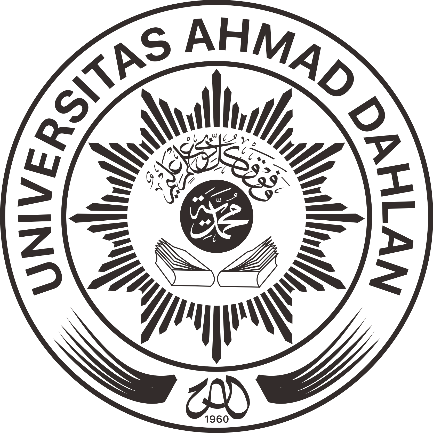
HALAMAN JUDUL **Dosen Pembimbing : Lisna Zahrotun, S.T., M.Cs.**

**PENGELOMPOKAN JUDUL PENELITIAN DOSEN MENGGUNAKAN METODE *SUFFIX TREE CLUSTERING***

**PROPOSAL SKRIPSI**

**Untuk Memenuhi Sebagian Persyaratan   
Mencapai Derajat Sarjana Komputer**

****

**Disusun Oleh:**

**Nova Anggraini**

**1500018196**

**PROGRAM STUDI TEKNIK INFORMATIKA**

**FAKULTAS TEKNOLOGI INDUSTRI**

**UNIVERSITAS AHMAD DAHLAN**

**YOGYAKARTA**

**2018**

**DAFTAR ISI**

Contents

[**HALAMAN JUDUL** i](#_Toc529822958)

[**DAFTAR TABEL** iii](#_Toc529822959)

[**DAFTAR GAMBAR** 4](#_Toc529822960)

[**DAFTAR LISTING** 5](#_Toc529822961)s

[**BAB I PENDAHULUAN** 1](#_Toc529822962)

[**1.1 Latar Belakang Masalah** 1](#_Toc529822963)

[**1.2 Identifikasi Masalah** 3](#_Toc529822964)

[**1.3 Batasan** **Masalah** 4](#_Toc529822965)

[**1.4 Rumusan Masalah** 4](#_Toc529822966)

[**1.5 Tujuan Penelitian** 5](#_Toc529822967)

[**1.** **6 Manfaat Penelitian** 5](#_Toc529822968)

[**BAB II KAJIAN PUSTAKA** 5](#_Toc529822969)

[**2.1 Kajian Hasil Penelitian Terdahulu** 6](#_Toc529822970)

[**2. 2 Landasan Teori** 1](#_Toc529822971)

[**2.2.1 Clustering** 1](#_Toc529822972)

[**2.2.2 Riset** 1](#_Toc529822973)

[**2.2.3 Python** 2](#_Toc529822974)

[**2.2.4 Text Mining** 4](#_Toc529822975)

[**2.2.5 Metode Suffix Tree Clustering** 5](#_Toc529822976)

[**BAB III METODELOGI PENELITIAN** 7](#_Toc529822977)

[**DAFTAR PUSTAKA** 1](#_Toc529822978)

# **DAFTAR TABEL**

**Tabel 1 1.: Tabel Penelitian Dosen**…………………………………………………………………………………2

# **DAFTAR GAMBAR**

# **DAFTAR LISTING**

# **BAB I**

**PENDAHULUAN**

## **1.1 Latar Belakang Masalah**

Universitas Ahmad Dahlan (UAD) adalah salah satu Perguruan Tinggi Muhammadiyah yang berada di Provinsi Daerah Istimewa Yogyakarta. Universitas Ahmad Dahlan memiliki kantor Lembaga Penelitian dan Pengembangan (LPP). LPP UAD merupakan unsur pelaksana tingkat universitas yang mempunyai tugas mengkoordinasikan, memonitor pelaksanaan kegiatan penelitian, menyelenggarakan kolokium hasil penelitian, dan mengembangkan bidang penelitian yang dilakukan oleh dosen-dosen UAD serta pusat-pusat studi maupun oleh Pusat Pengembangan UAD. Pengembangan lain yang dilakukan adalah pengendalian penggunaan sumber daya, pengembangan dan peningkatan mutu penelitian, mempublikasikan hasil-hasil penelitian, dan mengembangkan *output* penelitian sampai menghasilkan produk yang bernilai. Dalam peningkatan kualitas peneliti, LPP secara berkala melaksanakan kegiatan *upgrade* atau peningkatan kapasitas meneliti dan menyusun publikasi. Selain itu, LPP juga memberikan pelayanan informasi kegiatan penelitian baik yang dikelola didalam maupun lembaga diluar UAD baik dari lembaga pemerintah, swasta maupun masyarakat. Sebagai lembaga di Perguruan Tinggi Muhammadiyah, LPP juga melaksanakan kegiatan-kegiatan penelitian kebijakan yang bersifat pengembangan institusional dan organisasi otonom (ortom) Persyarikatan Muhammadiyah (Widodo, Rustiawan and Dkk, 2016).

Sampai saat ini yang menjadi sarana untuk mempublikasikan penelitian dosen ditangani langsung oleh Lembaga Penelitian dan Pengembangan. Berdasarkan hasil wawancara dengan Bapak Drh. Asep Rustiawan, M.S selaku Sekretaris Lembaga Penelitian dan Pengembangan (LPP), LPP telah memiliki sistem yang dapat dilihat oleh pihak dosen untuk mengetahui hasil penelitiannya. Pengarsipan judul penelitian dosen dilakukan pada dua tempat, yaitu di *website* portal.uad.ac.id dalam bentuk *database* dan di Kantor Lembaga Penelitian dan Pengembangan (LPP) dalam bentuk *file excel.* Pengolahan yang dilakukan Lembaga Penelitian dan Pengembangan UAD hanya menyimpan dan mempublikasikan belum mengelompokkannya. LPP tidak mengetahui secara akurat judul-judul penelitian dosen dikarnakan dari pihak LPP belum mengelompokkan judul-judul penelitian dosen berdasarkan bidang penelitian dan LPP juga belum mengetahui dosen-dosen yang telah melakukan penelitian dengan judul yang sama. Data-data judul penelitian dosen yang telah ada dapat diidentifikasi kemiripan judul penelitiannya yang dihasilkan dari pengelompokan judul penelitian dosen publikasi LPP. Sehingga setiap dosen yang akan melakukan penelitian bisa mencari judul atau topik penelitian berdasarkan judul-judul penelitian sesuai kategori penelitian tahun-tahun sebelumnya. Untuk menentukan judul penelitian maka perlu dilakukan pengolahan data dengan cara mengelompokkan judul penelitiannya. Data penelitian dosen yang diperoleh dari Lembaga Penelitian dan Pengembangan (LPP) UAD dari tahun 2015-2017 sebanyak 623 judul penelitian. Seperti pada Tabel 1.1. Penelitian Dosen.

**Tabel 1.1 : Tabel Penelitian Dosen**

|  |  |  |
| --- | --- | --- |
| No. | Tahun | Jumlah Penelitian Dosen |
| 1. | 2015 - 2016 | 268 |
| 2. | 2016 - 2017 | 355 |
| Total | | 623 |

Pengelompokan berdasarkan topik telah banyak dilakukan pada penelitian sebelumnya. Dalam pengembangan Algoritma *Suffix Tree Clustering* (STC) dapat digunakan pada pengelompokan dari hasil mesin pencari, selain itu STC juga sebagai *phrase* dasar dalam pembentukan *cluster* sehingga memungkinkan terjadinya *overlapping cluster*. Hal ini didasarkan bahwa setiap dokumen bisa memiliki lebih dari satu topik berdasarkan *phrase* yang terdapat pada dokumen tersebut. Kelebihan lain dari *Suffix Tree Clustering* adalah menyimpan semua *phrase* yang adasecara terstruktur untuk menunjukkan tingkat kemiripan dokumen (Tuwohingide *et al.*, 2016).

Penelitian ini akan menerapkan metode *Suffix Tree Clustering* sehingga diharapkan dapat melakukan pengelompokan judul penelitian dosen yang lebih akurat.

## **1.2 Identifikasi Masalah**

Berdasarkan uraian dari latar belakang masalah diatas, maka didapatkan beberapa identifikasi masalah sebagai berikut:

1. Kesulitan dalam mencari judul penelitian terdahulu berdasarkan kategori bidang penelitian.
2. Sistem yang ada baru sampai tahap menyimpan dan mempublikasikan.

## **1.3 Batasan** **Masalah**

Agar pembahasan tidak terlalu meluas, perlu adanya pembatasan masalah sebagai berikut:

1. Data yang digunakan dalam penelitian ini adalah data yang diambil dari Lembaga Penelitian dan Pengembangan UAD yaitu data judul penelitian dosen UAD dari tahun 2015 hingga tahun 2017.
2. Pengelompokan judul penelitian dosen tersimpan dalam bentuk file *excel* dan pengolahan yang dilakukan dalam penelitian ini hanya mengelompokan judul penelitian berdasarkan bidang penelitiannya.

## **1.4 Rumusan Masalah**

Berdasarkan uraian latar belakang masalah diatas, maka diperoleh rumusan masalah. Adapun rumusan masalah dalam penelitian ini, antara lain yaitu:

1. Bagaimana menerapkan metode *Suffix Tree Clustering* untuk mengelompokkan judul penelitian dosen berdasarkan bidang penelitiannya.
2. Melakukan evaluasi dengan menghitung keakuratan hasil pengelompokan STC dengan data judul penelitian dosen UAD.

## **1.5 Tujuan Penelitian**

Sesuai dengan rumusan masalah yang telah diuraikan diatas, maka penelitian ini mempunyai tujuan yaitu:

1. Menerapkan metode *Suffix Tree Clustering* untuk pengelompokan judul penelitian dosen.
2. Menguji metoode *Suffix Tree Clustering* untuk mengetahui nilai akurasi dalam mengelompokan judul penelitian dosen.

## **1.** **6 Manfaat Penelitian**

Manfaat yang ingin diperoleh dari penelitian ini adalah:

1. Dapat membantu pihak LPP dalam melakukan proses pengelompokan judul penelitian dosen.
2. Sebagai pengetahuan bagaimana cara kerja metode *Suffix Tree Clustering* dalam mengelompokkan judul penelitian dosen.

# **BAB II**

**KAJIAN PUSTAKA**

## **2.1 Kajian Hasil Penelitian Terdahulu**

Pada bagian ini akan dipaparkan mengenai kajian terdahulu yang disajikan sebagai bahan acuan dasar dalam penelitian. Selain kajian terdahulu, akan dijelaskan tentang kajian teori yang mendukung penelitian ini.

Penelitian ini membahas mengenai *Suffix Tree Clustering* dengan *WordNet* dan *Named Entity Recognition* untuk pengelompokan dokumen dilakukan oleh (Wijoyo *et al.*, 2017). Penelitian tersebut mengembangkan metode baru STC terhadap pengelompokan dokumen. Metode yang dikembangkan memiliki beberapa tahapan, yaitu preprocessing dokumen dengan mengekstraksi *named entity* serta melakukan deteksi sinonim berdasarkan *WordNet.* Tahap kedua adalah pembobotan *term* dengan tfidf dan neridf. Tahap ketiga adalah melakukan *clustering* dokumen dengan menggunakan *Suffix Tree Clustering*. Metode STC mampu membuktikan dapat meningkatkan kualitas *clustering* dokumen menjadi lebih baik.

Penelitian yang dilakukan oleh (Sutadi Heru, 2016) membahas mengenai metode *Suffix Tree Clustering* yang dilakukan terhadap pengelompokan dokumen *history* mesin pencari. Dalam penelitian tersebut menghasilkan pengelompokan dokumen untuk dilakukan klasifikasi data, sehingga dapat membuktikan bahwa unsur penting dalam pengolahan data terutama data – data mentah yang diinginkan untuk suatu bentuk pola – pola kumpulan data. Akan tetapi pada penelitian ini belum dapat menangani lebih dari satu Bahasa untuk pengelompokan hasil pencarian.

Penelitian yang dilakukan oleh (Tuwohingide *et al.*, 2016) juga membahas metode *Suffix Tree Clustering* terhadap pengelompokan dokumen web berbahasa indonesia. Dalam penelitian ini bertujuan untuk mengelompokan dokumen web berbahasa Indonesia berdasarkan *pharse* pada proses pemilihan *base cluster*. Metode yang digunakan pada penelitian ini mampu meningkatkan hasil pengelompokan dokumen yang semula menggunakan *scoring pharse* dengan kombinasi *document frequency* dan panjang *pharse* yang terlabeli pada STC tradisional. Metode tersebut juga berhasil mempertahan kan *overlapping cluster* yang merupakan kelebiha STC.

Hal tersebut juga dikuatkan dengan yang dilakukan oleh (Tri, Soedijono and Armadyah, 2015) bahwa metode STC dapat diterapkan pengelompokan abstrak skripsi. Hal tersebut dibuktikan dengan pengujian metode *Suffix Tree Clustering* dan *Singular Value Decomposite* yang dapat digunakan dalam abstrak menggunakan kata kunci menjadi tiga kelompok yaitu Komputer Akutansi, *Ebusines*, dan Sistem Informasi, hal tersebut mendapatkan nilai keakurasian sebesar 87,67%.

Setelah dipaparkan mengenai kajian terdahulu yang disajikan sebagai bahan acuan dalam penelitian maka menurut penelitian (Wijoyo *et al.*, 2017), (Sutadi Heru, 2016), (Tuwohingide *et al.*, 2016), (Tri, Soedijono and Armadyah, 2015) yang mampu memaparkan tentang metode *suffix tree clustering* (STC), memaparkan bahwa metode algoritma STC dapat digunakan dalam menentukan pola yang terbukti dapat membantu proses pengelompokan data serta dapat melakukan klasifikasi data sesuai dengan kategori atau jenis data yang digunakan, selain itu metode STC dapat menghasilkan nilai akurasi yang cukup tinggi dalam melakukan pengelompokan data yang sesuai dengan klasifikasi data tersebut. Dari kesimpulan penelitian metode STC, maka topik yang akan dikembangkan dalam penelitian ini menggunakan metode STC sebagai penerapan dalam mengelompokan judul penelitian dosen di ruang lingkup Universitas Ahmad Dahlan.

Tabel 2.1 Tabel review jurnal

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Penelitian** | **Peneitian 1** | **Penelitian 2** | **Penelitian 3** | **Penelitian 4** | **Yang diteliti** |
| Penulis | Satrio Hadi Wijoyo1, Admaja Dwi Herlambang2, Fahrur Rozi3, Septiyan Andika Isanta4 | Heru Sutadi | Desmin Tuwohingide1, Mika Parwita2, Agus Zainal Arifin3, Diana Purwitasari4 | Lina Tri Andaru1, Bambang Soedijono W2, Armadyah Amborowati3 | Nova Anggraini |
| Judul Naskah | Optimasi *Suffix Tree Clustering* dengan *Wordnet* dan *Named Entity Recognition* untuk Pengelompokan Dokumen | Implementasi *Suffix Tree Clustering* untuk Pengelompokan Dokumen yang telah di Akses Melalui Mesin Pencarian Google | Efisiensi *Phrase Suffix Tree* dengan *Single Pass Clustering* untuk Pengelompokan Dokumen Web Berbahasa Indonesia | Pengelompokan Abstrak Skripsi Menggunakan Metode *Suffix Tree Clustering* dan *Singular Value Decomposition* | Pengelompokan Judul Penelitian Dosen Menggunakan Metode *Suffix Tree Clustering* |
| Nama Jurnal | Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) | Jurnal Ilmiah INFOTEK | JURNAL TEKNOLOGI TECHNOSCIENTIA | Seminar Nasional Teknologi Informasi dan Multimedia | - |
| ISSN | p : 2355-7699  e : 2528-6579 | 2502-6968  (Media Cetak) | 1979-8415 | 2302-3805 | - |
| Volume | 4 | 1 | 8 | - | - |
| Nomor | 4 | 1 | 2 | - | - |
| Tahun Penelitian | 2017 | 2016 | 2016 | 2015 | 2018 |
| Metode | *Suffix Tree Clustering* dengan *Wordnet* dan *Named Entity Recognition* | *Suffix Tree Clustering* | *Suffix Tree* dengan *Single Pass Clustering* | *Suffix Tree Clustering* dan *Singular Value Decomposition* | *Suffix Tree Clustering* |
| Kesimpulan | Penelitian ini menunjukkan bahwa metode yang diusulkan dapat melakukan pengelompokan dokumen dengan sangat baik. WordNet (synonym detction) dapat mendeteksi kata yang beda penulisan tapi sama makna. NER dapat mendeteksi dokumen yang mempunyai entitas. Selain itu, WordNet dan NER dapat digunakan untuk optimasi clustering dokumen menggunakan STC. | 1. Pengelompokan tree clustering terdiri dari beberapa prosedur diantara nya, Documen cleaning, steming data, identifikasi base cluster, dan pengkombinasian base cluster  2. Untuk menerapka metode STC ada beberapa langkah di antara nya dengan pencarian shared phrase untuk semua berita yang dikoleksi,dan mengkobinasikan dua base cluster kedalam satu cluster  3. Bahasa pemrograman yang di bangun dalam pengelompokan metode STC yaitu visual basic 2008. | Metode yang diusulkan menghasilkan nilai F-Measure yang lebih tinggi dibandingkan dengan STC tradisional. Efisiensi phrase pada proses pemilihan base cluster dengan kombinasi perhitungan tf dan df mampu meningkatkan hasil pengelompokan dokumen yang semula menggunakan scoring phrase dengan kombinasi df dan panjang phrase yang terlabeli pada STC tradisional. Metode yang diusulkan juga berhasil mempertahankan overlapping cluster yang merupakan kelebihan STC. | Akurasi yang sesuai diperoleh pada 100 data abstrak dengan menggunakan reduksi dimensi 30% yaitu akurasi 87.67% dengan hasil 54 data pada kelompok Sistem Informasi Perusahaan/ Instansi , 14 data terkelompok dalam Ebusines dan 26 data terkelompok dalam Komputer akutansi (pengujian recall adalah 0.33 dan presisisebesar 0.37). | Data yang digunakan dalam penelitian ini adalah data Penelitian Dosen UAD tahun 2015 sampai tahun 2017. Hasil dari penelitian ini berupa sebuah sistem yang dapat menampilkan hasil pengelompokan judul penelitian dosen berdasarkan bidang penelitiannya dan diuji dengan tingkat akurasi. |

## **2. 2 Landasan Teori**

### **2.2.1 Pengelompokan**

Pengelompokan data menjadi sejumlah kategori juga dapat dilakukan menggunakan metode *clustering*. Berbeda dengan metode klasifikasi, *clustering* mengelompokkan data hanya berdasarkan fitur-fitur yang ada pada data tersebut. Berdasarkan sifat tersebut, *clustering* tidak memerlukan data latih yang telah diketahui kelasnya. Oleh karena itu, proses pembelajaran pada *clustering* bersifat mandiri, yang sering disebut dengan istilah ***unsupervised learning.***

Tujuan utama dari proses *clustering* adalah membagi sekumpulan data menjadi sekumpulan grup (*cluster)* sehingga data-data dalam satu *cluster* memiliki banyak kemikiripan namun berbeda dengan data-data yang berada pada *cluster* lainnya. Kemiripan antar data dapat dihitung menggunakan berbagai metode pengukuran jarak.

Teknik *clustering* memiliki banyak kegunaan dalam kehidupan sehari-hari. Mesin pencarian informasi, seperti Google dan Yahoo, dapat mengelompokkan hasil pencarian sehingga pengguna dapat menelusuri hasil pencarian dari kelompok yang relevan saja. Hal ini dapat mengurangi waktu pencarian dan memperbesar peluang mendapatkan hasil yang sesuai. (Adinugroho and Sari, 2018).

### **2.2.2 Penelitian**

Penelitian adalah suatu kegiatan ilmiah dalam memecahkan masalah dengan cara sistematis yang telah ditetapkan untuk mencapai tujuan yang telah dirumuskan.

Sebagian orang menganggap bahwa metode penelitian terdiri dari berbagai teknik penelitian, dan sebagian lagi menyamakan metode penelitian dengan teknik penelitian. Tetapi yang jelas, metode atau teknik penelitian apa pun yang kita gunakan, baik kuantitif ataupun kualitatif, haruslah sesuai dengan kerangka teoretis yang kita asumsikan.(Muh. Fitrah and Dr. Luthfiyah, 2017).

### **2.2.3 Python**

(Enterprise, 2016) menyatakan bahwa Python merupakan bahasa pemrograman yang bersifat interpretative. Dibandingkan dengan Bahasa pemrograman lainnya, Python termasuk mudah dipelajari karena merupakan salah satu Bahasa high level programming. Python dapat diperoleh dan dipergunakan secara bebas oleh siapapun bahkan bagi developer yang menggunakan Bahasa pemrograman Python untuk kepentingan komersial.

Python merupakan pemrograman multiplatform. Itu artinya, Python dapat dijalankan di berbagai platform sistem operasi yaitu Linux / Unix, Windows, Mac OS X, Java Virtual Machine, OS /2, Amiga, Palm, Symbian (untuk produk-produk Nokia). Dengan begitu, baik pengguna Macintosh, Windows, maupun Linux dapat membuat berbagai jenis aplikasi menggunakan Python tanpa masalah.

Python dikembangkan oleh Guido VanRossum (programmer kelahiran Belanda) pada tahun 1990 di CWI, Amsterdam sebagai kelanjutan dari bahasa pemrograman ABC. Nama Python dipilih oleh Guido sebagai nama bahasa ciptaannya karena kecintaan Guido pada acara televsi Monty Python’s Flying Circus. Saat ini perkembangan python terus dilakukan oleh sekelompok programmer yang dikoordinasi oleh Gudio dan Python Software Foundation. Sekarang, distribusi Python sudah mencapai versi 2.6.1 dan versi 3.0. beberapa fitur dan kelebihan yang dimiliki Python adalah :

1. Memiliki koleksi kepustakaan yang banyak. Itu artinya, telah tersedia modul-modul ‘siap pakai’ untuk berbagai keperluaan.
2. Memiliki struktur Bahasa yang jelas, sederhana, dan mudah dipelajari.
3. Berorientasi objek.
4. Memiliki sistem pengelolaan memori otomatis (garbage collection) seperti halnya Java.
5. Bersifat Modular sehingga mudah dikembangkan dengan menciptakan modul-modul baru, baik dibangun dengan Bahasa Python maupun C/C++.

### **2.2.4 Text Mining**

*Text mining* dapat didefinisikan secara luas sebagai proses pengetahuan intensif dimana pengguna berinteraksi dengan koleksi dokumen dari waktu ke waktu dengan menggunakan rangkaian alat analisis. Dengan cara yang serupa dengan *data mining*, *text mining* berusaha mengekstrak informasi yang bermanfaat dari sumber data melalui identifikasi dan eksplorasi pola yang menarik (Feldman dan Sanger, 2006).

Pada *text mining* hal yang harus diperhatikan yaitu pada pengumpulan dokumen yang dapat berupa pengelompokan dokumen berbasis *text*. Sebagian besar *text mining* ditujukan untuk menemukan pola pada koleksi dokumen yang berjumlah besar (Feldman dan Sanger, 2007). Langkah-langkah yang dilakukan dalam *text mining* yaitu :

* + 1. *Text Preprocessing*

Pada tahap ini yaitu pada *data cleaning* dan *cleansing*, termasuk menghilangkan *noise* pada data atau mengatasi data yang hilang (Adiwijaya, 2006). *Text preprocessing* pada dokumen merupakan tugas penting selama proses penambangan text seperti penyaringan dari data yang tidak valid (Katariya dan Chaudhari, 2015).

* + 1. *Feature Selection*

Pada tahap ini yaitu melakukan proses *tokenizing* yaitu proses penguraian deskripsi yang semula berupa kalimat-kalimat menjadi kata-kata dan menghilangkan delimiter-delimiter seperti tanda titik(.), koma(,), spasi dan karakter angka yang ada pada dokumen (Weiss *dkk.*, 2005). Kemudian melakukan proses *stopwords removal* yaitu menghilangkan kosakata yang bukan merupakan ciri atau kata unik dari suatu dokumen misalnya “di”, “oleh”, “pada”, “sebuah”, “karena” dan lain sebagainya (Kurniawan dkk, 2012). Setelah dilakukan proses *stopwords removal* kemudian tahap *stemming* yaitu proses pemetaan dan penguraian berbagai bentuk dari suatu kata menjadi bentuk kata dasarnya. Tujuannya adalah menghilangkan imbuhan-imbuhan baik itu berupa *prefix, suffix,* maupun *conflict* yang ada pada setiap kata.

### **2.2.5 Metode Suffix Tree Clustering**

Inti dari suatu hasil pencarian yang menerapkan clustering adalah penggunaan algoritma clustering. Algoritma Suffix Tree Clustering (STC) memiliki dua kunci utama, yaitu :

1. Menggunakan phrase sebagai dasar pembentukan clusternya.
2. Menggunakan suatu definisi cluster sederhana.

Suffix tree clustering memiliki dua langkah utama, langkah pertama adalah pencarian shared phrase untuk semua dokumen berita yang dikoleksi. Hal tersebut dapat disebut shared phrase sebagai phrase cluster atau base claster, yang ditemukan dengan menggunakan suatu struktur data yang dinamakan suffix tree. Dalam langkah kedua, kita mengkombinasikan base cluster-base antar dua base cluster ke dalam suatu cluster. Penggabungan antar dua base cluster didasarkan pada jumlah dokumen yang melakukan overlap diantara kedua base cluster tersebut. Suatu phrase yang dimaksud dalam konteks algoritma ini adalah urutan satu atau lebih kata. Suffix tree clustering memiliki empat langkah utama sumber (Arifin *et al.*, 2008) yaitu :

1. Pembersihan dokumen
2. Steaming data
3. Indentifikasi base cluster
4. Kombinasi base cluster

Bebrapa karakteristik yang membuat suffix tree clustering cocok digunakan untuk pengelompokkan dokumen. Pertama adalah mangenerate clustercluster untuk pengelompokkan dokumen berdasarkan phrase. Phrase juga bermanfaat untuk membangun uraian dan keakuratan deskripsi dari cluster-cluster. Kedua, tidak tergantung pada model data. Hal itu mengasumsikan hanya dokumendokumen dengan topik yang sama yang akan memiliki shared phrase. Ketiga, STC memperolehkan adanya overlaping cluster. Hal itu sangat penting untuk menghindari pembatasan bahwa setiap dokumen hanya memiliki satu cluster saja, karena setiap kita jumpai satu dokumen mempunyai lebih dari satu topik dan dengan begitu terdapat kemiripan yang lebih dari satu kelompok dokumen. Keempat, STC menggunakan defenisi cluster yang sederhana. Semua dokumen yang berisi salah satu phrase cluster akan menjadi anggota dari cluster tersebut (Arifin *et al.*, 2008).

STC menggunakan phrase untuk mendeteksi kemiripan antar dokumen. STC menggunakan suffix tree untuk mengidentifikasi phrase. Fitur yang membuat suksesnya STC sebagai algoritma clustering adalah adanya overlaping cluster. Kualitas cluster yang terbentuk dari algoritma STC ini akan menurun jika tanpa menggunakan multiword phrase dan tidak memperbolehkan adanya overlaping cluster.

# **BAB III**

**METODELOGI PENELITIAN**

**3.1 Objek Penelitian**

Penelitian “Pengelompokan Judul Penelitian Dosen Menggunakan Metode *Suffix Tree Clustering*” mengambil obyek penelitian dari Lembaga Penelitian dan Pengembangan Universitas Ahmad Dahlan berupa data judul Penelitian Dosen UAD dari tahun 2015 hingga tahun 2017 dengan atribut Peneliti Utama, Prodi, Fakultas, Judul Penelitian, dan Tahun dalam bentuk *file excel*.

**3.2 Metode Pengumpulan Data**

**3.2.1 Wawancara**

Wawancara dilakukan dengan cara tanya jawab secara langsung dengan Drh. Asep Rustiawan, M.S selaku Sekretaris Lembaga Penelitian dan Pengembangan Universitas Ahmad Dahlan terkait pengelompokan yang dilakukan oleh pihak LPP UAD.

**3.2.2 Data *Private***

Metode ini dimaksudkan untuk mendapatkan data *private* berupa judul penelitian dosen Universitas Ahmad Dahlan berupa *file excel* dari Sekretaris LPP UAD.

**3.2.3 Studi Literatur**

Pengumpulan data dilakukan dengan studi literatur dari berbagai macam buku, artikel, publikasi ilmiah dan *e-book* untuk mempelajari :

1. Mekanisme *text mining* dalam mengelompokan data.
2. Mekanisme pengelompokan data menggunakan metode *Suffix Tree Clustering.*

**3.3 Spesifikasi Kebutuhan**

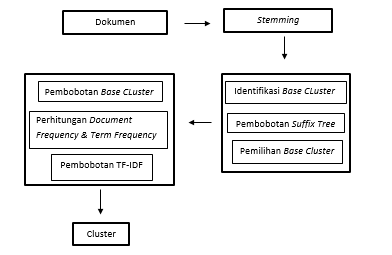
Spesifikasi kebutuhan dibagi menjadi dua, yaitu perangkat keras dan perangkat lunak. Spesifikasi yang dibutuhkan dalam melakukan penelitian ini adalah sebagai berikut:

* + 1. **Kebutuhan Perangkat Keras (Hardware)**

1. Processor minimal Intel Core Inside
2. Memori 1 GB
3. Kapasitas Hardisk minimal 100 GB
4. Monitor 14” HD LCD
   * 1. **Kebutuhan Perangkat Lunak (Software)**
5. Editor Bahasa pemrograman *Python*, *Microframework Flask,* dan *Sublime Text* 3.
6. Database Server yang digunakan yaitu MongoDB dan Database Client yang digunakan yaitu Robomongo.
7. Web Browser seperti Google Chrome, Mozilla Firefox, Opera Mini, dan sebagainya.
8. Modul bahasa pemrograman Python seperti Anaconda
9. Library Pemrograman yang digunakan suffix tree, python suffix tree, pysuffix dan sebagainya.

**3.4 Tahapan Pengembangan Sistem**

Tahapan *text mining* yang akan diterapkan pada penelitian ini adalah dimulai dengan *input* berupa *load* data teks, lalu data teks masuk ke tahapan *Document Cleaning, Stemming, Identifikasi Base Cluster, dan Kombinasi Base Cluster.* Kemudian pemahaman dari Algoritma *Suffix Tree Clustering* dilanjutkan dengan implementasi algoritma tersebut ke dalam pengelompokan Judul Penelitian Dosen. Adapun *Flowchart* Alur Proses *Clustering* Judul Penelitian Dosen sebagai berikut:



Gambar 3.1. Tahapan Pengembangan Sistem.

**3.4.1 Mengumpulan Data (*Collect Data*)**

Data yang dikumpulkan adalah Judul Penelitian Dosen dalam bentuk *file excel* yang didapatkan dari pihak Lembaga Penelitian dan Pengembangan Universitas Ahmad Dahlan dari tahun 2015 hingga 2017 dengan 623 data.

**3.4.2 Tahapan Penguraian Text**

1. ***Document Cleaning***

Tahap awal dalam algoritma *Suffix Tree Clustering.* Pada tahap ini, dokumen yang telah didapatkan akan dibersihkan dan dipersiapkan untuk tahap selanjutnya. Proses untuk mempersiapkan dokumen meliputi proses pembersihan dokumen, proses analisa leksikal teks, proses penghapusan *stopword* dan proses *stemming.*

1. ***Stemming***

Dalam morfologi kata Bahasa Inggris dikenal adanya tiga imbuhan yaitu awalan (*prefix),* sisipan dan akhiran (*suffixs).* Pada penelitian ini, *stemming* digunakan untuk penghilangan / pemotongan awalan dan akhiran dari kata-kata yang sudah dibersihkan (*cleaning).* Contoh : “Instrumen”, “Pengukuran”, “Kesadaran”, “Kanker”, “Payudara”, “Di”, “Yogyakarta” setelah melalui tahap *stemming* menjadi “Instrumen”, “Ukur”, “Sadar”, ”Kanker”, “Payudara”, “Yogyakarta”.

1. **Identifikasi *Base Cluster***

Tahap identifikasi *base cluster* merupakan tahap terpenting dalam algoritma *suffix tree clustering,* karena pada tahap ini akan menghasilkan cluster-cluster dasar. Pembentukan *base cluster* dilakukan dengan cara menemukan *share pharse* antar dokumen. Untuk menemukan *share pharse* digunakan struktur data *suffix tree.* Dengan menggunakan struktur data ini, maka setiap dokumen akan direpresentasikan menjadi suatu kalimat. Untuk menemukan *base cluster* dapat dilakukan dengan cara membuat suatu *invert* dari *phrase* untuk semua dokumen.

1. **Kombinasi *Base Cluster***

Tahap ini digunakan untuk menangani *overlapping cluster.* Dalam tahapini, *phrase* tidak dipertimbangkan. Sebelum melakukan kombinasi antar *base cluster,* terlebih dahulu harus menghitung nilai *similiaruty* antar *base cluster* yang didasarkan pada jumlah dokumen yang *overlap.* Adanya *overlapping* dokumen ini didasarkan karena dokumen memiliki lebih dari satu topik sehingga dokumen dapat memiliki lebih dari satu *phrase* yang dishare.

1. ***Algoritma Suffix Tree Clustering***

*Algoritma Suffix Tree Clustering* merupakan salah satu metode dalam pengelompokan dokumen. Pencarian *shared phrase* untuk semua dokumen yang dikoleksi. Hal tersebut dapat disebut *shared phrase* sebagai *phrase cluster* atau *base cluster,* yang ditemukan dengan menggunakan struktur data yang dinamakan *suffic tree.* Suatu *phrase* yang dimaksud dalam konteks algoritma ini adalah urutan satu atau lebih kata. *Suffix tree clustering* ini adalah urutan satu atau lebih kata.

**3.5 Perancangan**

Perancangan dilakukan untuk mengetahui lebih detail tahapan - tahapan yang dilakukan untuk menghasilkan program yang diinginkan. Tahap – tahapannya adalah sebagai berikut:

**3.5.1 Perancangan Basis Data**

Basis Data yang akan digunakan dalam program ini adalah MongoDB 3.4. Perancangan basis data digunakan untuk menentukan data apa saja yang akan masuk ke dalam *database* dan digunakan untuk mendukung berbagai rancangan sistem*.* Data tersebut diantaranya :

1. Data File, untuk menyimpan identitas file yang telah di*upload.*
2. Data Judul, untuk menyimpan setiap Judul Penelitian Dosen yang belum mengalami pemrosesan.
3. Data *Tokenizing*, untuk menyimpan data kata dari Judul Penelitian Dosen yang sudah mengalami tokenisasi.
4. Data *Stemming,* untuk menyimpan setiap kata dasar yang telah melalui proses *stemming.*
5. Data *Filter*ing, untuk menyimpan data hasil stemming dan menghilangkan kata penghubung dalam dokumen tersebut.
6. Data *Group,* untuk menyimpan data Judul Penelitian Dosen yang dikelompokan berdasarkan bidang penelitian.
7. Data Kelompok, untuk menyimpan data judul hasil penglompokan.

**3.5.2 Perancangan Antar Muka (*Interface)***

Perancangan Antarmuka meliputi tampilan yang akan digunakan dalam program yang berisi struktur menu dan tampilan *user.* Menu – menu yang disediakan antara lain halaman awal, halaman *load*  data, halaman *text mining,* halaman hasil, dan halaman grafik.

**3.5.3 Implementasi**

Tahapan implementasi merupakan tahapan untuk mengimplementasikan perancangan dengan menggunakan algoritma *Suffix Tree Clustering* ke dalam system yang akan dibuat dapat dimengerti oleh pengguna (*user).* Pada tahapan pengimplementasian dalam Bahasa pemrograman *Phyton* akan menghasilkan program untuk mengelompokan judul penelitian dosen. Tahap implementasi menerapkan tahapan *text mining* ke dalam aplikasi web. Aplikasi akan diimplementasikan dengan Bahasa pemrograman *Python, Microframework Flask,* HTML dan Jinja2.

**3.5.4 Pengujian**

Setelah system selesai dibuat maka tahapan selanjutnya adalah pengujian terhadap sistem. Pengujian ini bertujuan untuk menguji apakah system sudah berjalan sesuai keinginan atau belum. Pengujian terhadap system dilakukan dengan metode *Suffix Tree* untuk *clustering.* Tahap pengujian dengan *Suffix Tree* untuk mengetahui seberapa akurat yang dihasilkan dari pengelompokan Judul Penelitian Dosen menggunakan metode *Suffix Tree Clustering.*

# **DAFTAR PUSTAKA**

Adinugroho, S. and Sari, Y. A. (2018) *IMPLEMENTASI DATA MINING MENGGUNAKAN WEKA*. Pertama. Malang: UB Press.

Arifin, A. Z. *et al.* (2008) ‘KLASIFIKASI ONLINE DOKUMEN BERITA DENGAN MENGGUNAKAN ALGORITMA SUFFIX TREE CLUSTERING’.

Adiwijaya, I. (2006) ‘Text Mining dan Knowledge Discovery’, *Kolokium bersama komunitas datamining Indonesia & soft-computing Indonesia*, pp. 1–9. Available at: http://web.ipb.ac.id/~ir-lab/pdf/tm (text summarization).pdf.

Enterprise, J. (2016) *Trik Cepat Menguasai Pemrograman Python*. Jakarta: GRAMEDIA.

Feldman, R. and Sanger, J. (2006) *The Text Mining Handbook*. doi: 10.1017/CBO9780511546914.

Feldman, R. and Sanger, J. (2007) *The text mining handbook: advanced approaches in analyzing unstructured data*, *Imagine*. doi: 10.1179/1465312512Z.00000000017.

Katariya, N. P. and Chaudhari, M. S. (2015) ‘Text Preprocessing for Text Mining Using Side Information’, 3, pp. 3–7.

Kurniawan, B., Effendi, S. and Sitompul, O. S. (2012) ‘Klasifikasi Konten Berita Dengan Metode Text Mining’, *Jurnal Dunia Teknologi Informasi*, 1(1), pp. 14–19. Available at: http://download.portalgaruda.org/article.php?article=58993&val=4123.

Muh. Fitrah, M. P. and Dr. Luthfiyah, M. A. (2017) *Metodelogi Penelitian; Penelitian Kualitatif, Tindak Kelas & Studi Kasus*. Pertama. Sukabumi: CV Jejak.

Sutadi Heru (2016) ‘IMPLEMENTASI SUFFIX TREE CLUSTERING UNTUK PENGELOMPOKAN DOKUMEN YANG TELAH DI AKSES MELALUI MESIN PENCARIAN GOOGLE’, 1(1), pp. 49–54.

Tri, A. L., Soedijono, W. B. and Armadyah, A. (2015) ‘PENGELOMPOKAN ABSTRAK SKRIPSI MENGGUNAKAN METODE SUFFIX TREE CLUSTERINGDANSINGULAR VALUE DECOMPOSITION’, *Teknik Informatika*, pp. 6–7.

Tuwohingide, D. *et al.* (2016) ‘EFISIENSI PHRASE SUFFIX TREE DENGAN SINGLE PASS CLUSTERING UNTUK PENGELOMPOKAN DOKUMEN WEB BERBAHASA INDONESIA’, *JURNAL TEKNOLOGI TECHNOSCIENTIA*, 8(2), pp. 133–140.

Weiss, S. M. *et al.* (2005) *Text Mining : Predictive Methods fo Analyzing Unstructered Information*.

Widodo, Rustiawan, A. and Dkk, S. (2016) ‘Rencana Induk Penelitian (RIP) Universitas Ahmad Dahlan Tahun 2016-2021’, pp. 4–5. Available at: http://lpp.uad.ac.id/wp-content/uploads/2016/12/RIP-UAD-2016-2017-KOMPLIT.pdf (Accessed: 4 September 2018).

Wijoyo, S. H. *et al.* (2017) ‘OPTIMASI SUFFIX TREE CLUSTERING DENGAN WORDNET DAN NAMED ENTITY RECOGNITION UNTUK PENGELOMPOKAN DOKUMEN’, *Jurnal Teknologi Informasi dan Ilmu Komputer*, 4(4), p. 263. doi: 10.25126/jtiik.201744400.