1. In your own words, describe what vector embeddings are and what they are useful for?

   Vector embeddings are mathematical representations of objects, concepts, or entities in a multi-dimensional space, where similar items are positioned closer to each other. These embeddings capture the relationships and similarities between different items, making them a powerful tool in various machine learning tasks.

   vector embeddings convert complex and abstract entities, such as words, images, or even entire documents, into numerical vectors. Each dimension of these vectors corresponds to a specific feature or characteristic, and the proximity of vectors in the space reflects the similarity or relationship between the corresponding entities.

   The key advantage of vector embeddings lies in their ability to capture intricate relationships between items in a data space, facilitating more efficient and accurate machine learning models across diverse domains.

   Vector embeddings find utility in a wide range of applications, including natural language processing (NLP), computer vision, and recommendation systems. In NLP, word embeddings help represent words in a way that preserves semantic relationships, allowing algorithms to understand and work with language more effectively. In computer vision, image embeddings enable the comparison and recognition of visual patterns. Recommendation systems benefit from embeddings by identifying similarities between user preferences and content.

2. What do you think is the best distance criterion to estimate how far two embeddings (vectors) are from each other? Why?

   The choice often depends on the characteristics of the data and the specific goals of the application. Cosine similarity is popular for text-based applications because it captures the angle between vectors, emphasizing direction rather than magnitude.

   Euclidean distance could be relevant in the case of a model that classifies images of handwritten digits. In this scenario, each image is represented as a feature vector, and Euclidean distance could be used to measure the difference between two images. For example, if you have two feature vectors corresponding to two images of digits A and B, the Euclidean distance between them could be calculated to measure the separation

   cosine similarity is a versatile and widely used metric for assessing similarity between vectors, especially in NLP tasks where capturing semantic relationships is essential. Its emphasis on direction makes it a valuable tool for measuring similarity in various applications, including document retrieval, text clustering, and recommendation systems.

3. Let us build a Q&A (question answering) system! 😀For this, consider the following steps:

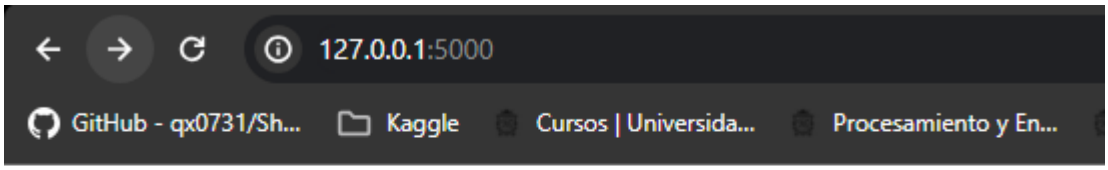
   c. Implement the embedding generation logic. Which tools and approaches would help you generate them easily and high-level?

   answer: To implement embedding generation logic for a Q&A system, can leverage pre-trained language models and libraries that provide easy integration with popular deep learning frameworks Such as Hugging Face Transformers Library.

   d. For every question asked by the user, return a sorted list of the N chunks/pieces in your text that relate the most to the question. Do results make sense?

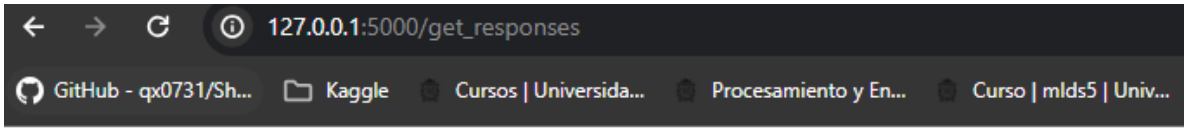
   Answer:

   Two types of files were used. An archive is a book called cognitive capitalism, the answers that were obtained when asking the questions are not of good quality but they try to provide an answer. For example, the question was asked about what cognitive capitalism is and the answer was the following:



Question Answering System

Enter your question: Que es el capitalismo cogniti Submit

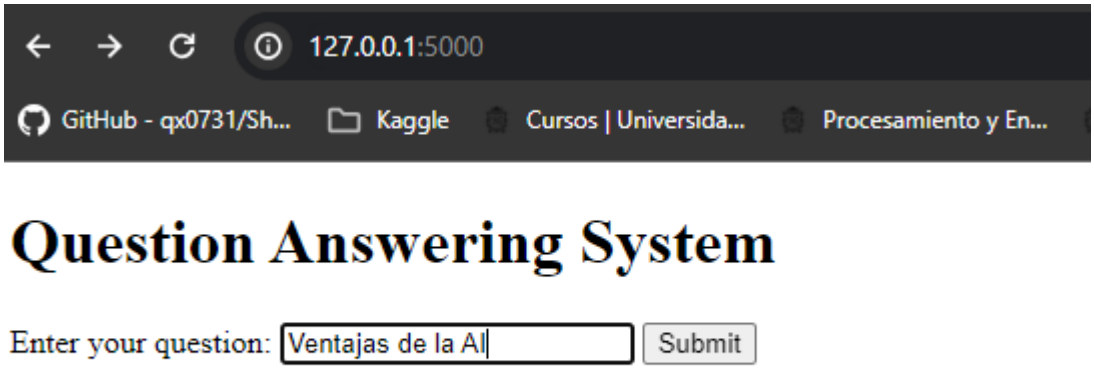


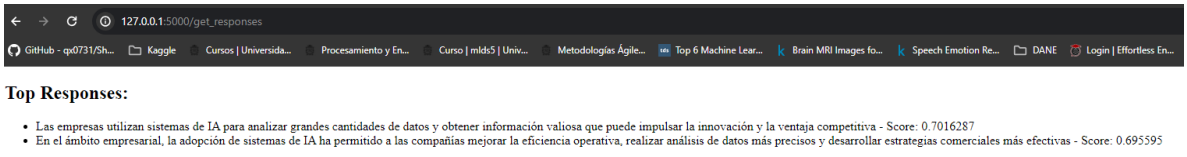
Top Responses:

- 134 Capitalismo cognitivo propio valor del conocimiento - Score: 0.77599066
- Riqueza, propiedad, libertad y renta en el capitalismo cognitivo - Score: 0.7720105

This may occur because the book used contains quite technical and complex language.

The results improve when using a text that is not so technical and more informative about artificial intelligence in the world and Colombia. For this example, the question was asked about the advantages of AI and these were their answers:





4. What do you think that could make these types of systems more robust in terms of semantics and functionality?

- Better Preprocessing:

Use advanced text preprocessing techniques to handle different language nuances, including stemming, lemmatization, and handling of synonyms.
Implement techniques to handle spelling errors and typos.

- Semantic Understanding:

Incorporate semantic role labeling and entity recognition to better understand the relationships between different parts of a sentence.

- Ensemble Models:

Combine predictions from multiple models or techniques to create an ensemble model. This can enhance robustness by reducing the impact of individual model weaknesses.

- Better Dataset:

A larger and better structured data set. That is, having different types of questions and answers.

- Integration of External Knowledge:

Integrate external knowledge bases or ontologies to supplement the understanding of the system. This can enhance the system's ability to answer questions beyond the scope of its training data.