Enhancing Query Retrieval Precision Through Optimized Embedding Text Selection

Alex P. Wang
i80 Solutions
Nov 16, 2024

Abstract

This paper explores strategies to optimize embedding texts for semantic search, focusing on the impact of normalization, synonyms, alternate phrasing, and typo handling. Using text-embedding-3-large with ChromaDB, we analyze test results to provide recommendations for creating embeddings that enhance retrieval accuracy. Key findings demonstrate the importance of text standardization, incorporation of contextual synonyms, and query preprocessing while cautioning against embedding literal typos. These conclusions are supported by detailed test results, illustrating how semantic models handle variations in user input.

Introduction

Semantic search systems use vector-based representations to match user queries to relevant information based on intent and meaning rather than exact keyword matches. Embedding models such as text-embedding-3-large generate multi-dimensional vectors where similar meanings are represented as closely related points in the vector space. For instance, "breakfast time" and "when is breakfast served" produce vectors that align closely despite differences in phrasing.

However, the effectiveness of semantic search depends heavily on how embedding texts are constructed. Factors such as normalization, inclusion of synonyms, and handling of typos can influence retrieval accuracy. This study investigates these factors using a series of tests, providing evidence-based recommendations for embedding text creation.

Methodology

To conduct this study, an embedding model and a vector database were utilized to enable semantic search and evaluate query retrieval precision. Embedding models convert textual data into high-dimensional vector representations, capturing semantic meaning in a format that allows for effective comparison and matching (Mikolov et al., 2013). A vector database is essential for storing these embeddings and performing similarity searches efficiently, enabling the alignment of user queries with relevant embedding texts based on their vector representations (Chhabra, 2023).

For this study, the text-embedding-3-large model, which generates 3072-dimensional embeddings, was selected. This model is among the most advanced in capturing nuanced semantic relationships, making it particularly suitable for tasks requiring high precision and contextual understanding (Open AI Embeddings). To manage and query these embeddings, we utilized **ChromaDB**, a robust vector database designed for efficient storage and retrieval of high-dimensional data (ChromaDB). ChromaDB was chosen for its reliability and support for cosine similarity (Wikipedia)—a standard metric for evaluating semantic similarity—ensuring consistent and accurate results. These tools allowed us to rigorously test and analyze various embedding text selection strategies.

This study was designed to evaluate how different text selection strategies influence semantic similarity and query retrieval performance. Each test case explores a distinct aspect of embedding text optimization, examining its impact on accuracy and consistency across a variety of user query scenarios.

The four areas were chosen for their direct relevance to improving semantic search retrieval accuracy in practical settings. Each represents a distinct aspect of user query variability:

- **Text Normalization** ensures consistent formatting.
- **Synonyms** address natural variability in word choice.
- **Typos** account for errors in user input.
- **Natural Conversational Language** matches real-world phrasing.

1. Text Normalization

Description:

- This test examined the impact of lowercasing all text and removing unnecessary punctuation from embedding texts and queries.
- The goal was to determine whether normalization improves retrieval consistency by reducing variations caused by formatting inconsistencies.

Reason for Inclusion:

- **Real-World Relevance**: Users often enter queries with mixed casing or punctuation errors.
- **Literature Reference**: Normalization is a well-documented preprocessing technique in natural language processing (NLP) to enhance model performance.

Expected Outcome:

- Normalized text should reduce the semantic variability introduced by capitalization and punctuation.

Other Possibilities:

- Exploring more advanced normalization techniques, such as lemmatization or stemming, was not included due to the semantic nature of embeddings, where word context is preserved better than in keyword-based methods. These could be subjects for future studies.

2. Synonyms and Alternate Phrasing

Description:

- Embedding texts were expanded to include synonyms and alternate phrasings for common queries, e.g., "breakfast time" and "what time is breakfast."

Reason for Inclusion:

- **User Query Variability**: Users often use synonyms or rephrase their queries in natural language.
- **Literature Reference**: The semantic capabilities of embeddings rely on capturing related meanings (Mikolov et al., 2013).

Expected Outcome:

- Embeddings with synonyms and alternate phrasings should reduce distance scores for varied queries, improving retrieval accuracy.

Other Possibilities:

- Contextual embeddings trained on specific domains could improve synonym recognition. This study used pre-trained embeddings, which may lack domain-specific context. Future studies could evaluate fine-tuned embeddings.

3. Common Typos

Description:

- Tests were conducted to evaluate how well the model handled common user typos, e.g., "breakfat time" instead of "breakfast time."

Reason for Inclusion:

- **Prevalence of Errors**: Typographical errors are common in user queries, particularly in mobile typing environments.
- **Robustness Testing**: Assessing whether embeddings inherently tolerate such errors is critical for system reliability.

Expected Outcome:

- Embeddings should match closely to the correct text without explicitly including typos, leveraging semantic similarity to handle minor variations.

Other Possibilities:

- Explicitly embedding typos was excluded due to concerns over database size and redundancy. Query preprocessing, such as typo correction using spelling correction libraries (SymSpell), could complement embeddings and will be explored in future work.

4 . Concise, Conversational Language

Description:

- Embedding texts were written in conversational, natural language to mimic real-world query phrasing, e.g., "how do I get WiFi?" instead of technical or verbose descriptions.

Reason for Inclusion:

- **Real-World Usability**: Users often phrase queries naturally, expecting responses in a conversational tone.
- **Literature Reference**: Conversational AI systems emphasize natural language for user-friendliness (Radford et al., 2018).

Expected Outcome:

- Embedding texts written in conversational language should align closely with user queries, improving retrieval precision.

Other Possibilities:

- Including formal or technical phrasings for specialized domains (e.g., medical terminology) could improve retrieval in professional settings. This study prioritized everyday conversational queries to focus on hospitality use cases.

This methodology leverages state-of-the-art embedding models and focuses on practical, real-world variability in user input. By addressing key challenges through normalization, synonyms, typos, and conversational language phrasing, the study aims to provide actionable insights for improving semantic search precision in hospitality applications. Future work will expand on contextual and domain-specific adaptations to further enhance the robustness of semantic retrieval systems.

Results and Analysis

1. Text Normalization

Objective: To determine if text normalization (e.g., converting to lowercase, removing punctuation) improves retrieval accuracy by reducing variability caused by formatting differences.

Test Results:

1. **Query**: "Wi-Fi"
    - **Distance Scores**:
        - 0.0 for 'Wi-Fi'
        - 0.134 for 'wi-fi'
        - 0.240 for 'WiFi'
        - 0.312 for 'wifi'
2. **Query**: "breakfast hours"
    - **Distance Scores**:
        - 0.0 for 'breakfast hours'
        - 0.052 for 'Breakfast Hours'
3. Query: breakfast time?
    - **Distance Scores**:
        - 0.08813, {'embedding_text': 'breakfast time'}
        - 0.18912, {'embedding_text': 'what time is breakfast'}
        - 0.20969, {'embedding_text': 'breakfast serving time'}
        - 0.24073, {'embedding_text': 'what are breakfast times'}

4. Query: breakfast time
    - **Distance Scores**:
        - 0.0, {'embedding_text': 'breakfast time'}
        - 0.15923, {'embedding_text': 'breakfast serving time'}
        - 0.17118, {'embedding_text': 'what time is breakfast'}
        - 0.18837, {'embedding_text': 'breakfast hours'}
        - 0.20495, {'embedding_text': 'what are breakfast times'}

Findings:

- Case differences (e.g., "Wi-Fi" vs. "wifi") led to significant variations in similarity scores, with differences of up to 0.3, even when the semantic meaning was identical.
- Removing punctuation (e.g., question marks) slightly reduced score variability, leading to improved retrieval consistency.
- These results indicate that inconsistencies in formatting can negatively affect semantic search precision.

Conclusion: Text normalization improves retrieval accuracy by ensuring closer matches for semantically identical terms that may differ due to minor formatting inconsistencies. Normalization reduces noise caused by capitalization and punctuation variations, allowing embeddings to focus on the core semantic content of the text.

Recommendation:

- Apply normalization to both user query text and embedding text in the vector database.
    - **For User Queries**: Normalize inputs by converting to lowercase and removing unnecessary punctuation before converting to embedding for searching.
    - **For Embedding Texts**: Store embeddings of normalized text (lowercase, punctuation-free) in the vector database.

This ensures uniformity across all texts, improving the model's ability to retrieve consistent and relevant results regardless of input formatting differences.

2. Synonyms and Alternate Phrasing

Objective: To evaluate the impact of including synonyms and alternate phrasings in embedding texts.

Test Results:

1. **Query:** "breakfast time"
    - Distance Scores:
        - 0.0 for 'breakfast time'
        - 0.242 for 'what time is breakfast'
        - 0.253 for 'breakfast hours'
        - 0.280 for 'when is breakfast'
2. **Query:** "internet access"
    - Distance Scores:
        - 0.0 for 'internet access'
        - 0.072 for 'internet connection'
        - 0.143 for 'how to get online'

Findings:

- Embedding synonyms such as 'breakfast hours' alongside 'breakfast time' reduced distance scores for varied queries.
- Alternate phrasings (e.g., 'how to get online' for 'internet access') improved query flexibility without adding significant redundancy.

Conclusion: Including synonyms and alternate phrasings expands coverage for semantically similar queries, ensuring higher retrieval accuracy.

Recommendation: Embed a diverse set of synonyms and natural language variations for key terms.

3. Typo Handling

Objective: To assess whether literal typo embeddings are necessary.

Test Results:

1. **Query:** "breakfat time" (typo: missing "s")

    - Distance Scores:
        - 0.073 for 'breakfast time'
        - 0.253 for 'breakfast hours'
2. **Query:** "interent access" (typo: "interent" instead of "internet")
    - Distance Scores:
        - 0.110 for 'internet access'
        - 0.297 for 'wifi access'

Findings:

- Typos like "breakfat time" resulted in low distance scores when matched with correct embeddings, indicating the model's semantic tolerance.
- Embedding literal typos increased redundancy without significant improvement in retrieval accuracy.

Conclusion: Embedding literal typos is unnecessary as semantic models can tolerate minor textual errors effectively.

Recommendation: Preprocess queries to correct common typos rather than embedding them.

4. Concise, Conversational Language

Objective: To evaluate the effectiveness of conversational language phrasing in embeddings.

Test Results:

1. **Query:** "how do I connect to wifi?"
    - Distance Scores:
        - 0.0 for 'how do I connect to wifi'
        - 0.128 for 'wifi access'
2. **Query:** "breakfast time"
    - Distance Scores:
        - 0.0 for 'breakfast time'

        - 0.242 for 'what time is breakfast'

3. Query: is there free wifi
    - Distance Scores:
        - 0.0, {'embedding_text': 'is there free wifi'}
        - 0.09686, {'embedding_text': 'is there wifi'}
        - 0.13923, {'embedding_text': 'is there wi-fi'}
        - 0.17553, {'embedding_text': 'is wifi free here'}

Findings: Embedding conversational phrases (e.g., 'how do I connect to wifi') improved matching for natural language queries while maintaining compact embeddings.

Conclusion: Natural language embeddings align well with conversational user inputs, enhancing retrieval accuracy.

Recommendation: Use concise, user-friendly sentences that reflect typical query patterns.

Conclusion and Recommendations

This study demonstrates that optimizing embedding text selection through normalization, synonyms, typo preprocessing, and natural language phrasing significantly enhances semantic search performance. By addressing key challenges in query variability, these strategies ensure closer semantic alignment and improved retrieval accuracy.

Recommendations:

1. **Normalize Embedding Texts**:
    - Convert to lowercase and remove unnecessary punctuation in both user queries and embedding texts to improve consistency and minimize formatting-related variability.
2. **Incorporate Synonyms and Phrasing**:
    - Add embeddings for common synonyms and natural language variations to cover a broader range of user query styles and phrasings.
3. **Preprocess Queries for Typos**:
    - Implement query preprocessing to correct common misspellings and typos before embedding, reducing the need for redundant embeddings.

4. **Focus on Natural Language**:
    - Use concise, conversational sentences in embedding texts to mimic real-world user input and enhance semantic alignment.

By implementing these recommendations, semantic search systems can achieve more robust and accurate query matching, delivering an improved user experience.

Future Work

Future work should explore advanced areas, such as multilingual query handling, domain-specific fine-tuning, and dynamic embedding strategies, to further refine retrieval precision and broaden the applicability of semantic search systems in diverse contexts.

- **Context Length Variations**:
    - Longer embedding texts might better capture user intent, while shorter ones might work for keyword queries. This study used medium-length phrases to balance granularity and efficiency.
    - **Future Work**: Analyze the impact of embedding text length on retrieval outcomes.
- **Multilingual Queries**:
    - This study focused on English embeddings. Multilingual embeddings (Conneau et al., 2020) could be explored to extend applicability to international audiences.
- **Domain-Specific Fine-Tuning**:
    - Embeddings were generated using general-purpose pre-trained models (e.g., text-embedding-3-large). Fine-tuning for hospitality-specific queries could improve relevance.
- **User Context Modeling**:
    - Incorporating user history or preferences to contextualize search results was beyond the study's scope. Future studies could integrate personalized embeddings to enhance relevance.

References

- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
- Chhabra, M. (2023). A Comprehensive Guide to Vector Databases: The Future of AI-Driven Data Retrieval. Medium.
- Open AI. (n.d.). Embeddings. OpenAI API Documentation.
- DataCamp. (n.d.). ChromaDB Tutorial: Step-by-Step Guide.
- Wikipedia. (n.d.). Cosine similarity.
- Jurafsky, D., & Martin, J. H. (2021). Speech and Language Processing. Stanford University.
- SymSpell. (n.d.). GitHub repository.
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI.
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:2004.09813.