1. Pendahuluan

1.1 Latar Belakang

Information Retrieval (IR) secara umum merupakan suatu teknik untuk menemukan informasi di dalam kumpulan dokumen atau di dalam media-media lainnya dengan memberikan *query* atau *keyword* berupa teks, suara, gambar atau bentuk-bentuk lainnya. Penerapannya yang paling sering dijumpai adalah *search engine* atau mesin pencari.

Untuk meningkatkan jumlah dokumen yang diperoleh salah satunya adalah dengan menggunakan pemotong kata berimbuhan (*stemmer*). *Stemmer* merupakan salah satu alat bantu paling sederhana dalam bidang IR. *Stemmer* digunakan untuk mendapatkan kata dasar atau bentuk yang lebih umum dari suatu kata sehingga mengurangi variasi kata pada dokumen-dokumen. Dengan demikian dokumen yang diinginkan akan semakin banyak diperoleh.

Sebagai contoh, dokumen yang mengandung kata-kata berimbuhan "pendapat", "pendapatan", "didapat" dan sebagainya akan dirujuk oleh satu kata dasar yang sama yaitu "dapat". Namun beberapa kata berimbuhan yang mempunyai kata dasar yang sama, memiliki makna yang berbeda, sehingga kurang tepat apabila menyamakan seluruh variasi kata tersebut kepada kata dasarnya dengan menggunakan *stemmer*. Contohnya kata "pendapat" dengan "pendapatan". Meskipun memiliki kata dasar yang sama, keduanya memiliki makna yang sangat berbeda.

Selain masalah perbedaan makna di atas, terdapat juga masalah yang terkait dengan jenis koleksi dokumen yang dapat mempengaruhi makna kata. Misalnya, kata "membintangi" dan "bintang". Pada jenis koleksi dokumen astronomi kata "membintangi" tidak mempunyai makna yang sama dengan kata "bintang". Sebaliknya pada jenis koleksi dokumen perfilman kedua kata ini bermakna sama yaitu permainan film.

Permasalahan tersebut melatarbelakangi penulis mengangkat *corpus-based stemmer,* yang selanjutnya akan disebut *stemmer* SVC (*Statistics of Variant Co-occurrence*), yang tidak tergantung bahasa untuk menghindari penyamarataan makna variasi kata khususnya pada ahasa Indonesia, karena pada Bahasa Indonesia terdapat banyak variasi kata yang berakar pada kata dasar yang sama, namun memiliki perbedaan makna. Di samping itu, penulis mencoba memperbaiki efektifitas penggunaan *stemmer* Indonesia yang sudah ada dengan menggunakan statistik *co-occurrence* dari variasi kata.

1.2 Perumusan Masalah

Dengan mengacu pada latar belakang masalah diatas, maka permasalahan yang akan dibahas dan diteliti adalah :

1. Bagaimana konsep pengumpulan seluruh bentuk kata-kata unik di dalam sebuah koleksi dokumen menjadi sebuah *equivalence class*.
2. Bagaimana pemodelan dan penyimpanan koleksi dokumen pada sistem aplikasi.
3. Bagaimana membangun aplikasi *corpus-based stemmer* yang mampu meningkatkan akurasi makna kata dengan menggunakan *co-occurrence* dari variasi kata.
4. Bagaimana performansi aplikasi terhadap tingkat akurasi makna kata yang dicapai dilihat dari parameter *recall*, *precision*, dan *F-Measure* antara *baseline stemmer* dibandingkan dengan *corpus based stemmer*.

1.3 Tujuan

Berdasarkan rumusan masalah di atas, maka tujuan dari tugas akhir ini adalah :

1. Memodelkan dan menyimpan koleksi dokumen pada sistem aplikasi.
2. Membangun aplikasi *corpus-based stemmer* yang mampu meningkatkan akurasi makna kata dengan menggunakan *co-occurrence* dari variasi kata.
3. Menganalisis performansi aplikasi terhadap tingkat akurasi makna kata yang dicapai dilihat dari parameter *recall*, *precision*, dan *F-Measure* antara *baseline stemmer* dibandingkan dengan *corpus based stemmer*.

Batasan masalah yang digunakan dalam penelitian ini antara lain :

1. Koleksi dokumen yang digunakan adalah dokumen teks berbahasa Indonesia.
2. Sebagai *baseline stemmer* digunakan Algoritma Nazief Adriani yang di dalamnya telah menggunakan kamus.
3. Aplikasi yang dirancang bersifat *stand alone* atau berdiri sendiri.

1.4 Metodologi Penyelesaian Masalah

Metode yang digunakan dalam penyelesaian tugas akhir ini dengan langkah kerja sebagai berikut :

1. **Studi Literatur**
   Tahap ini merupakan tahap persiapan yang meliputi pengumpulan bahan-bahan pustaka dengan melakukan studi literatur sebagai referensi Tugas Akhir yang meliputi *Information Retrieval*, *corpus analysis*, Algoritma Nazief Adriani, dan topik lainnya yang mendukung penyusunan Tugas Akhir ini. Bahan pustaka ini akan digunakan sebagai dasar teori penyusunan Tugas Akhir.

2. **Pemahaman Sistem**
   Memahami sistem aplikasi yang akan dibangun yang meliputi pemodelan dan penyimpanan koleksi dokumen pada sistem aplikasi dan pemodelan lingkungan perangkat lunak.

3. **Analisis dan Perancangan Aplikasi**
   Menjabarkan *requirement*, serta analisis dan desain perangkat lunak yang akan dibangun dengan mengacu pada hasil pemahaman sistem dan studi literatur yang telah diperoleh sebelumnya. Daftar kebutuhan sistem, desain proses, desain model data, dan desain antar muka aplikasi didefinisikan pada tahap ini.

4. **Implementasi Sistem**
   Pembuatan perangkat lunak yang sesuai dengan analisis perancangan dimulai dengan pengumpulan bentuk kata-kata unik yang ada di dalam koleksi dokumen, kemudian membangun sejumlah kelompok kata-kata yang identik dengan menggunakan Algoritma Nazief Adriani, dilanjutkan dengan memperbaiki kelompok kata-kata tersebut dengan menggunakan metode *corpus-based stemming.*

5. **Pengujian Sistem**
   Pengujian perangkat lunak dilakukan dengan mengukur parameter *recall, precision,* dan *F-Measure* berdasarkan dokumen yang dihasilkan sistem antara Algoritma Nazief Adriani dibandingkan dengan Algoritma Nazief Adriani yang telah dimodifikasi dengan metode *corpus-based stemming*.

6. **Analisis Hasil**

   Mengevaluasi dan menganalisis tingkat keakuratan hasil informasi yang dihasilkan sistem berdasarkan nilai setiap parameter yang diperoleh dari skenario uji terhadap modifikasi Algoritma Nazief Adriani. Berdasarkan analisis yang diperoleh kemudian dirumuskan sebuah kesimpulan terhadap performa dan kinerja sistem.