

1. Pendahuluan

1.1 Latar Belakang

Pada saat ini, berbagai informasi dalam jumlah besar dapat diperoleh dengan mudah dari Internet. Tetapi, tidak mudah untuk mencari informasi yang berguna diantara sekian banyak sumber informasi yang didapatkan. Beberapa *search engine* dapat melakukan pencarian teks lengkap, tetapi hasilnya kurang memuaskan. Efisiensi pencarian dari jumlah dokumen yang sangat banyak sangat dipengaruhi oleh kualitas dari kata kunci yang diberikan oleh setiap dokumen. Oleh karena itu, dibutuhkan kata kunci yang berkualitas yang diambil secara otomatis di-ekstrak dari dokumen-dokumen agar mendapatkan nilai efisiensi yang tinggi.

Ekstraksi kata kunci adalah suatu metode untuk meng-identifikasi serangkaian kecil kata-kata, frase kunci, kata kunci, atau segmen kunci dari sebuah dokumen yang dapat menjelaskan makna dari sebuah dokumen. Karena kata kunci adalah unit terkecil yang dapat mengungkapkan makna dokumen, banyak aplikasi *text-mining* dapat mengambil keuntungan dari hal itu seperti *automatic indexing*, *automatic summarization*, *automatic classification*, *clustering*, *automatic filtering* dan lain sebagainya.

Tugas Akhir ini akan membahas mengenai "Ekstraksi Kata Kunci dengan menggunakan Conditional Random Fields". Conditional Random Fields (CRF) adalah model probabilistik untuk segmentasi dan pelabelan data sekuens. CRF menggunakan teknik yang berbeda dalam melakukan *preprocessing* dimana kalimat-kalimat yang ada akan disegmentasi dan diberi label sesuai *part-of-speech* (POS) dari masing-masing kata yang dihasilkan. Fitur lain yang bisa diekstrak dari sebuah dokumen adalah TF-IDF dan *length-of-word* (Len). TF-IDF digunakan sebagai bobot/nilai dari kepentingan sebuah kata, baik dalam sebuah dokumen maupun dalam kumpulan dokumen (*corpus*). Nilai *precision*, *recall* dan *F-measure* digunakan untuk mengetahui performansi dari hasil penelitian ini. Hipotesis awal dari penelitian ini adalah dengan menggunakan metode Conditional Random Fields akan menghasilkan performansi yang baik.

1.2 Perumusan Masalah

Berdasarkan pada latar belakang diatas, permasalahan yang menjadi fokus pada tugas akhir ini diantaranya yaitu:

- a. Bagaimana cara mengimplementasikan ekstraksi kata kunci dengan menggunakan Conditional Random Fields?
- b. Bagaimana karakteristik pemodelan Conditional Random Fields dalam melakukan proses ekstraksi kata kunci?
- c. Seberapa baikah tingkat performansi yang bisa diperoleh dengan menggunakan Conditional Random Fields?

Sedangkan yang menjadi batasan masalah dalam tugas akhir ini diantaranya yaitu:

- a. File yang diinputkan hanya berupa dokumen teks.
- b. Dokumen teks yang digunakan dalam tugas akhir ini adalah artikel berbahasa Inggris.
- c. Dataset yang digunakan adalah koleksi dokumen dari INSPEC database yang didapat dari <https://github.com/snkim/AutomaticKeyphraseExtraction/blob/master/Hulth2003.tar.gz>
- d. Simulasi yang dibuat berbasis web menggunakan PHP dan basisdata MySQL.
- e. Tidak dilakukan stemming terhadap dokumen inputan.
- f. Aplikasi hanya melakukan *word* indexing dan tidak melakukan *phrase* indexing.

1.3 Tujuan

Adapun tujuan dari tugas akhir ini adalah :

- a. Mengimplementasikan Ekstraksi Kata Kunci dengan menggunakan Conditional Random Fields.
- b. Menganalisis Ekstraksi Kata Kunci dengan menggunakan Conditional Random Fields berdasarkan hasil pengukuran precision, recall dan F-measure.

1.4 Metodologi Penyelesaian Masalah

Metodologi penyelesaian masalah yang digunakan dalam menyelesaikan penelitian Tugas Akhir ini adalah:

- a. Studi literatur
Pencarian referensi dan sumber-sumber lain yang dapat digunakan sebagai acuan dalam pembuatan tugas akhir ini.
- b. Analisis dan Perancangan Perangkat Lunak
Pada tahap ini dilakukan proses analisis requirement perangkat lunak yang akan dibangun sehingga didapat gambaran mengenai sistem yang akan dibuat.
- c. Implementasi Sistem
Melakukan implementasi sistem sesuai dengan hasil analisis dan perancangan yang telah dilakukan di tahap dua. Tahap-tahap implementasinya antara lain :
 - Preprocessing.
Merubah dokumen teks tersebut menjadi *text chunking*. *Text chunking* adalah text yang telah disegmentasi dan diberi label *Part-of-speech* (POS). Kemudian melakukan ekstraksi fitur-fitur yang ada di dalam CRF model.
 - CRF model training.
Melakukan training data untuk menghasilkan file CRF model.
 - CRF labeling and keyword extraction.

- Menginputkan data testing kemudian melakukan ekstraksi kata kunci dengan menggunakan file CRF model yang telah dibuat.
- d. Analisis Hasil Implementasi
Menganalisis hasil implementasi aplikasi sehingga didapat data-data mengenai akurasi dari metode yang diimplementasikan.
 - e. Pembuatan Laporan
Merupakan tahapan pendokumentasian dari penelitian yang dikerjakan serta mengambil kesimpulan dari penelitian yang dikerjakan

1.5 Sistematika Penulisan

Tugas akhir ini disusun dengan sistematika penulisan sebagai berikut:

BAB I PENDAHULUAN

Berisi pemaparan mengenai latar belakang permasalahan, tujuan yang ingin dicapai dengan adanya penelitian ini, perumusan masalah, batasan masalah, metodologi tugas akhir, dan sistematika penulisan.

BAB II LANDASAN TEORI

Berisi uraian mengenai landasan teori yang akan digunakan, meliputi teori tentang Conditional Random Fields dan teori-teori lain yang berkaitan dengan penelitian tugas akhir ini

BAB III ANALISIS DAN PERANCANGAN SISTEM

Berisi tentang analisa dan perancangan terhadap sistem yang akan dibangun.

BAB IV IMPLEMENTASI DAN PENGUJIAN

Berisi implementasi dari hasil analisa dan perancangan sistem yang dibuat, serta pengujian sistem.

BAB V KESIMPULAN DAN SARAN

Berisi kesimpulan dan saran-saran untuk pengembangan lebih lanjut terhadap hasil penelitian ini.