1. Pendahuluan

1.1. **Latar Belakang**

Peningkatan jumlah dokumen dalam format teks yang cukup signifikan belakangan ini membuat proses pengelompokan dokumen (*document clustering*) menjadi penting. Pengelompokan dokumen bertujuan membagi dokumen kedalam beberapa kelompok (*cluster*) sehingga dokumen-dokumen yang mempunya tingkat kesamaan tinggi termasuk dalam *cluster* yang sama dan yang mempunyai mempunya kesamaan rendah termasuk dalam cluster yang berbeda.

Pada umumnya metode *clustering* teks menggunaan Algoritma *K-Means* [2]. *K-means* merupakan suatu metode memilih secara acak *k* buah data sebagai . Kemudian menempatkan data dalam *cluster* yang terdekat dihitung dari titik tengan *cluster* (*centroid*). Centroid baru ditentukan akan ditentukan bila semua data telah ditempatkan dalam *cluster* terdekat. Proses penentuan *centroid* dan penentuan data dalam *cluster* diulangi sampe nilai centroid konvergen. Namun kelemahan K-means adalah pada saat jumlah data set banyak. Pada jumlah data set yang banyak membutuhkan banyak iterasi hingga mencapai nilai yang konvergen sehingga membutuhkan waktu yang lama. Untuk mengatasi kelemahan tersebut telah ditemukan metode *Canopy Clustering[1]*. Maka dalam tugas akhir ini menggunakan metode *Canopy Clustering* dengan *K-means* untuk teks clustering.

Metode *Canopy Clustering* ini sering digunakan untuk membagi data set yang besar memnajdi beberapa kelompok (canopy) sebelum dilakukan teknil *clustering* yang lebih ketat, seperti *K-means*. Penghitungan jarak terdekat dihitung pada data yang ada dalam tiap *canopy* menggunakan *Euclidean Distence[3]*. Setiap dokumen dapat berada dibeberapa *l* dan menimal 1 buah *canopy. Clustering* selanjutnya digunakan K-means untuk setiap *canopy*- nya.

1.2. **Perumusan Masalah**

Berdasarkan latar belakang yang telah diuraikan, dapat diambil rumusan masalah sebagai berikut :

1. Bagaimana melakukan *pengelompokan* dokumen dengan menggunakan algoritma *Canopy Clustering* dengan *Kmeans*?
2. Bagaimana waktu proses teks *clustering* dengan algoritma *K-means* setelah menggunakan metode Canopy Clustering ?
3. Bagaimana performasi Metode *Canopy K-means* pada system teks *clustering* berdasarkan *confusion matriks* dengan parameter *F-measure* ?

1.3. **Batasan Masalah**

Adapun yang menjadi batasan-batasan masalah dalam pengerjaan tugas akhir ini adalah:

1. Tugas akhir ini menggunakan dokumen berita berbahasa Indonesia yang telah mengalami *text preprocessing* terlebih dahulu.

2. Tidak melakukan analisa terhadap *text preprocessing*.
3. Dokumen didapat dari artikel berita Berbahasa Indonesia.

1.4. **Tujuan**

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

1. Mengimplementasikan algoritma *Canopy Clustering* dalam mengelompokkan dokumen.
2. Mengetahui dan menganalisis waktu proses peng-*cluster* menggunakan Algoritma *K-means* setelah menggunakan metode *Canopy clutering*.
3. Menganalisa hasil implementasi metode *Canopy K-means* dalam teks *clustering* berparameter *F-measure* dilihat dari *precison* dan *recall*.

1.5. **Metodologi Penyelesaian**

Metodologi yang digunakan untuk menyelesaikan permasalahan dalam Tugas Akhir ini terdiri dari :

1. Studi literatur
Pencarian referensi dan sumber-sumber yang berhubungan dengan teks *clustering*.

2. Tahap pengumpulan data dan *preprocessing* data
Pada tahap ini akan dibangun model berkaitan dengan requirement-requirement yang dibutuhkan pada saat implementasi, mulai dari analisis kebutuhan, desain database, desain aplikasi/*interface*.

3. Tahap perancangan sistem
Pada tahap ini akan dibangun model berkaitan dengan requirement-requirement yang dibutuhkan pada saat implementasi, mulai dari analisis kebutuhan, desain database, desain aplikasi/*interface*.

4. Tahap Implementasi
Pada tahap ini, akan dilakukan implementasi sistem yang mampu mengklasifikasikan dokumen teks secara otomatis dengan menggunakan metode *K-means dan Canopy Clustering*.

5. Tahap Pengujian Sistem
Pada tahap ini akan dilakukan pengujian program yang telah diimplementasikan dengan melakukan memasukkan dokumen teks lalu melihat hasil klasifikasi.

6. Tahap Analisis Hasil Pengujian
Setelah sistem telah sempurna maka akan dianalisis hasil dari klasifikasi dan kestabilan dari sistem.

7. Tahap Pembuatan Laporan

Pada tahap ini, akan dilakukan penyusunan laporan akhir dan pengumpulan dokumentasi dengan mengikuti kaidah penulisan yang benar dan sesuai dengan ketentuan-ketentuan atau sistematika yang telah ditetapkan oleh institusi.