NEWS REPRESENTATION MENGGUNAKAN MATRIX METHOD

Istnain Faiz Alfajri, Jeason Ibrahim
[1] Fakultas Teknik, Pendidikan Teknik Informatika dan Komputer, Universitas Negeri Jakarta
[2] Fakultas Teknik, Pendidikan Teknik Informatika dan Komputer, Universitas Negeri Jakarta
istnainfaizalfajri@gmail.com[1], jeasonibrahim@gmail.com[2]

Abstract

This research is part of data mining, specifically within the realm of information retrieval and text mining. The focus of this study is to find approaches to retrieve relevant online news documents based on a specific threshold value and improve computer performance in retrieving a large number of relevant documents. In this regard, the author utilizes news articles from three popular news websites in Indonesia, namely tribunnews.com, detik.com, and liputan6.com. To search for relevant news documents, the author first establishes a threshold value by calculating the average similarity score of the documents used as test samples. This threshold value is then used to determine the similarity score for each document to be retrieved. The author also employs several techniques in this research process, such as text mining using configuration methods and document representation techniques using matrix methods and tala methods. Lastly, the author utilizes the cosine similarity method to determine the level of similarity between documents using matrix-based retrieval data. The research findings indicate that the approach using matrix methods and matrix compression processes yield good calculation results, making it applicable to a large number of documents.

Keywords: *Information Retrieval, Matrix Method, Text Mining.*

Abstrak

Penelitian ini merupakan bagian dari data mining, yaitu bagian information retrieval dan bagian text mining. Fokus dari penelitian ini adalah menemukan cara untuk mendapatkan kembali dokumen berita online yang relevan dengan nilai threshold tertentu, dan juga untuk meningkatkan kinerja komputer saat mengambil sejumlah besar dokumen yang relevan. Dalam hal ini penulis menggunakan berita dari tiga website berita yang cukup populer di Indonesia yaitu tribunnews.com, detik.com dan liputan6.com. Untuk mencari dokumen berita yang relevan, terlebih dahulu penulis menetapkan nilai threshold dengan menghitung rata-rata nilai kemiripan dokumen yang digunakan sebagai sampel uji. Nilai ambang ini kemudian digunakan untuk menentukan nilai kesamaan dari setiap dokumen yang akan digunakan. Penulis juga menggunakan beberapa teknik dalam proses penelitian ini, seperti text mining menggunakan metode konfigurasi dan teknik penyajian dokumen berita menggunakan metode matriks dan metode tala. Terakhir, penulis menggunakan metode cosine similarity untuk menentukan tingkat kemiripan antar dokumen dengan menggunakan data temu kembali berbasis matriks. Hasil penelitian menunjukkan bahwa pendekatan dengan menggunakan metode matriks dan proses kompresi matriks memberikan hasil perhitungan yang baik, sehingga dapat diterapkan pada dokumen dalam jumlah besar.

Kata Kunci: Information Retrieval, Matrix Method, Text Mining.

1. PENDAHULUAN

Website merupakan salah satu teknologi yang sedang dikembangkan untuk dengan cepat dan mudah menyebarkan informasi yang dapat diakses oleh banyak orang. Menurut Sekretaris Jenderal APJII (Asosiasi Penyelenggara Jasa Internet Indonesia), Henri Kasyfi Soemartono, penetrasi pengguna internet di Indonesia telah meningkat dari 64,8% pada tahun 2018 menjadi 73,7% pada tahun 2020. Ini berarti sekitar 196,7 juta orang di Indonesia menggunakan internet. Peningkatan ini disebabkan oleh faktor-faktor seperti penyebaran infrastruktur internet cepat dan merata melalui Palapa Ring, transformasi digital yang semakin masif karena pembelajaran *online*, dan kebijakan bekerja dari rumah akibat pandemi Covid-19 sejak Maret 2020.

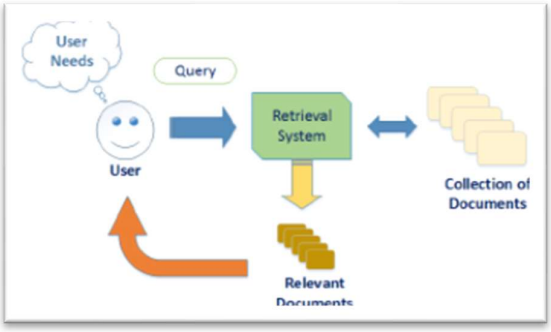
APJII telah melakukan survei pada tahun 2020 menggunakan teknik sampling seperti *probabilitas sampling*, *multistage*

random sampling, dan *area variant random sampling*. Survei ini menunjukkan bahwa mayoritas masyarakat Indonesia mencari informasi melalui media sosial, televisi, berita *online*, dan situs web resmi pemerintah. Hal ini menunjukkan bahwa orang-orang cenderung menggunakan perangkat digital seperti ponsel dan laptop untuk mengakses informasi. Situs web menjadi pilihan utama karena fleksibilitas dan aksesibilitasnya yang tinggi.

Pengembangan website sebagai media massa telah menghasilkan peningkatan drastis dalam jumlah artikel berita. Dalam pengamatan yang dilakukan pada tiga situs berita populer di Indonesia (Tribunnews.com, Detik.com, dan Liputan6.com) menggunakan teknik *scraping* dari Januari hingga Oktober 2021, ditemukan sebanyak 210.997 berita yang digunakan sebagai referensi oleh penulis. Meskipun jumlah beritanya sudah cukup, pembaca sering mengalami kesulitan dalam menemukan berita yang sesuai dengan apa yang mereka cari. Penggunaan judul artikel *clickbait* oleh media *online* menjadi salah satu penyebabnya, karena sering kali mengganggu rasa ingin tahu pembaca. Hal ini menyebabkan pembaca harus membaca berbagai dokumen berita dan memeriksa secara manual kebenarannya, yang memakan waktu dan sulit dilakukan.

Di sisi lain, saat ini metode pengambilan informasi umumnya menggunakan teknik TF-IDF. Namun, metode ini membutuhkan waktu pemrosesan yang lama dan tidak efisien ketika digunakan pada big data. Oleh karena itu, sebuah teknik alternatif menggunakan pendekatan matriks dan proses kompresi matriks diusulkan untuk meningkatkan kinerja komputasi dalam mengembalikan dokumen berita yang relevan. Teknik ini diharapkan dapat mempercepat proses komputasi meskipun digunakan pada dokumen-dokumen yang besar.

Information Retrieval adalah bidang ilmu yang mempelajari metode dan prosedur untuk mengambil informasi yang tersimpan dari berbagai sumber yang relevan dengan kebutuhan pengguna. Dalam pencarian data, terdapat berbagai jenis data seperti teks, tabel, gambar, video, dan suara. Tujuan utama *Information Retrieval* adalah memenuhi kebutuhan informasi pengguna dengan menemukan kembali dokumen-dokumen yang relevan atau mengurangi dokumen-dokumen yang tidak relevan dalam pencarian. Alur umum dari proses *Information Retrieval* dapat dilihat pada gambar.



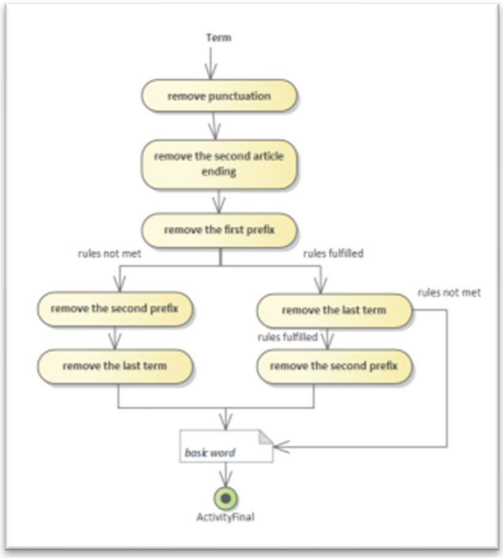
Gambar 1. *Flowchart Information retrieval*

Gambar 1 menggambarkan proses menampilkan dokumen yang relevan kepada pengguna dalam *Information Retrieval*. Proses dimulai dengan membaca kueri pertanyaan yang diajukan oleh pengguna. Selanjutnya, sistem menggunakan model *Information Retrieval* untuk memeriksa kueri terhadap dokumen yang tersimpan. Hasilnya kemudian dikembalikan kepada pengguna dalam bentuk dokumen yang relevan.

Text mining merupakan variasi dari sub-bidang data mining yang fokus pada penemuan pola atau karakteristik menarik dalam kumpulan data teks yang besar. Dalam proses *text mining*, langkah awal yang penting adalah *text preprocessing*. *Text preprocessing* adalah proses mengubah data teks yang tidak terstruktur menjadi data terstruktur yang lebih dapat diolah. Terdapat empat langkah utama dalam text preprocessing, yaitu *case-folding* (merubah semua huruf menjadi huruf kecil), *tokenizing* (memecah teks menjadi token atau kata-kata), *filtering* (menghapus kata-kata yang tidak relevan seperti stop words), dan *stemming* (mengubah kata-kata menjadi bentuk dasarnya).

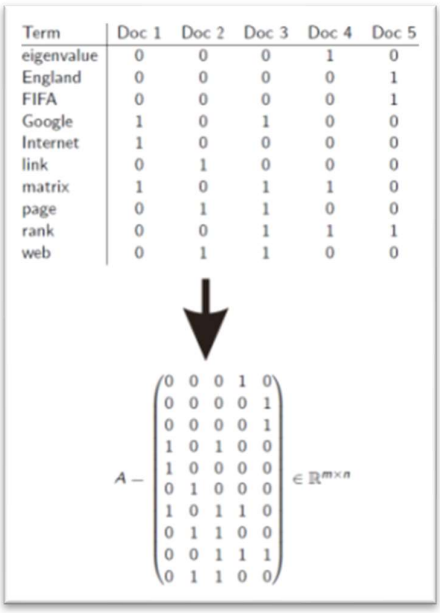
Belakangan ini, terdapat banyak teknik *stemming teks* yang membantu mendapatkan kata-kata penting dalam teks. Studi sebelumnya juga membandingkan beberapa algoritma stemming seperti algoritma Nazief dan Andriani, algoritma vega, algoritma Arifin, Setiono, dan algoritma tala. Namun, penulis menggunakan algoritma tala dalam penelitian

ini karena waktu komputasinya relatif lebih cepat, dengan akurasi lebih dari 75%. Gambaran proses dapat dilihat pada Gambar 2.



Gambar 2. *Flowchart Tala stemming algorithm.*

Teknik representasi dokumen digunakan untuk membantu menyederhanakan proses perhitungan data. Salah satu teknik representasi dokumen yang digunakan oleh penulis untuk mendukung penelitian adalah metode matriks sesuai dengan referensi Lars Elden dan Berkant Savas, Departemen Matematika, Universitas Linkoping, Swedia pada tahun 2012 tentang Data Mining menggunakan Metode Matriks. Teknik ini akan menjelaskan bagaimana dokumen berita yang digunakan akan direpresentasikan dalam format matriks mengikuti aturan yang ada. Ukuran matriks A, di mana untuk ukuran m ditentukan oleh jumlah kata penting yang disimpan, maka ukuran matriks A untuk ukuran n akan ditentukan oleh berapa jumlah total dokumen yang disimpan. Kemudian penentuan angka 1 dalam matriks jika dokumen mengandung kata tertentu. Sebaliknya, angka tersebut akan menjadi 0 jika tidak mengandung kata tertentu sesuai dengan semua baris kata penting dalam dokumen sampel sehingga akan menghasilkan representasi matriks A akhir A m×n dari dokumen berita dan pentingnya dokumen kata yang digunakan. Dalam proses penelitian ini, menggunakan 100 dokumen berita sampel, dihasilkan ukuran matriks akhir A 3760×100. Gambaran proses dan hasil dengan teknik ini dapat dilihat pada Gambar 3.



Gambar 3. Gambaran proses dan hasil dari metode matriks

Pembuatan matriks merupakan langkah awal untuk proses identifikasi vektor untuk setiap dokumen sampel; setelah menghasilkan matriks, kita dapat mendapatkan vektor dokumen masing-masing dari kolom matriks akhir yang dihasilkan. Sebagai contoh, berdasarkan Gambar 3, vektor dokumen 1 adalah $\vec{V}$ = (0,0,0,1,1,0,1,0,0,0) dan seterusnya.

Teknik penghitungan kemiripan antara dua objek dapat bervariasi. Salah satu teknik perhitungan yang digunakan oleh penulis untuk menghitung kemiripan dokumen dengan data pencarian adalah menggunakan Cosine Measure. Rumus yang digunakan Cosine Measure adalah:

$$Cos\ \alpha = \frac{A.B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i\, B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\ \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

A = Vektor A, yang akan dibandingkan kemiripannya

B = Vektor B, yang akan dibandingkan kemiripannya

A.B = hasil perkalian titik antara vektor A dan vektor B

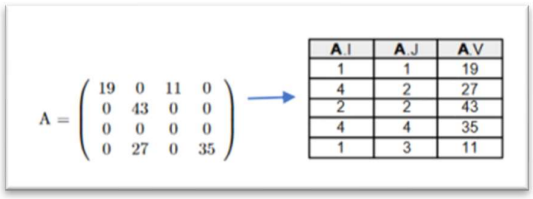
|A| = panjang vektor A

|B| = panjang vektor B

|A||B| = hasil perkalian silang antara |A| dan |B|

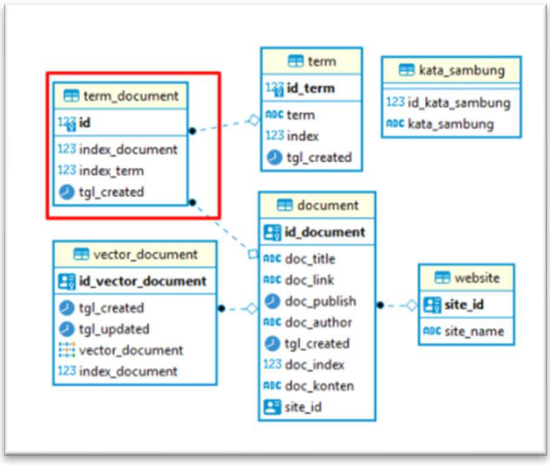
Matriks sparse adalah matriks di mana sebagian besar nilainya adalah 0 nol. Salah satu cara untuk mendeteksi bahwa matriks tersebut adalah matriks sparse biasanya ditentukan dari nilai kepadatannya (sparsity). Sebuah matriks dikatakan sebagai matriks sparse jika nilai kepadatannya > 0,5. Baru-baru ini, beberapa teknik telah muncul untuk menyimpan matriks sparse; salah satu teknik penyimpanan matriks sparse adalah representasi array (triples). Penulis akan menggunakan teknik ini untuk merepresentasikan dokumen dalam matriks, seperti yang terlihat pada Gambar 3. Penulis memilih teknik representasi array (triples) karena teknik ini merupakan yang paling populer dan mudah diimplementasikan. Teknik ini juga diharapkan dapat mengurangi media penyimpanan matriks dalam jumlah yang besar.

Cara paling sederhana untuk merepresentasikan matriks sparse adalah dengan format triples (atau koordinat). Untuk setiap A(i,j) ≠ 0, triple (i,j,A(i,j)) disimpan dalam memori. Setiap entri dalam triple biasanya disimpan dalam array yang berbeda, dan seluruh matriks A direpresentasikan sebagai tiga array A.I (indeks baris), A.J (indeks kolom), dan A.V (nilai numerik), seperti yang diilustrasikan pada Gambar.



Gambar 4. *Representasi tripel*

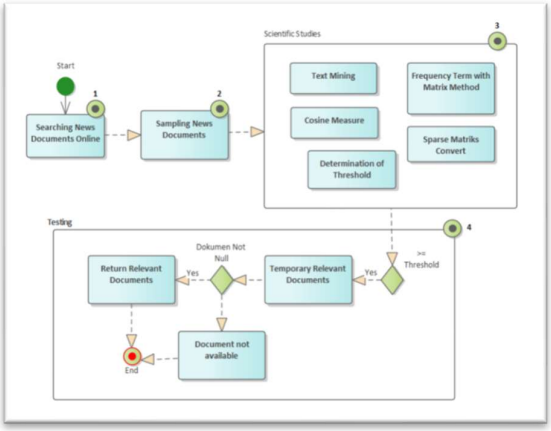
Penulis menggunakan konsep teknik ini untuk merepresentasikan istilah dokumen dalam basis data sehingga representasi vektor yang disimpan dalam basis data akan lebih efisien. Berikut adalah gambar relasi basis data yang dikembangkan menggunakan konsep representasi array (triples).



Gambar 5. Representasi istilah dokumen menggunakan konsep representasi array (tripel).

2. METODE PENELITIAN

Penelitian ini dilakukan dalam beberapa tahapan. Tahapan-tahapan dalam penelitian ini digambarkan dalam Gambar 6.



Gambar 6. Tahapan Penelitian

Gambar 6 menggambarkan tahapan-tahapan penelitian yang akan dilakukan oleh penulis. Langkah awal adalah menemukan kebutuhan akan dokumen berita yang digunakan sebagai sampel dalam penelitian ini. Dokumen berita sampel merujuk pada situs berita di Indonesia, yaitu Tribunnews.com, detik.com, dan liputan6.com. Peneliti mendapatkan dokumen tersebut dari penelitian

sebelumnya. Sampai saat ini, penulis telah mendapatkan 210.997 dokumen dan menggunakan data sampel eksperimental sebanyak 10 dokumen berita untuk perhitungan manual. Langkah kedua melakukan sampling acak terhadap 100 dokumen berita dari waktu publikasi berita terbaru.

Langkah selanjutnya, nomor 3, adalah peneliti melakukan studi literatur terkait dengan kebutuhan untuk menemukan dokumen yang relevan, seperti penerapan text mining dengan metode tala, penerapan cosine measure, kondisi untuk menentukan nilai ambang (threshold), dan lain-lain. Setelah langkah nomor 3 selesai, peneliti akan melanjutkan pengujian untuk melihat apakah telah mendapatkan dokumen yang relevan dari kata kunci pencarian

3. HASIL DAN PEMBAHASAN

3.1 Penentuan Nilai Ambang Dokumen Pencarian

Penentuan nilai minimum dari kesamaan kosinus antara dokumen dengan query diperlukan untuk memenuhi persyaratan dokumen dengan nilai kesamaan tertentu. Penulis menggunakan sampel 5 percobaan dengan menguji hasil kesamaan antara dokumen berita dan *query* menggunakan rumus Cosine Measure. Setiap uji coba menggunakan *query* dan dokumen sampel yang berbeda untuk memaksimalkan hasil akhir yang diperoleh. Nilai absolut kesamaan minimum dianggap sebagai persyaratan minimum untuk menyatakan bahwa dokumen berita mirip dengan *query*. Proses ini harus dilakukan terlebih dahulu untuk menentukan nilai ambang dokumen yang dikembalikan kepada pengguna. Berikut adalah nilai akhir dari hasil kesamaan yang diperoleh.

$$(threshold) = \frac{\left(\begin{matrix}0{,}155 + 0{,}196 + 0{,}137 + 0{,}201 \\ +0{,}123\end{matrix}\right)}{5} = \frac{0{,}812}{5} = 0{,}162$$

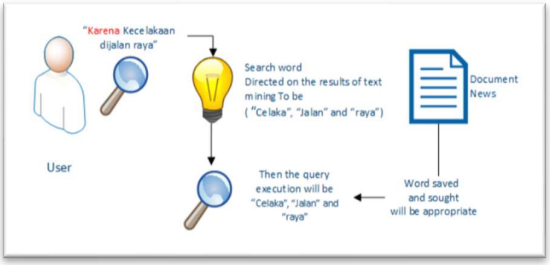
3.2 Ilustrasi Data

Pendekatan yang ditemukan untuk menemukan dokumen berita yang relevan dibagi menjadi beberapa bagian, yaitu pertama-tama merepresentasikan dokumen berita menjadi matriks Am×n untuk pemrosesan query, kemudian menghitung kesamaan vektor representasi query dengan matriks representasi dokumen berita. Terakhir, menentukan nilai ambang cosine measure untuk mengembalikan dokumen yang relevan dengan memilih nilai ambang dokumen pencarian untuk menghasilkan dokumen yang relevan dan mendapatkan semua dokumen koleksi yang memiliki kesamaan dengan konten berita lainnya.

Proses pertama adalah merepresentasikan dokumen berita ke dalam matriks berukuran Am×n. Proses ini digunakan untuk menampung kata-kata penting yang ada dan dokumen mana yang mengandung kata-kata tersebut, kemudian kita masukkan ke dalam format matriks Am×n. Hasil representasi matriks dapat dilihat pada Gambar 3.

Penulis menerapkan konsep representasi array (triples) dalam tabel term_document yang terdapat pada Gambar 3. Informasi dalam tabel term_document akan mempercepat proses komputasi dan penyimpanan karena hanya menyimpan data kata-kata yang hanya ada dalam dokumen tertentu. Sehingga ketika direpresentasikan dalam bentuk vektor dokumen, informasi dalam tabel term_document akan diterjemahkan ke dalam vektor yang nantinya akan berisi nilai 0 atau 1, yang berarti tidak termasuk kata-kata tertentu atau mengandung istilah-istilah khusus.

Proses selanjutnya adalah pemrosesan kueri yang dimulai dengan melakukan pemisahan kueri dalam proses pencarian dokumen. Ilustrasi dari pemrosesan kueri dapat dilihat pada Gambar.



Gambar 7. Ilustrasi dari pemrosesan kata menggunakan teknik text mining.

Berdasarkan Gambar 7, dapat disimpulkan bahwa proses ini mengatasi kesalahan makna kata sesuai dengan kata asli atau unsur-unsurnya. Kata-kata yang disimpan dalam database atau hasil text mining hanya menyimpan kata-kata penting atau kata-kata yang bukan konjungsi. Penulis tidak menggunakan konjungsi dalam bahasa Indonesia karena semua dokumen akan mengandung konjungsi ini untuk menghilangkan identitas unik dari setiap dokumen berita. Ketika pengguna memasukkan kata dengan konjungsi, kata tersebut akan dihancurkan. Oleh karena itu, diperlukan pemrosesan split query pada awal proses untuk mengarahkan kata pencarian yang sesuai dengan kata-kata yang disimpan dalam database.

Proses selanjutnya adalah menghitung representasi matriks dari dokumen berita dengan representasi vektor dari query. Penulis menggunakan teknik cosine measure untuk mencari kemiripan antara keduanya. Nilai ambang batas kemiripan dokumen berita ditemukan sebesar 0,162. Hal ini berarti bahwa ketika mengembalikan dokumen yang relevan, kondisi nilai cosine measure antara vektor representasi query dan matriks representasi dokumen berita harus lebih besar atau sama dengan >= 0,162. Oleh karena itu, dokumen-dokumen yang disediakan harus memenuhi persyaratan nilai ambang batas tersebut, sehingga diharapkan hasil rekaman yang dikembalikan benar-benar relevan atau sesuai dengan data pencarian. Uji coba ini telah dilakukan oleh penulis dan menunjukkan hasil yang menjanjikan dengan waktu eksekusi rata-rata fungsi di database untuk 100 dokumen dengan rata-rata 200-210 kata yang terdapat dalam dokumen. Hanya membutuhkan waktu 5-10 detik.

4. PENUTUP

Kesimpulan

Berdasarkan penelitian yang telah dilakukan, dapat disimpulkan beberapa pendekatan dan hasil yang diperoleh, yaitu:

- Berdasarkan hasil pengujian penentuan batas minimum cosine yang telah dilakukan, nilai kemiripan cosine minimum yang diperlukan untuk mendapatkan dokumen relevan adalah 0,162. Hal ini berarti bahwa ketika menampilkan dokumen berita yang relevan, diperlukan nilai ambang batas sebagai nilai acuan yang akan ditampilkan berdasarkan kemiripan antara dokumen tertentu dan query yang dimasukkan. Dengan demikian, saat menampilkan dokumen yang relevan, harus memiliki nilai cosine >= ambang batas (0,162).
- Penggunaan metode matriks menjadi efisien untuk menghitung kemiripan dengan menggunakan cosine measure, dan menggabungkannya dengan teknik penyimpanan matriks yang jarang terisi (sparse matrix) menggunakan representasi array (triples). Hal ini akan mempercepat proses komputasi dan mengurangi penggunaan ruang penyimpanan ketika proses implementasi dilakukan dalam program komputer. Pengujian ini telah dilakukan oleh penulis dan menunjukkan hasil yang menjanjikan dengan waktu eksekusi rata-rata fungsi di database untuk 100 dokumen dengan rata-rata 200-210 kata yang terdapat dalam dokumen. Hanya membutuhkan waktu 5-10 detik.

5. REFERENSI

[1] L. Eld, TANA07 : Data Mining using Matrix Methods, 2012.

[2] Alexa - Top Sites in Indonesia - Alexa. https://www.alexa.com/topsites/countries/ID (accessed Dec. 17, 2021).

[3] M. S. H. Simarangkir, Studi Perbandingan Algoritma - Algoritma Stemming Untuk Dokumen Teks Bahasa Indonesia, J. Inkofar, vol. 1, no. 1, pp. 40–46, 2017, doi: 10.46846/jurnalinkofar.v1i1.2.