Data Science and Engineering (DSE) Record, Volumn 5, issue 1.

Course Recommendation Model Based on Semantic Search of Job Description

Supawit Ongkariyapong[1] Dussadee Praserttitipong[2] and Wijak Srisujjalertwaja[3]

[1] Master's Degree Program in Data Science, Chiang Mai University, Chiang Mai, Thailand
[2,3] Department of Computer Science, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand
supawit_o@cmu.ac.th
dussadee.p@cmu.ac.th
wijak.s@cmu.ac.th

Abstract. This independent study emphasizes how integrating semantic search and curriculum analysis into course recommendation systems facilitates the alignment of academic education with user's wants. Nowaday, big data has become increasingly prominent in today's world which big data still increasing until becomes the massive volume. Consequently, it's becoming more difficult to accurately deliver the data that meet with user preferences. Therefore, implementing recommendation systems to filter data before delivering it to users can assist in meeting their needs effectively, through advanced natural language processing and semantic analysis techniques. This independent study has objective to enhance the recommendation system based on semantic search over traditional search. Moreover, users are navigated by course recommendation based on semantic search with better decision making.

Keywords: Semantic Search, Natural Language Processing (NLP), Course Recommendation, Job Description

1 Introduction

Big data has become increasingly prominent in today's world which big data still increasing until becomes the massive volume. One of the challenges that many individuals associate with big data is the difficulty in finding specific information within these vast datasets. Moreover, recommendation system will deliver and suggest ideas related to things that is relevant to the user's wants [1].

However, traditional recommendation system face limitations in comprehending the difference of terms, expressions, and the relationship between words. The inherent challenge arises from the potential confusion in the meanings of words, where a single word may carry multiple meanings. Conversely, various words may carry identical meanings. This lack of semantic understanding can lead search engines to confusion, resulting in inaccurate recommendations [2][3].

Semantic search gives the recommender systems to understand the context of search queries, allowing them to deliver intelligent and relevant results based on user queries [2].

Semantic search represents a new method of understanding and retrieving information that transcends the limitations of traditional recommendation systems [4].

This study aimed to develop a course recommendation model based on semantic search of job description which has a purpose to find the advantages of semantic search over traditional search, particularly within the critical domain of higher education where accuracy and context significantly impact information retrieval's effectiveness.

2 Literature Review

2.1 Natural Language Processing (NLP)

Natural Language Processing (NLP) exists as a subfield within artificial intelligence, centering on the computational interpretation of linguistics. This domain encompasses various aspects of understanding textual and audio data through the integration of statistically behaving machine learning methods [5]. Its focus lies in developing computational models aimed at resolving human interaction and understanding human language [6].

2.2 Traditional Search

Traditional search interprets queries by analyzing keywords or short strings of words. It is the traditional way of searching that relies on exact keyword matches and can often lead to irrelevant search results [7][8]. In traditional searching it is not considering about the different meanings the words can have infarct it will show all the matches possible [9].

2.3 Semantic Search

Semantic search will consider to relationship and meaning between words which is not only considered the key word. Furthermore, it makes the course recommender has more significant search results by assessing and understanding the search phrase [10].

2.4 Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF serves as a versatile tool in natural language processing, aiding in tasks that require the extraction of meaningful information from textual data. This statistical model calculates significance through two metrics: the TF-matrix, a two-dimensional matrix representing word frequency in documents, and the IDF, a one-dimensional matrix reflecting word rarity across the corpus [11].

$$tf_i = \frac{n_i}{\sum_k n_k} \quad (1) \qquad idf_i = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|} \quad (2)$$

TF, as shown in (1), represents the frequency of a term (word) within a document or a corpus. It indicates how many times a term occurs in a specific document relative to the total number of terms in that document. IDF is the measure of the importance of the term in the corpus as a whole. As defined in (2), IDF consists in measures the importance of a term within a corpus by calculating the logarithm of the inverse of the proportion of documents in the corpus that contain the term. [12].

2.5 Bidirectional Encoder Representations from Transformers (BERT)

BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. BERT capture rich contextual information for each token in a sequence, enhancing its ability to understand and generate language representations. As a result, with just one additional output layer BERT can be fine-tuned on a wide range of NLP tasks to use with the specific tasks [13].

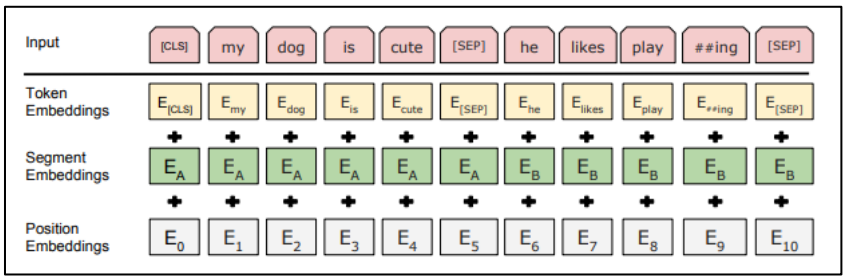

Figure 1. Bidirectional Encoder Representations from Transformers

BERT input representation is the sum of 3 parts as in Figure 1.

1) Token embeddings: Represent each word or token in a sequence as a numerical vector. It suggests that each word in the text has been assigned a unique numerical identifier, known as a token ID. It also contains two special tokens [CLS] at the start of the sequence and [SEP] at the end of the sequence.

2) Segment embeddings: This represents the segment or sentence embeddings. Each segment has its own embeddings separated by [SEP]. By providing separate embeddings for each segment and using [SEP] tokens to mark segment boundaries, these embeddings enable the model to capture the relationships between different parts of the text and perform tasks that require understanding across multiple segments.

3) Position embeddings: Play a crucial role in enabling sequence models to understand the order of tokens in a sequence. By encoding positional information, they help the model capture sequential relationships and dependencies

In practice, input embeddings also contain input/attention masks used to differentiate between actual tokens and padded tokens [13].

2.6 Distil BERT

A scaled-down, general-purpose version of BERT has been developed, boasting a 40% reduction in size and a 60% increase in speed, while still maintaining 97% of the original language understanding capabilities. This model, known as Distil BERT, introduces a method for pre-training a smaller yet efficient language representation model [14].

2.7 ROBERTa

This optimized BERT model implements a dynamic mask strategy, generating a masking pattern for each sequence fed into the model. This differs from BERT, where masking was performed once during data pre-processing, resulting in a single static mask [15].

2.8 Cosine-Similarity

The cosine similarity algorithm measure similarity by calculating the cosine angle between two vectors of vector A and vector B, these vectors can represent different types of data, including words, sentences, paragraphs, or entire documents as shown in (3) [16]. Cosine similarity will be used for this paper to find the similarity between job description and course description.

$$\cos\theta = \frac{A.B}{\|A\| \, \|B\|} \quad (3)$$

2.9 Accuracy

Model accuracy is calculated by dividing the total number of correct predictions made by the model by the overall number of predictions which shown in (4) [17].

$$\text{Accuracy} = \frac{\text{Correct predictions}}{\text{All predicitons}} \quad (4)$$

2.10 Research Related to Recommendation System

Zahra Abbasi-Mouda, Hamed Vahdat-Nejada, and Javad Sadri [18] introduce a context aware tourism recommendation system. It operates by extracting user preferences through semantic clustering and sentiment analysis of user's feedbacks. Through sentiment clustering, prevalent concepts in user reviews are pinpointed, refined further by sentiment analysis to discern preferences from less favored items. Additionally, the system extracts attraction features from aggregated user reviews, enabling the ranking of nearby attractions based on their similarity to user preferences. The result shown in precision, recall, and f-measure of one (Top1), three (Top3), and five (Top5) recommendations of tourist attractions.

Francisco García-Sáncheza, Ricardo Colomo-Palaciosb, and Rafael Valencia-Garcíaa [19] present social-semantic recommender system for advertisements which they provide advertisements to users who have similar preferences. They also use ontology on user's profile to find excerpt of the interest ontology ex. Arts, Business, Computer, etc. TF-IDF and wordnet are used to generate vector from user profile, which using User profile vector update. Each time the user makes a comment on the site or the user clicks onto an ad. the system will refreshes the information in the user's vector. For the result, they use 15 interconnected users were asked to check 150 advertisements divided by 10 ads with content related to each of the 15 general categories considered in the Interest Ontology. Then, show in precision, recall, and f-measure of the measuring of total relevant ads., total recommended ads. and correctly recommended ads.

Harsh Khatter et al. [20] developed a model with the objective of utilizing a movie dataset sourced from Kaggle and Wikipedia. They combine all the data features then transformed them into a similarity matrix. Following this transformation, they applied cosine similarity to determine the rank of similarity between movies. The model's recommendation was the movie with the highest cosine similarity score which compared to a selected movie.

Raghav Mehta and Shikha Gupta [21] represent hybrid recommender system model which is proposed to improve the result by using sentiment analysis and cosine similarity score. Data contained of movie ratings of Top 5000 movies from year 2008 to 2017 which is available on Kaggle and other is user tweets from MovieTweetings database. This model operates on two recommendation approaches: content-based and Collaborative Filtering. It uses TF-IDF to transform the data into vectors and applies cosine similarity to measure the similarity between movies. Additionally, the model incorporates user ratings as part of its recommendation process. Ultimately, the model provides recommendations for movies that are closely related to the user's selected movie.

3 Research Methodology

3.1 Research framework

The model uses both traditional search and semantic search, as depicted in Figure 2. The conceptual framework will be including data collection, data processing, convert data, ranking data, and evaluation.

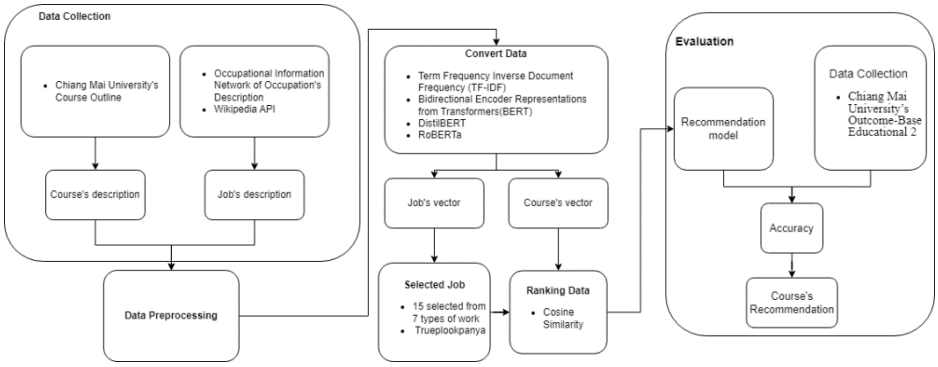

Figure 2. Conceptual Framework

3.2 Data

- Chiang Mai University's Course Outline
  - The model use course's description from this data set. The course recommendation model at Chiang Mai University will rely on a dataset which containing 29,029 courses and six key features.
- Occupational Information Network of Occupation's Description
  - Researchers select the data for the course recommendation model. This dataset consists of 1,017 jobs with 2 features which are title and description. Researchers use this dataset for course's description [22].
- Wikipedia API
  - Researchers connect course's description with Wikipedia API. For Wikipedia API, researchers will use only first paragraph of the searching job because the others are not related to their description [23].
  - For job's description, the researcher combines both descriptions from 2 datasets Wikipedia API and Occupational Information Network of Occupation's description by 85% setting score of SequenceMatcher.

- Chiang Mai University's Outcome-Base Educational 2
  - Chiang Mai University's Outcome-Base Educational 2 contains 1,214 rows with 3 features, which are faculty's name, faculty's major and occupation. Researchers use this dataset for evaluation with the output of course recommendation model.

3.3 Data Preprocessing

The researchers use the Natural Language Toolkit (NLTK) and Regular Expression operations (RE) to clean and preprocess data before imported it to the model. By remove punctuations, remove number, remove word less than 4 string, convert to lower alphabets, remove stop words, change word into word root form [24][25].

3.4 Convert Data

In this section, researcher covert data into vector by Term Frequency-Inverse Document Frequency (TF-IDF) for traditional Search and Bidirectional Encoder Representations from Transformers (BERT), DistilBERT and RoBERTa for semantic search.

3.5 Selected Job

Researchers select the sample of jobs and job's knowledge to perform in course's recommendation model by selected interesting career paths that the labor market needs under 15 jobs from 7 type of career paths in Trueplookpanya [26].

3.6 Ranking Data

After convert selected course dataset and job dataset into vectors. Then, researchers ranking the vector between course's vector and job's vector by cosine similarity. Cosine similarity will show the result that relate between course's vector and job's vector with 30 courses that relate with selected job.

3.7 Evaluation

Model evaluation will use accuracy to compare the result of course's recommendation model with Chiang Mai University's Outcome-Base Educational 2 dataset by count the number of faculty's major of course's recommendation then set the minimum threshold of faculty's major for comparing with Chiang Mai University's Occupation dataset. The models also show the courses that relate with the recommended faculty's majors.

For the threshold of faculty's major is setting by:

- if counts_df['count'].max() >= 6: filter the count's number of faculty's major is greater than or equal to 6.
- elif counts_df['count'].max() >= 4: filter the count's number of faculty's major is greater than or equal to 4.
- elif counts_df['count'].max() >= 3: filter the count's number of faculty's major is greater than or equal to 3.
- else: if the count's number of faculty's major is less than 3, shown the empty dataframe.

The researcher fine-tuned the numbers until get the number of 6, 4, and 3 as the model's threshold because testing other values resulted in lower accuracy. Moreover, these numbers are adjusted to be suitable for a limit of 30 courses in process 3.5 (Ranking data). The researcher attempted to use different numbers apart from the initial settings, start with 8 and 2. Example of models, accuracy of TFIDF decreased from 66.67% to 58.59% and accuracy of DistilBERT decrease from 73.33% to 67.22%.

When comparing the result of the model with Chiang Mai University's Outcome-Based Education 2, the result is 0%. This could indicate that the model either recommended the wrong major or did not recommend any major at all, possibly because the minimum threshold for the number of majors was not met.

4 Result

We compare the results of all models by accuracy score between traditional search and semantic search with course dataset and job dataset. The result will show the score and recommended course of selected jobs as following.

4.1 Term Frequency-Inverse Document Frequency (TF-IDF)

The performance of TF-IDF shown the result of model compare with Chiang Mai University's Outcome-Base Educational 2 which shown as the highest accuracy score is medical and health career's type which is 100% accuracy score. In contrast, transportation and logistics work career's type has the lowest score, which is 0%. The average accuracy of TF-IDF model is 62.22 % as shown in Table 1.

Table 1. Accuracy of Term Frequency-Inverse Document Frequency (TF-IDF)

Type of Career	Job/Knowledge	Accuracy (%)
IT Technology	Database Systems	33.34
	Game Programmer	0
Engineering	Electrical Engineer	100
	Construction Engineer	50
	Machine Design Engineer	50
Medical and Health	Physician	100
	Nursing	100
	Pharmacy	100
Marketing	Digital Marketing	50
	Marketing	50
Accounting and Finance	Accountant	100
	Auditor	50
Transportation and Logistics	Logistics Engineer	0
Production	Production Control Engineer	50
	Production Controller and Testing of Packaging	100
Average Accuracy		62.22

4.2 Bidirectional Encoder Representations from Transformers (BERT)

The performance of BERT shown the result of model compare with Chiang Mai University's Outcome-Base Educational 2 which shown as the highest accuracy score is engineering career's type which is 83.33% accuracy score. In contrast, IT technology career's type has the lowest score, which is 16.67%. The average accuracy of BERT model is 61.11% as shown in Table 2.

Table 2. Accuracy of Bidirectional Encoder Representations from Transformers (BERT)

Type of Career	Job/Knowledge	Accuracy (%)
IT Technology	Database Systems	33.34
	Game Programmer	0
Engineering	Electrical Engineer	100
	Construction Engineer	100
	Machine Design Engineer	50
Medical and Health	Physician	0
	Nursing	100
	Pharmacy	100
Marketing	Digital Marketing	50
	Marketing	50
Accounting and Finance	Accountant	100
	Auditor	50
Transportation and Logistics	Logistics Engineer	33.34
Production	Production control Engineer	50
	Production Controller and Testing of Packaging	100
Average Accuracy		61.11

4.3 Distil BERT

The performance of Distil BERT shown the result of model compare with Chiang Mai University's Outcome-Base Educational 2 which shown as the highest accuracy score is medical and health work career's type which is 100%. In contrast, IT technology career's type has the lowest score, which is 50%, marketing career's type which is 50% and transportation and logistics career's type which is 50%. The average accuracy of Distil BERT model is 73.33% as shown in Table 3.

Table 3. Accuracy of Distil BERT

Type of Career	Job/Knowledge	Accuracy (%)
IT Technology	Database Systems	66.67
	Game Programmer	33.34
Engineering	Electrical Engineer	100
	Construction Engineer	100
	Machine Design Engineer	50
Medical and Health	Physician	100
	Nursing	100
	Pharmacy	100
Marketing	Digital Marketing	50
	Marketing	50
Accounting and Finance	Accountant	100
	Auditor	50
Transportation and logistics	Logistics Engineer	50
Production	Production Control Engineer	50
	Production Controller and Testing of Packaging	100
Average Accuracy		73.33

4.4 RoBERTa

The performance of RoBERTa shown the result of model compare with Chiang Mai University's Outcome-Base Educational 2 which shown as the result of model compare with Chiang Mai University's Outcome-Base Educational 2 which shown as the highest accuracy score are engineering career's type which is 66.67% accuracy score medical and health career's type which is 66.67%. In contrast, production career's type has the lowest score, which is 25%. The average accuracy of Distil BERT model is 52.22% as shown in Table 4.

Table 4. Accuracy of RoBERTa

Type of Career	Job/Knowledge	Accuracy (%)
IT Technology	Database Systems	66.67
	Game Programmer	50
Engineering	Electrical Engineer	100
	Construction Engineer	50
	Machine Design Engineer	50
Medical and Health	Physician	0
	Nursing	100
	Pharmacy	100
Marketing	Digital Marketing	33.33
	Marketing	50
Accounting and Finance	Accountant	100
	Auditor	0
Transportation and Logistics	Logistics Engineer	33.33
Production	Production Control Engineer	50
	Production Controller and Testing of Packaging	0
Average Accuracy		52.22

4.5 Example of Course's Recommendation Model

Researcher show nursing job as sample of course's recommendation model of Distil BERT model which shown in table 5 which presents the course recommendation for nursing job, which shows the correct faculty's majors are NGM, NGP and NGC compared with Chiang Mai University's Occupation dataset.

Table 5. Course Recommendations of Nursing in Distil BERT model

Course	Faculty	courseno_en
Medical Nursing Professional Internship	Nursing	NGM
Pediatric Nursing Professional Internship	Nursing	NGP
Community Nursing Professional Internship	Nursing	NGC
Gerontological Nursing Practicum	Nursing	NGM
Medical Nursing Leadership Practicum	Nursing	NGM
Community Nursing Leadership Practicum	Nursing	NGC
Pediatric Nursing Leadership Practicum	Nursing	NGP
Pediatric and Adolescent Nursing Practicum	Nursing	NGP
Primary Medical Care	Nursing	NGC

5 Conclusion

In this independent study, researchers face the challenges of a recommendation system which traditional search has limitations in understanding the difference of relationship between words. So, the result of the recommendation system will not show answers that have different words but have the same meaning. By comparing these approaches, researchers aim to develop a more effective recommendation system to capable of capturing the subtleties of language and providing accurate and comprehensive results to users.

Consequently, researchers selected TF-IDF for traditional search and BERT, Distil Bert, and RoBERTa for semantic search to use with the collecting Chiang Mai University course description data of 29,029 courses. By model recommended course is based on 15 samples of job descriptions and job's knowledge that are generated from the Occupational Information Network of Occupation's Description and Wikipedia. The results show that TF-IDF has the highest average accuracy score of 62.22% in traditional search and Distil BERT has the highest average accuracy score of 73.33% in semantic search shown in table 64. The result will be given courses that are related to the selected job.

In applying, this model is designed to help users recommend courses for a selected job. Users can select only one job at a time. Once a job is selected, the model will recommend courses at Chiang Mai University and related majors within the university.

6 Discussion

Job descriptions cannot be effectively generated solely based on information from the Wikipedia API. The job descriptions must also align with the descriptions provided by the Occupational Information Network (O*NET). However, aligning the descriptions from both sources to achieve the desired length and semantic meaning poses a challenge, as it requires ensuring that both sources contain the same words of the selected job. Hence, the semantic search may yield suboptimal results if the input sentence lacks semantic coherence.

Another factor that contributes to this problem is the insufficient quantity of course descriptions which cannot provide enough information for the semantic model to understand the context effectively. Thus, the course recommendation model may underperform due to the lack of suitable data.

In future work, we plan to collect additional data on job descriptions and course descriptions for both traditional search models and semantic search models which this information improves the model's ability to understand context, particularly in the semantic search model. As the semantic model develops to the next version This expanded data set will contribute to further improved performance.

In the job description section, we face the challenge of integrating data from both the Wikipedia API and the O*Net dataset. To address this, we can use synonyms to help the model recognize different words with the same meaning. This approach may enhance the model's performance in detecting synonymous terms, resulting in improved job descriptions for the recommendation model.

References

1. S. N. Mohanty, J. M. Chatterjee, S. Jain, A. A. Elngar, and P. Gupta, Eds., “Recommender System with Machine Learning and Artificial Intelligence: Practical Tools and Applications in Medical, Agricultural and Other Industries,” 1st ed. Wiley, 2020, doi: 10.1002/9781119711582.
2. A. Malve and P. M. Chawan, “A Comparative Study of Keyword and Semantic based Search Engine,” International Journal of Innovative Research in Science, Engineering and Technology, vol. 4, no. 11, 2015, doi: 10.15680/IJIRSET.2015.0411039.
3. R. R. Zebari, S. R. M. Zeebaree and K. Jacksi, “Impact Analysis of HTTP and SYN Flood DDoS Attacks on Apache 2 and IIS 10.0 Web Servers,” The 2018 International Conference on Advanced Science and Engineering (ICOASE), pp. 156–161, 2018
4. T. Pirouz, “A Beginner’s Guide to Semantic Search,” Lumar, https://www.lumar.io/blog/best-practice/a-beginners-guide-to-semantic-search/, 2023, Nov, 10

5. J. Li, X. Chen, E. Hovy and D. Jurafsky, “Visualizing and understanding neural models in NLP,” Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, ArXiv, vol. abs/1506.01066, 2016
6. M. Singla, “Advances in Computer Science,” AkiNik Publications, vol. 7, pp. 1-25, 2020, doi:10.22271/ed.book.784.
7. V. Caruana, “Semantic search: the next big thing in search engine technology,” Algolia, https://www.algolia.com/blog/product/semantic-search-the-next-big-thing-in-search-engine-technology/, 2023, December, 28
8. E. Ho, T. Tran and H. Wandawa, “AI-powered search vs. conventional search: Which is better?,” Boost Commerce, https://boostcommerce.net/blogs/all/ai-powered-search-vs-conventional-search#:~:text=Conventional%20search%20is%20the%20traditional,lead%20to%20irrelevant%20search%20results., 2023, October, 3
9. R. Rubini and R. M. Chezian, “An Analysis on Search Engines: Techniques and Tools,” International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, no. 9, 2014
10. R. Guha, R. Macook and E. Miller, “Semantic search,” The Web Conference, 2003, doi: 10.1145/775152.775250.
11. A. K. Singh and M. Shashi, “Vectorization of text documents for identifying unifiable news articles,” International Journal of Advanced Computer Science and Applications, vol. 10, no. 7, doi: 10.14569/ijacsa.2019.0100742.
12. M. Chiny, M. Chihab, O. Bencharef and Y. Chihab, “Netflix recommendation system based on TF-IDF and cosine similarity algorithms,” The 2nd International Conference on Big Data, Modelling and Machine Learning, 2021, doi: 10.5220/0010727500003101.
13. J. Devlin, M. W. Chang, K. Lee, K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, ArXiv, vol. abs/1810.04805, pp. 4171-4186, 2019
14. V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” ArXiv, vol. abs/1910.01108, 2020
15. W. Shengq, K. Huaizhen, L. Chao, H. Wanli, Q. Lianyong and W. Hao, “. Service Recommendation with High Accuracy and Diversity,” Wireless Communications and Mobile Computing, 2020
16. L. Yunxiang, X. Qi, and T. Zhang, “Research on Text Classification Method based on PTF-IDF and Cosine Similarity,” Journal of Information and Communication Engineering, vol. 6, no. 1, pp. 335-338, 2020
17. DataRobot, “The value of model accuracy,” DataRobot, https://www.datarobot.com/blog/the-value-of-model-accuracy/#:~:text=Model%20accuracy%20is%20defined%20as,certainly%20not%20the%20only%20way., 2024, March, 27.
18. Z. A. Mouda, H. V. Nejada, and J. Sadri, “Tourism Recommendation system based on semantic clustering and sentiment analysis,” Expert Systems with Applications, vol. 167, 2021, doi: 10.1016/j.eswa.2020.114324.
19. F. G. Sáncheza, R. C. Palaciosb, and R. V. Garcíaa, “A social-semantic recommender system for advertisements,” Information Processing and Management, vol. 57, no. 2, 2020, doi: 10.1016/j.ipm.2019.102153.
20. H. Khatter, N. Goel, N. Gupta and M. Gulati, “Movie recommendation system using cosine similarity with sentiment analysis,” 2021 Third International Conference on Inventive

21. R. Mehta and S. Gupta, “Movie recommendation systems using sentimentanalysis and cosine similarity,” International Journal for Modern Trends in Science and Technology, vol. 7, no. 1, pp. 16-22, 2021, doi: 10.46501/ijmtst0701004.
22. O*NET, “About O*NET,” O*NET, https://www.onetcenter.org/overview.html., 2024, March,27.
23. Wikipedia, “Job Description,” Wikipedia The Free Encyclopedia, https://en.wikipedia.org/wiki/Main_Page, 2024, Mar, 22.
24. E. Loper, S. Bird, and E.Klein , “Natural Language Processing with Python,” 1st ed. O’Reilly Media, 2009
25. Python Software Foundation, “re — Regular expression operations,” Python Software Foundation, https://docs.python.org/3/library/re.html., 2024, March, 27.
26. Trueplookpanya, “Interesting career paths that the labor market needs,” Trueplookpanya, https://www.trueplookpanya.com/knowledge/content/93991., 2024, March, 27.