Semantic Search for Information Retrieval using Large Language Models

August 18, 2023

COMP 4905 - Honours Project

Project Supervisor: Dr. Majid Komeili

Author: Saurabh Gummaraj Kishore

1 Abstract

The goal for this project is to develop a website to facilitate collaboration among AI researchers and problem owners. This project enhances the website's search functionality by implementing advanced Natural Language Processing (NLP) techniques, specifically tokenization methods, semantic and natural language search. The problem statement centers around building an intelligent search solution powered by OpenAI's text embedding model, utilizing NLP techniques to improve query comprehension and outcome relevance. The text embedding solution involves leveraging word embeddings to enhance search results by considering contextual and semantic relationships between words. Evaluation demonstrates that OpenAI's text embedding model outperforms alternative options. Future work includes expanding data sources, refining search algorithms, and exploring additional NLP methods to create a versatile, domain-specific text retrieval engine. This project not only enhances user experience but also paves the way for collaborative AI research across various domains.

2 Acknowledgements

I would like to express my sincere gratitude to my project advisor, Dr. Majid Komeili, for their invaluable guidance and insightful discussions throughout the course of this project. His guidance, knowledge and forward thinking have been an inspiration to me and ensured that I always found innovative solutions, and challenged me to improve my solution at every step.

I would also like to thank the Associate Director Undergrad, Mark Lanthier, and my undergraduate advisors, Edina Storfer and Anum Saeed. Both of whom have gone through great lengths to facilitate my enrollment into the Honours Project course.

I extend my gratitude to the Senior Systems Administrator of School of Computer Science, Andrew Miles, for providing me with the necessary computing resources required to complete this project.

I would like to acknowledge Carleton University and the School of Computer Science for providing me with the opportunity and space to complete this project.

I extend my appreciation to the developers at OpenAI, The Python Software Foundation, HuggingFace, ReactJS, TailWindCSS, Render.com, Flask and Firebase. Their libraries and tools have been used extensively throughout this project and would not be possible to complete without their support.

Lastly, I would like to thank my family, friends, and colleagues who motivated, supported, guided me throughout the duration of this project.

3 Table of Contents

Contents

List of Figures

List of Tables

4 Motivation

Given the surge in research related to artificial intelligence and intelligent systems, effective collaboration among researchers within specific subdomains of artificial intelligence is becoming increasingly crucial. The website is designed to bridge this need by facilitating connections between researchers from various faculties at Carleton University and problem owners, all with the goal of promoting collaboration in research.

The ongoing development of the website employs an innovative web scraping strategy using Beautiful Soup, a Python tool for parsing HTML documents. This technique enables the extraction of data from diverse Carleton faculty pages. The collected information is stored in JavaScript Object Notation (JSON) format within Firebase's real-time database. This dataset encompasses details like researchers' contact information, research interests, publications, and their respective affiliations. The website's front end was built using React, TailWindCSS, and a few DaisyUI components to create an easy-to-use user interface.

With the vision of creating a platform for problem owners to connect with researchers, the website must effectively perform as a comprehensive database with an intelligent search function. This entails robust search and querying capabilities that can interpret user intent and provide relevant outcomes. The initial state of the website employed a rudimentary search algorithm, specifically the exact-match keyword search. The objective of this project was to enhance the effectiveness of the search function. This can be achieved through numerous sophisticated search algorithms such as fuzzy search, page-rank search (Google's primary search algorithm), natural-language search, semantic search, and more. Given the maturity of research and application in natural language processing (NLP), this project will focus on improving the website's search functionality through NLP techniques such as natural-language search and semantic search [Rajput, 2020].

Semantic search transcends the limitations of exact keyword matches by comprehending each search query's context and underlying meaning. Implementing a semantic search algorithm can help identify correlated concepts and synonyms, resulting in more pertinent outcomes aligned with user intent. It heavily relies on word embeddings, a fundamental NLP concept, to grasp queries and deliver similar results that might be phrased differently. Word embeddings represent words as numerical vectors in a multidimensional space, where the angle or distance between vectors signifies the semantic proximity of words [Sieg, 2019].

The project's objective revolves around enhancing the website's existing search capabilities by integrating semantic search and natural language search algorithms. This will significantly enrich user experience and empower problem owners and researchers to more effectively collaborate and advance ethical AI research.

5 Methodology

5.1 System Architecture

A web application can be categorized into two primary functionalities: the front end and the back end. The front end encompasses the visual components that users interact with, while the back end is responsible for managing the website's structure, data, and logic.

The design approach for this project was strongly influenced by the inherent nature of the final web application. A critical requirement was to ensure compatibility with a React front-end application while delivering prompt and precise search results. To achieve this, a backend Application Programming Interface (API) was developed, enabling seamless communication between the front end and back end.

Another influence on the software architecture stemmed from the work of a previous student who created a web scraper. This tool extracted information from diverse faculty web pages, including names, departments, contact details, and research interests. The collected data was then transformed into JSON format and uploaded to Firebase's Realtime Database, accessible via API calls. These calls encompass actions such as creating, reading, updating, or deleting data—collectively referred to as CRUD actions. This client-server interaction entails the client sending requests to the server, which responds with the required data or an error message.

This laid the foundation for developing a solution. With parsed data residing in Firebase's Realtime Database, the ReactJS-based front end, and the need for a versatile backend API, Python emerged as the language of choice. Python's capabilities extended to seamless Firebase communication, a diverse array of data manipulation tools, Machine Learning (ML), Artificial Intelligence (AI), and Natural Language Processing (NLP) libraries. It also offered simple backend server hosting through the Flask web framework.

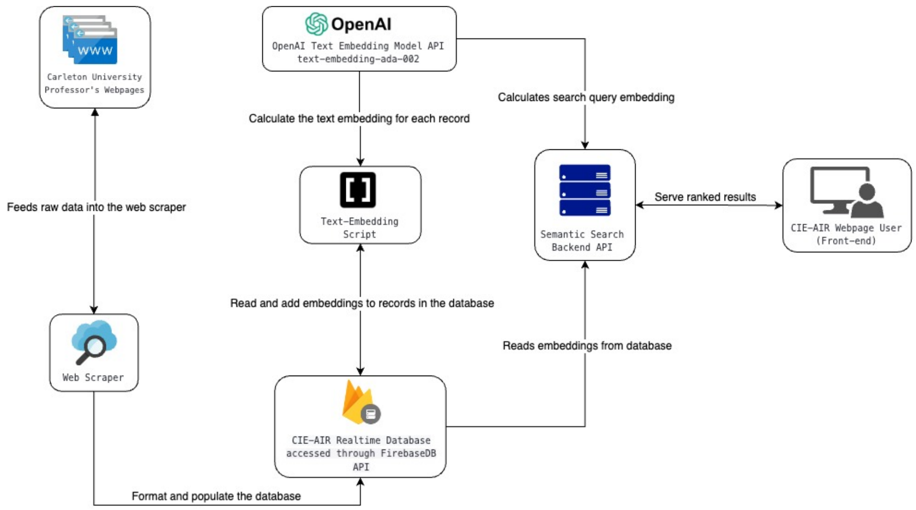

Figure 1: Semantic Search Data Flow and API Diagram

5.1.1 Tokenization

Tokenization, a fundamental process, involves dissecting text into smaller units called tokens. These tokens can be individual words, subwords, or even characters. For instance, consider the sentence "The quick fox jumped over the lazy dog." In this context, each word constitutes a token, and within each word, individual letters serve as tokens. Consequently, text is parsed into paragraphs, paragraphs into sentences, sentences into words, and words into letters. This breakdown facilitates preprocessing, enabling us to focus solely on the pertinent content. The tokenization at the character level assumes particular significance, as it empowers the removal of punctuation and extraneous symbols. Tokenizing at the word level, on the other hand, permits the identification of keywords while excluding insignificant words such as articles (a, the) and prepositions. The division of paragraphs into sentences aids in grasping the holistic narrative by identifying meaningful phrases. As part of preprocessing, both our corpus and search query underwent punctuation and stop word removal prior to executing exact match searches. The ensuing search methods encompassed exact match searches for the entire query, bigram exact match searches, and trigram exact match searches. The latter two techniques fall under the umbrella of tokenization's subcategory known as n-grams [Pai, 2020]

5.2 N-Grams

In natural language processing, n-grams are contiguous sequences of n items, where an ”item” can refer to a word, character, or even a subword. N-grams capture local linguistic patterns and contextual information within a text [Pai, 2020]. They are very useful for identifying the keywords and phrases within a text. The ”n” in n-grams represents the number of items in the sequence. A sentence like ”The quick brown fox jumps over the lazy dog” serves as an illustrative example. Depending on the chosen ”n” value, the following n-grams emerge:

1. Unigrams (1-grams):
   - ”The”, ”quick”, ”brown”, ”fox”, ”jumps”, ”over”, ”the”, ”lazy”, ”dog.”
2. Bigrams (2-grams):
   - ”The quick”, ”quick brown”, ”brown fox”, ”fox jumps”, ”jumps over”, ”over the”, ”the lazy”, ”lazy dog.”
3. Trigrams (3-grams):
   - ”The quick brown”, ”quick brown fox”, ”brown fox jumps”, ”fox jumps over”, ”jumps over the”, ”over the lazy”, ”the lazy dog.”

The utility of n-grams extends across numerous NLP tasks, including information retrieval. A devised algorithm facilitated the culling of bigrams and trigrams from the tokenized search query, which underwent preprocessing involving stop word elimination and punctuation removal. This iterative approach heightened the precision of exact match searches by accounting for partial query matches within the corpus.

5.3 Semantic Search Algorithm and Text Embedding Models

Having established the system architecture overview, the focus shifted to devising an effective search algorithm. Thorough research was conducted to explore potential solutions for refining the search functionality. The primary goal was to enhance the search bar’s ability to comprehend user intent.

This involved utilizing a trained NLP model for query analysis and gaining a comprehensive understanding of the corpus—the data being searched. Once the user's intent was deciphered, the challenge lay in locating relevant information. This was achieved by parsing JSON data to identify exact, partial, and related matches based on the query. Ultimately, the search results need to be presented in a manner that is ranked according to query relevance.

Initially, a plan was formulated to use SpaCy, an open-source NLP Python library, for text embedding. This approach entailed leveraging pre-trained word-vectorization methods to calculate embeddings for the corpus. However, further exploration revealed that SpaCy utilized Word2Vec, a technique dating back to 2013 [Mikolov et al, 2013]. Given the rapid evolution of AI/ML and NLP research in the past decade, this represented a limitation.

Upon deeper investigation, two promising text embedding models emerged: the OpenAI text-embedding-ada-002 model and the HuggingFace instructor-xl model [Su et al, 2022]. The decision-making process on which model to employ was facilitated by the Massive Text Embedding Benchmark (MTEB). The MTEB indicated that the instructor-xl model ranked seventh, while the text-embedding-ada-002 model secured the thirteenth position overall [Muennighoff et al, 2022].

To make a more informed choice, a minimum viable solution was implemented using research interest data for each faculty member from Firebase's Realtime Database as the corpus. Embeddings were calculated and stored locally for each research interest using both models. Fifteen test queries were generated to establish a baseline for evaluation.

The embeddings for each query and corpus, represented as numerical vectors, allowed relevance calculation using the cosine similarity equation. The below image (Figure 2) represents a multidimensional space where two vectors "Hi world" and "Hello world" lie. Using the cosine of the angle between the two vectors, we can interpret the distance between them or in other words, how closely related the two vectors are. This produces a numeric value between 0 and 1, where 1 signifies an

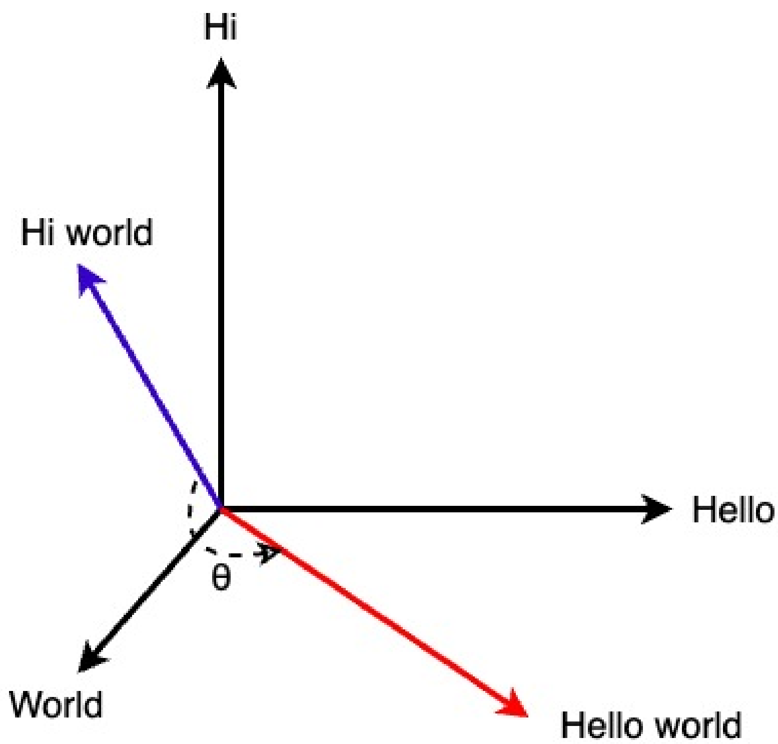

Figure 2: Visual Representation of a Three Dimensional Word Embedding

exact match, and 0 suggests no relevance between the vectors.

Cosine similarity is calculated using the equation:

$$\cos\theta = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

where:

$$A = \text{Vector } A B = \text{Vector } B \|A\| = \text{Magnitude of vector } A \|B\| = \text{Magnitude of vector } B$$

5.4 Ranking Algorithm

Once a model was identified as the superior solution for our domain-specific task, a ranking algorithm was developed to provide the most pertinent information as top suggestions, prior to transmitting our results from the back end to the front end. This was accomplished through a hierarchy of rules: Exact matches were given priority over semantically similar ones. A tri-gram exact match took precedence over semantically similar search results for search phrases containing more than three words. To illustrate, consider the query "NLP for mining software repositories." The search process would begin with an exact match search for the entire query, followed by each bigram: "NLP for","for mining", "mining software", "software repositories"; followed by each trigram: "NLP for mining," "mining for software," and "mining software repositories." After exhausting the exact match search, the outcomes of the semantically similar search were appended to the results. This method ensures that only the most pertinent search results appear at the forefront.

Ranking Hierarchy for the search query "NLP for mining software repositories:

1. Results for an exact match that contain the entire query: "NLP for mining software repositories"
2. Show results for the bigram exact match:
   (a) Results for the first bigram: "NLP for"
   (b) Results for the second bigram: "for mining"
   (c) Results for the third bigram: "mining software"
   (d) Results for the fourth bigram: "software repositories"
3. Show results for the trigram exact match:
   (a) Results for the first trigram: "NLP for mining"
   (b) Results for the second trigram: "for mining software"
   (c) Results for the third trigram: "mining software repositories"
4. Show results for the semantic search

5.5 Continuous Integration and Continuous Deployment (CI/CD) Pipelines

In order to ensure that the website's search functionality works seamlessly, a back end server built using the Flask Python web framework would be hosted and deployed on the cloud-based micro-web service called Render. Render facilitates API hosting, enabling multiple endpoints to communicate with the back-end server through a single URL. On of the advantages of using a micro web service such as Render is that our code base can sit on a local machine while our deployment is over the cloud. This allows us to use version control tools such as git/GitHub to manage the different development, staging and production code without breaking our connection between the back end the front end. Anytime we commit new code to our production branch on GitHub, Render will automatically deploy the new code to the production server. Lastly, in order to update the embeddings on the database, a python script will be run when any of the faculty member's research interest changes. This ensures that the corpus is up to date.

6 Results

6.1 Testing Results

A methodical approach was employed to assess the accuracy of the two distinct models selected. The specifics of this assessment can be found in the methodology section of this report, and the table below (Table 1). For every model, the top five search results were individually categorized as relevant or irrelevant, and assigned a score of 1 or 0 respectively. Each model had the potential to achieve a maximum score of 5 for any given query. A cumulative count was then conducted for both models, and the resulting accuracy percentage was calculated.

Query	**HuggingFace Instructor-xl Model**	**OpenAI text-embedding-ada-002 Model**
Algorithms	5	5
Biomedical Technology	5	4
Computational Theory	4	4
Computer Systems	5	4
Data Integrations and Data Analysis	5	5
Digital and Wireless Communication	5	5
Embedded Systems	4	3
Machine Learning	4	4
Network Security	5	5
Robotics	4	4
Social Media	1	3
Visualization	3	4
Web Development	3	5
Find me someone who can help with networks and software designs	5	5
Find me someone who can help with UI design	4	5
Total Score	**62**	**65**

Table 1: Comparison of HuggingFace Instructor-xl and OpenAI text-embedding-ada-002 Models

The instructor-xl model achieved a score of 62 out of 75, while the text-embedding-ada-002 obtained a score of 65 out of 75. Contrary to the MTEB leaderboard ranking, which positioned HuggingFace's instructor-xl text embedding model higher than OpenAI's text-embedding-ada-002 model, the domain-specific test demonstrated a different outcome.

	HuggingFace Instructor-xl model	**OpenAI text-embedding-ada-002 Model**
Massive Text Embedding Benchmark (MTEB) Accuracy	61.59%	60.99%
Website Testing Accuracy	82.67%	86.67%

Table 2: Comparison of Accuracy between the HuggingFace and OpenAI Text Embedding Models

This demonstrated that OpenAI's text embedding model outperformed the HuggingFace text embedding model with a 4% greater accuracy. This improvement can be primarily attributed to one of the limitations identified during the testing of the instructor-xl model. The instructor-xl model exhibited a tendency to assign higher significance to research interests left blank by a faculty member, leading to a Type I error or a false positive. In contrast, OpenAI's model displayed effective performance by largely disregarding research interests lacking content. This phenomenon was particularly evident in the search query "Social Media" (Refer to Table 13 in Appendix A).

6.2 Limitations and Future Work

The current implementation is by no means flawless. The essence of crafting a search algorithm is rooted in the understanding that attaining perfection is an ongoing pursuit, given the potential for consistently enhancing search result accuracy. Three primary limitations can be discerned: constraints stemming from the data source, constraints stemming from the text embedding model, and constraints arising from the dependencies.

6.2.1 Limitations Arising from the Data Source

Data serves as the principal driving component for solutions implementing AI/ML and NLP. The efficacy of a solution is intricately tied to the quality of the data it operates upon. In this project, one of the primary constraints on achieving an exceptionally efficient solution emanates from the data source itself. If a faculty member fails to update their webpage, users of the website might encounter outcomes that do not accurately reflect their current research interests.

One potential approach to mitigate this challenge involves broadening our web scraping scope. Expanding beyond faculty webpages, we could incorporate data retrieval from recent publications, personal web pages of professors, and Google Scholar profiles. This strategic expansion would enhance the reliability and currency of the data underpinning our solution.

6.2.2 Limitations Arising from the Text Embedding Model

The query "Social Media" yielded results featuring professors who had not specified any research interests on their Carleton SCS faculty web page. This outcome can be attributed to the text embedding model giving greater importance to vacant research interest fields. This highlights a significant concept known as the black-box effect. As developers, when we employ such models, it's imperative to recognize that we may lack comprehensive insight into their internal workings. Consequently, every model carries its distinct limitations that may not always be fully comprehensible.

A plausible approach to address this concern involves leveraging multiple text embedding models. By scrutinizing their behaviors, we can construct a tailored text embedding pipeline that gives us the flexibility to assign greater weight to one model over another. This strategy helps us to navigate the intricacies of these models and mitigate their limitations effectively.

6.2.3 Limitations Arising from the Dependencies

Another limitation stems from the diverse array of tools, software, and libraries employed throughout this project. While everything currently operates smoothly, there remains the possibility that one of the dependencies could encounter issues due to version incompatibility, or some other reason. Furthermore, a distinct limitation originates from our chosen API hosting service, Render. The utilization of Render's free tier entails that our API server may undergo periodic downtime during prolonged periods of inactivity.

A potential remedy for both these constraints involves consistent website maintenance and the consideration of upgrading from the free tier. Such measures would serve to prevent downtime in the backend API server, ensuring uninterrupted functionality and a seamless user experience.

7 Future Work

In light of the progress achieved through this project, several avenues for future work present themselves, some of which are geared towards enhancing our solution, while others extend beyond the project's scope, considering the diverse range of use cases that a generic solution could provide.

To start with, addressing the limitations stemming from the data source and broadening our web scraping scope would significantly enhance our solution, simultaneously simplifying the process for professors and providing more accurate results for users. Additionally, improving the accuracy of search results by employing multiple text embedding models and constructing a tailored hybrid embedding pipeline holds the potential for a superior solution. This endeavor also entails comprehending the behaviors, strengths, and limitations of each text embedding model. To amplify the user experience, the integration of advanced Natural Language Processing techniques like sentiment analysis, knowledge graphs, reinforcement learning, and Named Entity Recognition (NER) could facilitate conversational-level querying.

The problem statement at the core of this project can be distilled into a concise term: "Semantic text retrieval." This implies a wide spectrum of potential use cases for such a solution. The prospect of developing a versatile model that serves as a blueprint for creating custom, domain-specific text retrieval engines may assist a lot of different industries.

8 References

Jatnika, D., Bijaksana, M. A., & Suryani, A. A. (2019). Word2vec model analysis for semantic similarities in english words. Procedia Computer Science, 157, 160-167.

Keita, Z. (2023, June 9). Leveraging text embeddings with the openai API: A practical guide. DataCamp. https://www.datacamp.com/tutorial/introduction-to-text-embeddings-with-the-open-ai-api

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2022). MTEB: Massive text embedding benchmark. arXiv preprint arXiv:2210.07316.

Pai, A. (2020, May 24). What is tokenization in NLP? here's All you need to know. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/05/what-is-tokenization-nlp/

Rajput, A. (2020, September 3). Semantic search engine using NLP. Medium. https://medium.com/analytics-vidhya/semantic-search-engine-using-nlp-cec19e8cfa7e

Reimers, N. (2022, February 9). OpenAI GPT-3 text embeddings - really a new state-of-the-art in dense text embeddings?. Medium.

Sieg, A. (2019, November 13). Text similarities: Estimate the degree of similarity between two texts. Medium. https://medium.com/@adriensieg/text-similarities-da019229c894

Su, H., Kasai, J., Wang, Y., Hu, Y., Ostendorf, M., Yih, W. T., ... & Yu, T. (2022). One embedder, any task: Instruction-finetuned text embeddings. arXiv preprint arXiv:2212.09741.

A Appendix - Detailed Testing Tables

Search Query	HuggingFace instructor-xl Model		OpenAI text-embedding-ada-002 Model	
	Name	**Research Interest**	**Name**	**Research Interest**
Algorithms	Jörg-Rüdiger Sack	Algorithms. Computational Geometry. Data Structures. and GIS	Alina Shaikhet	Algorithms. Computational Geometry. Visualization.
	Anil Maheshwari	Design. Analysis and Implementation of Algorithms for Problems arising in Computational Geometry. Graph Theory. Discrete Mathematics. and Data Science.	Robert Collier	Genetic Algorithms. Computing Education
	Alina Shaikhet	Algorithms. Computational Geometry. Visualization.	Jörg-Rüdiger Sack	Algorithms. Computational Geometry. Data Structures. and GIS
	Robert Collier	Genetic Algorithms. Computing Education	Alexa Sharp	Algorithms. Algorithmic Game Theory. Network and Graph Theory.
	Alexa Sharp	Algorithms. Algorithmic Game Theory. Network and Graph Theory.	Pat Morin	Algorithms and combinatorics of Planar and Near-Planar Graphs. Graph Drawing. Computational Geometry

Table 3: Search Result Comparison - Query: "Algorithms"

Search Query	HuggingFace instructor-xl Model		OpenAI text-embedding-ada-002 Model	
	Name	Research Interest	Name	Research Interest
Biomedical Technology	Dwight Deugo	Eclipse. Beacon Technology. Health informatics and Systems	Adrian Chan	Non-invasive sensor systems. Biomedical signal processing. Biomedical image processing. Machine learning,Pattern classification,Accessibility
	Adrian Chan	Non-invasive sensor systems. Biomedical signal processing. Biomedical image processing. Machine learning,Pattern classification,Accessibility	Rafik Goubran	Digital signal processing and its applications in biomedical engineering and audio processing.
	James Green	My research focus has been in the following areas: Bioinformatics. proteomics. and prediction of protein structure. function. interaction. and post-translational modificatio. Machine intelligence. pattern classification. data minin. Development of novel assistive technology and device. I also have interest in: medical informatics. robotics. machine vision. and microprocessor application. Current projects include. Species-specific prediction of microRNA from genomic sequence or transcriptomics dat. Prediction of protein-protein interactions from sequenc. Development of various novel assistive devices for persons who are disabled and the elderl. Hardware acceleration of bioinformatics algorithms using GPGPU. including real-time mass spectrometr. Identification of post-translational modifications in proteins. including sumoylation. glycosylation. and hydroxylation.	Dwight Deugo	Eclipse. Beacon Technology. Health informatics and Systems
	Yuu Ono	Sensors development and applications. Biomedical monitoring/diagnosis/characterization. Ultrasonic imaging and acoustic microscope,	Yuu Ono	Sensors development and applications. Biomedical monitoring/diagnosis/characterization. Ultrasonic imaging and acoustic microscope,
	Andy Adler	Medical Image Processing. Applications for Monitoring lung and heart activity. Image reconstruction models to account for errors. Applications for Monitoring lung and heart activity. Image reconstruction models to account for errors. Monitoring lung function in mice. Modelling of Airways physiology. Monitoring lung function in mice. Modelling of Airways physiology. Face Recognition systems. Issues in Biometric Application Design and Integration. Security Implications of biometrics. Face Recognition systems. Issues in Biometric Application Design and Integration. Security Implications of biometrics. Applications for Monitoring lung and heart activity. Image reconstruction models to account for errors. Monitoring lung function in mice. Modelling of Airways physiology. Face Recognition systems. Issues in Biometric Application Design and Integration. Security Implications of biometrics	Junfeng Wen	Artificial Intelligence and Machine Learning

Table 4: Search Result Comparison - Query: "Biomedical Technology"

Search Query	HuggingFace instructor-xl Model		OpenAI text-embedding-ada-002 Model	
	Name	**Research Interest**	**Name**	**Research Interest**
Computational Theory	Alexa Sharp	Algorithms. Algorithmic Game Theory. Network and Graph Theory.	Saban Alaca	Number Theory.
	Kevin Cheung	Combinatorial Optimization. Integer and Logic Programming. Mathematical Programming Computation.	Ayse Alaca	Number Theory and Analysis.
	Anil Maheshwari	Design. Analysis and Implementation of Algorithms for Problems arising in Computational Geometry. Graph Theory. Discrete Mathematics. and Data Science.	Brett Stevens	Design Theory. Gray Codes. De Bruijn Sequences and Universal Cycles. Combinatorial Game Theory. Graph Theory. Mathematical Biology. Philosophy of Mathematics.
	Nicola Santoro	Algorithms. Communication Networks. Data Structures. Discrete Mathematics. Distributed Computing. Distributed and Real-time Systems. Evolutionary Computing. Graph Theory. Mobile Computing. Network Computing. Parallel and Distributed Computing. Robotics. Theory of Computing. and Wireless Communication	Robert Collier	Genetic Algorithms. Computing Education
	Doron Nussbaum	Algorithms. Computational Geometry. Computer Graphics. Computer Vision. Data Structures. Distributed Computing. GIS. Geographic Information Systems. Medical Computing. Parallel Computing. Parallel and Distributed Computing. Robotics and Machine Vision. and Robotics	Alina Shaikhet	Algorithms. Computational Geometry. Visualization.

Table 5: Search Result Comparison - Query: "Computational Theory"

Search Query	HuggingFace instructor-xl Model		OpenAI text-embedding-ada-002 Model	
	Name	**Research Interest**	**Name**	**Research Interest**
Computer Systems	Mengchi Liu	Database Systems. Intelligent Information Systems. Web Technologies	Mengchi Liu	Database Systems. Intelligent Information Systems. Web Technologies
	Anil Somayaji	Computer Security. Operating Systems. Intrusion Detection. Complex Adaptive Systems. Artificial Life	Anil Somayaji	Computer Security. Operating Systems. Intrusion Detection. Complex Adaptive Systems. Artificial Life
	Babak Esfandiari	Agent-based systems. Network computing. Applications of artificial intelligence. Object-oriented design and languages. Network management and supervision.	Junfeng Wen	Artificial Intelligence and Machine Learning
	Shikharesh Majumdar	Current. Parallel and Distributed Systems. Performance Modelling and Performance Evaluation of Computer Systems. Operating Systems. Middleware. Resource Management on Clouds and Grids. Resource Management on Wireless Sensor Networks. Web Services and Service Oriented Architecture. Recent Past. High Performance CORBA-based Middleware Systems. Management of multiple processes and parallel I/O. Scalable Web Servers.	Wilf LaLonde	Animation Systems. Interactive 3D Games. Object-Oriented Systems. Parsers. and Computer Graphics
	Wilf LaLonde	Animation Systems. Interactive 3D Games. Object-Oriented Systems. Parsers. and Computer Graphics	Robert Biddle	Human Computer Interaction. Computer Security. Computer Games. Agile Software Development

Table 6: Search Result Comparison - Query: "Computer Systems"

Search Query	HuggingFace instructor-xl Model		OpenAI text-embedding-ada-002 Model	
	Name	Research Interest	Name	Research Interest
Data Integrations and Data Analysis	Jason Nielsen	Functional Data Analysis. Longitudinal Data Analysis. Mixture Models. Computational Statistics/Numerical Analysis.	Jason Nielsen	Functional Data Analysis. Longitudinal Data Analysis. Mixture Models. Computational Statistics/Numerical Analysis.
	Ahmed El-Roby	Question answering over knowledge graphs. information integration in linked data. open data integration. incorporating non-expert users in data integration. knowledge graph mining. and fact checking using heterogeneous data.	Shirley E. Mills	Data Mining. Applied Statistics.
	Shirley E. Mills	Data Mining. Applied Statistics.	Yuhong Guo	Machine Learning. Artificial Intelligence. Natural Language Processing. Computer Vision. Bioinformatics. Data Analysis
	Patrick Farrell	Categorical Data Analysis. Biostatistics. Sampling. Applied Statistics.	Ahmed El-Roby	Question answering over knowledge graphs. information integration in linked data. open data integration. incorporating non-expert users in data integration. knowledge graph mining. and fact checking using heterogeneous data.
	Yuhong Guo	Machine Learning. Artificial Intelligence. Natural Language Processing. Computer Vision. Bioinformatics. Data Analysis	Junfeng Wen	Artificial Intelligence and Machine Learning

Table 7: Search Result Comparison - Query: "Data Integrations and Data Analysis"

Search Query	HuggingFace instructor-xl Model		OpenAI text-embedding-ada-002 Model	
	Name	Research Interest	Name	Research Interest
Digital and Wireless Communication	Sreeraman Rajan	Sensors and sensor systems (biomedical, defence and security applications). Compressive sensing. Signal processing (including biomedical signal processing, statistical signal processing, adaptive signal processing). Machine learning. Pattern classification.	Amir Banihashemi	Digital and Wireless Communications. Coding and Information Theory. Signal Processing and Algorithms
	Anil Somayaji	Computer Security. Operating Systems. Intrusion Detection. Complex Adaptive Systems. Artificial Life	Ian Marsland	General area of interest: Wireless digital communications (stationary and mobile). Noncoherent receiver design. Error control coding (convolutional and turbo codes). Applications of iterative decoding
	Mohamed Ibnkahla	Wireless sensor networks. Internet of Things (IoT). Cognitive radio networks. Adaptive signal processing. Reconfigurable networks. Sensor integration. Radio frequency identification (RFID) systems. Applications in smart homes, security, safety, healthcare, smart grid, renewable energies and green society, smart cities, intelligent transportation systems, environment monitoring, retail and logistics. and Dr. KC Collins and Digital Sound and Audio. Virtual and Augmented Reality. Computer Networks (Planning, Design, Architecture). Internet of Things (IoT)		
	Thomas Kunz	Internet of Things (Smart Grid/Smart City). Mobile computing systems and the mobile Internet. Mobile ad-hoc networks. Wireless sensor networks. Tactical radio networks. Adaptive mobile applications. Software-defined networking. Network function virtualization	Rafik Goubran	Digital signal processing and its applications in biomedical engineering and audio processing.
	Chung-Horng Lung	Software Engineering: Software Architecture. Service-Oriented Computing. Generative Programming. Self-Managing Systems. Software Reengineering. Networks: Software Defined Networking (SDN). Traffic Engineering. Quality of Service. Content-Based Routing. Wireless Communications. Ad-Hoc Networks. Sensor Networks. and Network-Based Control Systems. Cloud Computing and Distributed Computing: Internet of Things (IoT). Big Data Analytics. Resource Management. Web Services. Realtime Concurrent Software.	Roshdy Hafez	Broadband Wireless Access. Wireless Applications in Transportation Management and Safety. Wireless LAN's and Ad-Hoc networking. Wireless Sensor Networks. Integrated Fiber/Wireless local loop

Table 8: Search Result Comparison - Query: "Digital and Wireless Communication"

Search Query	HuggingFace instructor-xl Model		OpenAI text-embedding-ada-002 Model	
	Name	**Research Interest**	**Name**	**Research Interest**
Embedded Systems	Sreeraman Rajan	Sensors and sensor systems (biomedical, defence and security applications). Compressive sensing. Signal processing (including biomedical signal processing, statistical signal processing, adaptive signal processing). Machine learning. Pattern classification.	Mengchi Liu	Database Systems. Intelligent Information Systems. Web Technologies
	Anil Somayaji	Computer Security. Operating Systems. Intrusion Detection. Complex Adaptive Systems. Artificial Life	Anil Somayaji	Computer Security. Operating Systems. Intrusion Detection. Complex Adaptive Systems. Artificial Life
	Mohamed Ibnkahla	Wireless sensor networks. Internet of Things (IoT). Cognitive radio networks. Adaptive signal processing. Reconfigurable networks. Sensor integration. Radio frequency identification (RFID) systems. Applications in smart homes, security, safety, healthcare, smart grid, renewable energies and green society, smart cities, intelligent transportation systems, environment monitoring, retail and logistics.	Junfeng Wen	Artificial Intelligence and Machine Learning
	Thomas Kunz	Internet of Things (Smart Grid/Smart City). Mobile computing systems and the mobile Internet. Mobile ad-hoc networks. Wireless sensor networks. Tactical radio networks. Adaptive mobile applications. Software-defined networking. Network function virtualization	Dwight Deugo	Eclipse. Beacon Technology. Health informatics and Systems
	Chung-Horng Lung	Software Engineering: Software Architecture. Service-Oriented Computing. Generative Programming. Self-Managing Systems. Software Reengineering. Networks: Software Defined Networking (SDN). Traffic Engineering. Quality of Service. Content-Based Routing. Wireless Communications. Ad-Hoc Networks. Sensor Networks. and Network-Based Control Systems. Cloud Computing and Distributed Computing: Internet of Things (IoT). Big Data Analytics. Resource Management. Web Services. Real-time Concurrent Software.	Wilf LaLonde	Animation Systems. Interactive 3D Games. Object-Oriented Systems. Parsers. and Computer Graphics

Table 9: Search Result Comparison - Query: ”Embedded Systems”

Search Query	HuggingFace instructor-xl Model		OpenAI text-embedding-ada-002 Model	
	Name	Research Interest	Name	Research Interest
Machine Learning	Junfeng Wen	Artificial Intelligence and Machine Learning	Junfeng Wen	Artificial Intelligence and Machine Learning
	Yuhong Guo	Machine Learning. Artificial Intelligence. Natural Language Processing. Computer Vision. Bioinformatics. Data Analysis	Yuhong Guo	Machine Learning. Artificial Intelligence. Natural Language Processing. Computer Vision. Bioinformatics. Data Analysis
	Majid Komeili	I am interested in interpretable machine learning. deep neural networks. and transfer learning. The goal is to create ML models with explainability in mind or to develop methods that can decipher existing black-box ML models. I am also interested in how ML models can learn to perform new tasks with a limited amount of labeled data; a capability that human is very good at.	Robert Collier	Genetic Algorithms. Computing Education
	Matthew Holden	My primary research interest is in Surgical Data Science. I use machine learning for time series data to improve outcomes in surgery and surgical training. based on sensor information collected in the operating room or simulation environment. Due to the diversity in patients and the cost of adverse outcomes. I advocate for the incorporation of domain knowledge into machine learning for surgery. My work facilitates real-time decision support in the operating room. performance assessment and coaching. and increased surgical efficiency.	Shirley E. Mills	Data Mining. Applied Statistics.
	Howard Schwartz	Adaptive and Learning Systems. Multi Robot Learning and Adaptation. Cooperative Control of Multiple Autonomous Robots. Control of robot manipulators. Adaptive Control. System Identification.	Majid Komeili	I am interested in interpretable machine learning. deep neural networks. and transfer learning. The goal is to create ML models with explainability in mind or to develop methods that can decipher existing black-box ML models. I am also interested in how ML models can learn to perform new tasks with a limited amount of labeled data; a capability that human is very good at.

Table 10: Search Result Comparison - Query: ”Machine Learning”

Search Query	HuggingFace instructor-xl Model		OpenAI text-embedding-ada-002 Model	
	Name	Research Interest	Name	Research Interest
Network Security	David Barrera	Computer and Network Security. Internet of Things Security. Operating Systems Security.	David Barrera	Computer and Network Security. Internet of Things Security. Operating Systems Security.
	AbdelRahman Abdou	Internet Measurements. Computer Systems and Network Security. SSL/TLS. Secure DNS. Authentication. Secure BGP. Secure Internet Geolocation. Software Defined-Network Security.	AbdelRahman Abdou	Internet Measurements. Computer Systems and Network Security. SSL/TLS. Secure DNS. Authentication. Secure BGP. Secure Internet Geolocation. Software Defined-Network Security.
	Paul Van Oorschot	Authentication and Identity Management. System Security. Software Security. Application and Web Security. Network Security. Smartphone Security. Usability and Security. Passwords. Public-Key Infrastructure. SSL/TLS.	Paul Van Oorschot	Authentication and Identity Management. System Security. Software Security. Application and Web Security. Network Security. Smartphone Security. Usability and Security. Passwords. Public-Key Infrastructure. SSL/TLS.
	Dr. Wei Shi	Network Technology. Data Science. Computer Networks (Planning. Design. Architecture). Cloud. Fog. and Edge Computing. Internet of Things (IoT). Ad-hoc and Wireless Sensor Networks. Network security (Threat modelling. blockchain. usable security). Data Privacy. Big Data Analytics	Dr. F. Richard Yu	Network Technology. Data Science. Multimedia Streaming. Computer Networks (Planning. Design. Architecture). Cloud. Fog. and Edge Computing. Internet of Things (IoT). Software Defined Networks (SDN). Ad-hoc and Wireless Sensor Networks. Network security (Threat modelling. blockchain. usable security). Data Privacy. Data Modelling. Machine Learning and Artificial Intelligence. Connected/Autonomous Vehicles

Table 11: Search Result Comparison - Query: ”Network Security”

Search Query	HuggingFace instructor-xl Model		OpenAI text-embedding-ada-002 Model	
	Name	Research Interest	Name	Research Interest
Robotics	Howard Schwartz	Adaptive and Learning Systems. Multi Robot Learning and Adaptation. Cooperative Control of Multiple Autonomous Robots. Control of robot manipulators. Adaptive Control. System Identification.	Junfeng Wen	Artificial Intelligence and Machine Learning
	Mark Lanthier	Dr. Lanthier's research interests focus on robotics. artificial life and computational geometry. He has worked with numerous students on research projects involving robotic simulation and coordination. robotic path planning and robotic sensing and design.	Howard Schwartz	Adaptive and Learning Systems. Multi Robot Learning and Adaptation. Cooperative Control of Multiple Autonomous Robots. Control of robot manipulators. Adaptive Control. System Identification.
	Peter X. Liu	Interactive networked systems: teleoperation. telerobotics and telehaptics with applications to telemedicine. Haptics with applications to medical simulations. Robotics. control and intelligent Systems. Context-aware smart networks. Wireless sensor networks	Mark Lanthier	Dr. Lanthier's research interests focus on robotics. artificial life and computational geometry. He has worked with numerous students on research projects involving robotic simulation and coordination. robotic path planning and robotic sensing and design.
	Doron Nussbaum	Algorithms. Computational Geometry. Computer Graphics. Computer Vision. Data Structures. Distributed Computing. GIS. Geographic Information Systems. Medical Computing. Parallel Computing. Parallel and Distributed Computing. Robotics and Machine Vision. and Robotics	Peter X. Liu	Interactive networked systems: teleoperation. telerobotics and telehaptics with applications to telemedicine. Haptics with applications to medical simulations. Robotics. control and intelligent Systems. Context-aware smart networks. Wireless sensor networks
	Mohamed Ibnkahla	Wireless sensor networks. Internet of Things (IoT). cognitive radio networks. adaptive signal processing. reconfigurable networks. sensor integration. radio frequency identification (RFID) systems. Applications in smart homes. security. safety. healthcare. smart grid. renewable energies and green society. smart cities. intelligent transportation systems. environment monitoring. retail and logistics.	Yuhong Guo	Machine Learning. Artificial Intelligence. Natural Language Processing. Computer Vision. Bioinformatics. Data Analysis

Table 12: Search Result Comparison - Query: "Robotics"

Search Query	HuggingFace instructor-xl Model		OpenAI text-embedding-ada-002 Model	
	Name	Research Interest	Name	Research Interest
Social Media	Alan Tsang	Computational Social Choice. Voting Theory. Social Network Simulations. Algorithmic Fairness. Fair Allocation	Junfeng Wen	Artificial Intelligence and Machine Learning
	Safaa Bedawi	.	Dr. KC Collins	Digital Sound and Audio. Virtual & Augmented Reality. Computer Networks (Planning. Design. Architecture). Internet of Things (IoT)
	Dave McKenney	.	Jerome Talim	Web Development. Distributed Systems. Internet of things. Performance Evaluation. Resources Sharing. Routing optimization
	Leila Mostaßo-Guidolin	.	Mengchi Liu	Database Systems. Intelligent Information Systems. Web Technologies
	Darryl Hill	.	Samuel Ajila	Software Requirements Engineering; Model Driven Approach & Aspect Oriented Design; Cloud Resource Provisioning and Management; Software Deployment in the Cloud and Multitenancy; Big Data Analytics: Software Repository. Multi-media. and Networks Radio Traffic Data.

Table 13: Search Result Comparison - Query: "Social Media"

Search Query	HuggingFace instructor-xl Model		OpenAI text-embedding-ada-002 Model	
	Name	**Research Interest**	**Name**	**Research Interest**
Visualization	Alina Shaikhet	Algorithms. Computational Geometry. Visualization.	Oliver Van Kaick	Geometric Modeling. Processing. and Visualization Shape Analysis
	Oliver Van Kaick	Geometric Modeling. Processing. and Visualization Shape Analysis	Alina Shaikhet	Algorithms. Computational Geometry. Visualization.
	Fateme Rajabiyazdi	Information visualization. Human-computer interaction. Health data visualization. Shared display interactions	David Mould	Non-Photorealistic Rendering. Image Stylization. Texture Synthesis. Procedural Modeling. Nonlinear Storytelling
	Donald Bailey	.	Fateme Rajabiyazdi	Information visualization. Human-computer interaction. Health data visualization. Shared display interactions
	Dave Campbell	.	Michael Weiss	.

Table 14: Search Result Comparison - Query: "Visualization"

Search Query	HuggingFace instructor-xl Model		OpenAI text-embedding-ada-002 Model	
	Name	**Research Interest**	**Name**	**Research Interest**
Web Development	Jerome Talim	Web Development. Distributed Systems. Internet of things. Performance Evaluation. Resources Sharing. Routing optimization	Jerome Talim	Web Development. Distributed Systems. Internet of things. Performance Evaluation. Resources Sharing. Routing optimization
	Mengchi Liu	Database Systems. Intelligent Information Systems. Web Technologies	Mengchi Liu	Database Systems. Intelligent Information Systems. Web Technologies
	Robert Biddle	Human Computer Interaction. Computer Security. Computer Games. Agile Software Development	Robert Biddle	Human Computer Interaction. Computer Security. Computer Games. Agile Software Development
	Dr. Chris Joslin	Human-Computer Interaction. Game Design, Development, & Interactions. Virtual & Augmented Reality. Animation. Virtual Environments. and Simulation. Video/Image Processing and Compression. Medical Imaging and Applications. Multimedia Streaming. Machine Learning & Artificial Intelligence. Connected/Autonomous Vehicles	Dr. Robert Teather	Human-Computer Interaction. Game Design, Development, & Interactions. Virtual & Augmented Reality
	Paul Van Oorschot	Authentication and Identity Management. System Security. Software Security. Application and Web Security. Network Security. Smartphone Security. Usability and Security. Passwords. Public-Key Infrastructure. SSL/TLS.	Samuel Ajila	Software Requirements Engineering; Model Driven Approach & Aspect Oriented Design; Cloud Resource Provisioning and Management; Software Deployment in the Cloud and Multi-tenancy; Big Data Analytics: Software Repository. Multi-media. and Networks Radio Traffic Data.

Table 15: Search Result Comparison - Query: "Web Development"

Search Query	HuggingFace instructor-xl Model		OpenAI text-embedding-ada-002 Model	
	Name	**Research Interest**	**Name**	**Research Interest**
Find me someone who can help me with networks and software designs	Chung-Horng Lung	Software Engineering: Software Architecture. Service-Oriented Computing. Generative Programming. Self-Managing Systems. Software Reengineering. Networks: Software Defined Networking (SDN). Traffic Engineering. Quality of Service. Content-Based Routing. Wireless Communications. Ad-Hoc Networks. Sensor Networks. and Network-Based Control Systems. Cloud Computing and Distributed Computing: Internet of Things (IoT). Big Data Analytics. Resource Management. Web Services. Realtime Concurrent Software.	Dr. Marc St-Hilaire	Network Technology. Computer Networks (Planning. Design. Architecture). Cloud. Fog. and Edge Computing. Internet of Things (IoT). Software Defined Networks (SDN). Ad-hoc and Wireless Sensor Networks
	Dr. Marc St-Hilaire	Network Technology. Computer Networks (Planning. Design. Architecture). Cloud. Fog. and Edge Computing. Internet of Things (IoT). Software Defined Networks (SDN). Ad-hoc and Wireless Sensor Networks	Chung-Horng Lung	Software Engineering: Software Architecture. Service-Oriented Computing. Generative Programming. Self-Managing Systems. Software Reengineering. Networks: Software Defined Networking (SDN). Traffic Engineering. Quality of Service. Content-Based Routing. Wireless Communications. Ad-Hoc Networks. Sensor Networks. and Network-Based Control Systems. Cloud Computing and Distributed Computing: Internet of Things (IoT). Big Data Analytics. Resource Management. Web Services. Realtime Concurrent Software.
	Babak Esfandiari	Agent-based systems. Network computing. Applications of artificial intelligence. Object-oriented design and languages. Network management and supervision.	Babak Esfandiari	Agent-based systems. Network computing. Applications of artificial intelligence. Object-oriented design and languages. Network management and supervision.
	Changcheng Huang	Stochastic Control in Computer Networks. Modelling and Simulation Techniques. Resource Optimization in Wireless Networks. Reliability Mechanisms for Optical Networks. Network Protocol Design and Implementation Issues	Dr. Jie Gao	Network Technology. Virtual & Augmented Reality. Computer Networks (Planning. Design. Architecture). Cloud. Fog. and Edge Computing. Internet of Things (IoT). Software Defined Networks (SDN). Ad-hoc and Wireless Sensor Networks. Machine Learning & Artificial Intelligence. Connected/Autonomous Vehicles
	Dr. Jie Gao	Network Technology. Virtual & Augmented Reality. Computer Networks (Planning. Design. Architecture). Cloud. Fog. and Edge Computing. Internet of Things (IoT). Software Defined Networks (SDN). Ad-hoc and Wireless Sensor Networks. Machine Learning & Artificial Intelligence. Connected/Autonomous Vehicles	Samuel Ajila	Software Requirements Engineering; Model Driven Approach & Aspect Oriented Design; Cloud Resource Provisioning and Management; Software Deployment in the Cloud and Multitenancy; Big Data Analytics: Software Repository. Multi-media. and Networks Radio Traffic Data.

Table 16: Search Result Comparison - Query: "Find me someone who can help me with networks and software designs"

Search Query	HuggingFace instructor-xl Model		OpenAI text-embedding-ada-002 Model	
	Name	**Research Interest**	**Name**	**Research Interest**
Find me someone who can help me with UI design	Fateme Rajabiyazdi	Information visualization. Human-computer interaction. Health data visualization. Shared display interactions	Dr. Audrey Girouard	Human-Computer Interaction. Game Design. Development. & Interactions. Wearables
	Dr. Audrey Girouard	Human-Computer Interaction. Game Design. Development. & Interactions. Wearables	Dr. Robert Teather	Human-Computer Interaction. Game Design. Development. & Interactions. Virtual & Augmented Reality
	Nadine Moacdieh	Display clutter. data overload. eye tracking metrics and methods for usability testing and adaptive interfaces. healthcare informatics. interruption management. perception and performance	Dr. Ali Arya	Human-Computer Interaction. Game Design. Development. & Interactions. Wearables. Virtual & Augmented Reality. Animation. Virtual Environments. and Simulation. Educational Technologies
	Dr. Ali Arya	Human-Computer Interaction. Game Design. Development. & Interactions. Wearables. Virtual & Augmented Reality. Animation. Virtual Environments. and Simulation. Educational Technologies	Fateme Rajabiyazdi	Information visualization. Human-computer interaction. Health data visualization. Shared display interactions
	Dr. Robert Teather	Human-Computer Interaction. Game Design. Development. & Interactions. Virtual & Augmented Reality	Robert Biddle	Human Computer Interaction. Computer Security. Computer Games. Agile Software Development

Table 17: Search Result Comparison - Query: "Find me someone who can help me with UI design"