BAB I

PENDAHULUAN

1.1 Latar Belakang

Penerapan teknologi informasi di berbagai bidang telah berdampak pada kebutuhan akan informasi yang juga semakin meningkat. Sehingga jumlah informasi terus mengalami peningkatan secara signifikan setiap waktunya, tidak terkecuali informasi yang berbentuk teks atau dokumen teks. Dan sebagai pengguna informasi, melakukan pengunduhan informasi pada media internet dan pertukaran informasi adalah suatu hal yang biasa dilakukan. Namun, tanpa pengorganisasian dokumen secara berkala akan menyebabkan dokumen pada komputer terus bertambah jumlahnya dan mengalami penumpukan.

Dengan bertumpuknya dokumen teks tersebut, menyebabkan proses pencarian suatu dokumen menjadi sulit, sehingga sangat boros waktu dan tidak efisien. Selain itu untuk mencari dokumen dengan tema yang sama atau yang memiliki kesamaan isi sekalipun juga sulit. Salah satu solusi untuk masalah ini adalah teknik *clustering* dokumen atau yang dikenal juga dengan pengelompokkan dokumen. *Clustering* telah lama diterapkan untuk meningkatkan efektifitas dalam proses pencarian. Prinsip dari *clustering* adalah membagi suatu kumpulan dokumen menjadi beberapa kluster/kelompok dokumen, sehingga dokumen-dokumen yang memiliki tingkat kesamaan tinggi dapat berada pada satu kluster yang sama.

Ada banyak metode yang dapat digunakan dalam *clustering* dokumen, di antaranya metode *Single Linkage Clustering*, metode *Complete Linkage Clustering*, metode *Group Average Agglomerative Clustering*, metode *Centroid Clustering* (Christopher D. Manning dkk, 2009), metode *Bisecting K-Means* (Steinbach dkk, 2009), metode *K-Means* (Mc Queen, 1967) dan beberapa metode lainnya. Metode-metode ini dapat digunakan dalam melakukan *clustering* terhadap dokumen, termasuk dokumen teks.

Dari beberapa metode di atas, salah satu metode *clustering* yang sering digunakan untuk *clustering* dokumen adalah *K-Means* (Mc Queen, 1967). Metode

yang dikembangkan oleh Mc Queen ini memiliki kelebihan yakni mampu mengelompokkan dokumen dalam jumlah yang besar dengan waktu komputasi yang cepat dan efisien. Prinsip dari metode ini adalah mempartisi suatu koleksi dokumen menjadi beberapa kluster dan menentukan *centroid* (titik pusat awal) secara acak.

Pada penelitian ini, penulis mencoba melakukan *clustering* dokumen teks yang diawali dengan mengekstrak *fingerprint* dari masing-masing dokumen yang ada pada suatu koleksi dokumen dengan menggunakan algoritma *biword winnowing* (Muhammad Ridho, 2013), dimana *fingerprint* adalah kumpulan nilai-nilai *hash* yang kemudian dibandingkan untuk mendeteksi kesamaan isi dokumen. Lalu dilakukan proses pembentukan vector dokumen berdimensi frekuensi kemunculan *fingerprint* dari masing-masing dokumen tersebut, dan dilanjutkan dengan proses *clustering* dokumen menggunakan metode *K-Means*. Diharapkan dengan melakukan *clustering* dokumen teks berdasarkan *fingerprint* yang merupakan *feature*/ciri dari dokumen, dapat bekerja dengan baik dalam proses *clustering* dokumen teks, sehingga dapat menghasilkan kluster dengan dokumen-dokumen yang memiliki tingkat kesamaan yang tinggi di dalamnya.

1.2 Rumusan Masalah

Berdasarkan penjelasan pada latar belakang di atas, maka dapat diuraikan rumusan masalah pada penelitian ini, yaitu bagaimana mengelompokkan dokumen teks berdasarkan *fingerprint*-nya sebagai *feature* dari dokumen.

1.3 Batasan Masalah

Batasan masalah dalam tugas akhir ini yaitu :

1. Hanya menggunakan korpus/koleksi dokumen yang telah penulis sediakan sebagai bahan percobaan.
2. Jumlah kluster ditentukan oleh *user*.
3. Menggunakan fungsi *cosine similarity* untuk menentukan similaritas antar *centroid* dan dokumen.

1.4 Tujuan Penelitian

Tujuan yang ingin dicapai dalam pembuatan tugas akhir ini, yaitu melakukan *clustering* terhadap dokumen teks berdasarkan *fingerprint biword winnowing* dengan menggunakan metode *K-Means*.

1.5 Sistematika Penulisan

Berikut merupakan rencana susunan sistematika penulisan laporan tugas akhir yang akan dibuat. Sistematika laporan tugas akhir ini meliputi :

Bab I Pendahuluan

Bab ini merupakan bagian yang akan menjelaskan hal-hal seperti latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, dan sistematika penulisan laporan tugas akhir.

Bab II Landasan Teori

Pada bab ini dibahas mengenai pustaka atau literatur yang digunakan dalam pengerjaan tugas akhir.

Bab III Metodologi Penelitian

Bab ini berisi tentang tahapan-tahapan yang dilakukan untuk menyelesaikan permasalahan pada tugas akhir ini, seperti tahapan penelitian, identifikasi masalah, studi pustaka, analisa aplikasi, pengujian, kesimpulan dan saran.

Bab IV Analisa dan Perancangan

Bab ini berisi tentang analisa dari penelitian yang dilakukan dalam tugas akhir ini. Seperti analisa penggunaan *fingerprint* dokumen, analisa pembentukan *vector* dokumen, dan analisa *clustering* dengan menggunakan *K-Means*.

Bab V Implementasi Dan Pengujian

Bab ini berisi tentang implementasi dan pengujian dari *clustering* dokumen pada beberapa koleksi dokumen.

Bab VI Penutup

Bab ini berisi kesimpulan dan saran mengenai hasil analisa dan hasil pengujian yang telah dilakukan terhadap *clustering* dokumen teks.