Accredited SINTA 5
72/E/KPT/2024

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

ANALISIS KURIKULUM SISTEM INFORMASI DAN KEBUTUHAN IDUKA MENGGUNAKAN TEKNIK NATURAL LANGUAGE PROCESSING (NLP)

Renata H. Salalu[1*], Wahyudin Hasyim[2], Ir. Hilmansyah Gani, S.Kom,M.Kom[3], Tri Pratiwi Handayani[4], [1,2,3,4]Prodi Sistem Informasi, Universitas Muhammadiyah Gorontalo
Jl. Prof. Dr. H. Mansoer Pateda, Desa Pentadio Timur, Kec. Telaga Biru, Kab. Gorontalo, Gorontalo 96181. Indonesia
E-mail: renatahsalalu@gmail.com

Abstrak - Renata H. Salalu Analisis Kurikulum Sistem Informasi dan Kebutuhan Iduka Menggunakan Teknik *Natural Language Processing* (NLP) dari iklan kerja dan kurikulum sistem informasi menggunakan Metode TF-IDF dan *cosine similarity*. Skripsi, Program Studi Sistem Informasi, Fakultas Sains dan Ilmu Komputer, Universitas Muhammadiyah Gorontalo. Dibimbing oleh, Bapak Wahyudin Hasyim, M.Kom selaku pembimbing 1 dan Bapak Ir. Hilmansyah Gani, S.Kom,M.Kom selaku pembimbing 11.
Perkembangan teknologi informasi yang pesat menyebabkan kebutuhan kompetensi di Industri, Dunia Usaha, dan Dunia Kerja (IDUKA) terus berubah. Kondisi tersebut menuntut perguruan tinggi untuk melakukan evaluasi kurikulum agar kompetensi lulusan tetap selaras dengan kebutuhan industri. Penelitian ini bertujuan untuk menganalisis tingkat keselarasan kurikulum Program Studi Sistem Informasi Universitas Muhammadiyah Gorontalo dengan kebutuhan IDUKA menggunakan teknik Natural Language Processing (NLP). Data penelitian berupa dokumen kurikulum Program Studi Sistem Informasi dan data lowongan kerja yang diperoleh melalui proses web scraping pada platform JobStreet. Tahapan penelitian meliputi text preprocessing (cleaning, case folding, tokenizing, dan stopword removal), pembobotan Term Frequency-Inverse Document Frequency (TF-IDF), serta pengukuran tingkat kemiripan menggunakan Cosine Similarity.Hasil penelitian menunjukkan bahwa mata kuliah BI/Big Data memperoleh skor keselarasan tertinggi sebesar 46, diikuti Pemrograman Berbasis Web sebesar 44, dan Data Mining sebesar 43, yang termasuk dalam kategori relevan tinggi. Mata kuliah Sistem Informasi Manajemen memperoleh skor 36, Keamanan Jaringan 35, dan Sistem Operasi 34, yang termasuk dalam kategori relevan sedang. Sementara itu, Pemrograman Aplikasi Bergerak memperoleh skor 30, Analisis dan Perancangan Sistem Informasi 29, Sistem Basis Data 15, dan Audit Sistem Informasi 12, yang termasuk dalam kategori relevan rendah. Hasil tersebut menunjukkan bahwa sebagian mata kuliah telah sesuai dengan kebutuhan IDUKA, namun beberapa mata kuliah masih memerlukan pembaruan materi dan kompetensi agar lebih mengikuti perkembangan teknologi dan kebutuhan industri.Penelitian ini menyimpulkan bahwa teknik NLP yang dikombinasikan dengan metode TF-IDF dan Cosine Similarity mampu memberikan gambaran objektif mengenai tingkat keselarasan kurikulum dengan kebutuhan IDUKA. Hasil penelitian diharapkan dapat menjadi dasar dalam evaluasi dan pengembangan kurikulum Program Studi Sistem Informasi sehingga menghasilkan lulusan yang lebih kompeten dan sesuai dengan tuntutan dunia kerja.
Kata Kunci: *Natural Language Processing (NLP), TF-IDF, Cosine Similarity, Kurikulum, IDUKA, Sistem Informasi, JobStreet.*

Abstract - The rapid development of information technology has continuously changed the competency requirements of Industry, Business, and the Workforce (IDUKA). This condition requires higher education institutions to evaluate their curricula to ensure that graduates' competencies remain aligned with industry demands. This study aims to analyze the alignment between the Information Systems curriculum at Universitas Muhammadiyah Gorontalo and the competency requirements of IDUKA using Natural Language Processing (NLP) techniques. The research data consist of the Information Systems curriculum documents and job vacancy data collected through web scraping from the JobStreet platform. The research stages include text preprocessing (cleaning, case folding, tokenizing, and stopword removal), Term Frequency–Inverse Document Frequency (TF-IDF) weighting, and similarity measurement using the Cosine Similarity method.The results indicate that the Business Intelligence (BI)/Big Data course achieved the highest alignment score of 46, followed by Web Programming with a score of 44 and Data Mining with a score of 43, all of which fall into the highly relevant category. Management Information Systems obtained a score of 36, Network Security scored 35, and Operating Systems scored 34, which are categorized as moderately relevant. Meanwhile, Mobile Application Programming obtained a score of 30, Systems Analysis and Design scored 29, Database Systems scored 15, and Information Systems Audit scored 12, all of which are classified as low

relevance. These findings indicate that several courses have been aligned with the competency requirements of IDUKA. However, some courses still require updates to their learning materials and competencies to keep pace with technological advancements and industry needs.This study concludes that NLP techniques combined with the TF-IDF and Cosine Similarity methods can provide an objective assessment of the alignment between the curriculum and IDUKA requirements. The findings are expected to serve as a basis for evaluating and improving the Information Systems curriculum, thereby producing graduates who are more competent and better prepared to meet the demands of the workforce.
Keywords: Natural Language Processing (NLP), TF-IDF, Cosine Similarity, Curriculum, Industry, Business, and Workforce (IDUKA), Information Systems, JobStreet.

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)

Accredited SINTA 5
72/E/KPT/2024

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

1.1 LATAR BELAKANG

Dalam era digital saat ini, perkembangan teknologi informasi (TI) berlangsung sangat cepat dan berdampak langsung pada perubahan kebutuhan keterampilan di Industri, Dunia Usaha, dan Dunia Kerja (IDUKA). Munculnya berbagai teknologi baru menyebabkan tuntutan kompetensi tenaga kerja terus berubah dalam waktu yang relatif singkat, sehingga lulusan perguruan tinggi dituntut untuk selalu mengikuti perkembangan tersebut (Ningsih, 2024). Kondisi ini menunjukkan pentingnya evaluasi kurikulum secara manual sering terkendala oleh kurangnya objektivitas dan keterbatasan waktu, terutama karena volume data kompetensi yang besar untuk dibandingkan antara dokumen sistem informasi dan lowongan kerja di IDUKA. Karenanya, Dalam penerapan, Program Studi Sistem Informasi telah menerapkan kurikulum yang diarahkan pada pengembangan kompetensi inti mahasiswa di bidang teknologi informasi. Kurikulum tersebut dirancang untuk membekali mahasiswa dengan pemahaman teknis yang mencakup pengembangan perangkat lunak, pengelolaan basis data, analisis sistem informasi, serta infrastruktur jaringan. Mata kuliah inti seperti Algoritma dan Pemrograman, Sistem Basis Data, dan Rekayasa Perangkat Lunak berperan sebagai landasan utama dalam membentuk keterampilan teknis (*hard skills*) mahasiswa sebagai persiapan memasuki dunia kerja (Niken Rahma Diasri et al., 2025).

Di sisi lain, perkembangan pada Industri, Dunia Usaha, dan Dunia Kerja (IDUKA) menunjukkan perubahan kebutuhan yang berlangsung sangat secara cepat. Dunia industri tidak lagi hanya menuntut penguasaan konsep dasar, tetapi juga membutuhkan tenaga kerja pada peran tertentu seperti *Full-stack Developer*, Data *Analyst*, dan *Cybersecurity Specialist*, yang disertai dengan penguasaan *framework* modern seperti *Laravel, React Web* (Putra et al., 2025).

Perbedaan antara kompetensi yang tercantum dalam dokumen kurikulum sistem informasi dan kebutuhan keterampilan yang dipublikasikan oleh IDUKA melalui platform lowongan kerja sering kali menimbulkan kesenjangan yang berdampak pada lulusan sistem informasi. Oleh karena itu, kurikulum Program Studi Sistem Informasi perlu dievaluasi secara berkelanjutan agar selaras dengan kebutuhan pasar kerja. Berdasarkan kondisi tersebut, penelitian ini bertujuan untuk menggabungkan data kurikulum Sistem Informasi dengan data kebutuhan keterampilan IDUKA menggunakan pendekatan *Natural Language Processing* (*NLP*). Selanjutnya, metode *Trem Frequency – Inverse Document Frequency* (TF-*IDF*) digunakan untuk memberi bobot pada kata-kata dalam dokumen kurikulum dan dokumen kebutuhan industri (IDUKA), sehingga dapat dihitung tingkat kemiripannya dengan menggunakan *Cosine Similarity.* sehingga hasil analisis dapat dimanfaatkan sebagai dasar dalam merumuskan rekomendasi pengembangan kurikulum yang lebih relevan dan tepat sasaran.

penerapan teknik *NLP* berbasis *TF-IDF* dan *Cosine Similarity* sangatlah penting. Metode ini memfasilitasi pemetaan kompetensi secara otomatis, efisien, dan presisi, sehingga output analisisnya dapat dijadikan landasan ilmiah bagi Program Studi Sistem Informasi guna menyelaraskan kurikulum dengan tuntutan pasar kerja saat ini.

Kurikulum yang diterapkan di perguruan tinggi diharapkan mampu membekali mahasiswa dengan pengetahuan dan keterampilan yang sesuai dengan kebutuhan IDUKA. Namun, dalam pelaksanaannya masih sering ditemukan adanya kesenjangan antara kompetensi yang diajarkan di perguruan tinggi dan keterampilan yang dibutuhkan oleh dunia kerja. Salah satu permasalahan yang muncul adalah belum adanya kejelasan mengenai apakah Kurikulum Program Studi Sistem Informsi telah mencakup keterampilan yang paling dibutuhkan oleh IDUKA.

Penelitian ini menggunakan pendekatan Teknik *NLP* dengan metode *TF-IDF* digunakan untuk mengubah data teks kurikulum dan lowongan kerja menjadi bentuk numerik, sedangkan *Cosine Similarity* digunakan untuk

CC BY SA

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

mengukur tingkat kemiripan di antara keduanya. Melalui pendekatan ini, jika keterampilan tertentu seperti penggunaan *framework* Laravel banyak dibutuhkan oleh IDUKA namun belum tercantum dalam kurikulum, maka nilai kemiripan yang dihasilkan akan rendah. Hasil analisis ini akan memberikan gambaran objektif mengenai hubungan antara kurikulum Program Studi Sistem Informasi dengan kebutuhan IDUKA, sekaligus menjadi bahan evaluasi dan rekomendasi pengembangan kurikulum yang lebih relevan.

2.1 Landasan Teori

2.1.1 Analisis

Analisis dapat diartikan sebagai suatu tahapan yang dilakukan secara terstruktur untuk menelaah hubungan atau perbandingan antara dua objek berdasarkan kriteria tertentu dengan memanfaatkan data yang tersedia. Dalam ranah pendidikan tinggi, analisis digunakan untuk menelaah keterkaitan antara kurikulum perguruan tinggi dan kebutuhan dunia kerja secara lebih objektif (Rianto et al., 2022). Pendekatan berbasis data diperlukan agar proses evaluasi kurikulum tidak semata-mata didasarkan pada pandangan subjektif, melainkan didukung oleh informasi yang dapat dipertanggung jawabkan.

Selain itu, analisis juga dipahami sebagai upaya untuk mengkaji hubungan antara kompetensi yang termuat dalam kurikulum dengan kebutuhan Industri, Dunia Usaha, dan Dunia Kerja (IDUKA). Dengan memanfaatkan dokumen kurikulum serta data yang bersumber dari deskripsi lowongan kerja, proses analisis ini memberikan gambaran mengenai keterampilan yang dibutuhkan oleh dunia kerja dan keterampilan yang diajarkan di perguruan tinggi. Pendekatan tersebut membantu perguruan tinggi dalam memahami kondisi aktual kebutuhan keterampilan di lapangan.

Dalam penelitian ini, analisis digunakan sebagai landasan konseptual untuk mengkaji hubungan antara kurikulum Program Studi Sistem Informasi dan kebutuhan iduka berdasarkan data teks yang tersedia. Hasil kajian tersebut diharapkan dapat dimanfaatkan sebagai bahan pertimbangan dalam pengembangan kurikulum agar lebih selaras dengan dinamika kebutuhan dunia kerja.

CC BY SA

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

2.1.2 Kurikulum Program Studi Sistem Informasi

Kurikulum merupakan pedoman utama dalam penyelenggaraan pendidikan tinggi yang berfungsi untuk membentuk kompetensi lulusan sesuai dengan perkembangan ilmu pengetahuan, teknologi, dan kebutuhan dunia kerja. Pada Program Studi Sistem Informasi, kurikulum dirancang untuk membekali mahasiswa dengan keterampilan di bidang pemrograman, basis data, analisis dan perancangan sistem, serta manajemen teknologi informasi (Yahya et al., 2024).

Perkembangan teknologi informasi yang sangat cepat menyebabkan kebutuhan keterampilan di Industri, Dunia Usaha, dan Dunia Kerja (IDUKA) terus berubah. Kondisi ini menuntut adanya evaluasi kurikulum secara berkala agar materi pembelajaran tetap relevan dengan kebutuhan dunia kerja. Evaluasi kurikulum berbasis data menjadi pendekatan yang penting karena mampu memberikan hasil yang lebih objektif dibandingkan evaluasi yang bersifat subjektif.

Dalam penelitian ini, kurikulum dipandang sebagai dokumen teks yang merepresentasikan keterampilan lulusan. Oleh karena itu, kurikulum dapat dianalisis dan dibandingkan dengan data kebutuhan IDUKA yang diperoleh dari deskripsi lowongan kerja. Pendekatan ini sejalan dengan tujuan penelitian, yaitu menganalisis kurikulum Sistem Informasi dan kebutuhan IDUKA menggunakan teknik *Natural Language Processing (NLP)* sebagai dasar evaluasi dan pengembangan kurikulum yang lebih relevan (Sihombing, 2022).

2.1.3 Industri, Dunia Usaha, dan Dunia Kerja (IDUKA)

Industri, Dunia Usaha, dan Dunia Kerja (IDUKA) merupakan pihak yang berperan sebagai pengguna lulusan perguruan tinggi. Pada bidang Sistem Informasi, IDUKA memerlukan tenaga kerja yang memiliki perpaduan antara keterampilan teknis dan keterampilan pendukung.

Perkembangan teknologi informasi yang berlangsung secara cepat menyebabkan kebutuhan keterampilan iduka bersifat dinamis. Oleh karena itu, diperlukan proses analisis yang sistematis terhadap informasi kebutuhan iduka agar dapat diperoleh gambaran yang objektif mengenai kompetensi yang dibutuhkan di dunia kerja. Dalam penelitian ini, kebutuhan iduka direpresentasikan melalui data deskripsi lowongan kerja yang memuat berbagai keterampilan dan kompetensi yang dibutuhkan oleh industri. Data tersebut kemudian dianalisis sebagai data teks untuk mengidentifikasi pola keterampilan yang relevan dengan bidang Sistem Informasi.

2.2 Keterampilan (Skills) dalam Bidang Sistem Informasi

Keterampilan (*skills*) merupakan faktor utama yang menentukan kesiapan lulusan dalam menghadapi dunia kerja, khususnya pada bidang Sistem Informasi. Lulusan Sistem Informasi dituntut memiliki keterampilan teknis (hard skill) seperti pemrograman, pengelolaan basis data, dan analisis sistem,Karena mengikuti perkembangan Teknologi saat ini (Peng & Deng, 2022)

2.3 *Natural Language Processing* (*NLP*)

Natural Language Processing (*NLP*) dimanfaatkan untuk, salah satu subbidang dari kecerdasan buatan, difokuskan untuk menyelidiki hubungan kompleks antara komputer dan bahasa manusia. Sasaran utamanya adalah untuk memahami, menganalisis, dan memproduksi informasi berbasis teks dengan cara yang meniru keterampilan manusia. (Nurwanda et al., 2024).

2.3.1 *TF-IDF*

TF-IDF (*Term Frequency – Inverse Document Frequency*) digunakan untuk Mengekstrak kata kunci keterampilan yang paling relevan dari ribuan kata dalam iklan lowongan kerja, sehingga kata-kata umum seperti "dan" atau "yang" tidak mengganggu proses analisis (Ahmednor et al., 2025). Metode ini digunakan untuk

CC BY SA

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

mengetahui tingkat kepentingan suatu kata dalam sebuah dokumen dibandingkan dengan kumpulan dokumen lainnya. *TF-IDF* dari dua komponen utama, yaitu ***Term Frequency* (*TF*)** dan *Inverse Document Frequency* **(*IDF*).** *Term Frequency* (*TF)* menunjukkan seberapa sering suatu kata muncul dalam sebuah dokumen. Semakin sering kata tersebut muncul, maka nilai TF-nya semakin tinggi. Sementara itu, *Inverse Dokumen Frequency (IDF*) menunjukkan tingkat keunikan suatu kata dalam seluruh dokumen. Kata yang muncul di banyak dokumen akan memiliki nilai *IDF* yang rendah, sedangkan kata yang jarang muncul akan memiliki nilai *IDF* yang tinggi.

TF-IDF digunakan untuk menentukan bobot kepentingan sebuah kata dalam dokumen. Proses ini dilakukan melalui tiga tahapan perhitungan dan rumus nya:

1. *Trem Frequency (TF*): Mengukur seberapa sering suatu kata muncul dalam sebuah dokumen. Semakin sering kata tersebut muncul, semakin tinggi nilai *TF*-nya.

Rumus:

$$TF(t,d) = \frac{n_{t,d}}{\sum_k n_{k,d}} \quad (2.1)$$

Keterangan:

a. $n_{t,d}$: jumlah kemunculan kata t dalam dokumen d .
b. $\sum_k n_{k,d}$: Total seluruh kata dalam dokumen d

2. *Inverse Document Frequency* (*IDF*): Menghitung tingkat keunikan atau kepentingan kata dalam seluruh kumpulan dokumen. Kata yang jarang muncul di dokumen lain akan memiliki nilai *IDF* yang tinggi.

Rumus:

$$IDF(t) = \log\left(\frac{N}{df(t)}\right) \quad (2.2)$$

Keterangan:

a. N: Jumlah total dokumen dalam dokumen.
b. $df(t)$: Jumlah dokumen yang mengandung kata t .

3. Perhitungan *TF-IDF*: Bobot akhir didapatkan dengan mengalikan nilai TF dan IDF.

Rumus:

$$W_{t,d} = TF(t,d) \times IDF(t) \quad (2.3)$$

Keterangan:

a. $W_{t,d}$: Bobot akhir kata t pada dokumen d.

2.3.2 *Scraping*

Scraping adalah metode pengumpulan data yang dilakukan secara mandiri untuk memperoleh data kebutuhan Dunia Usaha dan Dunia Industri (IDUKA) dari tautan lowongan kerja online. Teknik ini memungkinkan peneliti untuk mengambil materi tekstual dari *platform* digital secara otomatis untuk kemudian diolah lebih lanjut (Adila, 2022).

2.3.3 *Cosine Similarity*

Cosine Similarity adalah Nilai hasil perhitungan. dengan metode ini berada dalam batas jangkauan 0 sampai 1, semakin menyentuh nilai 1 maka kedua vektor dianggap semakin mirip. Sebaliknya, nilai 0 menunjukkan bahwa tidak ada kemiripan antara kedua vektor tersebut (Rio Feriangga Kurniawan, 2022).

Untuk mengukur kemiripan antara kurikulum (Vektor A) dan kebutuhan IDUKA (Vektor B), digunakan perhitungan sudut kosinus dengan rumus:

CC BY SA

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

$$\text{Similarity}(A,B) = \cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (2.4)$$

Keteranga:

a. A_i : Bobot *TF-IDF* kata ke$-i$ pada dokumen Kurikulum.
b. B_i : Bobot *TF-IDF* kata ke$-i$ pada dokumen IDUKA (Lowongan Kerja).
c. $\sum A_i B_i$: *Dot Product* (perkalian titik) antara vektor A dan B.
d. $\sqrt{\sum A_i^2}$: Panjang (norma) dari vektor A.
e. $\sqrt{\sum B_i^2}$: Panjang (norma) dari vektor B.

Tahapan Perhitungan:

a. Vektorisasi: Mengubah teks menjadi angka berdasarkan hasil *TF-IDF*.
b. Perkalian Vektor: Mengalikan bobot kata yang sama antara kedua dokumen.
c. Normalisasi: Membagi hasil perkalian dengan panjang masing-masing vektor untuk mendapatkan nilai antara 0 (tidak mirip) hingga 1 (sangat identik).

2.4 Kebutuhan Perangkat Lunak

2.4.1 *Google Colab*

Google Colab adalah *platform* pemrograman berbasis *cloud* yang disediakan oleh *Google* untuk menjalankan kode *Python* secara langsung melalui *browser* tanpa perlu melakukan instalasi *Python* maupun *library* di komputer pengguna. Layanan ini berbentuk notebook interaktif yang menggabungkan penulisan kode, penjelasan teks, serta tampilan *output* dalam satu dokumen, sehingga sangat mendukung kegiatan analisis data, pembelajaran pemrograman, penelitian ilmiah, hingga pengembangan *machine learning* (Guntara, 2023).

Dengan *Google Colab*, seluruh proses komputasi dilakukan di *server Google*, sehingga pengguna dapat menjalankan program berat tanpa bergantung pada spesifikasi perangkat pribadi, bahkan dapat memanfaatkan akselerasi GPU atau TPU untuk mempercepat proses komputasi. Selain itu, *Colab* terintegrasi dengan penyimpanan daring sehingga *notebook* dan data dapat disimpan, dibagikan, dan diakses kembali dengan mudah. Oleh karena itu, *Google Colab* sering digunakan dalam penelitian akademik maupun proyek data science karena praktis, fleksibel, serta mendukung kolaborasi dan eksperimen pemrograman secara efisien.

2.5.2 *Platfrom Loker*

Platform lowongan kerja merupakan media berbasis *web* atau aplikasi digital yang digunakan sebagai sarana penghubung antara perusahaan sebagai penyedia pekerjaan dengan individu pencari kerja. Melalui platform ini, perusahaan menyampaikan informasi lowongan yang mencakup kualifikasi, tanggung jawab, serta keterampilan yang dibutuhkan.

2.4.3 *Python*

Python merupakan sebuah bahasa pemrograman tingkat tinggi yang dirilis pada tahun 1991. *Python* dirancang dengan tujuan utama untuk menghasilkan kode yang mudah dibaca dan dipahami oleh manusia, sehingga proses pengembangan perangkat lunak dapat dilakukan dengan lebih cepat dan efisien. Seiring perkembangannya, *Python* menjadi salah satu bahasa pemrograman yang sangat populer karena kesederhanaan sintaks, fleksibilitas penggunaan, serta dukungan komunitas global yang besar. Pengembangan *Python* saat ini dikelola oleh *Python Software Foundation* yang memastikan bahasa ini tetap bersifat terbuka dan dapat digunakan secara luas.

Python termasuk bahasa pemrograman yang bersifat *interpreted*, sehingga kode dapat dijalankan langsung

CC BY SA

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

tanpa proses kompilasi yang kompleks. Hal ini membuat *Python* sangat cocok digunakan dalam penelitian, eksperimen sistem, serta pengembangan aplikasi berbasis data. Selain itu, *Python* bersifat multifungsi karena dapat digunakan dalam berbagai bidang teknologi informasi, seperti pengembangan aplikasi *web*, analisis data, otomatisasi sistem, pengolahan citra, hingga kecerdasan buatan.

Dalam penelitian ini, *Python* dipilih karena memiliki banyak *library* pendukung yang relevan dengan kebutuhan pengolahan data dan teks, seperti *library* untuk *web scraping*, analisis data, serta pemrosesan bahasa alami. Keberadaan *library* tersebut memungkinkan proses pengumpulan data, pembersihan teks, analisis, hingga penyimpanan hasil dapat dilakukan dalam satu lingkungan pemrograman secara efisien. Oleh karena itu, *Python* banyak dimanfaatkan dalam bidang *Machine Learning, Deep Learning*, serta penelitian berbasis komputasi moderen (Santoso & Rochadiani, 2022).

2.4.4 *Natural Language Processing* (NLP)

Natural Language Processing (NLP) merupakan salah satu cabang dari Kecerdasan Buatan (Artificial Intelligence) yang mempelajari interaksi antara komputer dan bahasa manusia. NLP bertujuan agar komputer mampu membaca, memahami, menafsirkan, serta mengolah bahasa alami manusia, baik dalam bentuk teks maupun suara, sehingga informasi penting dari bahasa tersebut dapat dikenali dan diproses oleh sistem. Dalam bidang text mining, NLP berperan sebagai penghubung antara bahasa manusia dan pemahaman komputer. Teks yang bersifat tidak terstruktur, seperti deskripsi lowongan pekerjaan, tidak dapat langsung dipahami oleh mesin. Oleh karena itu, NLP digunakan untuk mengubah teks mentah menjadi data yang lebih terstruktur agar dapat diolah secara matematis menggunakan algoritma seperti TF-IDF dan *Cosine Similarity*.

Pada penelitian ini, NLP diterapkan pada proses ekstraksi informasi (*Information Extraction*), yaitu dengan cara sistem secara otomatis membaca deskripsi kualifikasi lowongan pekerjaan (IDUKA) dan deskripsi mata kuliah pada kurikulum, kemudian menyeleksi informasi yang berkaitan dengan kompetensi atau keterampilan teknis di bidang IT.

2.4.5 Tahapan Pra-pemrosesan Teks (*Text Preprocessing*)

Data teks mentah yang diperoleh dari proses ekstraksi (*scraping*) umumnya masih mengandung banyak noise atau gangguan, seperti tanda baca, penggunaan huruf kapital yang tidak konsisten, serta kata-kata yang tidak memiliki pengaruh penting terhadap analisis. Kondisi tersebut dapat memengaruhi dan menurunkan tingkat akurasi algoritma dalam melakukan perhitungan. Oleh karena itu, sebelum dilakukan proses pembobotan menggunakan TF-IDF, data teks perlu melewati beberapa tahapan preprocessing pada *Natural Language Processing* (NLP):

a. Text *Cleaning* (Pembersihan Teks) merupakan tahap awal dalam preprocessing yang bertujuan untuk membersihkan data teks dari elemen-elemen yang tidak memiliki makna atau tidak diperlukan dalam proses analisis. Pada tahap ini dilakukan penghapusan tanda baca, angka, karakter khusus, simbol seperti @, #, $, %, serta elemen format seperti spasi berlebih, enter ganda, dan tab. Setelah proses pembersihan selesai, teks akan menjadi lebih terstruktur dan hanya berisi kumpulan huruf atau kata yang siap digunakan pada tahapan analisis berikutnya
b. Tahapan *Case Folding*

   Case folding merupakan proses mengubah seluruh huruf pada dokumen menjadi huruf kecil (*lowercase*) agar format penulisan menjadi seragam. Pada sistem komputer, perbedaan huruf kapital dan huruf kecil dianggap sebagai kata yang berbeda.
c. Tahap *Tokenizing*

CC BY SA

Accredited SINTA 5
72/E/KPT/2024

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

Tokenizing atau tokenisasi merupakan proses memecah teks, baik berupa kalimat maupun paragraf, menjadi unit-unit kecil yang disebut token atau kata tunggal. Proses ini umumnya menggunakan spasi (*whitespace*) sebagai pemisah antar kata.

3.1 Objek Penelitian

Objek penelitian ini adalah data teks dokumen kurikulum Program Studi Sistem Informasi. Dan data lowongan kerja (*JobStreet*). Fokus utama penelitian adalah membandingkan tingkat kemiripan teks antara kedua sumber data tersebut untuk mengevaluasi relevansinya.

Kesesuaian kurikulum dengan kebutuhan Industri, Dunia Usaha, dan Dunia Kerja (IDUKA) telah diatur melalui beberapa standar baku yang ditetapkan pemerintah. Permenristekdikti Nomor 44 Tahun 2015 tentang Standar Nasional Pendidikan Tinggi mewajibkan setiap program studi merumuskan Capaian Pembelajaran Lulusan (CPL) yang unsur pengetahuan dan keterampilannya diperoleh melalui penalaran atas pengalaman kerja, sehingga kurikulum secara regulatif diwajibkan memuat keterkaitan dengan dunia kerja. Selain itu, Peraturan Presiden Nomor 8 Tahun 2012 tentang Kerangka Kualifikasi Nasional Indonesia (KKNI) menetapkan kesetaraan antara jenjang pendidikan dengan tingkat kualifikasi kerja nasional, yang mewajibkan setiap perguruan tinggi menyesuaikan capaian pembelajaran kurikulumnya agar selaras dengan kebutuhan kompetensi di dunia kerja. Pada tingkat kebijakan yang lebih operasional, Peraturan Presiden Nomor 68 Tahun 2022 tentang Revitalisasi Pendidikan Vokasi dan Pelatihan Vokasi menegaskan bahwa kurikulum harus terus disesuaikan dengan kebutuhan dunia kerja, sehingga lulusan yang dihasilkan benar-benar sesuai dengan apa yang dibutuhkan industri, bukan hanya berdasarkan apa yang selama ini diajarkan oleh prodi.

3.2 Metode Penelitian

Metode penelitian ini dilakukan dengan menerapkan teknik *NLP* menggunakan bahasa pemrograman *Python* dan *tools Google Colab*. Metode *TF-IDF* digunakan untuk mengubah data teks ke dalam bentuk numerik, sedangkan *Cosine Similarity* digunakan untuk menganalisis tingkat kemiripan antara kurikulum dan kebutuhan IDUKA.

3.3 Alur Penelitian



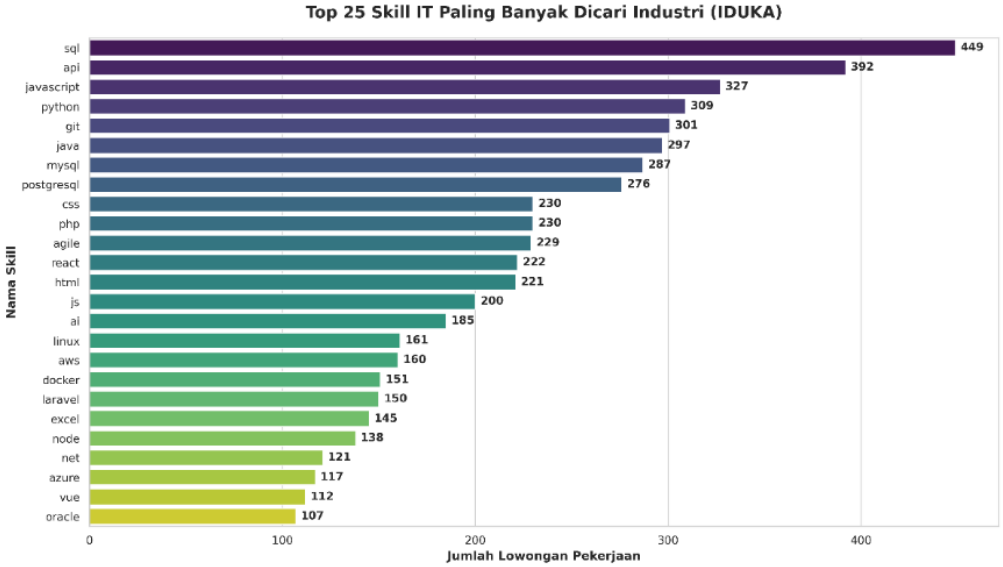
Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

3.3.1 Pengumpulan Data

Pengumpulan data dalam penelitian ini dilakukan dengan mengambil materi tekstual berupa data kurikulum dari dokumen resmi Program Studi Sistem Informasi sebagai representasi kompetensi akademik , serta data lowongan kerja dari *platform Jobstreet* melalui proses *web scraping* untuk mengidentifikasi keterampilan yang dibutuhkan oleh Industri, Dunia Usaha, dan Dunia Kerja (IDUKA) saat ini.(Hafid et al., 2025)

Selah dilakukan proses pengumpulan dan pengolahan data lowongan pekerjaan dari *platform Jobstreet* untuk merepresentasikan kebutuhan IDUKA, tahap selanjutnya adalah penyajian dokumen kurikulum Program Studi Sistem Informasi. Data ini mencakup rincian mata kuliah serta capaian pembelajaran yang akan dianalisis tingkat kemiripan yang dibutuhkan oleh iduka.

Gamabr Skill yang paling banyak dibutuhkan di iduka



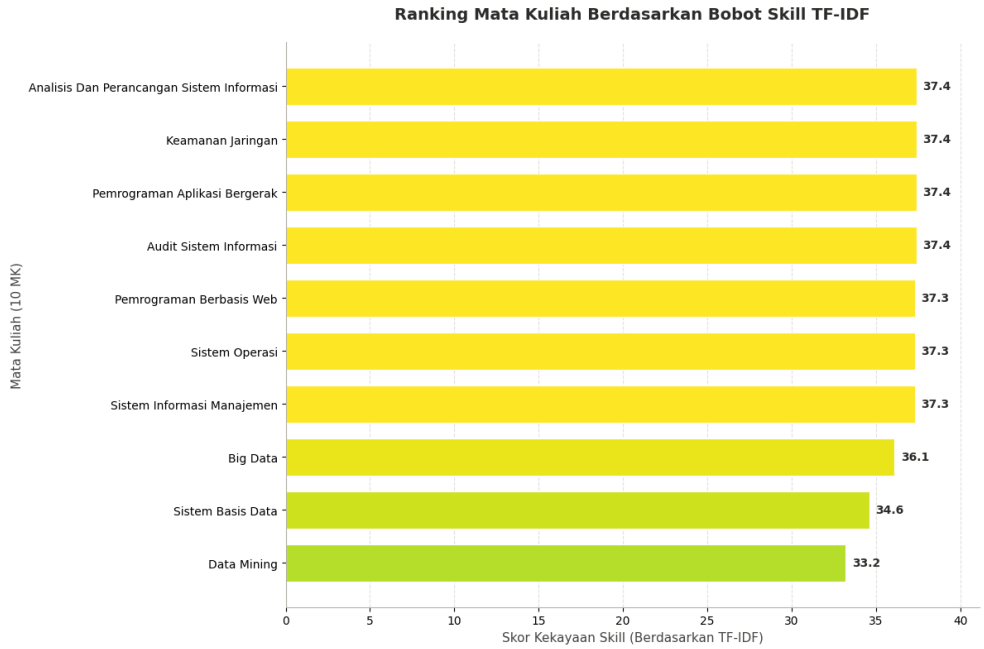
Data kurikulum Program Studi Sistem Informasi. Akan tetapi, tidak semua mata kuliah dijadikan data penelitian. Hanya mata kuliah yang berkaitan dengan bidang Teknologi Informasi (IT), seperti pemrograman, basis data, sistem informasi, jaringan komputer, dan teknologi informasi yang digunakan sebagai objek analisis. Sementara itu, mata kuliah umum di luar bidang IT tidak termasuk dalam proses penelitian. Data mata kuliah tersebut kemudian diolah untuk memperoleh skill atau kompetensi yang selanjutnya dibandingkan dengan kebutuhan industri (IDUKA) menggunakan metode TF-IDF dan *Cosine Similarity*.

CC BY SA

Accredited SINTA 5
72/E/KPT/2024

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

Gamabr Top Mata Kuliah si



3.3.2 Pra-pemrosesan Data

Pra-pemrosesan data merupakan serangkaian tahapan untuk memurnikan dan menormalisasi teks dari dokumen kurikulum Program Studi Sistem Informasi dan data lowongan kerja IDUKA dari *Jobstreet Cosine Similarity* agar siap dikonversi menjadi vektor numerik *TF-IDF* yang optimal bagi perhitungan. Berikut adalah urutan pelaksanaannya:

1. Data pertama dari Dokumen kurikulum langsung dibandingkan dengan data lowongan kerja. Dokumen kurikulum terlebih dahulu dianalisis untuk memperoleh data pokok bahasa dan sub pokok bahasa yang merepresentasikan kompetensi pembelajaran pada setiap mata kuliah. Data tersebut kemudian disusun dalam bentuk dataset dan digunakan sebagai dokumen kurikulum dalam proses analisis. Selanjutnya, data kurikulum dan data kebutuhan IDUKA yang diperoleh dari lowongan kerja diproses menggunakan metode TF-IDF untuk mengubah data teks menjadi representasi numerik. Hasil pembobotan TF-IDF kemudian dihitung menggunakan metode Cosine Similarity untuk mengukur tingkat kesesuaian antara kompetensi yang diajarkan dalam kurikulum dengan kompetensi yang dibutuhkan oleh IDUKA.

3.2 Data *Scrapping*

Data kedua dikumpul secara mandiri melalui proses *web scraping*. Proses ini dilakukan untuk memperoleh data kebutuhan Dunia, Usaha dan Dunia Industri (IDUKA) dari **link lowongan kerja online**. Data yang dikumpulkan berupa **deskripsi lowongan kerja**, yang mencakup informasi Skills.Tabel berikut merupakan contoh data lowongan kerja yang digunakan sebagai sumber informasi keterampilan yang dibutuhkan oleh IDUKA.

3.3.3 Pra-pemrosesan Data (*Preprocessing*)

1. Tahapan Data *Cleaning*

Pada bagian *cleaning text*, proses ini akan menghapus elemen-elemen yang tidak relevan atau dapat mengganggu analisis, seperti tanda baca, angka, karakter khusus, serta spasi berlebih. Selain itu, teks juga akan dinormalisasi dengan menghilangkan simbol-simbol yang tidak diperlukan, sehingga menghasilkan data yang lebih bersih dan siap digunakan.

CC BY SA

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

2. *Tahapan Case Folding*

Case Folding adalah untuk Mengubah seluruh huruf pada teks menjadi huruf kecil (*lowercase*) agar tidak terjadi perbedaan makna akibat perbedaan penulisan huruf besar dan kecil.

3. *Tahapan Stopword Removal*

Stopword Removal adalah tahapan dalam preprocessing teks pada *Natural Language Processing* (*NLP*) yang bertujuan untuk menghilangkan kata-kata umum yang sering muncul dalam dokumen tetapi tidak memiliki pengaruh signifikan terhadap makna teks**.** Tabel ini menunjukkan contoh kata umum yang dihilangkan pada tahap *stopword removal* dalam proses pra-pemrosesan data teks.

4. ***Tahapan Tokenizing***

Tokenizing adalah proses memecah teks menjadi unit-unit kecil yang disebut *token*, di mana setiap token umumnya berupa kata-kata individual dalam teks. Proses ini memudahkan analisis lebih lanjut terhadap data teks, seperti penghitungan frekuensi kata, pembobotan kata menggunakan *TF-IDF*, serta pencocokan kata antara kurikulum dan kebutuhan Dunia Usaha dan Dunia Industri (IDUKA).

3.3.4 Pelabelan Data

Tahap transformasi data merupakan proses representasi numerik pada dokumen kurikulum dan daftar kebutuhan kompetensi IDUKA yang telah diproses melalui tahap *NLP*. Setiap entitas teks dalam kurikulum akan dipetakan ke dalam ruang vektor menggunakan pembobotan *TF-IDF* untuk menghasilkan profil kompetensi universitas.

3.3.5 Pembobotan *TF-IDF*

Pembobotan TF-IDF merupakan metode dalam Teknik *NLP* yang digunakan untuk mengubah data teks menjadi representasi numerik berdasarkan tingkat kepentingan suatu kata di dalam dokumen. Dalam penelitian berjudul **"Analisis Kurikulum Sistem Informasi dan Kebutuhan IDUKA Menggunakan Teknik *NLP*"**, metode *TF-IDF* digunakan untuk memberikan bobot pada kata-kata keterampilan (*skills*) yang terdapat dalam dokumen kurikulum Program Studi Sistem Informasi dan dokumen kebutuhan IDUKA yang diperoleh dari data lowongan kerja.

Metode *TF-IDF* bekerja dengan mempertimbangkan dua komponen utama, yaitu tingkat kemunculan kata dalam sebuah dokumen dan tingkat keunikan kata tersebut dalam keseluruhan kumpulan dokumen. Kata yang sering muncul dalam satu dokumen namun jarang muncul pada dokumen lain akan memiliki bobot yang lebih tinggi, sehingga dianggap lebih representatif dalam menggambarkan isi dokumen tersebut.

Pembobotan *TF-IDF* dihitung dengan mengalikan nilai *Term Frequency* (*TF*) dan *Inverse Document Frequency* (*IDF*), dengan rumus sebagai berikut:

1. *Term Frequency* (*TF*)

$$TF(t,d) = \frac{\text{Jumlah kemunculan kata } t \text{ dalam dokumen } d}{\text{Total kata dalam dokumen } d}$$

Keterangan:

a. t= Kata (term)
b. d= dokumen
c. TF(t,d) = Mengukur frekuensi kemunculan kata tertentu (t) dalam dokumen (d).

CC BY SA

Accredited SINTA 5 72/E/KPT/2024

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

2. *Inverse Document Frequency (IDF)*

$$IDF(t) = \log\left(\frac{N}{df(t)}\right) \quad (3.1)$$

Keterangan:

a. N= Jumlah seluruh total dokumen
b. df(t) = jumlah dokumen yang mengandung kata term (t)
c. IDF(t) = tingkat keunikan kata t

3. TF-IDF

$$TF\text{-}IDF(t,d,D) = TF(t,d) \times IDF(t,D) \quad (3.2)$$

Keterangan:

a. $TF-IDF(t,d)TF\text{-}IDF(t,d)TF-IDF(t,d)$= bobot kata t pada dokumen d

Bobot akhir *TF-IDF* digunakan untuk merepresentasikan dokumen kurikulum dan kebutuhan IDUKA dalam bentuk vektor, yang selanjutnya menjadi input perhitungan *Cosine Similarity* untuk mengukur tingkat kesesuaian antara kurikulum Program Studi Sistem Informasi dan kebutuhan dunia kerja.

3.3.6 Pengukuran Kemiripan (*Cosine Similarity*)

Cosine Similarity digunakan untuk menghitung tingkat kemiripan antara dokumen kurikulum dan dokumen kebutuhan industri**,** yang sebelumnya telah diubah menjadi vektor angka menggunakan metode TF-IDF. Nilai *Cosine Similarity* berada pada rentang 0 sampai 1**,** di mana nilai yang semakin mendekati 1 menunjukkan tingkat kemiripan yang semakin tinggi antara kurikulum dan kebutuhan IDUKA menggunakan rumus sebagai berikut:

$$\text{similarity} = \cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (3.3)$$

Keterangan:

a. A_i : Bobot TF-IDF kata ke$-i$ pada dokumen Kurikulum.
b. B_i : Bobot TF-IDF kata ke$-i$ pada dokumen IDUKA (Lowongan Kerja).
c. $\sum A_i B_i$: *Dot Product* (perkalian titik) antara vektor A dan B.
d. $\sqrt{\sum A_i^2}$: Panjang (norma) dari vektor A.
e. $\sqrt{\sum B_i^2}$: Panjang (norma) dari vektor B.
f. i : jumlah term dalam kalimat
g. n : jumlah vektor

Algoritma ini digunakan untuk suatu metode yang digunakan untuk mengukur tingkat kemiripan antar teks berdasarkan dua vektor. Pengukuran ini dapat memberikan peringkat suatu dokumen sesuai dengan skor kemiripan yang dihasilkan Apabila total similarity yang didapatkan nol (0) maka dokumen yang diolah tidak memiliki

CC BY SA

kesamaan dan jika nilai yang didapatkan maksimal adalah 1 maka dokumen tersebut memiliki kemiripan.

Tabel 1. Contoh Hasil *Cosine Similarity*.

No	Rentang Nilai Kemiripan	Nilai *Similarity*	Interpretasi
1	Basis Data	0,85	Relavan
2	Pemrograman Web	0,72	Cukup Relavan
3	Jaringan Komputer	0,45	Tidak Relavan

3.3.7 Evaluasi

Evaluasi dilakukan untuk menilai tingkat kesesuaian antara kurikulum Program Studi Sistem Informasi dengan kebutuhan Industri, Dunia Usaha, dan Dunia Kerja (IDUKA) berdasarkan hasil pembobotan *TF-IDF* dan perhitungan *Cosine Similarity*. Tahap evaluasi ini bertujuan untuk memberikan gambaran objektif mengenai relevansi kompetensi yang diajarkan dalam kurikulum terhadap keterampilan yang dibutuhkan di dunia kerja berdasarkan data lowongan pekerjaan.

Hasil pembobotan *TF-IDF* menunjukkan bahwa keterampilan utama seperti pemrograman, basis data, dan *cybersecurity* memiliki bobot yang cukup dominan baik pada dokumen kurikulum maupun dokumen kebutuhan IDUKA. Hal ini menandakan bahwa terdapat kesamaan kompetensi inti yang menjadi fokus pembelajaran di Program Studi Sistem Informasi dengan kebutuhan industri saat ini. Namun, perbedaan nilai bobot *TF-IDF* juga menunjukkan adanya perbedaan tingkat penekanan kompetensi, di mana beberapa keterampilan tertentu lebih banyak dibutuhkan oleh IDUKA dibandingkan dengan yang tercantum dalam kurikulum.

Selanjutnya, hasil pengukuran *Cosine Similarity* menunjukkan bahwa mata kuliah Basis Data memiliki nilai *similarity* sebesar 0,85 yang termasuk dalam kategori sangat sesuai, sedangkan mata kuliah Pemrograman Web memperoleh nilai *similarity* sebesar 0,72 dan dikategorikan sesuai. Sementara itu, mata kuliah Jaringan Komputer memperoleh nilai *similarity* sebesar 0,45 yang masuk dalam kategori kurang sesuai. Berdasarkan hasil tersebut, dapat disimpulkan bahwa secara umum kurikulum Program Studi Sistem Informasi telah cukup relevan dengan kebutuhan IDUKA, namun masih diperlukan evaluasi dan pengembangan lebih lanjut pada beberapa mata kuliah agar kurikulum menjadi lebih adaptif terhadap perkembangan teknologi dan tuntutan dunia kerja.

4. HASIL PENELITIAN

4.1 Pengambilan Data (*Crawling* Data)

Pengumpulan data dalam penelitian ini diawali dengan merepresentasikan kebutuhan Industri, Dunia Usaha, dan Dunia Kerja (IDUKA) melalui *crawling* pada *platfrom Jobstreet*, dengan fokus *scraping* hanya data lowongan kerja yang relavan di bidang IT.

Proses *crawling* dijalankan dengan *Library Playwright* untuk otomatis ke *browser* serta pandas untuk menyusun dataset menjadi format terstruktur. Dataset dari kedua sumber kemudian di gabungkan menjadi data mentah tunggal, dan diproses lebih lanjut melalui pra-pemrosesan teks dengan teknik *Natural Language Processing* (NLP).

Langkah awal dalam pengumpulan data dimulai dengan proses persiapan lingkungan pengembangan berbasis *Google Colaboratory* untuk menjamin aksesibilitas sistem terhadap *platform Jobstreet*. Hal ini dilakukan

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

dengan menginisialisasi *library* Playwright yang dikonfigurasi agar dapat berfungsi sebagai mesin otomatis, di dalam lingkungan *virtual*, guna menavigasi halaman web secara dinamis. Prosedur dilanjutkan dengan tahap *crawling* data, di mana sistem melakukan penelusuran informasi pada *platform Jobstreet* dengan mengimplementasikan kata kunci pencarian (*search keyword*) seperti 'Bidang IT'. Proses ini bertujuan untuk menghimpun seluruh Kebutuhan IDUKA yang relavan secara sistematis, yang nantinya akan menjadi dataset primer dalam penelitian ini. Pengumpulan data dilaksanakan pada tanggal 8 mei 2026. Dari proses tersebut, peneliti berhasil mengekstraksi sebanyak 2.956 data mentah yang bersumber dari iklan lowongan kerja di *platform Jobstreet*. Setelah dilakukan pembersihan data (data cleaning) dengan menghapus data duplikat serta informasi yang tidak relavan, diperoleh sebanyak 1.981 data bersih yang valid dan siap digunakan sebagai dataset utama dalam proses analisis skill.

Seluruh data yang berhasil dihimpun melalui proses *crawling* secara otomatis dikonversi dan disimpan di *google drive*, dalam format berkas .csv (*Comma Separated Values*) guna menjamin struktur data yang konsisten. Dari rangkaian prosedur pengambilan data tersebut, diperoleh dataset mentah dengan total sebanyak 2.956 lowongan pekerjaan. Dataset mentah ini selanjutnya melewati tahap verifikasi dan seleksi data untuk memishkan data yang tidak relevan. Setelah data bersih diperoleh, dilakukan proses ekstraksi untuk mengetahui skill apa saja yang paling banyak dibutuhkan oleh iduka. Hasil pengolahan frekuensi skill tersebut:

4.2 Seleksi Data

Pada tahap ini, Dataset lowongan pekerjaan yang telah dikumpulkan melalui proses *web scraping* menjalani tahap seleksi dan verifikasi data, Proses seleksi difokuskan pada pembersihan data, guna mengeliminasi informasi lowongan yang serupa atau duplikat (data *redundant*).

Setelah dilakukan proses dataset utama yang terdiri dari 2.956 lowongan pekerjaan. Tahap validasi dilakukan untuk memastikan seluruh data bebas dari duplikat, sehingga jumlah dataset yang diproses terdiri dari 1.981. Selanjutnya, dataset tersebut menjalani proses pelabelan otomatis menggunakan metode *lexicon-based*. Implementasi pelabelan ini dilakukan di lingkungan *Google Colab* dengan bahasa pemrograman *Python*, mengacu pada kamus kompetensi (*skill lexicon*) yang telah diten

Tahapan berikutnya setelah integrasi dataset kurikulum dan data IDUKA dari *Google Drive* adalah pemetaan kompetensi secara otomatis. Langkah ini dilakukan untuk Data tersebut dikategorikan berdasarkan profil lulusan atau jenis keterampilan yang relevan. Pengelompokan ini memudahkan proses analisis untuk mengetahui apakah kata kunci dalam kurikulum sudah mencakup keterampilan yang dibutuhkan oleh IDUKA tertentu guna mempermudah analisis sinkronisasi. Dataset yang diolah meliputi dokumen kurikulum serta data lowongan kerja hasil *scraping* yang telah melewati proses *preprocessing* dan eliminasi duplikasi. Ekstraksi fitur dilaksanakan melalui lingkungan *Google Colab* menggunakan metode TF-IDF untuk pembobotan kata kunci, yang selanjutnya dianalisis tingkat relevansinya menggunakan algoritma *Cosine Similarity*. Implementasi teknis dari prosedur ini dipaparkan pada cuplikan kode berikut.

4.3 Pra-Pemrosesan Data (*Text Pre-Processing*)

Seluruh data teks yang telah dikumpulkan, baik dari dokumen kurikulum maupun dari data *Jobstreet*, perlu melewati tahap pembersihan agar informasi di dalamnya lebih berfokus pada kompetensi inti. Hal ini dilakukan untuk menyaring noise dan menyisakan kata kunci (*keywords*) penting yang merepresentasikan keahlian serta kualifikasi yang relevan dengan kebutuhan IDUKA. Tahap ini sangat krusial untuk memastikan bahwa data yang

CC BY SA

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

masuk ke tahap perhitungan kemiripan (*similarity*) adalah data yang berkualitas dan memiliki bobot informatif yang tinggi. Adapun rangkaian langkah untuk merapikan data tersebut meliputi:

4.3.1 Data Cleaning

Pada proses ini dilakukan pembersihan data dengan menyingkirkan elemen-elemen yang tidak memberikan makna atau pengaruh terhadap analisis kemiripan (*similarity*). Elemen yang dibersihkan meliputi simbol-simbol khusus dari dokumen kurikulum, tautan (URL) yang mungkin terbawa dari *scraping Jobstreet*, serta karakter non-alfabet lainnya. Angka dan tanda baca juga dibersihkan agar teks menjadi lebih konsisten. Berikut adalah implementasi *cleaning* pada *google colab* menggunakan *python*:

4.3.2 Case Folding

Tahap *case folding* dilakukan untuk menyeragamkan seluruh karakter agar seluruh teks menjadi huruf kecil (*lowercase)*. Langka ini sangat penting untuk menghindari ambiguitasfitur akibat perbedaan penulisan huruf kapilatyang sama . Berikut adalah implementasi *case folding* pada *google colab* menggunakan *python*.

4.3.3 Stopword Removal

Tahap ini dilakukan reduksi terhadap kata penghubung maupun partikel imbuhan yang di anggap sebagai gangguan (*noise*) dalam pemrosesan data. Dengan menghilangkan kata-kata umum seperti ‘di’, ‘ke’, ‘yang’ dan sejenisnya. Berikut implementasi filtering pada *google colab* menggunakan python.

4.3.4 Tokenizing

Pada tahap ini adalah proses memisahkan rangkaian kalimat menjadi kumpulan kata individu. Hal ini bertujuan untuk mempermudah identifikasi kesejangan keterampilan terkait sebelum masuk ke tahap pembobotan. Berikut implementasi *tokenizing* pada *Google Colab* menggunakan python.

4.5 Pembobotan TF-IDF

Pembobotan kata dalam penelitian ini menggunakan teknik TF-IDF untuk mengubah data teks hasil *preprocessing* dokumen kurikulum dan lowongan kerja industri menjadi nilai numerik. Penggunaan TF-IDF bertujuan untuk memberikan bobot yang lebih tinggi pada istilah kunci atau kompetensi teknis yang spesifik dan mengurangi bobot kata-kata umum yang kurang informatif. Transformasi ini sangat penting karena perhitungan *Cosine Similarity* memerlukan input berupa nilai numerik untuk mengukur tingkat keselarasan antara mata kuliah yang diajarkan di perguruan tinggi dengan kebutuhan keahlian di dunia industri. Berikut adalah implementasi pembobotan TF-IDF menggunakan bahasa pemrograman *Python*.

1) Hitung TF(Trem Frequency)

$$\mathrm{TF}(t, d) = \frac{\text{Jumlah kemunculan kata t dalam dokumen d}}{\text{Total kata dalam dokumen d}}$$

Tabel 2. Hasil *Term Frequency*

Trem	D1	D2	D3	D4	D5	D6
php	1	0	0	0	0	0
java	1	0	0	0	1	0
python	1	0	0	0	0	0
mysql	1	0	1	0	0	0
laravel	1	0	0	0	0	0

CC BY SA

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

spring	0	0	0	1	0	0
sql	1	0	1	0	0	0
linux	1	0	0	0	0	0
golang	1	0	0	0	0	0
c++	1	0	0	0	0	1
aws	0	0	0	0	0	1
mongodb	0	0	0	1	0	0
html	0	1	0	0	0	1
css	0	1	0	0	0	0
javascript	0	1	0	0	0	0
react	0	0	0	0	1	0
angular	0	1	0	0	0	0
vue	0	1	0	0	0	0
sql	1	0	1	0	0	0
mysql	0	0	0	0	0	1
postgresql	0	1	0	0	0	0
api	0	1	0	1	0	0
figma	0	0	1	0	0	0
odoo	0	0	1	0	0	0
ruby	0	0	1	0	0	0
excel	0	0	1	0	0	0
graphql	0	0	1	0	0	0
docker	0	0	0	0	0	1
tableu	0	0	1	0	0	0
jira	0	0	1	0	0	0
spring	0	0	0	1	0	0
nginx	0	0	0	1	0	0
apache	0	0	1	0	0	0
react	0	1	0	0	0	1
typescript	0	0	0	0	1	0
git	0	0	0	1	0	0
ai	0	1	0	0	0	1
kubernetes	0	0	0	0	1	0
terraform	0	0	0	1	0	0
jenkins	0	0	0	1	0	0
django	0	0	0	0	1	0
azure	0	0	0	0	1	0
swift	0	0	0	0	1	0
firabase	0	0	0	0	1	0
agile	0	0	0	1	0	0

CC BY SA

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

scrum	0	0	0	1	0	0
oracle	0	0	0	1	0	0
github	0	0	0	0	1	0
kotlin	0	0	1	0	0	0

Tabel *term frequency* digunakan untuk menghasilkan kata kunci keterampilan industri pada setiap dokumen lowongan kerja yang dianalisis. Setiap baris merepresentasikan term atau skill yang diperoleh dari data industri, sedangkan kolom D1 sampai D6 menunjukkan dokumen lowongan kerja yang digunakan dalam penelitian. Nilai 1 menandakan bahwa term muncul pada dokumen tertentu, sementara nilai 0 menunjukkan bahwa term tidak ditemukan pada dokumen tersebut. Matriks ini selanjutnya digunakan sebagai dasar dalam proses pembobotan TF-IDF serta perhitungan cosine similarity untuk mengetahui tingkat relavan antara kebutuhan industri dan kurikulum Program Studi Sistem Informasi.

2) Hitung DF (Document Frequency)

DF menghitung jumlah kata dalam Skill

Tabel 3. Hasil *Document Frequency*

Trem	DF
Php	1
Java	2
Python	1
Mysql	2
Laravel	1
Spring	1
Sql	1
Linux	1
Golang	1
c++	2
Aws	1
Mongodb	1
Html	1

Css	1
Javascript	1
React	1
Angular	1
Vue	1
Sql	1
Mysql	2
Postgresql	1
Api	2
Figma	1
Odoo	1
Ruby	1
Excel	1
Graphql	1
Docker	1
Tableu	1
Jira	1
Spring	1
Nginx	1
Apache	1
React	2
Typescript	1
Git	1
Ai	2

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

Kubernetes	1
Terraform	1
Jenkins	1
Django	1
Azure	1
Swift	1
Firabase	1
Agile	1
Scrum	1
Oracle	1
Github	1
Kotlin	1

Tabel *Document Frequency* (DF) digunakan untuk menunjukkan jumlah kemunculan term dalam dokumen atau kata kunci keterampilan industri. Kolom term menunjukkan daftar skill yang diperoleh dari data lowongan kerja, sedangkan kolom DF menunjukkan jumlah dokumen tempat term tersebut muncul. Semakin besar nilai DF, maka semakin sering term ditemukan pada berbagai dokumen. Sebaliknya, nilai DF yang kecil menunjukkan bahwa term hanya terdapat pada sedikit dokumen. term dalam keseluruhan dokumen yang dianalisis.

3) Hitung IDF (Inverse Dokument Frequency)

$$\mathrm{IDF(t)} = \log\left(\frac{\mathrm{N}}{\mathrm{df(t)}}\right) \qquad (4.1)$$

Dimana : N = jumlah dokumen atau skill (N=6) DF(t)= jumlah dokumen atau skill yang mengandung kata t

$$\mathrm{IDF(t)} = \log\left(\frac{6}{\mathrm{df(t)}}\right) \qquad (4.2)$$

Tabel 4. Hasil *Inverse Document Frequency*

Trem	DF	IDF
php	1	0.7782

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

java	2	0.4771
python	1	0.7782
mysql	2	0.4771
laravel	1	0.7782
spring	1	0.7782
sql	1	0.7782
linux	1	0.7782
golang	1	0.7782
c++	2	0.4771
aws	1	0.7782
mongodb	1	0.7782
html	1	0.7782
css	1	0.7782
javascript	1	0.7782
react	1	0.7782
angular	1	0.7782
vue	1	0.7782
mysql	2	0.4771
postgresql	1	0.7782
api	2	0.4771
figma	1	0.7782
odoo	1	0.7782
ruby	1	0.7782
excel	1	0.7782

CC BY SA

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

graphql	1	0.7782
docker	1	0.7782
tableu	1	0.7782
jira	1	0.7782
spring	1	0.7782
nginx	1	0.7782
apache	1	0.7782
react	2	0.4771
typescript	1	0.7782
git	1	0.7782
ai	2	0.4771
kubernetes	1	0.7782
terraform	1	0.7782
jenkins	1	0.7782
django	1	0.7782
azure	1	0.7782
swift	1	0.7782
firabase	1	0.7782
agile	1	0.7782
scrum	1	0.7782
oracle	1	0.7782
github	1	0.7782
kotlin	1	0.7782

CC BY SA

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

Tabel DF-IDF digunakan untuk menunjukkan tingkat kepentingan setiap term atau kata kunci keterampilan industri pada dokumen yang dianalisis. Kolom DF menunjukkan jumlah dokumen yang memuat *term*, sedangkan kolom IDF menunjukkan bobot penting term berdasarkan frekuensi kemunculannya. Semakin sedikit term muncul pada dokumen, maka nilai IDF akan semakin besar sehingga term dianggap lebih spesifik dan penting. Nilai IDF ini digunakan dalam proses pembobotan TF-IDF sebelum dilakukan perhitungan *cosine similarity* untuk mengukur tingkat kesesuaian antara kebutuhan industri dan kurikulum Program Studi Sistem Informasi.

4) Hitung TF-IDF

$$\mathrm{TF} - \mathrm{IDF}(t, d) = \mathrm{TF}(t, d) \times \mathrm{IDF}(t) \qquad (4.3)$$

Dimana : TF (t,d) = adalah frekuensi kemunculan kata t dikomentar t

$\mathrm{IDF}(t) =$ adalah nilai IDF dari kata t

Tabel 5. Hasil *Term Frequency - Iverse Document Frequency*

Trem	D1	D2	D3	D4	D5	D6
php	0.7782	0	0	0	0	0
java	0.4771	0	0	0	0.4771	0
python	0.7782	0	0	0	0	0
mysql	0.4771	0.4771	0	0	0	0
laravel	0.7782	0	0	0	0	0
spring	0	0	0	0.7782	0	0
sql	0.7782	0	0	0	0	0
linux	0.7782	0	0	0	0	0
golang	0.7782	0	0	0	0	0
c++	0.4771	0	0	0	0	0.4771
aws		0	0	0	0	0.7782
mongodb	0	0	0	0.7782	0	0
html	0	0	0	0	0	0.4771
css	0	0.7782	0	0	0	0
javascript	0	0	0	0	0.7782	0
react	0	0	0	0	0.7782	0
angular	0	0.7782	0	0	0	0
vue	0	0.7782	0	0	0	0
sql	0	0.7782	0	0	0	0
mysql	0.4771	0.4771	0	0	0	0
postgresql	0	0.7782	0	0	0	0
api	0	0.4771	0	0.4771	0	0
figma	0	0	0.7782	0	0	0
odoo	0	0	0.7782	0	0	0
ruby	0	0	0.7782	0		0

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

sap	0	0	0.7782	0	0	0
excel	0	0	0.7782	0	0	0
graphql	0	0	0.7782	0	0	0
docker	0	0		0	0	0.7782
tableu	0	0	0.7782	0	0	0
jira	0	0	0.7782	0	0	0
spring	0	0	0	0.7782	0	0
nginx	0	0	0	0.7782	0	0
apache	0	0	0.7782	0	0	0
react	0	0.4771	0	0	0	0.4771
typescript	0	0	0	0	0.7782	0
git	0	0	0	0.7782	0	0
ai		0.4771	0		0	0.4771
kubernetes	0	0	0	0	0.7782	0
terraform	0	0	0	0.7782	0	0
jenkins	0	0	0	0.7782	0	0
django	0	0	0	0	0.7782	0
azure	0	0	0	0	0.7782	0
swift	0	0	0	0	0.7782	0
firabase	0	0	0	0	0.7782	0
agile	0	0	0	0.7782	0	0
scrum	0	0	0	0.7782	0	0
oracle	0	0	0	0.7782	0	0
gthub	0	0	0	0.7782	0	0
kotlin	0	0	0.7782	0	0	0

Tabel DF-IDF digunakan untuk menunjukkan tingkat kepentingan setiap term atau kata kunci keterampilan industri dalam dokumen yang dianalisis. Kolom DF menunjukkan jumlah dokumen yang memuat *term*, sedangkan kolom IDF menunjukkan bobot penting term berdasarkan frekuensi kemunculannya. Term yang jarang muncul akan memiliki nilai IDF lebih besar karena dianggap lebih spesifik, sedangkan term yang sering muncul memiliki nilai IDF lebih kecil. Nilai IDF digunakan dalam proses pembobotan TF-IDF sebelum dilakukan perhitungan *cosine similarity* untuk mengetahui tingkat kesesuaian antara kebutuhan industri dan kurikulum Program Studi Sistem Informasi.

CC BY SA

Accredited SINTA 5
72/E/KPT/2024

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

Tabel 6. Sampel Data Kurikulum SI

Trem	D1
D1	dbms,erd,modeling,normalization,crud,sql ,query,linux,Bash Scripting,Virtualization,Cloud Computing
D2	System Security,RAID,Bash Scripting,sdlc,agile,scrum,waterfall,uml,Prototyping,Cyber Security,Information Security
D3	Security,Information Security,Cyber Security,html,css,javascript,php,mysql,sql,crud,Bootstrap,AJAX,JSON
D4	Network Security,Firewall,sql,Wireshark,Cryptography,Zero Trust Security,Wireless Security,ISO 27001
D5	JSON,Web Security,Deployment,mysql,php,javascript,html,css,flutter,dart,rest api,linix,agile,scrum,uml
D6	Network Security,agile,Cyber Security,firewall,sap,Encryption,Computer Networks,crm,erp,sql

5) Hitung TF (Trem Frequency)

$$\mathrm{TF}(\mathrm{t},\mathrm{d}) = \frac{\text{Jumlah kemunculan kata t dalam dokumen d}}{\text{Total kata dalam dokumen d}}$$

Tabel 7. Hasil *Term Frequency*

Kata Term	D1	D2	D3	D4	D5	D6
dbms	1	0	0	0	0	0
erd	1	0	0	0	0	0
modeling	1	0	0	0	0	0
normalization	1	0	0	0	0	0
sql	1	0	1	1	0	0
crud	1	0	1	0	0	0
query	1	0	0	0	0	0
linux	1	0	0	0	1	0
bash scripting	1	1	0	0	0	0
virtualization	1	0	0	0	0	0

CC BY SA

cloud computing	1	0	0	0	0	0
system securytt	1	0	0	0	0	0
raid	0	1	0	0	0	0
agile	0	1	0	0	1	1
scrum	0	1	0	1	0	0
waterfall	0	1	0	0	0	0
uml	0	1	0	0	1	0
prototyping	0	1	0	0	0	0
cyber security	0	1	1	0	1	0

Tabel TF-IDF kurikulum digunakan untuk menampilkan bobot setiap term atau kata kunci pada dokumen kurikulum yang dianalisis. Kolom D1 sampai D6 menunjukkan dokumen kurikulum, sedangkan nilai pada tabel menggambarkan bobot kemunculan term pada masing-masing dokumen. Bobot TF-IDF digunakan untuk mengetahui tingkat kepentingan suatu term dalam merepresentasikan isi dokumen kurikulum. Semakin besar nilai bobot term, maka semakin penting term tersebut pada dokumen. Hasil pembobotan ini kemudian digunakan dalam perhitungan cosine similarity untuk mengukur tingkat kesesuaian antara kurikulum dan kebutuhan industri.

6) Hitung DF (*Document Frequency*)

DF menghitung jumlah kata dalam Skill.

Tabel 8. Hasil Document Frequency

Kata Term	Df
Dbms	1
Erd	1
modeling	1
Normalization	1
Sql	3
Crud	2
Query	1
Linux	2
bash scripting	2
Virtualization	1

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

cloud computing	1
system securytt	1
Raid	1
Agile	3
Scrum	2
Waterfall	1
Uml	2
Prototyping	1
cyber security	3
information security	2
Security	1
Html	2
Css	2
Javascript	2
Php	2
Mysql	2
Bootstrap	1
Ajax	1
Json	2
network security	2
Firewall	2
Wireshark	1
Cryptogtaphy	1
zero trust security	1
wireless security	1
iso 27001	1
rest api	1
web security	1
deployment	1
flutter	1
dart	1
agile	2
sap	1
encryption	1
komputer nerwork	1
crm	1

CC BY SA

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

erp	1
html	**1**

Tabel TF-IDF kurikulum digunakan untuk menampilkan bobot setiap term atau kata kunci pada dokumen kurikulum yang dianalisis. Kolom D1 sampai D6 menunjukkan dokumen kurikulum, sedangkan nilai pada tabel menggambarkan bobot kemunculan term pada masing-masing dokumen. Bobot TF-IDF digunakan untuk mengetahui tingkat kepentingan suatu term dalam merepresentasikan isi dokumen kurikulum. Semakin besar nilai bobot term, maka semakin penting term tersebut pada dokumen. Hasil pembobotan ini kemudian digunakan dalam perhitungan *cosine similarity* untuk mengukur tingkat kesesuaian antara kurikulum dan kebutuhan industri.

7) Hitung IDF (Inverse Dokument Frequency)

$$\mathrm{IDF(t)} = \log\left(\frac{\mathrm{N}}{\mathrm{df(t)}}\right) \qquad (4.4$$

Dimana : N = jumlah dokumen atau skill (N=6) DF(t)= jumlah dokumen atau skill yang mengandung kata t

$$\mathrm{IDF(t)} = \log\left(\frac{6}{\mathrm{df(t)}}\right) \qquad (4.5$$

Tabel 9. Hasil *Inverse Document Frequency*

Kata Term	Df	IDF
Dbms	1	0.7782
Erd	1	0.7782
modeling	1	0.7782
Normalization	1	0.7782
Sql	3	0.3010
Crud	2	0.4771
Query	1	0.7782
Linux	2	0.4771
bash scripting	2	0.4771
Virtualization	1	0.7782
cloud computing	1	0.7782
system securytt	1	0.7782
Raid	1	0.7782
Agile	3	0.3010
Scrum	2	0.4771
Waterfall	1	0.7782
Uml	2	0.4771
Prototyping	1	0.7782
cyber security	3	0.3010
information security	2	0.4771
Security	1	0.7782

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

Html	2	0.4771
Css	2	0.4771
Javascript	2	0.4771
Php	2	0.4771
Mysql	2	0.4771
Bootstrap	1	0.7782
Ajax	1	0.7782
Json	2	0.4771
network security	2	0.4771
Firewall	2	0.4771
Wireshark	1	0.7782
Cryptogtaphy	1	0.7782
zero trust security	1	0.7782
wireless security	1	0.7782
iso 27001	1	0.7782
rest api	1	0.7782
web security	1	0.7782
deployment	1	0.7782
flutter	1	0.7782
dart	1	0.7782
agile	2	0.4771
sap	1	0.7782
encryption	1	0.7782
komputer nerwork	1	0.7782
crm	1	0.7782
erp	1	0.7782
html	1	0.7782

8) Hitung TF-IDF

$$TF - IDF(t,d) = TF(\mathrm{t,d}) \times IDF(\mathrm{t}) \qquad (4.6)$$

Dimana : TF (t,d) = adalah frekuensi kemunculan kata t dikomentar t

IDF(t) = adalah nilai IDF dari kata t

Tabel 10. Hasil *Term Frequency - Inverse Document Frequency*

Kata Term	D1	D2	D3	D4	D5	D6
Dbms	0.7782	0	0	0	0	0
Erd	0.7782	0	0	0	0	0

CC BY SA

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

Modeling	0.7782	0	0	0	0	0
Normalization	0.7782	0	0	0	0	0
Sql	0	0	0.3010	0.3010	0	0
Crud	0	0	0.4771	0	0	0
Query	0.7782	0	0	0	0	0
Linux	0.4771	0	0	0	0.4771	0
bash scripting	0.4771	0	0	0	0	0
Virtualization	0.7782	0	0	0	0	0
cloud computing	0.7782	0	0	0	0	0
system securytt	0.7782	0	0	0	0	0
Raid	0	0.7782	0	0	0	0
Agile	0	0	0	0	0.3010	0.3010
Scrum	0	0	0	0	0	0
Waterfall	0	0.7782	0	0	0	0
Uml	0	0	0	0	0	0
Prototyping	0	0.7782	0	0	0	0
cyber security	0	0	0.3010	0	0	0
information security	0	0	0.4771	0	0	0
Security	0	0	0.7782	0	0	0
Html	0	0	0.4771	0	0	0
Css	0	0	0.4771	0	0	0
Javascript	0	0	0.4771	0	0	0
Php	0	0	0.4771	0	0	0
Mysql	0	0	0.4771	0	0	0
Bootstrap	0	0	0.7782	0	0	0
Ajax	0	0	0.7782	0	0	0
Json	0	0	0.4771	0	0	0
network security	0	0	0	0	0	0
Firewall	0	0	0	0	0	0.4771
Wireshark	0	0	0	0.7782	0	0
Cryptogtaphy	0	0	0	0.7782	0	0
zero trust security	0	0	0	0.7782	0	0
wireless security	0	0	0	0.7782	0	0
iso 27001	0	0	0	0.7782	0	0
rest api	0	0	0	0	0.7782	0
web security	0	0	0	0	0.7782	0

CC BY SA

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

deployment	0	0	0	0	0.7782	0
flutter	0	0	0	0	0.7782	0
dart	0	0	0	0	0.7782	0
agile	0	0	0	0	0	0
sap	0	0	0	0	0	0.7782
encryption	0	0	0	0	0	0.7782
komputer nerwork	0	0	0	0	0	0.7782
crm	0	0	0	0	0	0.7782
erp	0	0	0	0	0	0.7782
html	0	0.7782	0	0	0	**0**

Berdasarkan hasil perhitungan TF-IDF pada kedua data, diperoleh nilai bobot untuk setiap kata pada dokumen D1 sampai dengan D6. Nilai bobot ini menujukkan tingkat kepentingan kata dalam masi-masing dokumen. Hasil pembobotan tersebut selanjutnya digunakan pada tahapan proses analisis.

4.5.1 Perhitungan Manual *Cosine Similarity*

Untuk memberikan pemahaman yang lebih jelas mengenai proses kerja algoritma dalam membandingkan kompetensi mahasiswa dengan kebutuhan industri, dilakukan simulasi perhitungan manual *Cosine Similarity* pada bidang Data Analisis. Berdasarkan hasil pembobotan TF-IDF yang diperoleh melalui *Google Colab*, dipilih beberapa kata kunci *(term)* yang menunjukkan kesamaan antara data kurikulum dan data IDUKA, (*term*) dominan yang merepresentasikan irisan data Kurikulum (A) dan data IDUKA (B) dengan rumus sebagai berikut:

$$\text{similarity} = \cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \qquad (4.7)$$

1. Perhitungan Perkalian Titik (*Dot Product*)

Pada tahap ini, setiap bobot kata yang sama dari kedua dokumen dikalikan, lalu hasilnya dijumlahkan secara total. Nilai ini menunjukkan seberapa banyak irisan kompetensi yang ada.

Tabel 11. Perhitungan Manual *Cosine Similarity*

Skill	Bobot Iduka (Ai)	Bobot Kurikulum (Bi)	Ai x Bi	Ai2	Bi2
php	0.7782	0.4771	0.3713	0.6056	0.2276
java	0.4771	0	0	0.2276	0
python	0.7782	0.7782	0.6056	0.6056	0.6056
mysql	0.4771	0.7782	0.3713	0.2276	0.6056
laravel	0.7782	0.4771	0	0.6056	0.2276
spring	0.7782	0	0	0.6056	0
sql	0.7782	0.7782	0.6056	0.6056	0.6056
linux	0.7782	0.4771	0.3713	0.6056	0.2276

CC BY SA

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

golang	0.7782	0	0	0.6056	0
c++	0.4771	0	0	0.2276	0
aws	0.7782	0	0	0.6056	0
mongodb	0.7782	0.7782	0.6056	0.6056	0.6056
html	0.7782	0.7782	0.6056	0.6056	0.6056
css	0.7782	0.7782	0.6056	0.6056	0.6056
javascript	0.7782	0.4771	0.3713	0.6056	0.2276
react	0.7782	0	0	0.6056	0
angular	0.7782	0	0	0.6056	0
vue	0.7782	0	0	0.6056	0
postgresql	0.7782	0	0	0.6056	0
Api	0.4771	0.7782	0	0.2276	0.6056
figma	0.7782	0	0	0.6056	0
odoo	0.7782	0	0	0.6056	0
ruby	0.7782	0	0	0.6056	0
excel	0.7782	0	0	0.6056	0
graphql	0.7782	0	0	0.6056	0
docker	0.7782	0	0	0.6056	0
tableu	0.7782	0	0	0.6056	0
Jira	0.7782	0	0	0.6056	0
nginx	0.7782	0	0	0.6056	0
apache	0.7782	0	0	0.6056	0
typescript	0.7782	0	0	0.6056	0
Git	0.7782	0	0	0.6056	0
Ai	0.4771	0	0	0.2276	0
kubernetes	0.7782	0	0	0.6056	0
terraform	0.7782	0	0	0.6056	0
jenkins	0.7782	0	0	0.6056	0
django	0.7782	0	0	0.6056	0
azure	0.7782	0	0	0.6056	0
swift	0.7782	0	0	0.6056	0
firabase	0.7782	0	0	0.6056	0
agile	0.7782	0.4771	0	0.6056	0.2276
scrum	0.7782	0.7782	0.6056	0.6056	0.6056
oracle	0.7782	0	0	0.6056	0
firewall	0.7782	0.7782	0.6056	0.6056	0.6056
cybersecurity	0	0.3010	0	0	0.0906
dbms	0	0.7782	0	0	0.6056
cryptography	0	0.7782	0	0	0.6056
mysql	0	0.4771	0	0	0.2276

CC BY SA

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

github	0.7782	0	0	0.6056	0
zero trust security	0	0.7782	0	0	0.6056
kotlin	0.7782	0	0	0.6056	0
wireless security	0	0.7782	0	0	0.6056
iso 27001	0	0.7782	0	0	0.6056
rest api	0	0.7782	0	0	0.6056
web security	0	0.7782	0	0	0.6056
deployment	0	0.7782	0	0	0.6056
flutter	0	0.7782	0	0	0.6056
dart	0	0.7782	0	0	0.6056
sap	0	0.7782	0	0	0.6056
encryption	0	0.7782	0	0	0.6056
komputer nerwork	0	0.7782	0	0	0.6056
crm	0	0.7782	0	0	0.6056
erp	0	0.7782	0	0	0.6056
dbms	0	0.7782	0	0	0.6056
erd	0	0.7782	0	0	0.6056
modeling	0	0.7782	0	0	0.6056
normalization	0	0.7782	0	0	0.6056
query	0	0.7782	0	0	0.6056
crud	0	0.4771	0	0	0.2276
bash scripting	0	0.4771	0	0	0.2276
virtualization	0	0.4771	0	0	0.2276
cloud computing	0	0.7782	0	0	0.6056
system securyti	0	0.7782	0	0	0.6056
raid	0	0.7782	0	0	0.6056
json	0	0.7782	0	0	0.6056
json decoding	0	0.7782	0	0	0.6056
sqlite	0	0.7782	0	0	0.6056
Total			6.8381	25.9675	23.3351

2. Menghitung Magnitudo IDUKA (A)

$$|A| = \sqrt{\sum A_i^2}$$
$$|A| = \sqrt{25.9676}$$
$$|A| = 5.0058$$

3. Menghitung Magnitudo Kurikulum (B)

$$|B| = \sqrt{\sum B_i^2}$$

$$|B| = \sqrt{23.3351}$$

$$|B| = 4.8306$$

CC BY SA

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

4. Hitung Hasil *Cosien Similarity*

$$Similarity = \frac{\sum(A_i \times B_i)}{|A| \times |B|}$$

$$Similarity = \frac{6.8381}{5.0958 \times 4.8306}$$

$$Similarity = \frac{6.8381}{24.6161}$$

$$Similarity = 0.2778$$

Setelah dilakukan perhitungan dengan menggunakan metode TF-IDF dan *Cosine Similarity*, diperoleh nilai akhir yang menunjukkan tingkat relevan antara kurikulum Program Studi Sistem Informasi dengan kebutuhan IDUKA. Untuk memberikan representasi data yang lebih informatif, hasil evaluasi tersebut dikelompokan ke dalam matrix yang dikategorikan berdasarkan tiga tingkat relevansi, yaitu: Tinggi, Sedang, dan Rendah.

4.5.2 Hasil *Cosine Similarity*

Tahap pengukuran kemiripan merupakan inti dari analisis untuk mengevaluasi kesesuaian antara kurikulum dengan kebutuhan iduka. Setelah seluruh data teks dari dokumen kurikulum dan lowongan kerja dikonversi menjadi representasi numerik menggunakan metode TF-IDF, sistem kemudian menghitung tingkat kemiripan antar kedua vektor tersebut menggunakan algoritma *Cosine Similarity*.

Metode ini mengukur sudut kosinus antara dua vektor dalam ruang multidimensi. Nilai hasil perhitungan ini berada dalam rentang jangkauan 0 sampai 1. Interpretasi skor tersebut adalah sebagai berikut:

a. Nilai mendekati 1: Menunjukkan bahwa kedua vektor (kurikulum dan kebutuhan iduka) dianggap semakin mirip atau memiliki kesesuaian yang tinggi.
b. Nilai mendekati 0: Menunjukkan bahwa tidak ada kemiripan atau keterkaitan kompetensi antara kedua dokumen tersebut.

$$\text{similarity} = \cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \qquad (4.6)$$

Keterangan:

a. A_i : Bobot TF-IDF kata ke−i pada dokumen Kurikulum
b. B_i : Bobot TF-IDF kata ke−i pada dokumen IDUKA (Lowongan Kerja)
c. $\sum A_i B_i$: *Dot Product* (perkalian titik) antara vektor A dan B
d. $\sqrt{\sum A_i^2}$: Panjang (norma) dari vektor A.
e. $\sqrt{\sum B_i^2}$: Panjang (norma) dari vektor B.
f. i : jumlah term dalam kalimat
g. n : jumlah vektor

Dalam implementasinya pada data TF-IDF, bagian pembilang *(Dot Product)* bekerja dengan cara mengalikan

CC BY SA

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

bobot istilah yang sama antara dokumen kurikulum dan dokumen IDUKA. Apabila suatu skill (misalnya "SQL") diajarkan di mata kuliah dan juga secara eksplisit disyaratkan oleh lowongan pekerjaan, maka hasil perkalian ini akan memperbesar nilai pembilang. Sebaliknya, apabila materi yang diajarkan tidak ditemukan sama sekali dalam *tech-stack* yang dicari oleh IDUKA, maka hasil perkalian silangnya akan bernilai nol.

Selanjutnya, nilai tersebut dibagi dengan bagian penyebut (perkalian norma vektor dari kedua dokumen). Tahap pembagian ini berfungsi sebagai proses normalisasi. Hal ini sangat krusial karena teks gabungan dari ribuan lowongan pekerjaan IDUKA memiliki panjang yang jauh lebih besar dibandingkan dengan deskripsi tunggal sebuah mata kuliah. Melalui pembagian ini, panjang dokumen tidak akan mendistorsi hasil, sehingga nilai kemiripan akan selalu stabil berada pada skala 0 hingga 1.

Algoritma ini memberikan peringkat pada suatu dokumen sesuai dengan skor kemiripan yang dihasilkan. Apabila total *similarity* yang didapatkan nol (0), maka dokumen yang diolah tidak memiliki kesamaan. Namun jika nilai yang didapatkan maksimal adalah 1, maka dokumen tersebut memiliki kemiripan yang mutlak.

Tabel 12.Hasil *Cosine Similarity*

No	Rentang Nilai Kemiiripan	Nilai Similarity	Interpretasi
1	Big Data	0.46	Relavan tinggi
2	Pemrograman Berbasis Web	0.44	Relavan tinggi
3	Data Mining	0.43	Sangat Tinggi
4	Sistem Informasi Manajemen	0.36	Relavan Sedang
5	Keamanan Jaringan	0.35	Relavan Sedang
6.	Sistem Operasi	0.34	Relavan Sedang
7	Pemrograman Aplikasi Bergerak	0.30	Relavan Rendah
8	Analisis Dan Perancangan Sistem Informasi	0.29	Relavan Rendah
9	Sistem Basis Data	0.15	Relavan Rendah
10	Audit Sistem Informasi	0.12	Relavan Rendah

Berdasarkan hasil pengolahan data menggunakan teknik *Natural Language Processing* (NLP) melalui pembobotan TF-IDF dan penghitungan *Cosine Similarity*, diperoleh nilai representasi kemiripan tekstual antara dokumen kurikulum mata kuliah dengan dokumen kebutuhan industri (IDUKA). Sebagai berikut:

a. Big Data memperoleh nilai *similarity* tertinggi, yaitu sebesar 46, sehingga termasuk dalam kategori relevan tinggi. Hasil ini menunjukkan bahwa materi pembelajaran yang berkaitan dengan pengolahan, penyimpanan, dan analisis data dalam jumlah besar memiliki tingkat kesesuaian yang sangat baik dengan kompetensi yang dibutuhkan oleh IDUKA.
b. Pemrograman Berbasis Web memperoleh nilai *similarity* sebesar 44 dan berada pada kategori relevan tinggi. Nilai tersebut mengindikasikan bahwa materi pembelajaran, teknologi, serta bahasa pemrograman web yang diajarkan telah sesuai dengan kebutuhan kompetensi yang banyak dibutuhkan di dunia industri.

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

c. Data Mining memiliki nilai *similarity* sebesar 43 dan termasuk dalam kategori relevan tinggi. Hal ini menunjukkan bahwa kompetensi dalam proses penggalian dan analisis informasi dari data telah selaras dengan kebutuhan industri yang semakin berorientasi pada pemanfaatan data untuk mendukung pengambilan keputusan.
d. Sistem Informasi Manajemen memperoleh nilai *similarity* sebesar 36 dan termasuk dalam kategori relevan sedang. Hasil tersebut menunjukkan bahwa materi yang diajarkan telah cukup sesuai dengan kebutuhan IDUKA, terutama dalam aspek pengelolaan sistem informasi dan proses bisnis organisasi.
e. Keamanan Jaringan memperoleh nilai *similarity* sebesar 35 dan berada pada kategori relevan sedang. Nilai ini menunjukkan bahwa kompetensi dasar mengenai keamanan jaringan telah memenuhi sebagian kebutuhan industri, meskipun masih diperlukan penyesuaian terhadap perkembangan ancaman siber dan teknologi keamanan yang terus berkembang.
f. Sistem Operasi memiliki nilai *similarity* sebesar 34 dan termasuk dalam kategori relevan sedang. Hasil ini menunjukkan bahwa materi mengenai sistem operasi masih memiliki keterkaitan dengan kebutuhan industri, terutama sebagai kompetensi dasar yang mendukung penguasaan teknologi informasi.
g. Pemrograman Aplikasi Bergerak memperoleh nilai *similarity* sebesar 30 dan termasuk dalam kategori relevan rendah. Nilai tersebut menunjukkan bahwa tingkat kesesuaian kompetensi yang diajarkan masih terbatas, sehingga diperlukan pembaruan materi, khususnya terkait penggunaan *framework* dan teknologi pengembangan aplikasi bergerak yang banyak digunakan di industri.
h. Analisis dan Perancangan Sistem Informasi memperoleh nilai *similarity* sebesar 29 dan berada pada kategori relevan rendah. Hasil ini mengindikasikan bahwa materi pembelajaran masih perlu disesuaikan dengan pendekatan pengembangan perangkat lunak yang saat ini lebih banyak diterapkan di industri, seperti metode Agile maupun Scrum.
i. Sistem Basis Data memperoleh nilai *similarity* sebesar 15 dan termasuk dalam kategori relevan rendah. Nilai tersebut menunjukkan bahwa materi yang diajarkan masih memiliki tingkat kesesuaian yang terbatas dengan kebutuhan IDUKA, sehingga diperlukan penguatan pada penerapan teknologi basis data yang banyak digunakan dalam dunia kerja.
j. Audit Sistem Informasi memperoleh nilai *similarity* terendah, yaitu sebesar 12, dan termasuk dalam kategori relevan rendah. Hasil ini menunjukkan bahwa kompetensi audit sistem informasi relatif jarang muncul sebagai persyaratan utama pada lowongan pekerjaan yang dianalisis, sehingga tingkat kesesuaiannya dengan kebutuhan IDUKA masih rendah.

2. Menentukan Nilai *Cosine Similarity* (Big Data)

$$CS = \frac{\sum(A_i \times B_i)}{\sqrt{\sum A_i^2} \times \sqrt{\sum B_i^2}}$$

$$CS = \frac{1.5825}{\sqrt{3.8612} \times \sqrt{3.0280}}$$

$$CS = \frac{1.5825}{1.9650 \times 1.7401}$$

CC BY SA

$$CS = \frac{1.5825}{3.4193} = 0.46$$

Tabel 13. Hasil Cosine Similarity (Basis Data)

Term (Kata Kunci)	Bobot Iduka (Ai)	Bobot Kurikulum (Bi)	AixBi	Ai2	Bi2
Mongodb	0.7782	0	0	0.6056	0
cloud computing	0	0.7782	0	0	0.6056
Aws	0.7782	0	0	0.6056	0
Azura	0.7782	0	0	0.6056	0
Tableu	0.7782	0	0	0.6056	0
Mongodb	0.7782	0.7782	0.6056	0.6056	0.6056
Python	0.7782	0.7782	0.6056	0.6056	0.6056
Nosql	0	0.7782	0	0	0.6055
Api	0.4771	0.7782	0.3713	0.2276	0.6056
Total			1.5825	3.8612	3.0280

3. Menentukan Nilai *Cosine Similarity* (Pemrograman Berbasis Web)

$$CS = \frac{\sum(A_i \times B_i)}{\sqrt{\sum A_i^2} \times \sqrt{\sum B_i^2}}$$

$$CS = \frac{3.3019}{\sqrt{11.5836} \times \sqrt{4.9220}}$$

$$CS = \frac{3.3019}{3.4035 \times 2.2186}$$

$$CS = \frac{3.3019}{7.5508} = 0.44$$

Tabel 14. Hasil *Cosine Similarity (*Pemrograman Berbasis Web*)*

Term (Kata Kunci)	Bobot Iduka (Ai)	Bobot Kurikulum (Bi)	AixBi	Ai2	Bi2
php	0.7782	0.4771	0	0.6056	0.2276
java	0.4771	0	0	0.2276	0
graphql	0.7782	0	0	0.6056	0
python	0.7782	0	0	0.6056	0
laravel	0.7782	0	0	0.6056	0
spring	0.7782	0	0	0.6056	0
html	0.7782	0.4771	0	0.6056	0.2276

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

css	0.7782	0.7782	0.6056	0.6056	0.6056
javascript	0.7782	0.4771	0	0.6056	0.2276
react	0.7782	0	0	0.6056	0
angular	0.7782	0	0	0.6056	0
vue	0.7782	0	0	0.6056	0
ruby	0.7782	0	0	0.6056	0
django	0.7782	0	0	0.6056	0
typescript	0.7782	0	0	0.6056	0
graphql	0.7782	0	0	0.6056	0
rest api	0	0.7782	0	0	0.6056
api	0.4771	0	0	0.2276	0
golang	0.7782	0	0	0.6056	0
sql	0.7782	0.7782	0.6056	0.6056	0.6056
jquery	0	0.7782	0	0	0.6056
json	0	0.7782	0	0.0000	0.6056
html	0.7782	0.7782	0.6056	0.6056	0.6056
mysql	0.4771	0.7782	0	0.2276	0.6056
Total			3.3019	11.5836	4.9220

4. Menghitung Nilai *Cosine Similarity* (Data Mining)

$$CS = \frac{\sum(A_i \times B_i)}{\sqrt{\sum A_i^2} \times \sqrt{\sum B_i^2}}$$

$$CS = \frac{0.6055}{\sqrt{2.8775} \times \sqrt{0.6962}}$$

$$CS = \frac{0.6055}{1.6964 \times 0.8344}$$

$$CS = \frac{0.6055}{1.4154} = 0.43$$

5. Tabel 15. Hasil *Cosine Similarity* (Data Ming)

Term (Kata Kunci)	Bobot Iduka (Ai)	Bobot Kurikulum (Bi)	AixBi	Ai2	Bi2
ai	0.4771	0	0	0.2276	0
python	0.7782	0	0	0.6056	0
tableu	0.7782	0	0	0.6056	0

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

excel	0.7782	0	0	0.6056	0
sql	0.7782	0.3010	0.2342	0.6056	0.0906
mysql	0.4771	0.7782	0.3713	0.2276	0.6055
Total			0.6055	2.8776	0.6962

6. Menghitung Nilai *Cosine Similarity* (Sistem Informasi Manajemen)

$$CS = \frac{\sum(A_i \times B_i)}{\sqrt{\sum A_i^2} \times \sqrt{\sum B_i^2}}$$

$$CS = \frac{0.9769}{\sqrt{3.6336} \times \sqrt{2.0444}}$$

$$CS = \frac{0.9769}{1.9062 \times 1.4298}$$

$$CS = \frac{0.9769}{2.7255} = 0.36$$

7. Tabel 16. Hasil *Cosine Similarity* (Sistem Informasi Manajemen)

Term (Kata Kunci)	Bobot Iduka (Ai)	Bobot Kurikulum (Bi)	AixBi	Ai2	Bi2
Crm	0	0.7782	0	0	0.6056
Erp	0	0.7782	0	0	0.605595
excel	0.7782	0	0	0.6056	0
Odoo	0.7782	0	0	0.6056	0
tableu	0.7782	0	0	0.6056	0
Agile	0.7782	0.4771	0	0.6056	0.2276
scrum	0.7782	0.7782	0.6056	0.6056	0.6056
Jira	0.7782	0	0	0.6056	0
Total			0.9769	3.6336	2.0444

8. Menghitung Nilai *Cosine Similarity* (Keamanan Jaringan)

$$CS = \frac{\sum(A_i \times B_i)}{\sqrt{\sum A_i^2} \times \sqrt{\sum B_i^2}}$$

CC BY SA

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

$$CS = \frac{0.6055}{\sqrt{0.6056} \times \sqrt{4.9354}}$$

$$CS = \frac{0.6055}{0.7782 \times 2.2216}$$

$$CS = \frac{0.6055}{1.7288} = 0.35$$

Tabel 17. Hasil *Cosine Similarity* (Keamanan Jaringan)

Term (Kata Kunci)	Bobot Iduka (Ai)	Bobot Kurikulum (Bi)	AixBi	Ai2	Bi2
firewall	0.7782	0.7782	0.6056	0.6056	0.6056
cybersecurity	0	0.3010	0	0	0.0906
cryptography	0	0.7782	0	0	0.6056
zero trust security	0	0.7782	0	0	0.6056
wireless security	0	0.7782	0	0	0.6056
iso27001	0	0.7782	0	0	0.6056
web security	0	0.7782	0	0	0.6056
komputer network	0	0.7782	0	0	0.6056
system security	0	0.7782	0	0	0.6056
Total			0.6056	0.6056	4.9354

9. Menghitung Nilai *Cosine Similarity* (Sistem Operasi)

$$CS = \frac{\sum(A_i \times B_i)}{\sqrt{\sum A_i^2} \times \sqrt{\sum B_i^2}}$$

$$CS = \frac{0.9769}{\sqrt{3.0280} \times \sqrt{2.6500}}$$

$$CS = \frac{0.9769}{1.7401 \times 1.6279}$$

$$CS = \frac{0.9769}{2.8327} = 0.34$$

CC BY SA

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

Tabel 30. Hasil Nilai Cosine Similarity (Sistem Operasi)

Term (Kata Kunci)	Bobot Iduka (Ai)	Bobot Kurikulum (Bi)	AixBi	Ai2	Bi2
crm	0	0.7782	0	0	0.6056
erp	0	0.7782	0	0	0.6056
sap	0	0.7782	0	0	0.6056
odoo	0.7782	0	0	0.6056	0
tableu	0.7782	0	0	0.6056	0
agile	0.7782	0.4771	0	0.6056	0.2276
scrum	0.7782	0.7782	0.6056	0.6056	0.6056
jira	0.7782	0	0	0.6056	0
Total			0.9769	3.0280	2.6500

10. Menghitung Nilai *Cosine Similarity* (Pemrograman Aplikasi Bergerak)

$$CS = \frac{\sum(A_i \times B_i)}{\sqrt{\sum A_i^2} \times \sqrt{\sum B_i^2}}$$

$$CS = \frac{0.6056}{\sqrt{1.6664} \times \sqrt{2.4224}}$$

$$CS = \frac{0.6056}{1.2909 \times 1.5564}$$

$$CS = \frac{0.6056}{2.0092} = 0.30$$

Tabel 31. Hasil Nilai Cosine Similarity (Pemrograman Aplikasi Bergerak)

Term (Kata Kunci)	Bobot Iduka (Ai)	Bobot Kurikulum (Bi)	AixBi	Ai2	Bi2
swift	0.7782	0	0	0.6056	0
kotlin	0.4771	0	0	0.2276	0
flutter	0.7782	0.7782	0.6056	0.6056	0.6056
dart	0	0.7782	0	0	0.6056
java	0.4771	0	0	0.2276	0
json decoding	0	0.7782	0	0	0.6056
sqlite	0	0.7782	0	0	0.6056

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

Total	0.6056	1.6664	2.4224

11. Menghitung Nilai *Cosine Similarity* (Analisis Dan Perancangan Sistem Informasi)

$$CS = \frac{\sum(A_i \times B_i)}{\sqrt{\sum A_i^2} \times \sqrt{\sum B_i^2}}$$

$$CS = \frac{0.9769}{\sqrt{2.4224} \times \sqrt{4.6944}}$$

$$CS = \frac{0.9769}{1.5564 \times 2.1667}$$

$$CS = \frac{0.9769}{3.3722} = 0.29$$

Tabel 32. Nilai Cosine Similarity (Analisis Dan Perancangan Sistem Informasi)

Term (Kata Kunci)	Bobot Iduka (Ai)	Bobot Kurikulum (Bi)	AixBi	Ai2	Bi2
Dbms	0	0.7782	0	0	0.6056
Erd	0	0.7782	0	0	0.6056
Modelling	0	0.7782	0	0	0.6056
normalization	0	0.7782	0	0	0.6056
Query	0	0.7782	0	0	0.6056
Crud	0	0.4771	0	0	0.228
Agile	0.7782	0.4771	0.3713	0.6056	0.2276
Scrum	0.7782	0.7782	0.6056	0.6056	0.6056
Jira	0.7782	0	0	0.6056	0
Figma	0.7782	0	0	0.6056	0
rest api	0	0.7782	0	0	0.6056
Total			0.9769	2.4224	4.6944

12. Menghitung Nilai *Cosine Similarity* (Sistem Basis Data)

$$CS = \frac{\sum(A_i \times B_i)}{\sqrt{\sum A_i^2} \times \sqrt{\sum B_i^2}}$$

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

$$CS = \frac{0.6055}{\sqrt{3.8612} \times \sqrt{4.3298}}$$

$$CS = \frac{0.6055}{1.9650 \times 2.0808}$$

$$CS = \frac{0.6055}{4.0888} = 0.15$$

Tabel 33. Nilai Cosine Similarity (Sistem Basis Data)

Term (Kata Kunci)	Bobot Iduka (Ai)	Bobot Kurikulum (Bi)	AixBi	Ai2	Bi2
sql	0.7782	0.3010	0.2342	0.6056	0.0906
mysql	0.4771	0.7782	0.3713	0.2276	0.6056
postgresql	0.7782	0	0	0.6056	0
oracle	0.7782	0	0	0.6056	0
dbms	0	0.7782	0	0	0.6056
erd	0	0.7782	0	0	0.6056
modeling	0	0.7782	0	0	0.6056
normalization	0	0.7782	0	0	0.6056
query	0	0.7782	0	0	0.6056
crud	0	0.7782	0	0	0.6056
mongodb	0.7782	0	0	0.6056	0
graphql	0.7782	0	0	0.6056	0
firabase	0.7782	0	0	0.6056	0
Total			0.6055	3.8612	4.3298

13. Menghitung Nilai *Cosine Similarity* (Audit Sistem Informasi)

$$CS = \frac{\sum(A_i \times B_i)}{\sqrt{\sum A_i^2} \times \sqrt{\sum B_i^2}}$$

$$CS = \frac{0.2342}{\sqrt{1.2112} \times \sqrt{3.1186}}$$

$$CS = \frac{0.2342}{1.1005 \times 3.1186}$$

$$CS = \frac{0.2342}{1.9435} = 0.12$$

CC BY SA

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

Tabel 34. Nilai Cosine Similarity (Audit SIstem Informasi)

Term (Kata Kunci)	Bobot Iduka (Ai)	Bobot Kurikulum (Bi)	AixBi	Ai2	Bi2
iso27001	0	0.7782	0	0	0.6056
cybersecurity	0	0.7782	0	0	0.6056
system security	0	0.7782	0	0	0.6056
zero trust security	0	0.7782	0	0	0.6056
encryption	0	0.7782	0	0	0.6056
sql	0.7782	0.3010	0.2342	0.6056	0.0906
excel	0.7782	0	0	0.6056	0
Total			0.2342	1.2112	3.1186

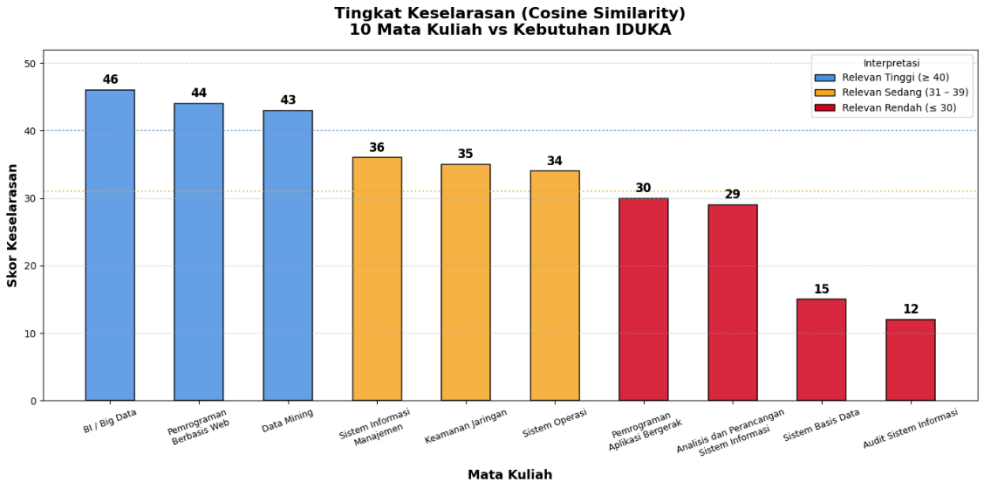
Berdasarkan hasil perhitungan Cosine Similarity, mata kuliah Big Data memperoleh nilai 46 dan termasuk dalam kategori relevan tinggi. Hasil ini menunjukkan bahwa kompetensi yang diajarkan memiliki tingkat kesesuaian yang sangat baik dengan kebutuhan IDUKA. Pemrograman Berbasis Web memperoleh nilai 44 dan termasuk dalam kategori relevan tinggi. Nilai tersebut menunjukkan bahwa materi pembelajaran telah sesuai dengan kompetensi yang banyak dibutuhkan oleh industri. Data Mining memperoleh nilai 43 dan termasuk dalam kategori relevan tinggi, yang menunjukkan bahwa kompetensi analisis dan pengolahan data memiliki tingkat keselarasan yang tinggi dengan kebutuhan IDUKA. Sistem Informasi Manajemen memperoleh nilai 36 dan Keamanan Jaringan memperoleh nilai 35. Kedua mata kuliah tersebut termasuk dalam kategori relevan sedang, sehingga telah memenuhi sebagian kebutuhan kompetensi industri, namun masih memerlukan pengembangan. Sistem Operasi memperoleh nilai 34 dan juga termasuk dalam kategori relevan sedang, yang menunjukkan bahwa materi pembelajaran cukup sesuai dengan kebutuhan IDUKA. Sementara itu, Pemrograman Aplikasi Bergerak memperoleh nilai 30, Analisis dan Perancangan Sistem Informasi memperoleh nilai 29, Sistem Basis Data memperoleh nilai 15, dan Audit Sistem Informasi memperoleh nilai 12. Keempat mata kuliah tersebut termasuk dalam kategori relevan rendah, sehingga diperlukan evaluasi dan pembaruan materi agar lebih selaras dengan kebutuhan industri. Secara keseluruhan, hasil penelitian menunjukkan bahwa Big Data, Pemrograman Berbasis Web, dan Data Mining memiliki tingkat keselarasan yang tinggi dengan kebutuhan IDUKA. Sistem Informasi Manajemen, Keamanan Jaringan, dan Sistem Operasi berada pada tingkat keselarasan sedang, sedangkan Pemrograman Aplikasi Bergerak, Analisis dan Perancangan Sistem Informasi, Sistem Basis Data, serta Audit Sistem Informasi memiliki tingkat keselarasan rendah sehingga perlu menjadi prioritas dalam pengembangan kurikulum.

CC BY SA

Accredited SINTA 5
72/E/KPT/2024

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

Gambar hasil cosine similarity antara kurikulum si dan kebutuhan iduka



Berdasarkan hasil evaluasi menggunakan metode Cosine Similarity, mata kuliah Big Data, Pemrograman Berbasis Web, Data Mining, Sistem Informasi Manajemen, dan Keamanan Jaringan memiliki tingkat keselarasan yang tinggi terhadap kebutuhan Industri, Dunia Usaha, dan Dunia Kerja (IDUKA). Hal ini menunjukkan bahwa materi yang diajarkan pada mata kuliah tersebut telah mencerminkan kompetensi yang banyak dibutuhkan dalam lowongan pekerjaan di bidang teknologi informasi. Sementara itu, mata kuliah Sistem Operasi, Pemrograman Aplikasi Bergerak, dan Analisis dan Perancangan Sistem Informasi berada pada kategori relevansi sedang. Hasil ini menunjukkan bahwa kompetensi yang diberikan masih memiliki keterkaitan dengan kebutuhan IDUKA, namun masih terdapat beberapa aspek yang perlu dikembangkan agar lebih sesuai dengan perkembangan teknologi dan kebutuhan industri saat ini.Adapun mata kuliah Sistem Basis Data dan Audit Sistem Informasi termasuk dalam kategori relevansi rendah, yang menunjukkan bahwa materi atau kompetensi yang terkandung pada kedua mata kuliah tersebut belum banyak muncul atau dibutuhkan berdasarkan data lowongan kerja yang dianalisis. Secara keseluruhan, hasil penelitian menunjukkan bahwa tingkat keselarasan antara kurikulum Program Studi Sistem Informasi dengan kebutuhan IDUKA bervariasi pada setiap mata kuliah. Oleh karena itu, diperlukan evaluasi dan pengembangan kurikulum secara berkala, khususnya pada mata kuliah yang memiliki tingkat relevansi sedang dan rendah. Pembaruan materi pembelajaran, penambahan kompetensi sesuai perkembangan teknologi, serta penyesuaian dengan kebutuhan industri diharapkan dapat meningkatkan keselarasan kurikulum sehingga lulusan memiliki kompetensi yang lebih sesuai dengan tuntutan dunia kerja.

5.1 Kesimpulan

Berdasarkan hasil penelitian mengenai analisis kurikulum Sistem Informasi dan kebutuhan Industri, Dunia Usaha, dan Dunia Kerja (IDUKA) menggunakan teknik *Natural Language Processing* (NLP), dapat ditarik kesimpulan sebagai berikut:

1. Berdasarkan penerapan teknik *Natural Language Processing* (NLP), penelitian berhasil mengidentifikasi berbagai keterampilan yang terdapat pada dokumen kurikulum Program Studi Sistem Informasi melalui tahapan *cleaning*, *case folding*, *stopword removal*, dan *tokenizing*.

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

Keterampilan tersebut berasal dari pokok bahasan dan subpokok bahasan mata kuliah yang kemudian digunakan sebagai representasi kompetensi kurikulum.

2. Hasil analisis terhadap data lowongan kerja yang diperoleh melalui *web scraping* pada platform JobStreet menunjukkan bahwa kebutuhan IDUKA didominasi oleh keterampilan pada bidang Business Intelligence/Big Data, pemrograman web, data mining, manajemen sistem informasi, keamanan jaringan, sistem operasi, pengembangan aplikasi bergerak, analisis sistem, basis data, dan audit sistem informasi.
3. Hasil perhitungan menggunakan metode TF-IDF dan Cosine Similarity menunjukkan bahwa tingkat keselarasan kurikulum dengan kebutuhan IDUKA berbeda pada setiap mata kuliah. BI/Big Data memperoleh skor 46, Pemrograman Berbasis Web 44, dan Data Mining 43 (relevan tinggi). Sistem Informasi Manajemen 36, Keamanan Jaringan 35, dan Sistem Operasi 34 (relevan sedang). Pemrograman Aplikasi Bergerak 30, Analisis dan Perancangan Sistem Informasi 29, Sistem Basis Data 15, dan Audit Sistem Informasi 12 (relevan rendah).
4. Nilai hasil Cosine Similarity menunjukkan bahwa kurikulum Program Studi Sistem Informasi telah memiliki tingkat keselarasan yang cukup baik terhadap kebutuhan IDUKA, terutama pada mata kuliah yang berkaitan dengan Big Data, Pemrograman Berbasis Web, dan Data Mining. Namun, mata kuliah dengan tingkat relevansi rendah masih memerlukan pembaruan materi dan kompetensi agar lebih sesuai dengan perkembangan teknologi dan kebutuhan industri. Dengan demikian, metode TF-IDF dan Cosine Similarity dapat digunakan sebagai indikator objektif dalam mengevaluasi dan mengembangkan kurikulum Program Studi Sistem Informasi.

CC BY SA

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

.

[1] Adila, N. (2022). *Implementasi web scraping untuk pengumpulan data jurnal pada website sinta skripsi*. *7*, 2478–2485.

[2]Ahmednor, M., Suhartono, & Imamudin. (2025). Klasifikasi Keterampilan Kerja Menggunakan Metode Tf-Idf Dan Decision Tree Pada Data Lowongan Kerja Linkedin. *Jurnal Aplikasi Dan Inovasi Ipteks "Soliditas" (J-Solid)*, *8*(1), 52–60. https://doi.org/10.31328/js.v8i1.7152

[3] Aini, N., Arif, M., Toyibah, Z. B., & Agustin, I. T. (2024). Implementasi Metode Hybrid Recommendation untuk Sistem Rekomendasi Mitra MSIB di Prodi Pendidikan Informatika. *Jurnal Ilmiah Edutic : Pendidikan Dan Informatika*, *10*(2), 171–180. https://doi.org/10.21107/edutic.v10i2.23920

[4] Al Rasyid, R., & Ningsih, D. H. U. (2024). Penerapan Algoritma TF-IDF dan Cosine Similarity untuk Query Pencarian Pada Dataset Destinasi Wisata. *Jurnal JTIK (Jurnal Teknologi Informasi Dan Komunikasi)*, *8*(1), 170–178. https://doi.org/10.35870/jtik.v8i1.1416

[5[Azmi, M. (2022). Analisis Tingkat Plagiasi Dokumen Skripsi Dengan Metode Cosine Similarity Dan Pembobotan Tf-Idf. *TEKNIMEDIA: Teknologi Informasi Dan Multimedia*, *2*(2), 90–95. https://doi.org/10.46764/teknimedia.v2i2.51

[6] Guntara, R. G. (2023). Pemanfaatan Google Colab Untuk Aplikasi Pendeteksian Masker Wajah Menggunakan Algoritma Deep Learning YOLOv7. *Jurnal Teknologi Dan Sistem Informasi Bisnis*, *5*(1), 55–60. https://doi.org/https://doi.org/10.47233/jteksis.v5i1.750

[7]Hafid, M., Firmansyah, N., & Budiman, S. N. (2025). *6713-Article Text-23241-1-10-20250130*. *4*, 674–682.

Niken Rahma Diasri, Arini Winur Baeti, & Ary Prabowo. (2025). Pengaruh Penerapan Algoritma Pemrograman Dalam Dunia Pekerjaan (Studi Kasus: Metode Deep Learning). *Jurnal CoSciTech (Computer Science and Information Technology)*, *6*(1), 21–31. https://doi.org/10.37859/coscitech.v6i1.8531

[8] Ningsih, S. R. (2024). Pengaruh Teknologi Terhadap Produktivitas Tenaga Kerja di Indonesia. *Benefit: Journal of Bussiness, Economics, and Finance*, *2*(1), 1–9. https://doi.org/10.70437/benefit.v2i1.341

[9]Noor Latifah, & Mahavita Nabila Syahputri. (2026). Analisis Pola Asosiasi Kompetensi Teknis pada Lowongan Kerja Artificial Intelligence Menggunakan Algoritma Apriori dan Visualisasi Network Graph. *Modem : Jurnal Informatika Dan Sains Teknologi.*, *4*(1), 65–74. https://doi.org/10.62951/modem.v4i1.747

[10] Nurwanda, N., Suarna, N., & Prihartono, W. (2024). Penerapan Nlp (Natural Language Processing) Dalam Analisis Sentimen Pengguna Telegram Di Playstore. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(2), 1841–1846. https://doi.org/10.36040/jati.v8i2.8469

[11] Peng, J., & Deng, C. (2022). Research on the relationship between college students' employability and IT skills training based on mixed research methods. *Frontiers in Psychology*, *13*, 1–14. https://doi.org/10.3389/fpsyg.2022.1054134

[13] Putra, F. P. E., Efendi, R. W., Tamam, A. B., & Pramadi, W. A. (2025). Tren dan Praktik Terbaik dalam Pengembangan Web Berbasis API : Kajian Literatur terhadap Framework Laravel dan React. *Infomatek*, *27*(1), 165–178. https://doi.org/10.23969/infomatek.v27i1.25122

[14] Rianto, J., Kamil, M., Putry, W. A., & Bahri, S. (2022). The Relevance of the Higher Education Curriculum in the Development of the World of Work and the Industrial Sector: The Case of Universitas Islam Syekh-Yusuf. *Society*, *10*(1), 110–125. https://doi.org/10.33019/society.v10i1.408

15] Rio Feriangga Kurniawan. (2022). Implementasi Text Mining Menggunakan Metode Cosine Similarity Untuk Klasifikasi Konten Berita Di Postingan Grup Facebook Info Lantas Dan Kriminal Pasuruan. *JAMI: Jurnal Ahli Muda Indonesia*, *3*(1), 9–17. https://doi.org/10.46510/jami.v3i1.41

[16] Santoso, H., & Rochadiani, T. H. (2022). Pelatihan Machine Learning Menggunakan Bahasa Pemrograman Python Bagi Karyawan PT. Yokogawa Indonesia. *Jurnal ABDINUS : Jurnal Pengabdian Nusantara*, *6*(2), 349–356. https://doi.org/10.29407/ja.v6i2.16018

[17] Setyawan, A., Caesario Septiadi, M., & Wicaksono, S. R. (2023). Perancangan Sistem Informasi Lowongan Pekerjaan dengan Pemetaan Dataset Jobstreet Indonesia. *Jurnal Teknosains Kodepena*, *03*(02), 12–19.

[18] Sihombing, D. O. (2022). Implementasi Natural Language Processing (NLP) dan Algoritma Cosine Similarity dalam Penilaian Ujian Esai Otomatis. *Jurnal Sistem Komputer Dan Informatika (JSON)*, *4*(2), 396. https://doi.org/10.30865/json.v4i2.5374

[19] Ula, K. R., & Faisal, M. (2025). Analisis Kesesuaian Lowongan Pekerjaan dan Kemampuan Pencari Kerja dengan Menggunakan Metode Cosine Similarity. *Techno.Com*, *24*(3), 617–631. https://doi.org/10.62411/tc.v24i3.12852

[20] Yahya, A. E., Yafooz, W. M. S., & Gharbi, A. (2024). Mapping Graduate Skills to Market Demands: A Holistic Examination of Curriculum Development and Employment Trends. *Engineering, Technology and Applied Science Research*, *14*(4), 14793–14800. https://doi.org/10.48084/etasr.7454