BAB I
PENDAHULUAN

1.1 Latar Belakang

Data yang diperoleh dari Prodi Informatika Universitas Sahid Surakarta selama periode tahun 2004-2021 telah meluluskan 329 alumni. Dengan jumlah alumni tersebut topik Skripsi dan Tugas Akhir yang dibuat beragam. Selama ini Skripsi mahasiswa ditata berdasarkan periode wisuda tanpa memperhatikan topik dari Skripsi dan Tugas Akhir, sehingga mahasiswa sering kesulitan menentukan topik karena keterbatasan pustaka dan dalam mencari daftar pustaka dari penelitian sebelumnya mahasiswa harus memilih satu per satu secara manual di perpustakaan.

Hal ini juga berpengaruh pada data yang digunakan dalam penelitian ini, karena keterbatasan pustaka maka data yang digunakan dalam penelitian ini terbatas pada Skripsi dan Tugas Akhir Prodi Informatika Universitas Sahid Surakarta dengan periode wisuda ke-21 sampai wisuda ke-28 dengan jumlah 143 data. Berdasarkan pemaparan diatas maka perlu adanya metode khusus yang dapat mengelompokkan topik Tugas Akhir dan Skripsi Prodi Informatika Universitas Sahid Surakarta.

Penelitian Skripsi dan Tugas Akhir dapat dikelompokkan berdasarkan kemiripan topik, objek maupun metode penelitian. Pengelompokan data penelitian berbentuk teks, dapat dilakukan dengan *text mining*. *Text mining* atau bisa juga disebut *Text Data Mining* (TDM) merupakan proses ekstrasi informasi dari dokumen-dokumen teks yang tidak terstruktur (*unstructured*). *Text mining* dapat digunakan untuk klasifikasi, *information extraction*, *information retrival* dan *clustering* dalam penelitian menggunakan *clustering*.

Clustering digunakan untuk mengolah data dalam jumlah banyak dan mengelompokkan dokumen yang belum terstruktur dengan baik. Salah satu teknik *clustering* yang sering digunakan untuk pengelompokkan data adalah *K-means clustering* karena algoritma ini cukup sederhana. Metode *K-Means* mampu mengelompokkan data dengan jumlah yang cukup besar dan waktu komputasi yang cukup singkat, namun hasilnya sangat bergantung dengan pusat awal *cluster*. Masalah tersebut dapat diatasi dengan mengkombinasikan dengan *Hierarchical*

clustering. Maksud dari kombinasi antara *Hierarchical clustering* dan *K-Means clustering* agar mendapatkan hasil *clustering* yang lebih baik. Jumlah dan pusat *cluster* akan ditentukan melalui *Hierarchical clustering*, kemudian *K-Means clustering* akan mengoptimalkan posisi *centroid* dengan melakukan hitungan berulang pada *centroid* dari tiap *cluster* sampai nilai *centroid* stabil atau batas *iterasi* tercapai.

1.2 Rumusan Masalah

Dari latar belakang yang telah disampaikan di atas, maka rumusan masalah yang dibahas dalam laporan ini adalah Bagaimana mengelompokkan topik skripsi dan tugas akhir agar mudah dalam pencarian pustaka menggunakan metode *Hierarchical k-means*?

1.3 Batasan Masalah

Batasan masalah dari penelitian ini meliputi :

1. Data yang digunakan adalah abstrak berbahasa Indonesia pada tugas akhir dan skripsi prodi Informatika
2. Periode data yang digunakan adalah lulusan wisuda ke-21 sampai ke-28 Universitas Sahid Surakarta dengan jumlah data 143.
3. Metode yang dipakai adalah *Hierarchical k-means*.
4. Evaluasi hasil *clustering* menggunakan *inter cluster similarity* dan *intra cluster similarity*.
5. Menggunakan bahasa pemrograman Phyton.

1.4 Tujuan Dan Manfaat

1.4.1 Tujuan

Mengelompokkan topik penelitian tugas akhir dan skripsi mahasiswa prodi Informatika Universitas Sahid Surakarta berdasarkan abstrak yang berbahasa Indonesia.

1.4.2 Manfaat

a. Bagi penulis

1. Sebagai syarat memenuhi tugas akhir.
2. Penulis dapat menerapkan ilmu pengetahuan yang telah didapat dari bangku perkuliahan untuk dapat mengelompokkan laporan penelitian tugas akhir dan skripsi mahasiswa prodi Informatika Universitas Sahid Surakarta.

b. Bagi Universitas Sahid Surakarta

Dapat mengetahui berbagai topik penelitian yang terbentuk berdasarkan hasil *clustering*.

c. Bagi Prodi

1. Hasil *clustering* bisa dijadikan sebagai acuan untuk meningkatkan kualitas pembelajaran prodi Informatika.
2. Memudahkan dalam pencarian pustaka dan menambah referensi penelitian.

1.5 Metode Penelitian

Penelitian ini menggunakan metode *Cross Industry Standard Process Model for Data Mining* (CRISP-DM), dengan tahapan sebagai berikut :

1. *Business Understanding*

   Penelitian ini dilakukan untuk membantu mahasiswa agar tidak kesulitan menentukan topik karena keterbatasan pustaka dan agar mahasiswa tidak mencari satu per satu daftar pustaka secara manual di perpustakaan. Situasi yang terjadi pada penelitian ini adalah semua skripsi dan tugas akhir program studi Informatika masih bercampur, tanpa ada pembeda antara topik satu dengan topik lainnya.

   Tujuan dari data mining adalah mengelompokkan topik tugas akhir dan skripsi program studi informatika Universitas Sahid Surakarta.

2. *Data Understanding*

   Pada tahap ini yang dilakukan adalah mengumpulkan data awal, mempelajari data untuk bisa mengenal data yang akan dipakai. Pengumpulan data dilakukan dengan wawancara dan observasi. Data diambil dari perpustakaan Universitas Sahid Surakarta melalui petugas perpustakaan. Data

yang diminta adalah tugas akhir dan skripsi prodi Informatika Universitas Sahid Surakarta periode wisuda ke-21 sampai ke-28 dengan abstrak berbahasa Indonesia.

3. *Data Preparation*

Data yang telah didapatkan diolah terlebih dahulu dengan menyiapkan data. Ada 2 tahap dalam proses persiapan data :

a. *Preprocessing data*

Data tugas akhir mahasiswa diolah dalam proses *preprocessing* dengan tujuan untuk memastikam data yang akan diolah pada proses selanjutnya adalah data yang baik. *Preprocessing* terdiri dari lima proses, yaitu *tokenization*, *case folding, filtering*, *normalisasi*, *stemming*, dan *stopword*.

b. TF-IDF (*Term Frequency-Inverse Document Frequency)*

Data skripsi dan tugas akhir yang sudah melalui proses *preprocessing*, kemudian dilakukan proses pembobotan *term* (kata). Pembobotan *term* menggunakan hasil dari proses *preprocessing*. Metode TF-IDF ini akan menghitung nilai TF (*Term Frequency*) dan nilai IDF (*Inverse Document Frequency*) pada setiap *term*. Hasil pembobotan *term* digunakan untuk menghitung bobot setiap dokumen dengan cara menjumlahkan semua bobot *term* pada suatu dokumen

4. *Modelling*

Tahap menentukan teknik data mining yang digunakan, pada penelitian ini menggunakan *clustering* dan pengerjaannya menggunakan bahasa pemrograman Python.

Proses *clustering* menggunakan kombinasi antara dua metode, yaitu *hierarchical clustering* dan *k-means clustering*. Proses dalam metode *hierarchical clustering* ini menggunakan bobot dari setiap dokumen. Hasil dari *hierarchical clustering* ini berupa dendogram, di mana akan dipotong seuai dengan batas *threshold* yang telah dipilih. Batas *threshold* tersebut diperoleh dengan mempertimbangkan keterkaitan antar dokumen. Hasil dari pemotongan dendogram yang berupa *cluster*, akan digunakan dalam algoritma k-means

untuk proses selanjutnya, yaitu mengelompokkan dokumen ke dalam beberapa *cluster*.

5. *Evaluation*

   Melakukan pengukuran tingkat akurasi pada model data yang dihasilkan pada tahap evaluasi. Evaluasi keakuratan pada penelitian ini menggunakan *Intra Cluster Similarity dan Inter Cluster Similarity.*

6. *Deployment*

   Tahap ini merupakan tahap penarikan kesimpulan dari hasil *clustering* data dan analisis hasil *clustering*. Selain itu juga akan dilakukan penulisan laporan dari hasil tahap evaluation sebagai bahan rekomendasi mengenai *clustering* dokumen laporan tugas akhir dan skripsi.

1.6 Sistematika Penulisan

Sistematika penulisan ini digunakan untuk mempermudah dalam mengetahui dan memahami isi atau uraian dari tiap-tiap bab penulisan. Sistematika penulisan ini terbagi dalam lima bab pembahasan yang dijelaskan secara singkat sebagai berikut:

BAB I PENDAHULUAN

Pada bab ini menjelaskan tentang latar belakang, perumusan masalah, batasan masalah, tujuan dan manfaat, metodologi penelitian dan sistematika penulisan.

BAB II LANDASAN TEORI

Pada bab ini berisi mengenai tinjuan pustaka, kerangka pemikiran dan teori pendukung yang dijadikan sebagai landasan landasan dalam penelitian.

BAB III METODE PENELITIAN

Menguraikan tentang gambaran objek penelitian, proses pengumpulan data, serta gambaran langkah-langkah yang dilakukan oleh penulis untuk melaksanakan dan menyelesaikan penelitian ini.

BAB IV HASIL DAN PEMBAHASAN

Bab ini berisi tentang bagaimana menyelesaikan masalah yang telah dirumuskan berdasarkan metode yang dipilih dan berusaha untuk mewujudkan tujuan, serta manfaat yang ingin diraih.

BAB V SIMPULAN DAN SARAN

Pada bab ini memuat kesimpulan serta saran-saran untuk melengkapi dan menyempurnakan penyusunan sekaligus akhir dari laporan skripsi.

DAFTAR PUSTAKA

LAMPIRAN