TUGAS AKHIR

OTOMATISASI PELABELAN KARYA TULIS ILMIAH KOLEKSI PERPUSTAKAAN UNPAR

Josep Cliff Jonathan

NPM: 2017730046

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN
2024

FINAL PROJECT

THE AUTOMATION OF SCIENTIFIC PAPER LABELING FOR UNPAR LIBRARY COLLECTION

Josep Cliff Jonathan

NPM: 2017730046

DEPARTMENT OF INFORMATICS
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES
PARAHYANGAN CATHOLIC UNIVERSITY
2024

LEMBAR PENGESAHAN

OTOMATISASI PELABELAN KARYA TULIS ILMIAH KOLEKSI PERPUSTAKAAN UNPAR

Josep Cliff Jonathan

NPM: 2017730046

Bandung, 12 Januari 2024

Menyetujui,

Pembimbing

Digitally signed by Veronica Sri Moertini

Dr. Veronica Sri Moertini

Ketua Tim Penguji

Digitally signed by Gede Karya

Gede Karya, M.T.

Anggota Tim Penguji

Digitally signed by Cecilia Esti Nugraheni

Dr.rer.nat. Cecilia Esti Nugraheni

Mengetahui,

Ketua Program Studi

Digitally signed by Lionov

Lionov, Ph.D.

PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa tugas akhir dengan judul:

OTOMATISASI PELABELAN KARYA TULIS ILMIAH KOLEKSI PERPUSTAKAAN UNPAR

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
Tanggal 12 Januari 2024

Josep Cliff Jonathan
NPM: 2017730046

ABSTRAK

Penelitian ini mengkaji tentang otomatisasi pelabelan karya ilmiah di perpustakaan Universitas Katolik Parahyangan (UNPAR) di Bandung, Indonesia. Latar belakang penelitian ini adalah tantangan yang dihadapi oleh perpustakaan dalam memberikan label atau tanda khusus pada setiap dokumen, terutama di era teknologi yang berkembang pesat. Tujuan penelitian ini adalah untuk mengotomatisasi pelabelan karya ilmiah berdasarkan jurusan dan topik yang terkait. Data yang digunakan dalam penelitian ini diambil dari situs Repositori UNPAR dengan menggunakan teknik *web scraping.* Data kemudian diproses dengan berbagai teknik pra-pemrosesan, seperti pembersihan data, tokenisasi, penghapusan stop words, dan lemmatisasi. Untuk pelabelan jurusan, digunakan algoritma klasifikasi *Multinomial Naive Bayes*, sedangkan untuk analisis topik, digunakan algoritma pemodelan topik *Latent Dirichlet Allocation* (LDA). Hasil penelitian menunjukkan bahwa algoritma klasifikasi *Multinomial Naive Bayes* dapat menghasilkan label jurusan yang cukup baik, dan algoritma pemodelan topik LDA dapat mengidentifikasi topik yang berbeda dalam setiap jurusan. . Namun, hasil yang didapatkan tidak begitu spesifik dan rinci karena satu kata kunci dapat memiliki lebih dari satu makna untuk jurusan yang berbeda. Kesimpulan dari penelitian ini adalah bahwa otomatisasi pelabelan karya ilmiah dapat membantu perpustakaan dalam memudahkan perpustakaan dalam pelabelan dan pengguna dalam menemukan topik-topik yang sesuai dengan karya tulis yang dicari.

Kata-kata kunci: Otomatisasi, Perpustakaan, Pemodelan topik, LDA, Naive Bayes Classifier, Klasifikasi

ABSTRACT

This research examines the automation of labeling scientific papers in the library of Parahyangan Catholic University (UNPAR) in Bandung, Indonesia. The background of this research is the challenge faced by the library in providing labels or special marks on each document, especially in the era of rapid technological development. The purpose of this research is to automate the labeling of scientific papers based on the department and related topics. The data used in this research were taken from the UNPAR Repository site using web scraping techniques. The data were then processed with various pre-processing techniques, such as data cleaning, tokenization, stop words removal, and lemmatization. For department labeling, the Multinomial Naive Bayes classification algorithm was used, while for topic analysis, the Latent Dirichlet Allocation (LDA) topic modeling algorithm was used. The results of the research show that the Multinomial Naive Bayes classification algorithm can produce good department labels, and the LDA topic modeling algorithm can identify different topics in each department. However, the results obtained are not very specific and detailed because one keyword can have more than one meaning for different departments. The conclusion of this research is that the automation of labeling scientific papers can help the library in facilitating the library in labeling and users in finding topics that match the papers they are looking for.

Keywords: Automation, Library, Topic modeling, LDA (Latent Dirichlet Allocation), Naive Bayes Classifier, Classification.

untuk saya

KATA PENGANTAR

Puji syukur kehadirat Tuhan Yang Maha Esa, atas berkat dan rahmat-Nya, saya dapat menyelesaikan tugas akhir dengan judul "OTOMATISASI PELABELAN KARYA TULIS ILMIAH KOLEKSI PERPUSTAKAAN UNPAR" sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer di Universitas Katolik Parahyangan.

Tugas akhir ini merupakan hasil dari pembelajaran dan pengalaman yang saya peroleh selama menempuh pendidikan di Jurusan Informatika Unpar angkatan 2017. Saya menyadari bahwa tugas akhir ini tidak akan terwujud tanpa bantuan dan dukungan dari berbagai pihak. Oleh karena itu, saya ingin menyampaikan ucapan terima kasih yang sebesar-besarnya kepada:

- Bapak, Ibu, dan Cici yang telah memberikan kasih sayang, doa, dan dukungan moral dan materi yang tidak pernah putus selama saya menempuh pendidikan.
- Ibu Maria Veronica, S.T., M.T. sebagai dosen pembimbing utama yang telah memberikan bimbingan, arahan, saran, dan kritik yang sangat berharga dalam penyusunan tugas akhir ini.
- Ibu Dr. Veronica Sri Moertini, Ir., M.T. sebagai dosen pembimbing kedua yang juga telah memberikan bimbingan, arahan, saran, dan kritik yang sangat berharga dalam penyusunan tugas akhir ini.
- Bapak Gede Karya, S.T., M.T. dan Ibu Dr. rer. nat Cecilia Esti Nugraheni, S.T., M.T. sebagai dosen penguji yang telah memberikan masukan dan evaluasi yang sangat berguna untuk penyempurnaan tugas akhir ini.
- Ka Putri, Ka Istoko, Mba Ratna, Katt, Rico, Monic, Sam, Andi, Alma, Jennison, Yani, Nadia, Yusuf, Jason, dan teman-teman lainnya yang telah memberikan semangat, motivasi, dan bantuan dalam berbagai hal selama saya menyelesaikan tugas akhir ini.
- Pihak-pihak lain yang tidak dapat saya sebutkan satu per satu, yang telah membantu saya dalam menyelesaikan tugas akhir ini.

Saya menyadari bahwa tugas akhir ini masih jauh dari sempurna dan masih banyak kekurangan. Oleh karena itu, saya sangat mengharapkan kritik dan saran yang membangun dari pembaca untuk perbaikan di masa depan. Semoga tugas akhir ini dapat bermanfaat bagi saya sendiri dan bagi orang lain yang berkepentingan.

Bandung, Januari 2024

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR KODE PROGRAM

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Pada era teknologi yang berkembang pesat, data yang dihasilkan juga semakin pesat. Kebanyakan dari data yang dihasilkan adalah teks yang tidak berstruktur dalam bentuk yang beraneka ragam, contohnya yaitu buku, halaman web, artikel, dan dokumen lainnya. Perpustakaan adalah tempat penyimpanan kumpulan buku, majalah, jurnal, dan bahan pustaka lainnya disimpan, diorganisir, dan dipinjamkan kepada pengguna. Perpustakaan biasanya berada di dalam suatu lembaga pendidikan seperti sekolah, universitas, kampus, dan lainnya. Perpustakaan di dalam universitas memiliki koleksi buku dan sumber daya informasi yang luas dan mendalam, dan terdiri dari berbagai macam ilmu yang terdapat di dalam universitas tersebut. Salah satu tantangan besar bagi organisasi seperti perpustakaan adalah untuk memberikan tanda khusus atau label pada setiap dokumen agar mempermudah pelanggan dalam mencari informasi yang sesuai.

Sebagai solusi dari tantangan tersebut penelitian ini bertujuan untuk mengotomatisasi pelabelan koleksi Perpustakaan Universitas Katolik Parahyangan (UNPAR) di Kota Bandung, Jawa Barat, Indonesia. Ratusan ribu koleksi Perpustakaan UNPAR terdiri dari buku, karya mahasiswa, karya dosen, penelitian, dan lainnya. Dalam studi kasus ini dikumpulkan sampel dari beberapa koleksi perpustakaan UNPAR dari berbagai jenis dan topik, dengan label jenis dan topik.

Koleksi pada Perpustakaan UNPAR biasanya memiliki label berdasarkan jenis dan topiknya atau yang biasa disebut katalog. Sebagai contoh, katalog dari buku "Filosofi Teras: Filsafat Yunani-Romawi Kuno untuk Mental Tangguh Masa Kini". Katalog untuk buku tersebut adalah "171.2 MAN f" yang menunjukkan bahwa buku tersebut termasuk dalam kategori filsafat umum dan diterbitkan oleh penerbit Mantra.

Meringkas teks dan mengkategorikan dokumen yang ada di perpustakaan dikenal sebagai katalog. Membuat ringkasan singkat tentang isi dokumen, seperti abstrak atau deskripsi, disebut meringkas teks. Proses ini menghasilkan abstrak yang dibuat secara manual. Pada penelitian ini, dokumen dikategorikan ke dalam jurusan tertentu dan penentuan topik atau *related terms* dari tiap dokumen

Ilmu Informasi merupakan ilmu yang mempelajari pengkodean informasi dengan menggunakan pendekatan matematika. Salah satu aplikasi dari ilmu ini adalah *citation index* yang berfungsi untuk mempelajari dan mengukur teks dan informasi menggunakan sebuah bilangan. *Natural Language Processing* (NLP) merupakan ilmu gabungan dari *Linguistic Science* dan Ilmu Informasi yang bertujuan untuk memodelkan bahasa alami manusia di komputer.

Dataset ini diperoleh dari situs web Repository UNPAR, juga dikenal sebagai UNPAR-IR. Tujuan dari penelitian ini adalah untuk melabeli karya tulis ilmiah dengan jurusan dan kata kunci, sehingga sivitas akademika dapat lebih mudah menemukan informasi yang mereka butuhkan. Hasil akhir dari penelitian ini adalah program yang dapat melakukan prediksi topik dan jurusan dengan input pdf. Program ini dievaluasi dengan pengujian fungsional dan eksperimental.

1.2 Rumusan Masalah

Rumusan masalah yang muncul berdasarkan deskripsi dan latar belakang yang sudah dibahas adalah sebagai berikut:

1. Bagaimana mengumpulkan *dataset* dari Perpustakaan UNPAR?
2. Bagaimana menyiapkan *dataset* yang diperoleh dari Perpustakaan UNPAR?
3. Bagaimana melakukan otomatisasi pelabelan jurusan dan topik untuk karya tulis ilmiah di Perpustakaan UNPAR?
4. Bagaimana cara membangun model klasifikasi jurusan dan model topik menggunakan LDA untuk pelabelan karya tulis ilmiah?
5. Bagaimana mengevaluasi keakuratan dari model klasifikasi dan model LDA yang digunakan untuk melabeli dokumen?
6. Bagaimana cara membangun perangkat lunak untuk meluncurkan model klasifikasi?

1.3 Tujuan

Tujuan dari penelitian ini adalah sebagai berikut:

1. Mengumpulkan data-data yang tersedia di koleksi Perpustakaan UNPAR.
2. Menyiapkan data berupa karya tulis ilmiah yang tersedia di laman Repositori UNPAR.
3. Membuat otomatisasi pelabelan jurusan dan topik untuk karya tulis ilmiah di koleksi Perpustakaan UNPAR.
4. Membangun model klasifikasi jurusan dan model topik menggunakan LDA untuk karya tulis ilmiah yang tersedia di Perpustakaan UNPAR.
5. Melakukan evaluasi keakuratan dari model klasifikasi dan model LDA yang digunakan untuk melabeli dokumen.
6. Membuat perangkat lunak berbasis *desktop* untuk meluncurkan model klasifikasi dan model otomatisasi label.

1.4 Batasan Masalah

Batasan masalah pada penelitian ini adalah:

1. Data yang diambil hanya mencakup repositori UNPAR.
2. Data yang dikumpulkan hanya mencakup tahun publikasi 2017 sampai tahun 2022.
3. Data yang telah terambil memiliki beberapa bahasa dan tipe. Pada penelitian ini, data yang dipakai hanya abstrak berbahasa indonesia.

1.5 Metodologi

Metodologi yang digunakan dalam penyusunan tugas akhir ini terdiri dari langkah-langkah sebagai berikut:

1. Studi literatur terkait pengolahan bahasa alami, klasifikasi teks, dan LDA.
2. Studi literatur mengenai penggunaan NLTK Python.
3. Melakukan pra-pemrosesan text sebagai tahap awal pengolahan *dataset* karya tulis ilmiah yang tersedia di UNPAR-IR.
4. Analisis fitur-fitur yang dapat membedakan jenis-jenis karya tulis ilmiah.
5. Eksperimen pembuatan model klasifikasi untuk pelabelan jenis karya tulis dengan fitur-fitur sesuai hasil analisis.
6. Mengevaluasi hasil pelabelan topik dan sub topik di topik tertentu.
7. Memperbaiki pelabelan dokumen.
8. Membuat perangkat lunak model sub topik di topik tertentu.
9. Menampilkan hasil pelabelan topik.

10. Menyelesaikan perangkat lunak dan melakukan peluncuran.
11. Melakukan analisis dari hasil pengujian dan eksperimen yang telah dilakukan.
12. Menyusun dokumen skripsi.

1.6 Sistematika Pembahasan

1. Bab 1 Pendahuluan
   Bagian pertama dari skripsi ini membahas mengenai latar belakang permasalahan, rumusan masalah, tujuan penelitian, batasan masalah, serta metodologi penelitian.
2. Bab 2 Landasan Teori
   Bab 2 membahas beberapa konsep dan teknik yang digunakan dalam pengolahan teks. Konsep-konsep tersebut meliputi teknik *Natural Language Processing* (NLP) untuk mengolah dan memproses teks, seleksi fitur untuk mengekstrak fitur dari data teks, evaluasi model teks, penambangan data, klasifikasi, evaluasi klasifikasi, serta penggunaan *Latent Dirichlet Allocation* (LDA) untuk melakukan klasifikasi topik pada teks.
3. Bab 3 Eksplorasi Teknologi
   Bab ini mencakup studi kasus kecil, hasil komputasi atau perhitungan manual yang dilakukan untuk mempelajari algoritma yang digunakan. Hasil eksplorasi penggunaan teknologi juga dicantumkan di sini, termasuk hasil pengujian menggunakan *library* yang diperlukan di lingkungan pemrograman Python.
4. Bab 4 Penambangan Data
   Bab ini mencakup deskripsi masalah yang ingin diselesaikan, cara pengumpulan data dan langkah pra-pemrosesan data untuk tujuan klasifikasi, serta hasil yang diperoleh dari proses penambangan data.
5. Bab 5 Perangkat Lunak untuk Meluncurkan Model Klasifikasi
   Bab ini berisi implementasi model, pengujian perangkat lunak, dan eksperimen berdasarkan rancangan yang telah dibuat.
6. Bab 6 Kesimpulan dan Saran
   Bab ini berisi kesimpulan dan saran dari seluruh penelitian yang telah dilakukan.