

Multilingual Semantic Search

Overview

The realm of information retrieval and search is plagued with the inherent complexity of language. Phrases often have diverse representations, utilizing synonyms, varied word orders, and distinct structures. This diversity in

representation is amplified when accounting for multiple languages, each with its unique grammar, semantics, and vocabulary. This variability leads to significant challenges in developing a universal, language-agnostic

search solution capable of accurately identifying semantic similarities between phrases, despite their disparate representations and linguistic origins.



Situation

Car sales and appraisal: In the context of car descriptions and specifications, a multilingual semantic search can identify mentions of specific car equipment, features, or models, enhancing the accuracy and relevance of search results. Recruitment: For job descriptions and CVs, this technology can precisely match skills listed, enabling more efficient and effective talent acquisition and job matching, even when documents are in different languages.

Solution

AI technology was utilized to create a solution leveraging compact and efficient language models, capable of generating multilingual embedding vectors. These vectors served as universal, language-agnostic representations of words and phrases, capturing their semantic essence. This allowed the discerning of semantic similarities between varied phrases across multiple languages, yielding accurate and semantically enriched search results, irrespective of linguistic differences.

Unlike many contemporary Large Language Models which demand significant computational resources and are typically cloud-based, our models are engineered to be small, resource-efficient, and capable of executing on local machines, ensuring high-speed, cost-efficient processing. This allows users to experience the benefits of enhanced semantic search without the necessity for extensive computational power or dependence on cloud-based solutions.

Keywords

semantic search
language-agnosticity
language modeling
big data
phrase normalization
multilingual
embedding
structured/unstructured data
multilingual
dictionary
synonym resolution
natural language processing (NLP)
information retrieval
semantic similarity
vector representation

Requirements

Ability to parse and analyze both structured and unstructured multilingual data from diverse sources.

Development of a comprehensive multilingual dictionary to facilitate accurate translations and synonym resolutions.

Capability to normalize phrases and expressions across different languages and linguistic structures.

Efficient handling and processing of big data volumes, ensuring scalability and performance.

Benefits and Results

- **Enhanced Accuracy:** By understanding the semantic essence of phrases, the solution delivers more accurate and relevant search results across languages.
- **Language-Agnosticity:** The language-agnostic nature of the solution ensures seamless integration and applicability across diverse linguistic landscapes.
- **Scalability:** Due to dynamic creation of vocabularies solution is inherently scalable and able to accommodate growing data volumes and complexities.
- **Increased Operational Efficiency:** Organizations can experience improved efficiency in information retrieval, reducing the time and effort involved in finding relevant information.
- **Versatile Applicability:** Its adaptability across various domains like car sales and recruitment illustrates its versatile applicability and potential to revolutionize multilingual information retrieval.