *Stemming, Regular Expression, Indonesian Stemming Reconstruction.*

(halaman ini sengaja dikosongkan)

KATA PENGANTAR

Puji syukur Penulis panjatkan kepada Allah SWT atas berkat dan rahmat-Nya, Penulis dapat menyelesaikan Tugas Akhir yang berjudul **Penerapan Penghapusan Pola Kata Dalam Rekonstruksi *Stemming* Bahasa Indonesia**. Pada kesempatan ini, Penulis ingin menyampaikan terima kasih dan penghormatan yang sebesar-besarnya kepada:

1. Bapak Sholichin dan ibu Ismiati tercinta yang telah mendukung segenap jiwa raga dan doa untuk selesainya kuliah dan Tugas Akhir penulis.
2. Bapak Firdaus Solihin dan Bapak Iwan Santosa atas semua bimbingan dan nasehat yang diberikan selama masa kuliah dan penyusunan Tugas Akhir. Penulis tidak dapat memberikan balasan apa pun selain ucapan terimakasih dan doa.
3. Segenap staf dan dosen pengajar di Teknik Informatika Fakultas Teknik Universitas Trunojoyo Madura (UTM).
4. Mas Ibda', mbak Itsna, adik Vina, mbak Sulis, mas Miko, Abyan, Vicky, Kayla dan seluruh keluarga yang selalu menemani di kala suka dan menghibur di kala duka.
5. Alfian Qomaruddin dan keluarga yang telah memberikan dukungan, semangat dan doa untuk selesainya Tugas Akhir Penulis.
6. Teman seperjuangan SOLITE'09 yang selalu menemani hari-hari di kampus Trunojoyo Madura dari semester Satu hingga semester 8.
7. Mas Muhammad Alif dan mas Muhammad Syaiful yang telah memberikan izin untuk pengembangan penelitiannya, juga atas semangat yang diberikannya.
8. Warga laboratorium Informatika, LPM SAINT, UKMF ITC, JARBIS Indonesia dan Cyber Buana yang telah memberikan pengalaman dan pengetahuan baik mengenai informatika atau pun organisasi.
9. Pak Pri dan Bu Pri kos Gili, Pak Kyai dan Bu Nyai PonPes Darul Asmail Husna, teman-teman kos Gili, Desy dkk, teman-teman kos Trunojoyo Asri, mbak Lely, Cak Kur dan teman-teman Cempedak Lima yang telah memberikan bantuan tak terhingga.

10. Dan semua pihak yang telah membantu dalam penyelesaian Tugas Akhir penulis yang tidak dapat penulis sebutkan satu persatu.

Penulis menyadari bahwa pengerjaan Tugas Akhir ini tidak luput dari berbagai kesalahan. Oleh karena itu, Penulis memohon maaf atas segala kekurangan yang ada. Dengan segala kerendahan hati, berbagai saran dan kritik yang membangun sangat Penulis harapkan dari rekan-rekan pembaca.

Bangkalan, 21 Mei 2013

P e n u l i s

DAFTAR ISI

	hal
HALAMAN JUDUL	Error! Bookmark not defined.
LEMBAR PENGESAHAN	Error! Bookmark not defined.
LEMBAR PERSEMBAHAN	Error! Bookmark not defined.
ABSTRAK.....	ix
KATA PENGANTAR	xi
DAFTAR ISI.....	xiii
DAFTAR GAMBAR	xv
DAFTAR TABLE	xvii
DAFTAR ISTILAH	xix
DAFTAR LAMPIRAN	xxi
PENDAHULUAN	Error! Bookmark not defined.
1.1. Latar belakang	Error! Bookmark not defined.
1.2. Perumusan masalah	Error! Bookmark not defined.
1.3. Tujuan dan manfaat	Error! Bookmark not defined.
1.4. Batasan masalah	Error! Bookmark not defined.
1.5. Metodologi penelitian.....	Error! Bookmark not defined.
1.6. Sistematika penulisan	Error! Bookmark not defined.
STUDI PUSTAKA.....	Error! Bookmark not defined.
2.1. Tinjauan Pustaka	Error! Bookmark not defined.
2.2. Landasan Teori	Error! Bookmark not defined.
2.2.1. Information Retrieval (IR).....	Error! Bookmark not defined.
2.2.2. Web Crawler	Error! Bookmark not defined.
2.2.3. Stemming.....	Error! Bookmark not defined.
2.2.4. Regular Expression dan Transition Graph.....	Error! Bookmark not defined.
2.2.5. Morfologi Bahasa Indonesia..	Error! Bookmark not defined.
PERANCANGAN SISTEM	Error! Bookmark not defined.
3.1. Analisis Permasalahan.....	Error! Bookmark not defined.
3.2. Penanganan Masalah	Error! Bookmark not defined.
3.3. Perancangan Sistem.....	Error! Bookmark not defined.

3.3.1.	Perancangan Sistem Secara Umum	Error! Bookmark not defined.
3.3.2.	Flowchart Proses Ambil Dokumen.....	Error! Bookmark not defined.
3.3.3.	Flowchart <i>Preprocessing, Filtering dan Tokenization</i>	Error! Bookmark not defined.
3.3.4.	Flowchart Proses <i>Enhanced Confix Stripping Porter</i>	Error! Bookmark not defined.
3.3.5.	Flowchart Proses <i>ECS</i>	Error! Bookmark not defined.
3.3.6.	Flowchart Proses <i>Porter Stemmer</i>	Error! Bookmark not defined.
3.3.7.	Flowchart Proses <i>Indonesian Stemming Reconstruction</i>	Error! Bookmark not defined.
3.4.	Perancangan Database	Error! Bookmark not defined.
3.4.1.	CDM (Conceptual Data Model) ..	Error! Bookmark not defined.
3.4.2.	PDM (Physical Data Model) .	Error! Bookmark not defined.
3.4.3.	Keterangan Tabel Pada Perancangan Database	Error! Bookmark not defined.
IMPLEMENTASI SISTEM.....		Error! Bookmark not defined.
4.1.	Batasan Implementasi.....	Error! Bookmark not defined.
4.2.	Lingkungan Implementasi.....	Error! Bookmark not defined.
4.3.	Pembuatan Program	Error! Bookmark not defined.
4.4.	Implementasi Program ...	Error! Bookmark not defined.
4.5.	Implementasi Ujicoba.....	Error! Bookmark not defined.
4.6.	Ujicoba Hasil Perbaikan Kata	Error! Bookmark not defined.
PENUTUP		Error! Bookmark not defined.
3.1.	Kesimpulan.....	Error! Bookmark not defined.
3.2.	Saran.....	Error! Bookmark not defined.
DAFTAR PUSTAKA.....		Error! Bookmark not defined.

DAFTAR LAMPIRAN.....	Error! Bookmark not defined.
DAFTAR RIWAYAT HIDUP.....	Error! Bookmark not defined.

DAFTAR GAMBAR

Gambar 2.1. Desain umum <i>web crawler</i>	9
Gambar 2.2. Transition Graph	22
Gambar 3.1. Perancangan Sistem	31
Gambar 3.2. Flowchart Proses Ambil/Input Dokumen	32
Gambar 3.3. Flowchart <i>Preprocessing</i>	32
Gambar 3.4. Flowchart <i>Filetering</i> dan <i>Tokenization</i>	33
Gambar 3.5. Flowchart <i>Enhanced Confix Stripping Porter Stemmer</i>	34
Gambar 3.6. Flowchart Proses <i>ECS</i>	35
Gambar 3.7. Flowchart <i>Porter Stemmer</i>	36
Gambar 3.8. Flowchart Proses <i>Indonesian Stemming</i> <i>Reconstruction</i>	37
Gambar 3.9. CDM Inputan Berita	38
Gambar 3.10. CDM 249 Berita Data Fadhila Tala	38
Gambar 3.11. CDM 50 Berita Data Crawler RISE	39
Gambar 3.12. PDM Inputan Berita	39
Gambar 3.13. PDM 249 Berita Data Fadhila Tala	40
Gambar 3.14. PDM 50 Berita Data Crawler RISE	40
Gambar 4.1. Halaman Index dan input dokumen	52
Gambar 4.2. Halaman Perbandingan Kata Hasil <i>Stemming</i>	52
Gambar 4.3. Halaman Perbandingan Banyak Kebenaran Kata dan Perbandingan Kecepatan.....	53
Gambar 4.4. Diagram uji coba dokumen bebas.....	56
Gambar 4.5. Diagram uji coba 50 dokumen Fadila Tala	57

Gambar 4.6. Diagram uji coba 100 dokumen Fadila Tala	57
Gambar 4.7. Diagram uji coba 300 dokumen Fadila Tala	58
Gambar 4.8. Diagram uji coba 50 dokumen RISE Crawler.....	58

DAFTAR TABLE

Tabel 2.1. Kombinasi Awalan Akhiran yang tidak diijinkan Nazief Andriani	10
Tabel 2.2. Cara Menentukan Tipe Awalan Untuk Kata Yang Diawali Dengan “te-”	10
Tabel 2.3. Jenis Awalan Berdasarkan Tipe Awalannya	11
Tabel 2.4. Aturan Pemenggalan Awalan Stemmer Nazief Andriani ...	12
Tabel 2.5. Aturan Untuk Inflectional Particle	14
Tabel 2.6. Aturan Untuk Inflectional Possesive Pronoun	14
Tabel 2.7. Aturan Untuk First Order Derivational Prefix	14
Tabel 2.8. Aturan untuk Second Order Derivational Prefix	15
Tabel 2.9. Aturan Untuk Derivational Suffix	16
Tabel 2.10. Modifikasi dan tambahan aturan Nazief Andriani	17
Tabel 2.11. Modifikasi aturan tabel 2.4. untuk algoritma ECS	19
Tabel 2.12. Beberapa operator regex yang digunakan dalam penelitian [8]	22
Tabel 3.1. Tabel Pemenggalan <i>Indonesian Stemming Reconstruction</i>	28
Tabel 3.2. Tabel <i>Morfologi</i> Bahasa Indonesia	30
Tabel 4.1. Hasil Ujicoba dokumen bebas	53
Tabel 4.2. Hasil Uji Coba 50 Dokumen Fadhila Tala	54
Tabel 4.3. Hasil Uji Coba 100 Dokumen Fadhila Tala 20813 kata	54
Tabel 4.4. Hasil Uji Coba 300 Dokumen Fadhila Tala 77081 kata	55
Tabel 4.5. Hasil Uji Coba 50 Dokumen Hasil Crawler RISE V.1. 13951 kata	55
Tabel 4.6. Tabel perbaikan penghapusan imbuhan tidak diperbolehkan	59
Tabel 4.7. Beberapa Perbaikan <i>overstemming</i>	59

Tabel 4.8. Rata-rata Hasil Ujicoba 60

DAFTAR ISTILAH

<i>CRAWL</i>	Adalah sebuah cara untuk menghimpun atau mengumpulkan beberapa website yang telah ditentukan
<i>CS</i>	Singkatan dari <i>Confix Stripping</i> yang merupakan penelitian pengembangan dari <i>stemming</i> Nazief dan Andriani
<i>ECS</i>	Singkatan dari <i>Enhanced Confix Stripping</i> yang merupakan penelitian pengembangan dari <i>CS</i>
<i>ECSP</i>	Singkatan dari <i>Enhanced Confix Stripping Porter</i> yang merupakan algoritma gabungan antara <i>ECS</i> dengan <i>Porter Stemmer</i> dari pengembangan penelitian tahun 2012
<i>IR</i>	Singkatan dari <i>Information Retrieval</i> yang merupakan suatu sistem yang mampu melakukan penyimpanan, pencarian, dan pemeliharaan informasi
<i>ISR</i>	Singkatan dari <i>Indonesian Stemming Reconstruction</i> yang merupakan algoritma pengembangan <i>stemming</i> khusus kata dalam bahasa Indonesia pada penelitian ini
<i>Morfologi</i>	Sebuah aturan penulisan kata
<i>PORTER</i>	Algoritma <i>stemming</i> yang dicetuskan oleh warga asing untuk bahasa Inggris, kemudian dikembangkan <i>Porter Indonesian</i>
<i>Reguler Expression</i>	Teori yang digunakan dalam pendeteksian teks dengan menggunakan formula tertentu
<i>RISE</i>	Singkatan dari <i>Research Indonesian Search Engine</i> yang merupakan penelitian <i>Search Engine</i> Indonesia dicetuskan tahun 2011 oleh Program Studi Teknik Informatika Universitas Trunojoyo Madura

Root Word

Kata dasar

Stemming

Pemisahan kata dari imbuhan

Term

Kata

DAFTAR LAMPIRAN

Lampiran 1. Ujicoba pada 50 dokumen Fadila Tala	65
Lampiran 2. Ujicoba pada 100 dokumen Fadila Tala	68
Lampiran 3. Ujicoba pada 300 dokumen Fadila Tala	74
Lampiran 4. Ujicoba pada 50 dokumen konten Crawler RISE	91
Lampiran 5. Ujicoba pada beberapa dokumen selama 12 kali	94

(halaman ini sengaja dikosongkan)