Lab No. 3

Machine Learning 2

By.

Jeison Arias

Juan Esteban Rojas

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

R:// Vector embeddings is a technique to convert words or sentences into numbers representation (vectors) which keep their meaning inside to let the computers understand and process the information contained into those numbers, so the position of each vector in the multidimensional space represents the meaning of the text/word. In this context, the closer 2 vectors are, most similarity exists between 2 texts.

They are useful because vectors let the machine learning algorithms to find patterns in natural language processing, recommendation systems, feeling analysis, text comparison, etc. That is because as the words, sentences or texts are represented as vectors may be reduced, analyzed, and compared with other vectors.

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

R:// The most suitable criterion corresponds to the cosine similarity, which is not based on the distance between vectors (magnitude) but measures the similarity through the angle between the vectors in a N-Dimensional space.

This method is effective for measuring the similarity between text embedding vectors due to its properties of magnitude independence and semantic preservation, as it focuses on the orientation of vectors while ignoring differences in magnitude. This method relies on simple operations between vectors, bringing advantages in terms of computational efficiency.

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

Run the Lab3.ipynb notebook and the Task5.html to type the question about the text. Once you run it, there will be deployed the closer embedding

3.c. Which tools and approaches would help you generate them easily and high-level?

It would be helpful to use pre-trained models which have been trained with large amounts of data and can capture complex semantic relationships. In this case hugging face can help providing such models with multiple approaches. Here you can find multiple models for Natural Language Processing (NLP) and being into NLP, there is able to filter for a most specific need such as text similarity.

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

Augment the training data to expose the model to a wider range of linguistic variations and ensure robustness to different expressions of the same concept.

Apply post-processing techniques to embeddings, such as dimensionality reduction or normalization, to enhance their interpretability and robustness.

In terms of semantics, as we could see in the results of task 3, the answer is just a piece of the original text due to its similarity with the question. But, in order to get a best answer (semantically) it could be combined with models like Retrieval Augmented Generation to retrieve a coherent answer.

1. Bonus points if deployed on a local or cloud server.

Run the Lab3.ipynb notebook and the Task5.html to type the question about the text.