Abstrak

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. Untuk melakukan pengelompokan tersebut, digunakan salah satu algoritma *clustering* yaitu Canopy Clustering. Canopy Clustering merupakan pengembangan dari K-means clustering. Algoritma ini dapat mengatasi permasalah yang terdapat pada K-means dalam masalah akurasi dan waktu proses untuk set data yang besar. Clustering dari nilai parameter T. Parameter ini berfungsi sebagai ukuran cluster pada pembentukan Canopy. Untuk mengukur similarity antar dokumen sebelum proses clustering digunakan Euclidean distance.

Pada tugas akhir ini cluster yang dihasilkan diukur akurasinya menggunakan precision, recall, dan F1-measure . Berdasarkan percobaan yang dilakukan bahwa *Canopy Clustering* dengan menggunakan K-means lebih tinggi tingkat akurasinya dan lebih sedikit waktu prosesnya dibandingkan dengan Algoritma *K-means* murni.

Kata kunci: Canopy Clustering, K-means , Clustering