Topik : **Pengaruh Preprocessing Teks terhadap Akurasi Sistem Temu Kembali Informasi dalam Pencarian Dokumen**

Latar Belakang

Memproses teks untuk akurasi dalam sistem pencarian informasi. Preprocessing adalah langkah penting dalam pemrosesan teks yang memungkinkan sistem mengekstrak informasi yang relevan dengan lebih efisien. Teknik prapemrosesan seperti tokenisasi, penghapusan stopword, stemming, dan lemmatization sangat penting untuk menganalisis input teks yang lebih terstruktur, yang memungkinkan mesin untuk lebih memahami hubungan antar kata.

Dalam sistem pencarian informasi, proses pencarian yang sebenarnya memiliki dampak yang signifikan terhadap kualitas data yang dimasukkan ke dalam model. Dokumen yang ditulis dengan buruk dapat mengandung banyak kata yang tidak relevan, bentuk kata yang beragam, dan konten yang sulit dipahami sehingga mengurangi efektivitas model dalam memahami konteks. Akibatnya, hasil pencarian menjadi tidak relevan dan menghambat proses pencarian.

Proyek penelitian ini akan menganalisis dan menilai keefektifan setiap metodologi preprocessing yang diterapkan pada keakuratan penyimpanan dokumen dalam sistem temu kembali informasi. Model-model seperti Word2Vec dan Term Frequency-Inverse Document Frequency (TF-IDF) akan digunakan untuk mengukur dampak dari teknik-teknik yang telah disebutkan di atas. Dengan membandingkan hasil data yang diproses dan yang tidak diproses, diharapkan penelitian ini akan memberikan wawasan tentang bagaimana preprocessing teks dapat meningkatkan relevansi dan akurasi hasil pada sistem pemindaian dokumen, serta mengidentifikasi teknik preprocessing yang paling efektif.

Masalah dan tujuan

- Masalah

Sistem temu kembali informasi sering kali menghasilkan hasil yang tidak sepenuhnya akurat atau relevan, terutama jika tidak ada langkah prapemrosesan yang efektif yang diterapkan pada data. Teks yang dikirimkan secara diam-diam ke model pencarian dapat berisi berbagai jenis kata, kata-kata yang tidak penting (stopwords), dan data yang berlebihan yang meningkatkan kemampuan model untuk memahami konteks secara lengkap. Sebagai contoh, kata-kata yang memiliki arti dasar yang sama, seperti "mengajar" dan "mengajarkan", dapat ditafsirkan sebagai entitas yang berbeda jika tidak dilakukan stemming atau lemmatisasi, yang menurunkan kualitas hasil penelitian.

Selain itu, artikel pengetahuan kata teks dalam dokumen yang berasal dari beberapa bidang sering kali membuat pencarian Model kurang ideal. Ketidakmampuan model untuk menganalisis kata-kata yang jelas relevan dari teks dapat menyebabkan pengguna kesulitan mengingat informasi yang mereka butuhkan, yang menurunkan produktivitas dan efisiensi waktu. Tanpa prapemrosesan yang tepat, sistem juga memiliki kemampuan untuk menghasilkan output komputer yang lebih besar tanpa mencapai peningkatan kinerja yang berarti.

- Tujuam

Penelitian ini bertujuan untuk memahami bagaimana teknik prapemrosesan teks dapat meningkatkan akurasi hasil dalam sistem informasi temu kembali. Dengan mengimplementasikan teknik-teknik seperti tokenisasi, penghapusan stopword, stemming, dan lemmatization, penelitian ini akan meningkatkan akurasi model dan relevansi hasil yang diperoleh. Secara khusus, tujuan dari penelitian ini adalah:

1. Tujuannya adalah untuk meningkatkan kualitas dan akurasi persiapan dokumen dengan mengevaluasi dan mengurangi dampak dari setiap teknik preprocessing, baik yang digunakan secara terpisah maupun dikombinasikan.
2. membandingkan Model Akurasi: Untuk mengidentifikasi perubahan akurasi dan relevansi hasil, bandingkan hasil penelitian antara data yang telah diproses (melalui pre-processing) dan data yang belum diproses.
3. Menentukan Teknik Preprocessing Terbaik: mengidentifikasi teknik Preprocessing atau kombinasi teknik yang paling efektif untuk meningkatkan kemampuan sistem dalam menangani informasi kembali. Hal ini akan memberikan informasi mengenai teknik preprocessing yang paling efektif digunakan untuk membuat sistem pencarian menjadi lebih efisien.

Sumber Dataset dan penjelasan

Sumber berasal dari : CISI (a dataset for Information Retrieval)
https://www.kaggle.com/datasets/dmaso01dsta/cisi-a-dataset-for-information-retrieval/data

Gambaran Umum dari Dataset

1. Dataset CISI (Koleksi Uji Temu Kembali Informasi) merupakan salah satu koleksi dataset standar yang digunakan untuk menguji dan mengevaluasi kemampuan sistem dalam mengambil informasi. Dataset ini terdiri dari kumpulan abstrak atau ringkasan artikel untuk penelitian dari bidang perpustakaan dan informasi dan berasal dari konferensi dan jurnal pada tahun 1960-an. CISI sangat membantu dalam mengurangi keefektifan algoritma pencarian dengan menyediakan struktur data yang mirip dengan kasus pencarian informasi di perpustakaan.
2. Dataset untuk Struktur: Sekitar 1.460 dokumen yang menyediakan informasi dalam format teks membentuk dokumen ini. Setiap dokumen memiliki beberapa fitur utama:
   ID Dokumen: Identifikasi unik untuk setiap dokumen.
   Judul: Judul artikel atau studi yang ditulis dengan baik.
   Abstrak / Isi Utama: Elemen utama yang menjelaskan ide utama atau tesis dari penelitian.
   Referensi: Rujukan atau daftar pustaka yang digunakan dalam dokumen yang disebutkan di atas (atau beberapa dokumen).
   Kumpulan Kueri: Dataset ini juga mencakup sekitar 112 query (pertanyaan) standar yang sering digunakan untuk menilai kinerja sistem saat mengambil informasi. Setiap pertanyaan memiliki ID dan deskripsi teks yang menjelaskan informasi spesifik yang dibutuhkan.
   Relasi Relevansi (Penilaian Relevansi): Selain dokumen dan pertanyaan, CISI menyediakan daftar dokumen yang relevan, yang berisi daftar dokumen yang benar-benar relevan dengan setiap pertanyaan. Hal ini sangat membantu untuk mengevaluasi model pencarian karena memungkinkan

Anda untuk menggunakan metrik evaluasi seperti F1-Score, Precision, dan Recall.

3. Penggunaan dan Manfaat dalam Penelitian: Dataset CISI sering digunakan untuk menilai berbagai metode dalam sistem temu kembali informasi, terutama dalam: mengevaluasi efektivitas teknik persiapan teks (seperti stemming dan penghilangan stopword) pada temuan penelitian. membandingkan kinerja beberapa algoritma pencarian, seperti Word Embeddings, BM25, dan TF-IDF. Melakukan eksperimen dalam klasifikasi dokumen dengan menggunakan metode machine learning.
4. Tantangan Penggunaan CISI: Terlepas dari kegunaan CISI dalam penelitian pencarian informasi, dataset ini berukuran kecil dan mungkin tidak dapat menangkap data terkini yang lebih komprehensif. Oleh karena itu, CISI sering digunakan sebagai baseline untuk analisis pertama atau model sederhana sebelum menerapkan model tersebut pada dataset yang lebih besar dan kompleks.

Alur dan tahapan experimen :

1. Pengumpulan dataset
   pada tahapan ini saya menggunakan dataset publik yang berasal dari Kaggle CISI (a dataset for Information Retrieval)
   https://www.kaggle.com/datasets/dmaso01dsta/cisi-a-dataset-for-information-retrieval/data
2. Prepocessing
   Menggunakan berbagai cara prepocessing pada dokumen dan kueri, seperti
   - Tokenisasi adalah proses mengubah teks menjadi istilah unit-unit atau token.
   - Penghapusan Kata Henti: Menggunakan kata-kata yang biasanya tidak menyampaikan informasi, seperti "di", "dan", dan "yang.
   - Lemmatisasi dan stemming: mengubah kata menjadi bentuk dasarnya untuk mengurangi variasi bentuk kata yang ada (misalnya, "belajar" dan "belajaran" menjadi "belajar"). Membandingkan hasilnya dengan data yang belum diproses memungkinkan seseorang untuk mengamati bagaimana prapemrosesan memengaruhi keakuratan hasil.

3. Fitur ekstraksi
   Menggunakan metode representasi teks seperti:
   Term Frequency-Inverse Document Frequency (TF-IDF): Menentukan jumlah kata berdasarkan frekuensi kemunculan dalam dokumen dan semua dataset.
   Penyematan kata, seperti Word2Vec, digunakan untuk menyematkan teks dalam format vektor berdasarkan konteks teks dalam kumpulan data. membandingkan dan membedakan representasi teks yang berbeda terkait dengan efektivitas pencarian.

4. Pengembangan model sistem temu kembali informasi
   Pengembangan dan evaluasi model informasi temu kembali yang dapat menilai relevansi dokumen berdasarkan query, termasuk: Algoritma pencarian seperti K-Nearest Neighbor (KNN), Support Vector Machine (SVM), atau metode pencarian berbasis vektor lainnya. menggunakan hasil preprocessing dari langkah sebelumnya dan membandingkan performa model dengan data yang sudah diproses dan data mentah.
5. Evaluasi kinerja model

Meningkatkan produktivitas sistem menggunakan metrik evaluasi yang relevan, seperti: F1-Score, Precision, dan Recall: Untuk menentukan keakuratan hasil dan relevansi dokumen yang dievaluasi oleh model. gunakan data relevansi dataset CISI untuk mengidentifikasi beberapa model yang efektif untuk mengidentifikasi dokumen yang relevan untuk setiap kueri.

6. Analisis hasil dan kesimpulan
   Menganalisis hasil kerja dari setiap teknik preprocessing dan membandingkannya untuk menentukan pendekatan yang paling efektif. Menyusun laporan dan kesimpulan mengenai dampak preprocessing terhadap akurasi sistem temu kembali informasi, serta memberikan saran untuk teknik preprocessing yang dapat digunakan pada sistem lain.
7. Pengembangan lebih lanjut
   Jika perlu, lakukan eksperimen lanjutan dengan menggunakan teknik atau model yang berbeda, atau uji sistem pada kumpulan data yang lebih besar untuk menguji generalisasi temuan penelitian.