


Introduction

In this paper, we offer an in-depth exploration of vector search using OpenAI embeddings and PineconeDB, specifically applied to a collection of NTRS (NASA Technical Reports Server) article titles. We aim to shed light on the current landscape of vector search within the realm of AI and demonstrate how it integrates into contemporary AI tech stacks. Specifically, we take as our problem that of the NASA researcher. Their current workflow may involve running multiple queries on public and commercial databases, downloading hundreds of papers, and reviewing the relevance of these papers based on their title, abstract, and other metadata. Synthesizing the results of these papers – while at the same time traversing different reporting standards and jargon across disciplines and publications (!) – is challenging, which is where our prototyped solution comes into focus.

 Research Assistant Prototype

Welcome to the NASA Research Assistant!

Enter text below to launch your research session 🚀

The impact of applied research and development on productivity

Submit

Recommended Articles from NTRS:

Score	Title
0.86	Collaborative Product Development in an R&D Environment
0.86	Some effects of time usage patterns on the productivity of engineers
0.86	Improving Customer Satisfaction in an R and D Environment
0.86	Impact of the International Space Station Research Results
0.86	Concurrent Engineering for the Management of Research and Development

Recommended Articles from arXiv.org:

[Are Happy Developers more Productive? The Correlation of Affective States of Software Developers and their self-assessed Productivity](#)

Authors: Daniel Graziotin, Xiaofeng Wang, Pekka Abrahamsson

Categories: cs.SE, cs.HC, D.2.8; K.6.3; H.1.2

Published: 2013-06-07T16:51:39Z

Description: For decades now, it has been claimed that a way to improve software developers' productivity is to focus on people. Indeed, while human factors have been recognized in Software Engineering research, few

Based on the provided information, here are the key findings and guidance for the user: ArXiv Articles: 1. [Title of ArXiv Article 1] - Authors: [Authors] - Published: [Published Date] - Summary: [Summary] 2. [Title of ArXiv Article 2] - Authors: [Authors] - Published: [Published Date] - Summary: [Summary] These ArXiv articles provide valuable insights into the research direction and interests. However, without the specific details of the articles, it is challenging to provide precise guidance. Here are some general suggestions: 1. Read the summaries of the ArXiv articles carefully to identify the key findings and research methodologies used. Pay attention to any novel approaches or significant results mentioned. 2. Consider the authors' expertise and track record in the field. This can help gauge the credibility and reliability of their research. Recommended NTRS Articles: 1. [Title of NTRS Article 1] - Score: [Score] - [Link to NTRS Article 1] 2. [Title of NTRS Article 2] - Score: [Score] - [Link to NTRS Article 2] These NTRS articles

A user enters text, which returns not only a result set of NTRS articles produced using the vector search machinations described more fully in the following sections, but also related papers from researchers outside of NASA via arXiv.org, with whom fruitful research partnerships could be formed based on common interest. This output is then used as input for OpenAlex API calls that return the entire body of work attributable to

an author of interest to the researcher in the context of their search. Finally, all of these elements, rendered dynamically within our prototype's interface, become the context for a ChatGPT prompt fed to version 3.5 via OpenAI's API, the output being a summary of the research surfaced for the user.

Methodology

Recent advancements in AI have seen a significant focus on deep neural networks applied to search, particularly the bi-encoder architecture. In this architecture, content such as queries and article titles is represented as dense vectors, often referred to as "embeddings." These dense retrieval models form the foundation for enhancing large language models (LLMs) for various AI tasks.

There has been recent debate regarding the need for a dedicated "vector store" in AI stacks, considering the substantial investments in existing infrastructure. With this in mind, we conducted a practical demonstration of vector search using OpenAI embeddings and PineconeDB. Our approach involved encoding a vast corpus of article titles from the NTRS and arXiv databases.

The key components of our methodology include:

Embedding Generation: We utilized the OpenAI ada2 model (via their API) to generate embeddings for article titles and queries. The ada2 model boasts an input limit of 8191 tokens and produces embeddings with 1536 dimensions. To efficiently handle a substantial amount of data, we made parallel API calls while adhering to the rate limit of 3500 calls per minute.

Indexing with PineconeDB: We indexed the resulting embedding vectors using PineconeDB. PineconeDB offers excellent support for cosine similarity scoring and top-k retrieval, which aligns perfectly with the task of facilitating vector search.

Datasets: Our demonstration encompassed two datasets: a collection of 100,000 NTRS article titles and approximately 1.1 million arXiv article titles.

Results

Source	Search Results
--------	----------------

<p>NTRS</p>	<ul style="list-style-type: none"> • Ageing Mechanisms and Control Specialists' Meeting on Life Management Techniques for Ageing Air Vehicles • Emerging and Future Computing Paradigms and Their Impact on the Research, Training, and Design Environments of the Aerospace Workforce • The Shock and Vibration Bulletin: Part 2 - Invited Papers, Space Shuttle Loads and Dynamics, Space Shuttle Data Systems, Shock Testing, Shock Analysis Space Shuttle Thermal Protection Systems • The impact of space research expenditures on urban and regional development • Software productivity improvement through software engineering technology
<p>Science Discovery Engine</p>	<ul style="list-style-type: none"> • SUP15.FINAL1 • ARSET - Remote Sensing for Monitoring Land Degradation and Sustainable Cities SDGs • Microsoft Word - sup-15rept.txt • ALGORITHM THEORETICAL BASIS DOCUMENT FOR MODIS PRODUCT MOD-27 OCEAN PRIMARY PRODUCTIVITY (ATBD-MOD-24) • Pick-and-eat Salad-crop Productivity, Nutritional Value, and Acceptability to Supplement the ISS Food System (Veg-05)
<p>Our Prototype</p>	<p>Suggested NTRS Articles:</p> <ul style="list-style-type: none"> • Collaborative Product Development in an R&D Environment • Some effects of time usage patterns on the productivity of engineers • Improving Customer Satisfaction in an R and D Environment • Impact of the International Space Station Research Results • Concurrent Engineering for the Management of Research and Development <p>Suggested arXiv Articles:</p> <ul style="list-style-type: none"> • Impact of Software Engineering Research in Practice: A Patent and Author Survey Analysis • Are Happy Developers more Productive? The Correlation of Affective States of Software Developers and their self-assessed Productivity • A measure of total research impact independent of time and discipline

Our demonstration exemplified the ease with which state-of-the-art vector search can be implemented within modern AI ecosystems. Here are some notable outcomes:

- **Efficiency:** Encoding 100,000 NTRS article titles took less than 15 minutes, showcasing the efficiency of the process. The cost of encoding was remarkably low at \$0.0001 per 1,000 token embeddings.
- **Integration:** Our approach seamlessly integrated OpenAI embeddings and PineconeDB into the AI stack. The process of generating embeddings was as straightforward as making API calls, and searching for dense vectors mirrored the principles of traditional text indexing.

Discussion

Vector search represents a pivotal advancement in AI search capabilities, particularly with the rise of dense retrieval models and embeddings. While the debate over the necessity of a dedicated "vector store" continues, our demonstration suggests that modern enterprises can benefit significantly from incorporating vector search into their AI stacks. Moreover, the coexistence of legacy systems with new AI architectures is a crucial consideration, emphasizing the need for seamless integration.

With that being said, The bottleneck here was that PineconeDB only allows "bottom-tier" subscribers to their service to create a single vector database, which is capped at 100,000 records. To increase beyond this poses a cost of at least \$70 per month. This might make open source alternatives (i.e., Postgres' pgvector database solution) more attractive, but there are technical aspects (i.e., pgvector produces scores using L2, or Euclidean, distance instead of cosine similarity) that must be considered.

Conclusion

Our demonstration of vector search using OpenAI embeddings and PineconeDB on NTRS article titles illustrates the practicality and efficiency of this approach within contemporary AI stacks. The integration of dense retrieval models and embeddings opens up exciting possibilities for enhancing search capabilities, and the debate over dedicated vector stores underscores the evolving landscape of AI infrastructure. Embracing vector search can empower NASA to leverage state-of-the-art AI techniques while accommodating legacy systems, ultimately advancing their AI capabilities in their pursuit of exploring the unknown.