



Hak cipta dan penggunaan kembali:

Lisensi ini mengizinkan setiap orang untuk menggubah, memperbaiki, dan membuat ciptaan turunan bukan untuk kepentingan komersial, selama anda mencantumkan nama penulis dan melisensikan ciptaan turunan dengan syarat yang serupa dengan ciptaan asli.

Copyright and reuse:

This license lets you remix, tweak, and build upon work non-commercially, as long as you credit the origin creator and license it on your new creations under the identical terms.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Temu kembali informasi (*information retrieval*) adalah sebuah proses menemukan kembali dokumen-dokumen relevan untuk memenuhi kebutuhan informasi para pengguna (Manning, et al, 2009). Kebutuhan informasi pengguna tersebut direpresentasikan dalam bentuk *query*. Pengambilan keputusan dalam menemukan kembali dokumen-dokumen relevan dilakukan dengan membandingkan kata-kata pada *query* dengan kata-kata yang berada pada dokumen yang dicari atau dengan mengestimasi tingkat relevansi dokumen tersebut dengan *query* dari pengguna.

Stemming adalah salah satu cara untuk meningkatkan performa *information retrieval* dengan cara mentransformasikan kata-kata dalam sebuah dokumen teks ke kata dasarnya (Agusta, 2009). Teknik *stemming* terbagi dalam empat kelompok, yaitu *Affix Removal*, *Successor Variety*, *Table Lookup*, dan N-gram (Frakes, 1992).

Dalam penelitian ini akan dibandingkan performa algoritma *Affix Removal*, *Successor Variety*, dan N-gram termodifikasi dalam *stemming* dokumen berbahasa Indonesia. Algoritma *Affix Removal* yang akan digunakan dalam penelitian ini adalah algoritma Nazief-Adriani, dan metode yang akan digunakan dalam algoritma *Successor Variety* adalah metode *Cutoff*.

Dalam pembuatan penelitian ini tentunya tidak lepas dari referensi jurnal-jurnal pendukung dari sumber lain yang telah melakukan penelitian sejenis. Penelitian-penelitian sejenis yang berhubungan dengan penelitian ini diantaranya:

- 1) Ledy Agusta (2009) yang membandingkan dua algoritma *affix removal*, yaitu algoritma Porter dan algoritma Nazief-Adriani yang menyimpulkan bahwa algoritma Nazief-Adriani memiliki prosentase keakuratan lebih besar walaupun membutuhkan waktu lebih lama.
- 2) Penelitian Riyad Al-Shalabi dan kawan-kawan (2005) yang membandingkan dua metode algoritma *Successor Variety*, yaitu metode *Cutoff* dan *Entropy* dan mendapatkan hasil bahwa metode *Cutoff* lebih bagus dibanding metode *Entropy*.
- 3) Penelitian B.P. Pande dan kawan-kawan (2013) dalam membandingkan performa metode algoritma N-gram yang diajukan dengan algoritma Porter dan mendapatkan hasil bahwa metode N-gram yang diajukan tidak kalah dengan algoritma Porter.

Dari ketiga penelitian yang telah disebutkan, timbul sebuah pemikiran untuk membandingkan performa ketiga algoritma tersebut dalam *stemming* teks berbahasa Indonesia.

1.2 Perumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, berikut rumusan masalahnya.

- 1) Bagaimanakah cara membandingkan ketiga algoritma tersebut?

- 2) Algoritma manakah yang lebih cocok untuk digunakan dalam *stemming* dokumen teks berbahasa Indonesia?

1.3 Batasan Masalah

Penelitian ini hanya membahas perbandingan performa algoritma Nazief-Adriani, *Successor Variety* metode *Cutoff*, dan algoritma N-gram termodifikasi. Variabel penelitian dibatasi pada dua parameter metode evaluasi Paice, yaitu *understemming index* dan *overstemming index*. Bahasa yang digunakan dalam dokumen yang di-*stemming* adalah bahasa Indonesia. Dokumen yang digunakan untuk *stemming* adalah dokumen berekstensi .txt.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah, maka penelitian ini bertujuan untuk membandingkan performa algoritma Nazief-Adriani, *Successor Variety* metode *Cutoff*, dan N-gram termodifikasi menggunakan parameter perbandingan *understemming index* dan *overstemming index* yang kemudian dari hasil tersebut akan ditentukan algoritma manakah yang lebih cocok digunakan untuk *stemming* dokumen, khususnya teks berbahasa Indonesia.

1.5 Manfaat Penelitian

Manfaat dalam penelitian ini antara lain mengetahui algoritma manakah di antara ketiga algoritma yang dibandingkan yang lebih cocok untuk digunakan

dalam *stemming* dokumen teks berbahasa Indonesia sehingga dapat meningkatkan hasil temu kembali informasi (*information retrieval*).

Bagi peneliti sendiri penelitian ini berguna untuk mempelajari beberapa materi baru, yakni beberapa algoritma *stemming* yang belum pernah diketahui sebelumnya, metode evaluasi algoritma *stemming*, dan sebagainya.

1.6 Sistematika Penulisan

Secara umum, skripsi ini disusun dan dibagi dalam 5 (lima) bab dengan sistematika penulisan sebagai berikut.

Bab I: Latar Belakang

Pada bab ini berisi mengenai latar belakang masalah, perumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.

Bab II: Landasan Teori

Pada bab ini dibahas uraian teori-teori yang berhubungan dengan algoritma-algoritma dan metode evaluasi yang digunakan.

Bab III: Metodologi dan Perancangan Sistem

Bab ini menjelaskan tentang metode-metode yang digunakan dalam penelitian dan juga mengenai perancangan sistem aplikasi perbandingan performa algoritma *stemming*.

Bab IV: Pengujian dan Pembahasan

Bab ini memaparkan pengujian yang telah dilakukan berikut penjelasannya.

Bab V: Kesimpulan dan Saran

Bab terakhir ini berisikan tentang kesimpulan dari seluruh bab-bab yang sudah dibahas serta saran untuk pengembangan penelitian selanjutnya

