PENGGUNAAN PEMODELAN TOPIK UNTUK PENCARIAN DOKUMEN TERMIRIP

Skripsi

Diajukan oleh:
Joseph Fernando Lim
71170141

PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA
YOGYAKARTA
2021

PENGGUNAAN PEMODELAN TOPIK UNTUK PENCARIAN DOKUMEN TERMIRIP

Skripsi

Diajukan kepada Program Studi Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana
Sebagai Salah Satu Syarat dalam Memperoleh Gelar
Sarjana Komputer

Disusun oleh

Joseph Fernando Lim

71170141

PROGRAM STUDI INFORMATIKA

FAKULTAS TEKNOLOGI INFORMASI

UNIVERSITAS KRISTEN DUTA WACANA

2021

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI SKRIPSI/TESIS/DISERTASI UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademika Universitas Kristen Duta Wacana, saya yang bertanda tangan di bawah ini:

Nama	: Joseph Fernando Lim
NIM	: 71170141
Program studi	: Informatika
Fakultas	: Teknologi Informasi
Jenis Karya	: Skripsi

demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Kristen Duta Wacana **Hak Bebas Royalti Noneksklusif** (*None-exclusive Royalty Free Right*) atas karya ilmiah saya yang berjudul:

"PENGGUNAAN PEMODELAN TOPIK UNTUK PENCARIAN DOKUMEN TERMIRIP"

beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti/Noneksklusif ini Universitas Kristen Duta Wacana berhak menyimpan, mengalih media/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat dan mempublikasikan tugas akhir saya selama tetap mencantumkan nama kami sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Yogyakarta
Pada Tanggal : 27 Juli 2021

Yang menyatakan

(Nama Lengkap)
NIM.71170141

PERNYATAAN KEASLIAN SKRIPSI

Saya menyatakan dengan sesungguhnya bahwa skripsi dengan judul:

PENERAPAN PERMODELAN TOPIK UNTUK PENCARIAN DOKUMEN TERMIRIP

yang saya kerjakan untuk melengkapi sebagian persyaratan menjadi Sarjana Komputer pada pendidikan Sarjana Program Studi Informatika Fakultas Teknologi Informasi Universitas Kristen Duta Wacana, bukan merupakan tiruan atau duplikasi dari skripsi kesarjanaan di lingkungan Universitas Kristen Duta Wacana maupun di Perguruan Tinggi atau instansi manapun, kecuali bagian yang sumber informasinya dicantumkan sebagaimana mestinya.

Jika dikemudian hari didapati bahwa hasil skripsi ini adalah hasil plagiasi atau tiruan dari skripsi lain, saya bersedia dikenai sanksi yakni pencabutan gelar kesarjanaan saya.

Yogyakarta, 21 Juni 2021

JOSEPH FERNANDO LIM
71170141

HALAMAN PERSETUJUAN

Judul Skripsi	:	PENGGUNAAN PEMODELAN TOPIK UNTUK PENCARIAN DOKUMEN TERMIRIP
Nama Mahasiswa	:	Joseph Fernando Lim
NIM	:	71170141
Matakuliah	:	Skripsi (Tugas Akhir)
Kode	:	TIW276
Semester	:	Genap
Tahun Akademik	:	2020/2021

Telah diperiksa dan disetujui di Yogyakarta,

Pada tanggal 2 Juli 2021

Dosen Pembimbing I	Dosen Pembimbing II
digitally signed on 28 May 2021	Digitally signed by Gloria Virginia 12.04, 28 May 2021
Dr. Phil. Lucia Dwi Krisnawati, S.S., M.A.	Gloria Virginia, S.Kom., MAI., Ph.D

HALAMAN PENGESAHAN

SKRIPSI

PENGGUNAAN PEMODELAN TOPIK UNTUK PENCARIAN DOKUMEN TERMIRIP

Oleh: JOSEPH FERNANDO LIM / 71170141

Dipertahankan di depan dewan Penguji Skripsi
Program Studi Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana – Yogyakarta
Dan dinyatakan diterima untuk memenuhi salah satu syarat memperoleh gelar
Sarjana Komputer
Pada tanggal 15 Juni 2021

Yogyakarta, 2 Juli 2021

Mengesahkan,

Dewan Penguji:

1. Lucia Dwi Krisnawati, Dr. Phil.
2. Gloria Virginia, S.Kom., MAI, Ph.D.
3. Antonius Rachmat C., S.Kom.,M.Cs.
4. Danny Sebastian, S.Kom., M.M., M.T.

Digitally signed by Gloria Virginia
Tujuan: Penguji Skripsi
Joseph Fernando Lim (71170141)

Dekan	Ketua Program Studi
(Restyandito, S.Kom., MSIS., Ph.D.)	(Gloria Virginia, Ph.D.)

UCAPAN TERIMA KASIH

Puji syukur kepada Tuhan Yang Maha Esa dengan segala rahmat dan berkat-Nya, sehingga penulis dapat menyelesaikan skripsi yang berjudul "Penggunaan Pemodelan Topik Untuk Pencarian Dokumen Termirip". Penelitian ini dimaksudkan untuk memenuhi salah satu syarat kelulusan dan memperoleh gelar Sarjana Komputer di Universitas Kristen Duta Wacana Yogyakarta.

Dalam pelaksanaanya, penulis mendapat banyak bantuan, dukungan dan masukan dari pihak-pihak terkait. Oleh karena itu, pada kesempatan ini penulis ingin menyampaikan rasa terima kasih kepada:

1. Keluarga yang sudah memberikan dukungan dan motivasi kepada penulis sehingga skripsi dapat diselesaikan tanpa hambatan
2. Yusta Cindy Claresta yang selalu memberi semangat kepada penulis untuk mengerjakan skripsi
3. Ibu Dr. Phil. Lucia Dwi Krisnawati, S.S., M.A. dan Ibu Gloria Virginia, S.Kom., selaku Dosen Pembimbing Skripsi yang bersedia memberikan saran dan arahan selama masa penulisan laporan tugas akhir
4. Pihak-pihak lainnya yang tidak dapat disebut satu persatu yang telah ikut serta membantu proses penulisan laporan ini secara langsung atau tidak langsung

INTISARI

PENGGUNAAN PEMODELAN TOPIK UNTUK PENCARIAN DOKUMEN TERMIRIP

Sebuah dokumen sering kali sulit diproses oleh komputer karena jika setiap kata dijadikan fitur maka dimensi data akan sangat besar dan akan memperlambat atau bahkan tidak memungkinkan untuk diproses (Shahmirzadi, Lugowski, & Younge, 2018). Salah satu pendekatan untuk mengatasi masalah tersebut adalah pemodelan topik. Pemodelan topik sering kali digunakan untuk menemukan pola-pola yang ada pada dokumen. Daripada menggunakan setiap kata sebagai fitur, akan lebih efektif jika hanya menggunakan beberapa kata saja untuk merepresentasikan suatu dokumen.

Salah satu platform yang menyediakan produk pemodelan topik adalah Lexikat. Penelitian ini berfokus pada penggunaan topik yang dihasilkan Lexikat sebagai fitur untuk mencari dokumen termirip. Term dan bobot dari topik yang dihasilkan Lexikat akan digunakan sebagai fitur. Fitur dari dokumen ini akan diukur kemiripannya dengan fitur dari dokumen lain menggunakan *cosine similarity*. Penulis akan menggunakan dataset 20Newsgroups dan bereksperimen dengan algortima Lexikat dengan mengubah parameter dan pembobotan untuk pemodelan topik. Hasil menunjukkan bahwa pemodelan topik yang menggunakan pembobotan TF-IDF menghasilkan skor terbaik jika jumlah topik yang digunakan sebagai query hanya 1 yaitu *F-measure* sebesar 0.323 dan *break-even point* sebesar 0.307, sedangkan pembobotan TF menghasilkan skor terbaik jika jumlah topik yang digunakan sebesar 50 yaitu dengan *F-measure* 0.274, dan *break-even point* sebesar 0.265. Pencarian dokumen menggunakan query dari pemodelan topik Lexikat berhasil mengurangi waktu pengukuran kemiripan dokumen sebanyak 50%-90%.

Kata Kunci— TFIDF, Pemodelan Topik, Pencarian Dokumen, Kolokasi, Lexikat

DAFTAR ISI

Bab 1
Pendahuluan

1.1 Latar Belakang Masalah

Perkembangan teknologi dari waktu ke waktu, membuat masyarakat sangat dipermudah dalam berbagai hal. Salah satunya ialah untuk mendapatkan informasi terbaru. Informasi adalah hal yang sangat penting untuk masyarakat agar mendapat update terbaru demi kelangsungan hidup bermasyarakat. Dengan adanya teknologi, kecepatan untuk mendapatkan informasi tidak sampai berhari-hari, bahkan bisa didapat hanya dalam hitungan detik. Informasi dapat ditemukan dimana saja mulai dari mesin pencari, media sosial, media massa dan masih banyak lagi.Tetapi dengan banyaknya informasi yang ada, membutuhkan waktu yang cukup lama untuk mencerna informasi-informasi yang ada dengan cara manual. Contohnya disaat orang-orang selesai membaca buku di perpustakaan, mereka akan mencari buku yang berkaitan dengan buku yang mereka baca sebelumnya. Hal yang pertama mereka lakukan adalah mencari buku dengan judul yang mirip. Tapi, dikarenakan satu dua hal, judul dari sebuah buku bisa saja tidak mencerminkan isinya. Orang-orang tadi pun akan membaca satu persatu isi dari buku tersebut untuk mengetahui apakah isinya berkaitan. Proses ini tentunya akan memakan waktu yang lama, sehingga manusia memanfaatkan komputer dalam menyelesaikan tugas ini, yaitu mencari dokumen yang mirip atau berkaitan.

Sebuah dokumen sering kali sulit diproses oleh komputer karena jika setiap kata dijadikan fitur maka dimensi data akan sangat besar dan akan memperlambat atau bahkan tidak memungkinkan untuk diproses (Shahmirzadi, Lugowski, & Younge, 2018). Salah satu contohnya adalah TF-IDF, TF-IDF yang berkepanjangan *Term Frequency – Inverse Document Frequency* memanfaatkan setiap kata sebagai fitur, lalu setiap fitur diberi nilai berdasarkan frekuensi fitur itu pada dokumen dan frekuensi kemunculan pada *corpus*. Jika seorang *developer* ingin membuat sistem untuk mencari dokumen termirip menggunakan *cosine similarity* yang pada database memiliki 10000 dokumen dan setiap dokumen memiliki 80 – 200 kosa kata, maka skenario terburuk adalah setiap dokumen harus mengukur kemiripan fiturnya dengan dokumen lain yang berukuran 140 sebanyak 9999 kali. Proses ini kurang efektif dan akan memakan waktu yang lama.

Ada beberapa pendekatan untuk mengatasi jumlah kosa kata yang berlebihan. Salah satunya adalah pemodelan topik. Pemodelan topik digunakan untuk mendapatkan topik dari suatu dokumen dengan menggunakan algoritma. Dalam penelitian ini, pemodelan topik

menggunakan asumsi bahwa setiap dokumen terdiri dari beberapa topik, dan setiap topik terdiri dari beberapa kata yang memiliki keterkaitan secara semantic. Pemodelan topik sering kali digunakan untuk menemukan pola-pola yang ada pada dokumen. Daripada menggunakan setiap kata sebagai fitur, akan lebih efektif jika hanya menggunakan beberapa kata saja untuk merepresentasikan suatu dokumen.

Salah satu platform yang menyediakan produk pemodelan topik adalah Lexikat. Lexikat merupakan start-up yang memenangkan program *Graduate Research Innovation Programme* (GRIP) dari *National University Singapore* (NUS). Dari kerjasama dengan UKDW, Lexikat membuka lowongan magang dan memperbolehkan karyawan magangnya untuk melanjutkan topik skripsi dari produknya.

Penelitian ini berfokus pada penggunaan topik yang dihasilkan Lexikat sebagai fitur untuk mencari dokumen termirip. Term dan bobot dari topik yang dihasilkan Lexikat akan digunakan sebagai fitur. Fitur dari dokumen ini akan diukur kemiripannya dengan fitur dari dokumen lain menggunakan *cosine similarity*.

1.2 Perumusan Masalah

Berdasarkan latar belakang yang sudah diurai sebelumnya, maka penelitian ini berusaha menjawab permasalahan tentang kinerja sistem pencarian dokumen jika menggunakan topik yang digenerasikan dari Lexikat sebagai fitur dokumen jika dibandingkan dengan yang tidak menggunakan topik dari Lexikat

1.3 Batasan Masalah

Untuk lebih fokus pada masalah yang diteliti, maka dalam skripsi ini batasan masalah yang ditentukan adalah sebagai berikut:

1. Data yang digunakan untuk mengukur kinerja sistem merupakan dataset yang bernama 20 Newsgroups. Penulis menggunakan salah satu pustaka eksternal sklearn untuk mendapatkan akses ke dataset ini.
2. Untuk menyesuaikan keperluan algoritma Lexikat, setiap dokumen paling sedikit memiliki 150 kata.
3. Agar performa sistem dapat diukur menggunakan *precision* dan *recall*, diperlukan label/*ground-truth* untuk setiap dokumen. Label ditentukan berdasarkan kategori, dokumen yang berada pada kategori yang sama merupakan dokumen yang memiliki label yang sama.

4. Menggunakan topik yang dihasilkan Lexikat sebagai fitur, dan menggunakan *Relative Term Frequency – Inverse Document Frequency* sebagai pembobotan fitur.

1.4 Tujuan Penelitian

Tujuan penelitian ini adalah untuk membangun sistem pencarian dokumen termirip dengan memanfaatkan topik dari aplikasi Lexikat sebagai fitur dokumen. Selain itu, penelitian ini juga bertujuan untuk mengamati apakah topik yang dihasilkan Lexikat dapat digunakan sebagai fitur untuk mendeteksi kemiripan dokumen berdasarkan topik.

1.5 Metodologi Penelitian

Penelitian akan dilakukan dalam beberapa tahap yaitu:

a. Studi Literatur

Tahap ini dilakukan agar dapat memahami permasalahan yang sudah dirumuskan sebelumnya. Proses yang dilakukan pada tahap ini adalah studi literatur tentang sistem pencarian dokumen dan permodelan topik.

b. Mempersiapkan Data

Tahap ini terdiri dari pengumpulan dan praproses data

1. Pengumpulan Data

   Mengambil data 20 Newsgroups menggunakan library sklearn.

2. Praproses Data

   Tahap ini untuk menormalisasi dokumen agar dokumen tidak menangkap fitur-fitur yang kurang penting dalam mengukur kemiripan dokumen.

   i. Penghapusan URL, Karakter Spesial dan Email
   ii. Penghapusan kosa kata berdasarkan document frequency
   iii. Lemmatisasi

c. Pemodelan Topik

Pada tahap ini, setiap dokumen dimodelkan untuk didapatkan topiknya. Lalu peneliti akan mengubah parameter untuk mendapatkan hasil terbaik untuk digunakan sebagai fitur dokumen.

d. Mengukur Kemiripan Dokumen

Tahap ini mengukur kemiripan dokumen dengan menggunakan algoritma *cosine similarity*.

e. Evaluasi

Mengukur performa sistem dengan menggunakan presisi, *recall*, *F-Measure* dan kecepatan dalam memproses pemodelan topik dan mengukur kemiripan dokumen.

1.6 Sistematika Penulisan

Penelitian ini terbagi menjadi beberapa bab, yang di mana masing-masing memiliki tujuan tersendiri sebagai berikut:

Bab 1 Pendahuluan, berisi pengantar untuk penelitian. Pendahuluan terdiri dari latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, metode penelitian, dan sistematika penulisan.

Bab 2 Tinjauan Pustaka dan Landasan Teori, terdiri dari tinjauan pustaka dan landasan teori. Tinjauan pustaka berisi tentang penelitian-penelitian terdahulu yang berhubungan dengan judul yang sedang diteliti. Landasan teori berisi tentang konsep utama untuk memecahkan masalah, bersumber dari buku dan jurnal.

Bab 3 Perancangan Sistem, berisi tentang rancangan alur sistem pencarian dokumen dan implementasi berdasarkan teori-teori yang didapatkan di Bab 2.

Bab 4 Hasil dan Pembahasan, bab ini berisi hasil dari penelitian dan pembahasan dari hasil tersebut.

Bab 5 Kesimpulan dan Saran, bab ini berisi kesimpulan dari sistem yang telah dikembangkan pada penelitian ini.

BAB V
KESIMPULAN

5.1 Kesimpulan

Kesimpulan yang didapatkan dari penelitian yang sudah dijalankan selama pengerjaan tugas akhir ini adalah sebagai berikut:

1. Pencarian dokumen termirip menggunakan *query* dari pemodelan topik Lexikat mengurangi waktu yang dihabiskan untuk pengukuran kemiripan dokumen sebanyak 50%-90% dan berperforma jauh lebih baik dibandingkan dengan menggunakan semua fitur sebagai *query*.
2. Skenario praproses data yang cocok untuk korpus 20 newsgroups adalah *preprocessed* dan *preprocessed_lemma*.
3. Pemodelan topik dengan pembobotan TF-IDF menghasilkan performa yang lebih baik dari pemodelan topik pembobotan TF jika jumlah topik yang digunakan sebagai *query* kurang dari 10, dan sebaliknya.
4. Jumlah topik yang digunakan sebagai query menentukan kespesifikan proses pencocokan dokumen, termasuk penggunaan kolokasi yang berakibat pada kinerja sistem. Hal ini menunjukkan bahwa jumlah query dan kolokasi berbanding terbalik dengan performa sistem yang diukur menggunakan *precision*, *recall* dan *F-Measure*.

5.2 Saran

Dari penelitian ini, terdapat beberapa saran untuk pengembangan penelitian selanjutnya.

1. Data yang digunakan pada penelitian mayoritas merupakan dokumen pendek. Pada penelitian selanjutnya, dapat diujicobakan dataset yang terdiri dari dokumen berukuran panjang untuk menguji apakah panjang dokumen berpengaruh terhadap efektivitas query yang dihasilkan oleh Lexikat.
2. Pada penelitian ini stopwords yang digunakan didapati dari Lexikat. Akan lebih baik jika menggunakan stopwords yang disesuaikan dengan *corpus*.
3. Kolokasi pada penelitian ini dianggap sebagai fitur yang berdiri sendiri yang artinya fitur/kata pembentuk kolokasi tidak dihilangkan. Pada penelitian selanjutnya, dapat diujicobakan apakah kolokasi yang berdiri sendiri menghasilkan query yang lebih baik dibandingkan jika fitur/kata pembentuk kolokasi dihilangkan.

Daftar Pustaka

Anugrah, I. G., & Rosyid, H. (2019). Penerapan Information Retrieval Menggunakan Pemodelan Topik Pada Deskripsi Portal Multimedia . *Jurnal Nasional Komputasi dan Teknologi Informasi*.

Ariafandi, F. (2011). Implementasi Vector Space Model untuk Pencarian Ayat dalam Kitab Perjanjian Baru.

Christopher D. Manning, P. R. (2009). *An Introduction to Information Retrieval.* England.

David M. Blei, A. Y. (2003). Latent Dirichlet Allocation.

Imbar, R. V., Ayub, M., & Rehatta, A. (2014). Implementasi Cosine Similarity dan Algoritma Smith-Waterman untuk Mendeteksi Kemiripan Teks. *Jurnal Informatika*.

Lexikat. (2018). *lexikat.com*. Retrieved from Lexikat: https://lexikat.com

Myeong-Ha Hwang, d. (2017). Related Document Extraction based on Topic Modelling using Cloud System. *International Journal of Grid and Distributed Computing*.

Ramos, J. (2003). Using TF-IDF to Determine Word Relevance in Document Queries .

Shahmirzadi, O., Lugowski, A., & Younge, K. (2018). Text Similarity in Vector Space Models:A Comparative Study.

Thompson, V. U., Panchev, C., & Oakes, M. (2015). Performance Evaluation of Similarity Measures on Similar and Dissimilar Text Retrieval. *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015)* , 577-584.

Wood, T. (2019, 5 17). *What is the F-Score*. Retrieved from DeepAI: https://deepai.org/machine-learning-glossary-and-terms/f-score

Nastase, V. 2012. Introduction to Topic Model. ICL, University of Heidelberg.

Akwei, J. 2019. ContextBase – Topic Modeling