

SISTEM CADANGAN HOTEL BERASASKAN PENAPISAN KANDUNGAN

NUR KHAIRUNNADIA BINTI KHARUDIN

SABRINA TIUN

*Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia, 43600 UKM Bangi,
Selangor Darul Ehsan, Malaysia*

ABSTRAK

Setelah Malaysia pulih dilanda Covid-19, pelancong dari dalam dan luar negara semakin meningkat. Fakta ini dapat dilihat daripada Tourism Malaysia 2021 yang merekodkan sejumlah 3036117 penginap hotel di Malaysia pada suku ketiga manakala 11572476 penginap pada suku keempat. Mencari tempat penginapan yang sesuai merupakan sesuatu perkara yang penting bagi pelancong agar mereka dapat menikmati percutian mereka dengan gembira. Hal ini demikian kerana mereka boleh menjimatkan kos dan masa agar mereka dapat menikmati percutian dengan seronok. Perancangan dapat dilakukan dengan mudah jika mereka mempunyai banyak maklumat berkaitan hotel. Mengumpul maklumat berkaitan dengan tempat penginapan yang bersesuaian memerlukan banyak masa. Mereka perlu membandingkan harga, kemudahan and jenis bilik yang terbaik yang sesuai dengan kehendak mereka. Ini boleh dilihat apabila sesetengah pelancong lebih gemar dengan hotel yang lebih murah berbanding permandangan yang cantik dari bilik hotelnya dan sesetengah memilih yang sebaliknya. Semua ulasan and maklumat terhadap tempat tersebut akan mengambil masa yang lama untuk dibaca dengan teliti. Sebagai penyelesaiannya, sistem pencadang dapat membantu pengguna untuk membandingkan maklumat dengan mudah. Pelancong boleh membandingkan mengikut harga, penilaian daripada pelancong yang telah menggunakan perkhidmatan tersebut. Data diambil daripada laman sesawang Booking.com dengan menggunakan bahasa Python dan perpustakaan Selenium dan seterusnya data perlu diproses. Model sistem cadangan yang digunakan ialah sistem cadangan berasaskan penapisan kandungan. Terdapat dua langkah penting dalam sistem cadangan ini iaitu pengvektoran teks dan pengiraan persamaan. Teknik yang terlibat dalam pengvektoran teks adalah kekerapan istilah-kekerapan dokumen sonsang (TF-IDF) dan bag kata manakala untuk pengiraan persamaan adalah menggunakan teknik persamaan kosinus. Hasil daripada proses ini dilihat melalui antaramuka yang dihasilkan dengan menggunakan perpustakaan Flask. Sistem pencadang ini akan menggalakkan banyak pelancong untuk melancong di dalam Malaysia. Pelancong dapat melakukan perancangan dalam masa yang singkat dan meneroka perkara baru di tempat yang baru.

Kata kunci: Python, TF-IDF, Flask, beg kata, persamaan kosinus

PENGENALAN

Hotel merupakan salah satu tempat penginapan apabila kita pergi ke suatu tempat yang jauh dari rumah sama ada untuk bercuti atau bekerja di luar kawasan yang mengambil masa beberapa hari. Hotel bukan sahaja tempat untuk kita berteduh tetapi juga tempat yang memberikan keselesaan dengan kemudahan yang disediakan. Pada tahun 2019, dunia digemparkan dengan wabak Covid-19 yang menyebabkan perintah kawalan pergerakan dikuatkuasakan di Malaysia. Selepas wabak itu pulih, ramai orang merancang untuk pergi bercuti atau bekerja di luar kawasan selepas kuarantin dan bekerja dari rumah.

Berdasarkan laporan tahunan daripada Tourism Malaysia pada tahun 2021, sebanyak 281.2% peningkatan jumlah penginap hotel di Malaysia dari tempoh suku ketiga ke tempoh suku keempat. Ini menunjuk ramai pelancong dari dalam dan luar negara yang melancong di Malaysia yang telah menempah hotel di Malaysia. Banyak perkara yang perlu diambil kira semasa meneliti tempat penginapan termasuklah harga, lokasi tempat tinggal dan kemudahan yang disediakan. Tambahan pula, kebanyakan tempat penginapan ditempah secara atas talian dan semua maklumat berada diujung jari. Oleh itu, maklumat boleh diakses dengan mudah dan pemilihan boleh dilakukan. Akan tetapi, terdapat banyak tempat penginapan yang terdapat di sesetengah tempat dan mengambil masa yang lama untuk membaca satu persatu keterangan yang disediakan. Oleh itu, sistem cadangan dapat membantu dalam menyelesaikan masalah ini.

Sistem cadangan merupakan sistem yang meramalkan, mengecilkan skop pilihan untuk orang mencari sesuatu yang mereka mahu mengikut citarasa mereka. Cadangan yang diberikan adalah mengikut pelbagai kriteria termasuklah citarasa pengguna, persamaan genre dan beberapa faktor lain. Sistem cadangan ini memberikan manfaat kepada semua bidang termasuklah pelancongan seperti yang terdapat di laman sesawang Booking.com. Sistem cadangan yang terdapat di laman sesawang itu memberikan cadangan hotel-hotel yang hampir sama dengan citarasa pengguna.

Dalam projek ini, data yang digunakan adalah data hotel yang diambil daripada laman sesawang Booking.Com. Data hotel yang diambil adalah terhadap kepada hotel yang berada di Malaysia. Hal ini demikian kerana, dapat membantu pelancong yang berkunjung ke tempat yang berada di Malaysia. Pada fasa pertama, data diekstrak daripada halaman sesawang Booking.com. Pada fasa transformasi, data dibersihkan dan nilai data yang hilang dikendalikan. Seterusnya, data dimuat ke dalam sasaran pangkalan data iaitu dalam bentuk fail Comma-Separated Value (CSV). Selain itu, data profil pengguna juga digunakan. Data tersebut diambil semasa pengguna mengisi maklumat mereka semasa mendaftar akaun dan dimuatkan ke dalam fail CSV.

Sistem cadangan memberikan cadangan hotel dengan mengira persamaan antara hotel tersebut. Sistem yang digunakan adalah penapisan berasaskan kandungan atau Content-based Recommendation System. Dua elemen penting dalam model sistem cadangan iaitu pengvektoran teks dan pengiraan persamaan. Semasa pengvektoran teks, model mencari perkataan yang sama yang terdapat pada keterangan hotel tersebut manakala semasa pengiraan

persamaan pula adalah untuk mengira persamaan yang dalam teks tersebut. Cadangan yang dihasilkan daripada pergiraan persamaan tersebut secara susunan menurun. Kemudian, sistem diuji dengan menggunakan data maklum balas penilaian pengguna dan cadangan yang dihasilkan. Akhir sekali, hasil cadangan yang dibina ini dipaparkan melalui antaramuka di laman sesawang.

METODOLOGI KAJIAN

Metodologi yang digunakan dalam pembangunan projek ini ialah Kitaran Hayat Pembelajaran Mesin atau Machine Learning Life Cycle. Kitaran hayat ini dipilih kerana fokus utama projek ini iaitu membina sistem cadangan hotel yang lebih tepat dengan menitikberatkan pra-pemprosesan data dan ketepatan model. Tambahan pula, model yang digunakan merupakan salah satu model pembelajaran mesin. Oleh itu, Kitaran Hayat Pembelajaran Mesin sangat sesuai, mudah dan tepat untuk digunakan dalam projek ini.

Fasa penyediaan data

Fasa yang pertama adalah penyediaan data. Dalam projek ini, Terdapat dua dataset yang diperlukan untuk membina model sistem cadangan ini iaitu data hotel dan profil pengguna. Data hotel diekstrak daripada laman sesawang Booking.com dengan menggunakan perpustakaan Selenium dengan bantuan WebDriver. Bahasa yang digunakan adalah Python dan digunakan di persekitaran Visual Studio Code. Selain itu, data profil pengguna diambil ketika pengguna mendaftar akaun dan disimpan ke dalam fail CSV.

Fasa pemprosesan data

Fasa pemprosesan data adalah fasa yang penting untuk memastikan cadangan yang dikeluarkan adalah tepat. Pra-pemprosesan data adalah untuk memastikan data yang digunakan berkualiti dan bebas dari tanda yang tidak perlu. Sebelum melakukan pra-pemprosesan, mengenalpasti dataset yang digunakan adalah amat penting. Setelah itu, antara pra-pemprosesan yang dilakukan adalah mengendalikan data yang hilang dan bertindih, menukarkan semua huruf kepada huruf kecil, membuang tanda baca dan membuang kata henti.

Fasa melatih model

Fasa ini mempunyai dua bahagian iaitu pengvektoran teks dan pengiraan persamaan. Terdapat dua pengvektoran teks yang digunakan iaitu kekerapan istilah-kekerapan dokumen sonsang (TF-IDF) dan beg kata. Tujuan pengvektoran teks adalah untuk mencari kata kunci pada sorotan dan kemudahan pada dataset hotel dan profil pengguna.

Seterusnya, teknik pengiraan persamaan yang digunakan dalam projek ini adalah Persamaan Kosinus. Persamaan ini digunakan untuk mengukur kesamaan antara data hotel dan data profil pengguna selepas pengvektoran teks dilakukan. Cadangan dikeluarkan berdasarkan tahap kesamaan antara data hotel dan profil pengguna tersebut. Hanya 10 hotel dengan nilai persamaan tertinggi akan dipilih untuk disarankan kepada pengguna.

Fasa menguji model

Fasa penilaian menilai ketepatan dua pengvektoran teks, iaitu TF-IDF dan beg kata. Dalam fasa ini, sistem cadangan dinilai menggunakan metrik ketepatan (precision) dan skor F1. Penilaian dilakukan dengan membandingkan cadangan yang dihasilkan dengan data maklum balas.

Data maklum balas diperoleh apabila antaramuka sistem cadangan memaparkan cadangan kepada pengguna. Setiap hotel yang dicadangkan mempunyai butang suka atau benci agar pengguna boleh memilih sama ada mereka suka atau tidak pada cadangan tersebut. Data maklum balas disimpan dalam fail CSV yang digunakan semasa analisis penilaian. Dataset tersebut mengandungi id pengguna, nama hotel, dan tindakan sama ada pengguna suka atau tidak pada cadangan yang dipaparkan.

Pengiraan metrik penilaian adalah berdasarkan matriks kekeliruan dengan jumlah cadangan yang diberikan oleh sistem dan jumlah hotel yang disukai dan tidak disukai oleh pengguna. Matriks termasuk Positif Benar (TP), yang merupakan ramalan yang betul bagi kejadian positif; Negatif Benar (TN), ramalan yang betul bagi kejadian negatif; Positif Palsu (FP), ramalan kejadian positif yang salah (ralat Jenis I); dan Negatif Palsu (FN), ramalan kejadian negatif yang salah (ralat Jenis II). Ketepatan dan skor F1 dikira menggunakan formula yang diberikan.

Jadual 1 Matriks Kekeliruan

	Sebenar Positif	Sebenar Negatif
Jangkaan Positif	Positif Benar (TP)	Negatif Benar (TN)
Jangkaan Negatif	Positif Palsu (FP)	Negatif Palsu (FN)

$$Ketepatan = \frac{TP}{TP + FP}$$

$$Peringatan = \frac{TP}{TP + FN}$$

$$SkorF1 = 2 * \frac{Ketepatan * Peringatan}{Ketepatan + Peringatan}$$

Fasa pelaksanaan

Fasa yang terakhir adalah pembinaan antaramuka sistem cadangan tempat pelancongan ini. Antaramuka ini dibina untuk memudahkan interaksi pengguna dan model serta pengeluaran hasil sistem cadangan ini.

KEPUTUSAN DAN PERBINCANGAN

Sistem Cadangan Hotel Berasaskan Penapisan Kandungan telah berjaya dibangunkan dengan

dokumentasi lengkap. Sistem ini dibangun menggunakan Python dan perpustakaan Flask.

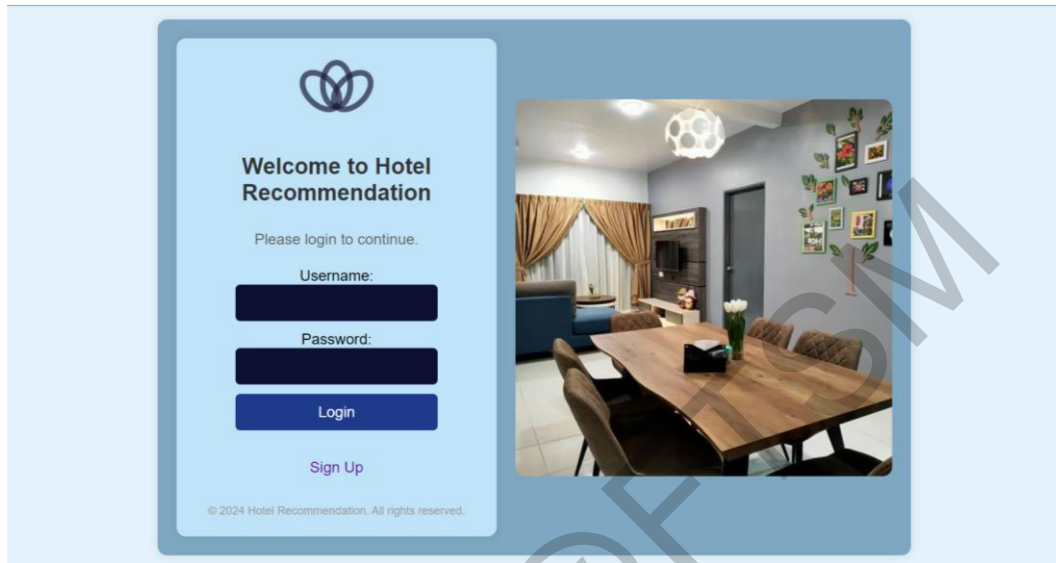
Apabila pengguna memasuki sistem, mereka akan melihat antaramuka log masuk. Jika belum mempunyai akaun, mereka perlu mendaftar dengan menekan butang 'Sign Up'. Pengguna akan dibawa ke antaramuka pendaftaran seperti dalam Rajah 1, di mana mereka perlu memasukkan nama dan mencipta kata laluan, serta mengesahkannya di bahagian 'Confirm Password'.

Rajah 1 Antaramuka Mendaftar Akaun I

Kemudian, pengguna perlu memilih sorotan dan kemudahan yang mereka cari dalam pemilihan hotel. Hal ini supaya cadangan yang dikeluarkan menepati citarasa pengguna. Setelah memilih sorotan dan kemudahan, pengguna boleh menekan butang 'Sign Up' untuk log masuk seperti yang ditunjukkan ditunjukkan di Rajah 2.

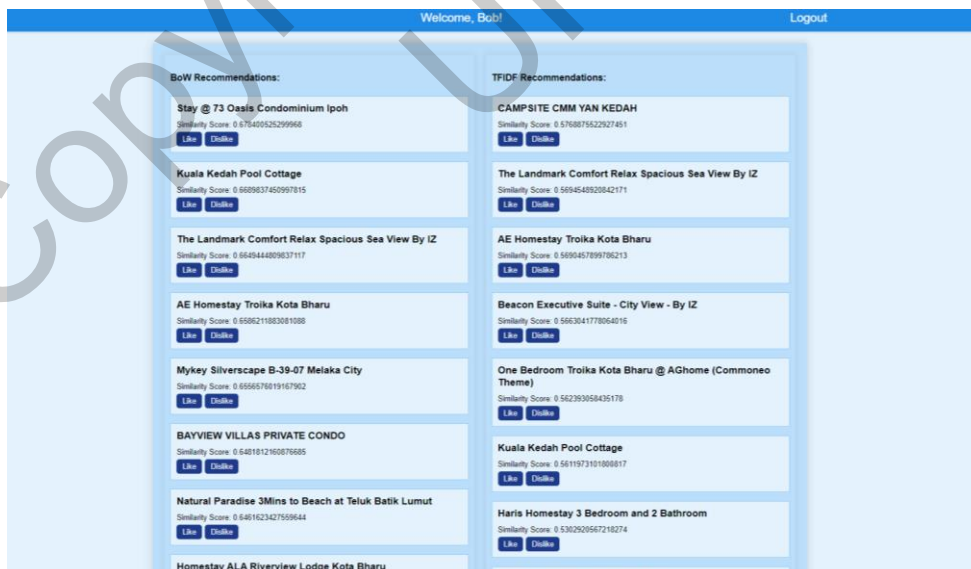
Rajah 2 Antaramuka Mendaftar Akaun II

Seterusnya, Rajah 3 menunjukkan antaramuka untuk pengguna log masuk ke akaun mereka. Untuk log masuk, pengguna memerlukan nama dan kata laluan. Setelah mengisi nama dan kata laluan pengguna boleh menekan butang 'login'.



Rajah 3 Antaramuka Log Masuk

Kemudian, skrin ada pergi ke halaman utama sistem. Terdapat dua cadangan yang dikemukakan di Rajah 4 iaitu cadangan daripada pengvektoran teks TF-IDF dan beg kata. pengguna perlu memberikan maklum balas terhadap cadangan hotel yang diberikan dengan menekan butang 'Like' atau 'Dislike'. Hal ini dapat membantu untuk mengenal pasti pengvektoran teks yang manakah sesuai untuk digunakan dalam sistem ini.



Rajah 4 Antaramuka Halaman Utama

Analisis Penilaian

Analisis penilaian dilakukan dengan cadangan yang dihasilkan oleh sistem cadangan berasaskan penapisan kandungan dengan data maklum balas yang diberikan oleh pengguna.

Terdapat dua analisis penilaian yang dilakukan dalam sistem ini iaitu ketepatan dan skor F1. Untuk menjalankan analisis penilaian, terdapat 20 pengguna yang telah memberikan maklum balas. Sebanyak 200 cadangan telah dikeluarkan. Berikut merupakan pengiraan ketepatan dan skor f1 bagi cadangan yang dikeluarkan.

Jadual 2

	Ketepatan	Skor F1
Beg Kata	0.8447	0.8860
TF-IDF	0.804	0.8919

Daripada formula, ketepatan dan skor F1 untuk pendekatan beg kata ialah 0.8448 dan 0.8860 manakala pendekatan TF-IDF menghasilkan purata skor F1 sebanyak 0.8919 dan purata ketepatan 0.8369. TF-IDF mempunyai skor F1 yang lebih tinggi, menunjukkan keseimbangan yang lebih baik antara ingat dan ketepatan, manakala BoW mempunyai ketepatan yang lebih tinggi sedikit, menunjukkan lebih sedikit positif palsu.

Oleh itu, TF-IDF boleh dianggap lebih baik secara keseluruhan kerana skor F1 yang lebih tinggi menunjukkan bahawa ia lebih seimbang dan berkesan dalam mengendalikan data yang mungkin tidak seimbang, walaupun ketepatan keseluruhannya sedikit lebih rendah daripada BoW.

Cadangan Penambahbaikan

Beberapa langkah penambahbaikan boleh diambil untuk meningkatkan sistem cadangan hotel ini. Pertama, meneroka penggunaan teknik pembelajaran mesin dan pembelajaran mendalam dapat meningkatkan pemahaman tentang citarasa pengguna dan menghasilkan cadangan yang lebih relevan. Kedua, meningkatkan teknik pengurusan data hilang dan pembersihan teks akan memastikan data yang digunakan adalah berkualiti tinggi dan komprehensif. Selain itu, melibatkan pengguna dalam proses penilaian untuk mendapatkan maklum balas langsung tentang kualiti cadangan dan membuat penambahbaikan berdasarkan maklum balas tersebut adalah penting. Akhir sekali, menggunakan sumber pengiraan yang lebih kuat dapat menangani set data yang lebih besar dan kompleks, mengurangkan masa pemrosesan, dan meningkatkan kecekapan model. Implementasi sistem yang lebih baik untuk mengendalikan data yang tidak seimbang dan mengoptimalkan algoritma untuk prestasi yang lebih baik juga dapat dipertimbangkan.

KESIMPULAN

Sistem Cadangan Hotel Berasaskan Penapisan Kandungan ini berjaya dibangunkan menggunakan data yang diperolehi dari Booking.com. Pengvektoran teks yang terbaik untuk sistem cadangan hotel ini adalah teknik kekerapan istilah-kekerapan dokumen sonsang (TF-IDF) dan pengiraan persamaan iaitu persamaan kosinus. Ia menawarkan kemudahan bagi pelancong untuk membandingkan hotel berdasarkan harga, penilaian, dan kemudahan. Sistem ini memudahkan pelancong dalam membuat perancangan percutian mereka dengan lebih cepat dan efisien.

Kekuatan Sistem

Sistem ini mempunyai beberapa kekuatan yang menonjol. Pertama, penggunaan teknik pengvektoran teks yang canggih dan pengiraan persamaan memastikan cadangan yang diberikan adalah relevan dan tepat. Kedua, antaramuka pengguna yang intuitif membolehkan pengguna untuk mendaftar, log masuk, dan mendapatkan cadangan hotel dengan mudah. Ketiga, pengguna boleh memberikan maklum balas terhadap cadangan yang diberikan, yang membantu meningkatkan ketepatan sistem. Keempat, sistem ini membolehkan pelancong untuk memilih hotel berdasarkan pelbagai kriteria seperti harga dan kemudahan, menjadikan perancangan percutian lebih personal dan sesuai dengan keperluan individu.

Kelemahan Sistem

Namun, terdapat beberapa kelemahan dalam sistem ini. Keberkesanan sistem bergantung pada data yang diberikan oleh pengguna semasa pendaftaran. Data yang tidak lengkap atau tidak tepat boleh mengurangkan ketepatan cadangan. Proses pengvektoran teks dan pengiraan persamaan mungkin memerlukan sumber pengiraan yang tinggi, yang boleh menjadi cabaran jika set data besar dan kompleks. Selain itu, kualiti cadangan yang dihasilkan bergantung pada kualiti data yang diekstrak dari Booking.com. Data yang tidak bersih atau tidak lengkap boleh mempengaruhi keputusan cadangan. Akhirnya, sistem ini memerlukan sambungan internet yang stabil untuk berfungsi dengan baik, yang boleh menjadi kekangan bagi pengguna di kawasan dengan akses internet yang terhad.

PENGHARGAAN

Penulis kajian ini ingin ucapkan setinggi-tinggi penghargaan dan jutaan terima kasih kepada Prof. Madya Dr. Sabrina Tiun, penyelia penulis kajian ini yang telah memberi tunjuk ajar serta bimbingan untuk menyiapkan projek ini dengan jayanya.

Penulis kajian ini juga ingin mengucapkan terima kasih kepada semua pihak yang membantu secara langsung mahupun tidak langsung dalam menyempurnakan projek ini. Segala bantuan yang telah dihulurkan amatlah dihargai kerana tanpa bantuan mereka, projek ini tidak dapat dilaksanakan dengan baik. Semoga tuhan merahmati dan memberikan balasan yang terbaik.

RUJUKAN

- Chen, O. 2023. Scraping 'Booking.com' with Selenium & Formulating Pricing Strategies. *Medium*. Tersedia daripada: <https://medium.com/@chenatq/the-process-of-scraping-booking-com-with-python-selenium-57128ecfaf09>
- Dhanashree. 2024. Web Scraping with Selenium and Python in 2024: Examples & Code. *Nanonets Intelligent Automation, and Business Process AI Blog*. Tersedia daripada: <https://nanonets.com/blog/web-scraping-with-selenium/>

- Gaurav, P. 2023. Step by step Content-Based Recommendation System - Prateek Gaurav - Medium. *Medium*. Tersedia daripada: <https://medium.com/@prateekgaurav/step-by-step-content-based-recommendation-system-823bbfd0541c>
- Great Learning Team. 2024b. An Introduction to Bag of Words in NLP using Python | What is BoW?. *Great Learning Blog: Free Resources What Matters to Shape Your Career!*. Tersedia daripada: <https://www.mygreatlearning.com/blog/bag-of-words/>
- Iftikhar, U. 2023. Part 3: Exploring Content-Based Filtering in Recommendation Systems. *Medium*. Tersedia daripada: <https://umair-iftikhar.medium.com/part-3-exploring-content-based-filtering-in-recommendation-systems-836e5e2fe152>
- Kanstrén, T. 2023. A look at precision, recall, and F1-Score. *Towards Data Science*. Tersedia daripada: <https://towardsdatascience.com/a-look-at-precision-recall-and-f1-score-36b5fd0dd3ec> [27 Sep. 2023]
- Kristian Wahyudi, J. L., Ritchie Chandra, Abba Suganda Girsang. 2020. Hotel Content-Based Recommendation System. *Journal of Physics: Conference Series*
- Life cycle of Machine Learning - Javatpoint. (t.t.). *Javatpoint*. Tersedia daripada: <https://www.javatpoint.com/machine-learning-life-cycle>.
- Mboccenti. (t.t.). Recommender_system_project/04. Content-based Course Recommender System Using User Profile and Course Genres.ipynb at main · mboccenti/Recommender_system_project. *GitHub*. Tersedia daripada: https://github.com/mboccenti/Recommender_system_project/blob/main/04.%20Content-based%20Course%20Recommender%20System%20Using%20User%20Profile%20and%20Course%20Genres.ipynb.
- Miftahul Jannat Mokarrama, S. K., Mohammad Shamsul Arefin. 2020. A content-based recommender system for choosing universities. *TURKISH JOURNAL OF ELECTRICAL ENGINEERING & COMPUTER SCIENCES*
- Mokarrama, M. J., Khatun, S., & Arefin, M. S. 2020. A content-based recommender system for choosing universities. *Turkish Journal of Electrical Engineering and Computer Sciences/Elektrik*, 28(4), 2128–2142. Tersedia daripada: <https://doi.org/10.3906/elk-1911-37>
- Pham, K. 2022. What are Recommendation Systems? - Khang Pham - Medium. *Medium*. Tersedia daripada: <https://medium.com/@khang.pham.exxact/what-are-recommendation-systems-6bb5036042db>
- Rakesh, S. (t.t.). Movie recommendation system using content based filtering. *Al-Bahir Journal for Engineering and Pure Sciences*. Tersedia daripada: <https://bjeps.alkafeel.edu.iq/journal/vol4/iss1/7/>
- Sharma, D. 2022. Beautiful Soup vs. Scrapy vs. Selenium: Which Web Scraping Tool Should You Use?. *MUO*. Tersedia daripada: <https://www.makeuseof.com/beautiful-soup-vs-scrapy-vs-selenium/>
- Udofia, E. 2024. WebScraping: BeautifulSoup or Selenium? - etietop Udofia - Medium. *Medium*. Tersedia daripada: <https://medium.com/@udofiaetietop/web-scraping-beautifulsoup-or-selenium-3467edb3c0d9>

Welcome to Flask — Flask Documentation (3.0.X). (t.t.). *Flask*. Tersedia daripada: <https://flask.palletsprojects.com/en/3.0.x/>

What is a Recommendation System? (t.t.). *NVIDIA Data Science Glossary*. Tersedia daripada: [[\(https://www.nvidia.com/en-us/glossary/recommendation-system/#:~:text=A%20recommendation%20system%20\(or%20recommender,exponentially%20growing%20number%20of%20options\)](https://www.nvidia.com/en-us/glossary/recommendation-system/#:~:text=A%20recommendation%20system%20(or%20recommender,exponentially%20growing%20number%20of%20options)))]([https://www.nvidia.com/en-us/glossary/recommendation-system/#:~:text=A%20recommendation%20system%20\(or%20recommender,exponentially%20growing%20number%20of%20options](https://www.nvidia.com/en-us/glossary/recommendation-system/#:~:text=A%20recommendation%20system%20(or%20recommender,exponentially%20growing%20number%20of%20options))

Nur Khairunnadia Binti Kharudin (A188055)

Prof. Madya Dr. Sabrina Tiun

Fakulti Teknologi & Sains Maklumat

Universiti Kebangsaan Malaysia

Copyright@FTSM
UKM