BAB I

PENDAHULUAN

A. Latar Belakang

Dalam era teknologi informasi seperti saat ini, informasi berupa teks sudah tidak lagi selalu tersimpan dalam media cetak seperti kertas. Orang sudah mulai cenderung menyimpan informasi secara digital, karena lebih mudah dalam penyimpanan dan cepat. Tuntutan dari gerakan *anti global warming* juga mendukung penyimpanan informasi secara digital, baik untuk keperluan pribadi, perkantoran, atau bahkan pemerintahan. Selain itu, informasi sudah menjadi suatu komoditas yang dapat diperjualbelikan. Orang sampai kadang rela membayar banyak, baik dengan waktu maupun uang demi mendapatkan informasi yang mereka butuhkan, supaya mereka mendapatkannya dengan cepat. Tidak hanya kecepatan mendapatkan informasi, keakuratan dari informasi juga menjadi harapan semua orang.

Dengan banyaknya dokumen ber*form*at teks seperti jurnal, buku, dan berita yang sudah tersimpan secara digital, muncul permasalahan dimana informasi yang tadinya tersedia dengan baik menjadi kabur/hilang karena terlalu banyak dokumen/berkas yang tersimpan dalam media penyimpanan digital. Imbasnya, proses mencari informasi tertentu yang dibutuhkan dari berkas-berkas tersebut menjadi makin sulit dan lama.

Masalah lain akan terjadi ketika setiap dokumen tersebut ingin dikategorikan ke dalam kelas-kelas tertentu, karena harus dilihat, dibaca, dan dipahami isi tiap

dokumen dalam korpus. Setelah selesai membaca seluruh isi korpus, barulah bisa ditentukan kelas-kelas bagi dokumen dan membagi dokumen dalam kelas tersebut.

Sebuah kajian ilmu yang bernama *Information Retrieval* (IR) memunculkan beberapa metodologi yang memudahkan pencarian informasi dari sejumlah besar dokumen digital, salah satunya adalah dengan proses *clustering/classification*, yaitu pengelompokan data/berkas berbasis teks berdasarkan kemiripannya.

Beberapa metode *clustering* telah dikembangkan untuk mengelompokkan data terstruktur sejenis relational *database* seperti *k-Means*, *decision tree*, *Naïve Bayes*, dan sebagainya. Salah satu metode *clustering*, *k-Means*, terkenal simpel dan cepat dalam perhitungannya (Arthur, 2006), serta menjadi dasar pengembangan metode *clustering* yang lain (Kanungo, 2002; Bhatia, 2004; Pham, 2004, Mahdavi, 2008; Tarpey 2007). Metode *k-Means* yang dipadukan dengan pembobotan TF-IDF menjadi solusi untuk pengelompokan data tak terstruktur seperti dokumen teks secara otomatis. Karena TF-IDF sendiri juga merupakan metode yang populer dan memiliki hasil perhitungan yang cukup akurat (Ramos, 2010).

Penelitian ini bermaksud menggabungkan dan mengevaluasi kinerja perpaduan metode *k-Means* dan TF-IDF dalam proses *clustering* dokumen teks ke dalam suatu aplikasi *clustering* dokumen teks digital. Aplikasi ini diharapkan mampu melakukan klasifikasi secara otomatis bagi dokumen-dokumen dalam korpus.

B. Perumusan Masalah

Beberapa permasalahan yang dibahas dalam penelitian yang diajukan ini adalah:

1. Bagaimana implementasi proses pengelompokan (*clustering*) dokumen berformat teks yang dilakukan oleh metode *k-Means* yang dipadukan dengan pembobotan TF-IDF.
2. Bagaimana aplikasi mampu mengelompokkan dokumen berita dengan akurat berdasarkan kemiripan antar dokumen.
3. Penentuan nilai *threshold* yang paling cocok untuk aplikasi *clustering* dokumen teks dengan metode *k-Means*.
4. Evaluasi akurasi dari metode *k-Means clustering* menggunakan parameter *precision* dan *recall*.

C. Batasan Masalah

Beberapa batasan masalah yang ditetapkan dalam penelitian ini adalah:

1. Dokumen teks uji dalam penelitian ini terbatas hanya berekstensi *.txt* (*plain text*)
2. Proses *indexing* dokumen tidak melalui proses *stemming* dan normalisasi karena diharapkan aplikasi yang dibangun memiliki *precision* lebih tinggi.
3. Proses penghilangan *stopword* akan dilakukan untuk mengurangi kata-kata yang tidak mempengaruhi isi dokumen.

D. Keaslian Penelitian

Setelah dilakukan beberapa pengamatan terhadap beberapa buku, artikel, dan jurnal ilmiah yang sudah ada sebelumnya, belum ditemukan adanya penelitian yang secara khusus membahas tentang pengintegrasian pembobotan TF-IDF pada metode *clustering k-Means* pada aplikasi *clustering* dokumen teks.

E. Manfaat yang Diharapkan

Adapun manfaat-manfaat yang diharapkan dari penelitian yang dilakukan ini adalah:

1. Membuktikan bahwa metode *k-Means* yang dipadukan dengan pembobotan TF-IDF dapat digunakan untuk *clustering* pada dokumen tak terstruktur seperti dokumen teks.
2. Memberikan nilai *threshold* yang baik untuk penggunaan metode *k-Means* dengan pembobotan TF-IDF.
3. Memberikan informasi mengenai akurasi penggunaan metode *k-Means* dengan pembobotan TF-IDF untuk *clustering* dokumen teks.

F. Tujuan Penelitian

Penelitian dalam tesis ini bertujuan untuk:

1. Menguji keakuratan metode *k-Means clustering* dengan pembobotan TF-IDF dalam *clustering* dokumen teks.

2. Mengembangkan perangkat lunak pengelompokan dokumen teks untuk membantu penemuan kembali informasi yang hilang karena data yang terlalu banyak.
3. Mengidentifikasi nilai *threshold* yang optimal untuk diterapkan dalam implementasi metode *k-Means clustering* dengan pembobotan TF-IDF.

G. Sistematika Penulisan

Penulisan bagian utama tesis ini dibagi menjadi lima bagian, yaitu pendahuluan, tinjauan pustaka, metodologi penelitian, hasil penelitian dan pembahasan, kesimpulan dan saran.

1. Pendahuluan

   Bagian ini memuat latar belakang, perumusan masalah, batasan masalah, keaslian penelitian, manfaat yang diharapkan, tujuan penelitian, dan sistematika penelitian.

2. Tinjauan Pustaka

   Berisi mengenai tinjauan pustaka, yakni hasil-hasil penelitian terdahulu, landasan teori tentang *k-Means clustering* dan pembobotan TF-IDF, serta hipotesis mengenai penelitian yang dilakukan.

3. Metodologi Penelitian

   Dalam metodologi penelitian terdapat uraian terinci tentang: bahan atau materi penelitian, alat, langkah-langkah penelitian yang dilakukan, analisis, perancangan perangkat lunak, implementasinya, dan pengujian perangkat lunak tersebut.

4. Hasil Penelitian dan Pembahasan

   Bagian ini memuat hasil penelitian dan pembahasan terpadu. Pembahasan berisi mengenai analisis yang dilakukan terhadap hasil yang diperoleh, ditinjau secara utuh, baik secara kualitatif, kuantitatif, maupun normatif.

5. Kesimpulan dan Saran

   Bagian ini berisi kesimpulan yang merupakan pernyataan singkat dari hasil penelitian dan pembahasan, serta pembuktian kebenaran hipotesis. Selain itu, dituliskan saran yang ditujukan pada peneliti dalam bidang sejenis, yang ingin melanjutkan atau mengembangkan penelitian yang sudah diselesaikan.