JURNAL LARIK
LADANG ARTIKEL ILMU KOMPUTER
http://jurnal.bsi.ac.id/index.php/larik

Information Retrieval Pemetaan Peta Jalan Penelitian Perguruan Tinggi Berbasis Dokumen Publikasi Ilmiah Dosen

Rusda Wajhillah [1], Agung Wibowo [2]
[1]Sistem Informasi Akuntansi, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika
[2]Sistem Informasi, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika
Jl. Cemerlang No. 8 Sukakarya, Warudoyong Kota Sukabumi, Telp 0266-6251992
Rusda.rwh@bsi.ac.id ,Agung.awo@bsi.ac.id

Abstrak-- Kualitas penelitian perlu diarahkan dan diklasifikasikan untuk perbaikan. Roadmap perguruan tinggi harus sesuai dengan minat dan keahlian dari dosennya. Oleh karena itu, menjadi kewajiban setiap perguruan tinggi untuk membuat perencanaan strategis dan riset unggulan. Fakultas Teknologi Informasi di hampir semua perguruan tinggi telah menghasilkan banyak publikasi ilmiah. Dokumen publikasi karya ilmiah merupakan salah satu contoh dokumen tidak terstruktur. Isinya berupa gaya penulisan, sebagian besar ditentukan oleh bahasa pengarang. Umumnya, judul dokumen hanya menentukan jumlah kata maksimal. Tujuan utama dari information retrieval system adalah untuk menentukan kata kunci dokumen dari query yang diberikan oleh pengguna dalam sekelompok dokumen. Algoritma TF/IDF (Term Frequency – Inversed Document Frequency) dan Algoritma Vector Space Model adalah beberapa metode dari algoritma yang dapat memanfaatkan text mining dalam menganalisis tahapan sebagai opsi penentuan klasifikasi dokumen berdasarkan kata-kata solusi yang sering muncul pada judul dokumen. Makalah ini dapat membantu para pengambil keputusan untuk menentukan, menilai, mengadaptasi roadmap penelitian ke perguruan tinggi. Penggambaran model pohon menggunakan roadmap jangka panjang membuatnya lebih mudah dibaca dan dipahami.

Kata kunci: Information Retrieval, Peta Jalan Penelitian, Dokumen Publikasi

Abstract *- The quality of research needs to be directed and classified for improvement. A college roadmap must accordance interest and expertise from it lecturers. Therefore, be the duty of every college to create a strategic plan and pre-eminent research. Faculty of information technology in most all College has produced many scientific publications. Publication document of scientific papers is one example of unstructured documents. Its contents form of writing style, mostly defined by the author language. Generally, the document title only determined the maximum number of words. The main objective of the information retrieval system is to determine the documents keywords from the query provided by the user in a group of documents. TF/IDF Algorithm (Term Frequency – Inversed Document Frequency) and the Vector Space Model algorithm is several methods of the algorithm that can utilize on text mining in analysing phases as options document classification determination-based solutions words that often appear on the document title. This paper can help decision makers to determine, assess, adapt research roadmap to College. The depiction of a tree model using long-term roadmap makes it easier to read and understand.*

Keywords: Information Retrieval, Research Roadmap, Publication Papers

I. PENDAHULUAN

Perguruan tinggi, lembaga penelitian dan pengembangan, industri dan masyarakat menjadi pihak yang berkompeten menghasilkan invensi dan inovasi ilmu pengetahuan dan teknologi. Pada tahun 2015 terdapat 122 Perguruan Tinggi Negeri (PTN) dan 4.315 Perguruan Tinggi Swasta (PTS) di Indonesia, sehingga total menjadi 4.437 Perguruan Tinggi. Kekuatan untuk menyediakan sumber daya manusia yang memadai dan dapat menjadi universitas riset yang menghasilkan inovasi teknologi yang dibutuhkan oleh industri [1]. Kualitas penelitian perlu diarahkan dan diklasifikasikan untuk perbaikan. Peta jalan perguruan tinggi harus sesuai dengan minat dan keahlian dari dosennya. Oleh karena itu, menjadi kewajiban setiap perguruan tinggi untuk membuat perencanaan strategis dan riset unggulan. Fakultas Teknologi Informasi di hampir semua perguruan tinggi telah menghasilkan banyak publikasi ilmiah.

*Korenpondensi.
Alamat E-mail : jurnal.larik@bsi.ac.id.
Diterima : 10 Sept 2022; Direvisi 10 Nopember 2022; Diterima 27 Desember 2022
© 2022 Jurnal Larik.

Pada tahun 2017 Kemenristekdikti telah meluncurkan nama inisiasi Science and Technology Index (SINTA) [2]. SINTA hadir sebagai salah satu pengindeks publikasi dan sitasi karya ilmiah di Indonesia. Ini juga memiliki fungsi penting dalam mendorong kecenderungan publikasi bagi dosen dan peneliti di Indonesia. Dokumen publikasi karya ilmiah merupakan salah satu contoh dokumen tidak terstruktur. Isinya berupa gaya penulisan, sebagian besar ditentukan oleh bahasa pengarang. Umumnya, judul dokumen hanya menentukan jumlah kata maksimal.

Roadmap penelitian perguruan tinggi yang disusun dalam 'Rencana Induk Penelitian' (RIP). Demikian pula dengan penelitian rencana strategis kebijakan manajemen. RIP merupakan arahan titik-titik bidang penelitian dan penelitian unggulan suatu perguruan tinggi. RIP merupakan arah pengambilan kebijakan dan keputusan dalam pengelolaan lembaga penelitian pada suatu periode tertentu [3]. Renstra juga dapat menjadi acuan bagi dosen dalam mengembangkan makalah iptek. Itu harus sesuai dengan kebijakan sains institusi dan dibatasi.

Penelitian ini untuk mengklasifikasikan makalah penelitian dari tiga tahun terakhir. Menggunakan classifier Naïve Bayes tetapi sumber dataset hanya sembilan judul penelitian [4]. Penelitian untuk mengembangkan model evaluasi kinerja dosen yang valid dan reliabel [5]. Penelitian ini menggunakan objek yang sama dengan menggunakan kertas publikasi namun tidak memiliki korelasi langsung dengan RIP. Peneliti lain untuk mengetahui Penelitian Analisis Kebijakan Perencanaan di UIN SGD Bandung periode 2015-2019. Untuk melihat apa saja rekomendasi hasil analisis Policy Planning Research UIN SGD Bandung [6]. Penelitian ini untuk membuat Indikator khusus untuk menilai kapasitas dan kinerja para peneliti. Masih minim dan perlu diperbaiki. Penelitian ini bertujuan untuk mengembangkan indikator kunci untuk menilai kapasitas penelitian fakultas Teknik Universitas Diponegoro. perbaikan beberapa rekomendasi untuk meningkatkan kapasitas literatur penelitian [7].

Tujuan utama dari information retrieval system adalah untuk menentukan kata kunci dokumen dari query yang diberikan oleh pengguna dalam sekelompok dokumen. Informasi dari dokumen tidak akan ditemukan jika tidak sama dengan kata kunci yang diberikan [4]. Tujuan utama lainnya dari information retrieval adalah untuk membantu pengguna menemukan informasi yang relevan dan berguna dalam jumlah besar data teks. Informasi retrieval melibatkan teknik dan algoritma yang mengolah dan menganalisis data teks untuk memastikan bahwa informasi yang relevan dan berguna ditemukan dan disajikan dengan cepat dan efisien.

Berikut adalah beberapa tujuan utama dari information retrieval:

a. Menemukan informasi yang relevan: Information retrieval membantu pengguna menemukan informasi yang relevan dengan query pencarian mereka dengan cepat dan efisien.
b. Menyediakan akses yang cepat dan efisien: Information retrieval memastikan bahwa informasi yang relevan dapat diakses dengan cepat dan efisien, tanpa harus menelusuri seluruh data teks.
c. Meningkatkan efisiensi dan produktivitas: Information retrieval membantu pengguna menemukan informasi yang mereka cari lebih cepat dan membantu mereka bekerja dengan lebih efisien dan produktif.
d. Menyediakan hasil yang terurut dan relevan: Information retrieval memastikan bahwa informasi yang relevan ditemukan dan disajikan dalam urutan relevansi, membantu pengguna menemukan informasi yang mereka butuhkan dengan lebih mudah.
e. Memastikan privasi dan keamanan data: Information retrieval memastikan bahwa data pribadi dan rahasia tetap terlindung dan tidak dapat diakses oleh pihak yang tidak berwenang.

Information retrieval juga untuk memastikan bahwa pengguna dapat menemukan informasi yang relevan dan berguna dalam jumlah besar data teks dengan cepat dan efisien. Ini membantu mereka bekerja dengan lebih produktif dan memastikan privasi dan keamanan data.

Algoritma TF/IDF (Term Frequency – Inversed Document Frequency) dan Algoritma Vector Space Model adalah beberapa metode dari algoritma yang dapat memanfaatkan text mining dalam menganalisis tahapan sebagai opsi penentuan klasifikasi dokumen berdasarkan kata-kata solusi yang sering muncul pada judul dokumen. Penekanan penelitian ini pada pembobotan berdasarkan penelitian judul kalimat akan menghasilkan suatu klasifikasi berdasarkan kata-kata yang sering muncul.

TF-IDF (Term Frequency-Inverse Document Frequency) adalah ukuran statistik yang digunakan dalam pencarian informasi dan penambangan teks untuk mengevaluasi pentingnya kata dalam dokumen atau kumpulan dokumen. TF mewakili frekuensi kata dalam dokumen, sedangkan IDF

mewakili kebalikan dari frekuensi kata dalam korpus, yang menunjukkan kelangkaan kata tersebut. Produk TF dan IDF memberi bobot pada setiap kata dalam dokumen, yang mencerminkan kepentingan relatifnya. TF-IDF digunakan untuk menentukan relevansi dokumen dengan permintaan pencarian dan biasanya digunakan di mesin pencari, klasifikasi teks, dan sistem rekomendasi.

Vector Space Model (VSM) adalah suatu metode yang digunakan dalam information retrieval dan pemrosesan bahasa alami (NLP) untuk menentukan relevansi suatu dokumen terhadap query. VSM memandang setiap dokumen sebagai vektor spasial dalam ruang dimana setiap term adalah dimensinya. Ukuran vektor setiap term mencerminkan bobot term yang diperoleh dari term weighting. Query juga diterjemahkan menjadi vektor dan perbandingan antara vektor dokumen dan query digunakan untuk menentukan skor relevansi dan melakukan ranking dokumen.

Bahasa Indonesia memiliki struktur morfologis yang relatif rumit dibandingkan bahasa Inggris atau bahasa lainnya. Struktur morfologi bahasa Indonesia adalah bagian dari tata bahasa yang mempelajari tentang bentuk dan fungsi kata dalam bahasa Indonesia. Struktur morfologi membahas tentang pembentukan kata, perubahan bentuk kata, dan penggunaan awalan, akhiran, dan infleksi (perubahan bentuk kata) untuk memodifikasi makna kata. Struktur morfologi bahasa Indonesia meliputi:

1. Pembentukan kata dasar
2. Perubahan bentuk kata (infleksi)
3. Penggunaan awalan, akhiran, dan infleksi untuk memodifikasi makna kata
4. Pembentukan kata turunan dan kata majemuk
5. Penggunaan kata kerja, kata sifat, dan kata benda dalam bentuk tunggal dan jamak
6. Penggunaan kata depan, kata sambung, dan partikel kata.

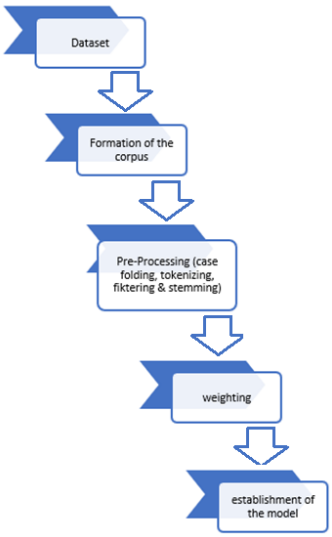
II. METODE PENELITIAN

Ada beberapa jenis metode penelitian yang umum digunakan, di antaranya:

1. Penelitian Kuantitatif: menggunakan data numerik dan analisis statistik untuk menjawab pertanyaan penelitian.
2. Penelitian Kualitatif: menggunakan teknik pengumpulan data seperti wawancara, observasi, dan analisis dokumen untuk memahami fenomena atau persepsi subjek penelitian.
3. Penelitian Campuran (Mixed-Method Research): menggabungkan elemen dari metode kuantitatif dan kualitatif dalam satu studi.
4. Penelitian Longitudinal: melacak perubahan dalam waktu pada suatu populasi atau individu.
5. Penelitian Cross-Sectional: membandingkan beberapa populasi atau individu pada waktu yang sama.
6. Penelitian Experimental: menguji hipotesis dengan mengontrol variabel dan membandingkan kelompok yang dipengaruhi dan kelompok kontrol.
7. Penelitian Survey: mengumpulkan data melalui kuesioner atau wawancara untuk memahami pandangan dan persepsi individu.

Pemilihan metode penelitian tergantung pada tujuan, sumber daya, dan keterbatasan penelitian. Setiap metode memiliki kelebihan dan kekurangan yang berbeda dan membutuhkan pemikiran yang cermat dan kreatif untuk memilih metode yang tepat untuk penelitian. Penelitian ini mengunakan lima tahapan.

Pada penelitian ini kami menggunakan jenis penelitian kuantitatif dimana penelitian ini disusun menjadi lima tahap. Tahapan penelitian ini dapat dilihat pada gambar 1. Tahap pertama adalah tahapan pengumpulan data, dilanjutkan ke tahap kedua berupa pembentukan corpus data berdasarkan dataset yang sudah ada. Tahap ketiga adalah proses “pre-processing”, secara umum proses ini seluruh koleksi corpus yang ada akan diformat ulang terkait menyamakan bentuk huruf (menjadikan huruf besar menjadi huruf kecil semua), membuang karakter simbol dan angka, mengubah kata berimbuhan menjadi kata dasar. Tahap keempat adalah tahap pembobotan, diakhiri dengan pemodelan yang dilaksanakan di tahap kelima.



Gambar 1. Tahapan penelitian

Fase satu, kumpulan data. Sampel dataset diambil dari daftar judul publikasi dosen dari dua perguruan tinggi swasta di Indonesia. Perguruan tinggi yang digunakan sebagai contoh adalah Universitas Bina Sarana Informatika Kampus Sukabumi. Dokumen tersebut tercatat di Ristekdikti dan terindeks SINTA. Daftar publikasi diambil dari google scholar, dengan jumlah total 108 data.

Tahap kedua adalah pembentukan korpus. Korpus adalah istilah yang digunakan untuk menggambarkan koleksi dokumen atau teks dalam suatu bahasa. Dalam informasi retrieval dan NLP, korpus digunakan sebagai dasar untuk melatih model atau menentukan relevansi dokumen terhadap query. Korpus biasanya terdiri dari banyak dokumen yang berbeda dan dapat berisi teks dalam berbagai format, seperti artikel berita, buku, atau tweet. Ukuran korpus dapat beragam, mulai dari beberapa dokumen hingga jutaan dokumen. Penggunaan korpus yang representatif dan besar sangat penting untuk memastikan hasil yang akurat dan berguna dalam aplikasi NLP. Data mentah yang diperoleh dari Google Scholar diubah menjadi format teks. Fase selanjutnya, Pra-pemrosesan. Pra-pemrosesan information retrieval adalah tahap yang dilakukan sebelum melakukan pencarian informasi, dimana data dokumen diolah dan disiapkan untuk diproses oleh algoritma pencarian informasi. Beberapa tahap pra-pemrosesan yang umum dilakukan meliputi:

1. Tokenisasi: memecah teks menjadi token atau kata-kata individu.
2. Stopword removal: menghapus kata-kata yang tidak memiliki makna seperti "dan", "atau", dll.
3. Stemming/Lemmatization: mengurangkan kata-kata menjadi bentuk dasar (stem) atau menjadi entri kata dalam kamus (lemmatization).
4. Term weighting: memberikan bobot numerik pada setiap term untuk mencerminkan relevansi terhadap dokumen.
5. Representasi dokumen: mengubah dokumen menjadi representasi numerik seperti vektor atau matrix yang dapat diproses oleh algoritma pencarian informasi.
6. Tahap pra-pemrosesan ini sangat penting karena mempengaruhi hasil akhir dari proses pencarian informasi dan memastikan bahwa data dokumen siap diproses oleh algoritma pencarian informasi.

Pada tahap ini dilakukan proses berikut:

- Case folding adalah tahap pra-pemrosesan dalam information retrieval dan NLP, dimana teks dikonversi menjadi huruf kecil (lowercase). Tujuan dari case folding adalah untuk mengatasi perbedaan bentuk huruf (uppercase/lowercase) yang tidak memiliki makna dan dapat mempengaruhi hasil akhir. Proses case folding meliputi beberapa tahap sebagai berikut:

a. Pembacaan teks: membaca semua teks dalam korpus.
b. Konversi huruf besar menjadi huruf kecil: setiap huruf besar dalam teks dikonversi menjadi huruf kecil.
c. Penyimpanan teks: teks yang sudah dikonversi menjadi huruf kecil disimpan untuk digunakan dalam tahap selanjutnya.
d. Case folding memastikan bahwa teks yang sama tidak dicatat sebagai teks yang berbeda karena perbedaan bentuk huruf. Ini juga memastikan bahwa hasil akhir dari pencarian informasi tidak terpengaruh oleh perbedaan bentuk huruf.

- Tokenizing: Tokenizing adalah tahap pra-pemrosesan dalam information retrieval dan NLP, dimana teks dibagi menjadi unit-unit yang lebih kecil, yang disebut sebagai "token". Tujuan dari tokenizing adalah untuk memecah teks menjadi unit-unit yang lebih manajemennya untuk mempermudah analisis dan pemrosesan.

Proses tokenizing meliputi beberapa tahap sebagai berikut:

a. Pembacaan teks: membaca semua teks dalam korpus.
b. Pemecahan teks: teks dibagi menjadi unit-unit lebih kecil, seperti kata, frase, atau kalimat.
c. Identifikasi token: setiap unit teks diberi label sebagai token.
d. Penyimpanan token: token disimpan untuk digunakan dalam tahap selanjutnya.
e. Tokenizing memastikan bahwa teks yang sama dicatat sebagai teks yang sama, tidak peduli bagaimana teks itu diformat. Ini juga mempermudah analisis dan pemrosesan teks karena memecah teks menjadi unit-unit yang lebih mudah diolah.

- Filtering: Filtering adalah tahap pra-pemrosesan dalam information retrieval dan NLP, dimana token yang tidak diperlukan dalam analisis dibuang. Tujuan dari filtering adalah untuk membuang token-token yang tidak memiliki makna dan hanya memperkeruh hasil akhir.

Proses filtering meliputi beberapa tahap sebagai berikut:

a. Pembacaan token: membaca semua token dalam korpus.
b. Penentuan token yang tidak diperlukan: token yang tidak memiliki makna atau hanya memperkeruh hasil akhir ditentukan.
c. Penghapusan token: token yang tidak diperlukan dibuang.
d. Penyimpanan token: token yang tersisa setelah filtering disimpan untuk digunakan dalam tahap selanjutnya.
e. Filtering memastikan bahwa token yang tidak diperlukan tidak mempengaruhi hasil akhir dari analisis dan membuang token-token yang memperkeruh hasil akhir. Ini juga memastikan bahwa token yang tersisa adalah token yang memiliki makna dan relevan untuk analisis.

contoh:' dalam', 'dan,'atau'. itu akan dihapus.

- Stemming: Stemming adalah tahap pra-pemrosesan dalam information retrieval dan NLP, dimana kata direduksi menjadi kata dasar (stem) untuk mempermudah analisis dan pemrosesan. Tujuan dari stemming adalah untuk mengatasi variasi bentuk kata yang tidak memiliki makna dan dapat mempengaruhi hasil akhir.

Proses stemming meliputi beberapa tahap sebagai berikut:

a. Pembacaan kata: membaca semua kata dalam korpus.
b. Identifikasi variasi bentuk kata: variasi bentuk kata yang tidak memiliki makna ditentukan.
c. Reduksi kata ke kata dasar: variasi bentuk kata direduksi menjadi kata dasar.
d. Penyimpanan kata dasar: kata dasar disimpan untuk digunakan dalam tahap selanjutnya.

Stemming memastikan bahwa kata yang sama dicatat sebagai kata yang sama, tidak peduli bagaimana kata itu diformat. Ini juga memastikan bahwa hasil akhir dari pencarian informasi tidak terpengaruh oleh variasi bentuk kata. Misalnya, dari fase pra-pemrosesan: Kalimat baku: 'Sistem Penunjang Keputusan dalam Membuat Rencana Induk Penelitian'

Case folding: Case folding adalah tahap pra-pemrosesan dalam information retrieval dan NLP, dimana teks dikonversi menjadi huruf kecil untuk mempermudah analisis dan pemrosesan. Tujuan dari case folding adalah untuk mengatasi perbedaan bentuk kapitalisasi yang tidak memiliki makna dan dapat mempengaruhi hasil akhir.

Proses case folding meliputi beberapa tahap sebagai berikut:

a. Pembacaan teks: membaca semua teks dalam korpus.
b. Konversi ke huruf kecil: semua teks dikonversi menjadi huruf kecil.
c. Penyimpanan teks: teks dalam bentuk huruf kecil disimpan untuk digunakan dalam tahap selanjutnya.

Case folding memastikan bahwa teks yang sama dicatat sebagai teks yang sama, tidak peduli bagaimana teks itu diformat. Ini juga memastikan bahwa hasil akhir dari pencarian informasi tidak terpengaruh oleh perbedaan bentuk kapitalisasi. Contoh hasil case folding: 'sistem penunjang keputusan dalam membuat rencana induk penelitian'

Tokenisasi, Tokenisasi adalah tahap pra-pemrosesan dalam information retrieval dan NLP, dimana teks dibagi menjadi token-token individu untuk mempermudah analisis dan pemrosesan. Tujuan dari tokenisasi adalah untuk memecah teks menjadi unit-unit yang lebih kecil dan lebih mudah diterima oleh mesin.

Proses tokenisasi meliputi beberapa tahap sebagai berikut:

a. Pembacaan teks: membaca teks dalam korpus.
b. Pemecahan teks menjadi token: teks dibagi menjadi token-token individu.
c. Penyimpanan token: token-token disimpan untuk digunakan dalam tahap selanjutnya.

Tokenisasi memastikan bahwa teks dapat dianalisis dengan lebih baik dan mempermudah pemrosesan oleh mesin. Ini juga memastikan bahwa informasi dalam teks dapat ditemukan dan dianalisis dengan lebih efisien. Contoh hasil proses tokenisasi: 'sistem'; 'penunjang'; 'keputusan'; 'dalam'; 'membuat'; 'rencana'; 'induk'; 'penelitian';

Penyaringan: penyaringan atau filtering adalah tahap pra-pemrosesan dalam information retrieval dan NLP, dimana token-token individu dalam teks dianalisis dan difilter untuk mempermudah analisis dan pemrosesan. Tujuan dari filtering adalah untuk menghilangkan token-token yang tidak berguna atau tidak memiliki makna dan dapat mempengaruhi hasil akhir.

Proses filtering meliputi beberapa tahap sebagai berikut:

a. Pembacaan token: membaca token-token individu yang sudah terbentuk.
b. Penentuan token yang tidak berguna: token-token yang tidak berguna atau tidak memiliki makna ditentukan.

c. Penghapusan token: token-token yang tidak berguna dihapus.
d. Penyimpanan token: token-token yang berguna disimpan untuk digunakan dalam tahap selanjutnya.

Filtering memastikan bahwa analisis dan pemrosesan teks tidak terpengaruh oleh token-token yang tidak berguna dan mempermudah pemrosesan oleh mesin. Ini juga memastikan bahwa hasil akhir dari pencarian informasi menjadi lebih akurat dan efisien. Contoh hasil penyaringan: 'sistem'; 'penunjang'; 'keputusan'; 'membuat'; 'rencana'; 'induk'; 'penelitian';

Stemming, Stemming adalah tahap pra-pemrosesan dalam information retrieval dan NLP, dimana token-token individu dalam teks direduksi menjadi bentuk dasar (atau "stem") untuk mempermudah analisis dan pemrosesan. Tujuan dari stemming adalah untuk mengatasi variasi bahasa dan mempermudah pencocokan antara teks yang berbeda.

Proses stemming meliputi beberapa tahap sebagai berikut:

a. Pembacaan token: membaca token-token individu yang sudah terbentuk.
b. Reduksi token menjadi bentuk dasar: token-token direduksi menjadi bentuk dasar atau stem.
c. Penyimpanan stem: stem disimpan untuk digunakan dalam tahap selanjutnya.

Stemming memastikan bahwa variasi bahasa dalam teks dapat ditemukan dan dianalisis dengan lebih efisien.

Contoh hasil proses stemming: 'sistem'; 'tunjang'; 'putus'; 'buat'; 'rencana'; 'induk'; 'neliti';

Tahap keempat adalah pembobotan, Pembobotan adalah proses penentuan bobot atau nilai numerik untuk setiap token atau dokumen dalam korpus dalam informasi retrieval dan NLP. Tujuan dari pembobotan adalah untuk mempermudah pemrosesan informasi dan menentukan relevansi token atau dokumen terhadap query pencarian.

Metode pembobotan yang umum digunakan meliputi Term Frequency (TF), Inverse Document Frequency (IDF), dan TF-IDF. Metode ini mempertimbangkan frekuensi token dalam dokumen dan frekuensi munculnya token dalam korpus. Nilai bobot yang lebih tinggi menunjukkan bahwa token atau dokumen lebih relevan dengan query pencarian.

Hasil akhir dari pembobotan digunakan dalam pencarian informasi dan analisis teks untuk menentukan relevansi dan memperbaiki hasil pencarian.

proses ini, hasil pra-pemrosesan diberi bobot berdasarkan jumlah kemunculan pada dokumen lain. Jumlah kemunculan kata dalam dokumen yang disimpan di TF. Nomor IDF dihitung berdasarkan jumlah seluruh temuan kata dalam dokumen dibagi dengan jumlah dokumen. Bobot ini disimpan ke variabel TF/IDF.

Contoh fase pembobotan, lihat tabel 1. Menunjukkan

Tabel 1. Term Weighting

	Doc 1	Doc 2	Doc 3	TF	IDF
sistem	1	6	10	17	0.499999

Term weighting adalah teknik yang digunakan dalam pencarian informasi dan pemrosesan bahasa alami (NLP) untuk menentukan pentingnya masing-masing kata dalam dokumen, kueri, atau kumpulan dokumen. Tujuannya adalah untuk menetapkan bobot numerik pada setiap istilah dalam dokumen untuk mencerminkan relevansi atau signifikansinya terhadap konten. Bobot ini digunakan dalam berbagai algoritma, seperti TF-IDF dan BM25, untuk menentukan peringkat dokumen berdasarkan kemiripannya dengan kueri. Kata “sistem” ditemukan 1 kata pada doc 1, 6 kata pada doc 2 dan 10 kata pada doc 3 maka TF adalah 1 + 6 + 10 = 17, untuk IDF 17/3 = 5.667.

III. HASIL DAN PEMBAHASAN

Tahap kelima, Pembentukan model adalah proses pembuatan representasi matematis dari data dalam NLP dan information retrieval. Model ini digunakan untuk menganalisis dan memproses data untuk memecahkan masalah yang berbeda.

Proses pembentukan model meliputi beberapa tahap sebagai berikut:

a. Persiapan data: data dalam teks harus diolah melalui proses pra-pemrosesan seperti tokenizing, stemming, dan filtering.
b. Pembobotan: data dalam teks harus diberikan bobot atau nilai numerik melalui proses pembobotan.
c. Pemodelan: data dalam teks diformat menjadi representasi matematis, seperti vektor, menggunakan algoritma seperti Vector Space Model.

d. Pelatihan model: model dalam tahap pemodelan harus dilatih menggunakan data latih untuk memperbaiki kinerjanya.
e. Evaluasi model: model harus divalidasi dan dievaluasi untuk memastikan akurasi dan kinerja.

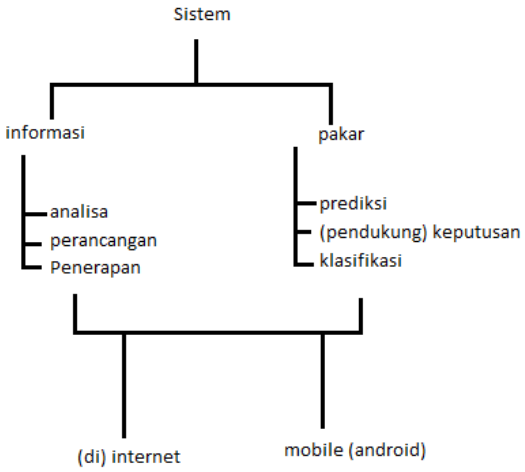
Pembentukan model memastikan bahwa data dalam teks dapat dianalisis dan diolah secara efisien dan akurat untuk memecahkan masalah yang berbeda, seperti pencarian informasi dan analisis sentimen. Model yang benar dapat memastikan hasil yang lebih baik dan akurat dalam memecahkan masalah.

Secara sederhana pembentukan model proses dilakukan dengan membuang kata dengan TF rendah dan tidak menemukan 'kamus' yang sesuai dengan visi, misi, dan roadmap Perguruan Tinggi. Hasil TF/IDF sampel dapat dilihat pada tabel 2.

Tabel 2. TF/IDF Universitas Bina Sarana Informatika Kampus Sukabumi

Documents	Term Before Process	Term After Process	Roadmap Term
94	446	50	11

Dokumen lengkap hasil pengolahan data dapat diakses di inarxiv.id/. Model istilah roadmap tree dapat dilihat pada gambar 2



Gambar 2. Pohon model

Pada gambar 2 menunjukkan bahwa kata yang paling banyak adalah kata ‘sistem’ diikuti dengan kata ‘informasi’ dan ‘pakar’. Berdasarkan pohon kata tersebut dapat disimpulkan bahwa dokumen penelitian dosen di Universitas Bina Sarana Informatika Kampus Sukabumi banyak membahas atau mengangkat judul mengenai ‘Sistem Informasi’ dan ‘Sistem Pakar’.

Sistem informasi atau Sistem pakar yang dibahas dalam dokumen penelitian dosen banyak diterapkan dalam bentuk web di ‘internet’ dan berupa aplikasi perangkat bergerak atau ‘mobile’ android.

IV. KESIMPULAN DAN SARAN

Roadmap tersebut sesuai dengan visi dan misi yang menjadi salah satu poin mata kuliah akreditasi. Roadmap perguruan tinggi dapat dipetakan berdasarkan publikasi seorang dosen. Makalah ini dapat membantu para pengambil keputusan untuk menentukan, menilai, mengadaptasi roadmap penelitian ke perguruan tinggi. Penggambaran model pohon menggunakan roadmap jangka panjang membuatnya lebih mudah dibaca dan dipahami. Penelitian ini dapat membangun model dengan menggunakan algoritma clustering atau klasifikasi.

V. REFERENSI

[1] D. Risbang, "Rencana Strategis 2015-2019," Ristekdikti, Jakarta, 2015.

[2] D. Risbang, "Direktorat Jenderal Penguatan Riset dan Pengembangan Kementerian Riset, Teknologi dan Pendidikan Tinggi Republik Indonesia," 30 January 2017. [Online]. Available: https://risbang.ristekdikti.go.id/publikasi/berita-media/menristek-rilis-portal-sinta-untuk-kembangkan-penelitian/. [Accessed 22 April 2019].

[3] L. P. U. N. Gorontalo, "Rencana Induk Penelitian (RIP) Universitas Negeri Gorontalo," Universitas Negeri Gorontalo, Gorontalo, 2010.

[4] A. Modjo, A. Mulyanto and L. Hadjaratie, "Klasifikasi Penelitian Berdasarkan Rencana Induk Penelitian Universitas Negeri Gorontalo Menggunakan Naive Bayes Classifier," Universitas Negeri Gorontalo, Gorontalo, 2014.

[5] T. H. Retnowati, D. Mardapi and Badrun, "Model evaluasi kinerja dosen: pengembangan instrumen untuk mengevaluasi kinerja dosen," Jurnal penelitian dan evaluasi Pendidikan, vol. 21, no. 2, p. 206 – 214, 2017.

[6] M. Asro and I. A. Halim, "ANALISIS PERENCANAAN KEBIJAKAN PENELITIAN

UIN SUNAN GUNUNG DJATI BANDUNG," Jurnal Perspektif, vol. 1, no. 2, pp. 29-52, 2017.

[7] A. Arvianto, M. S. Zekben, N. U. Handayani, P. W. Adi and S. Saptadi, "PENGEMBANGAN INDIKATOR KINERJA UTAMA UNTUK MENGUKUR KEBERLANJUTAN PENELITIAN PADA FAKULTAS TEKNIK UNIVERSITAS DIPONEGORO," Jurnal Muara Ilmu Ekonomi dan Bisnis, vol. 2, no. 2, pp. 263-274, 2018.

[8] K. D. Putung, A. Lumenta and A. Jacobus, "Penerapan Sistem Temu Kembali Informasi Pada Kumpulan Dokumen Skripsi," E-journal Teknik Informatika, vol. 8, no. 1, pp. 18-23, 2016.