FOCUS: Find Out Characters Under Specification

- **Team name: Whatever**
- **Team ID: 25**
- **Team members**
  - 110550036 張家維
  - 110550014 吳權祐
  - 110550100 廖奕瑋

Abstract

Our final project revolves around the evaluation and potential enhancement of various word embedding models. The models under consideration are Word2Vec, GloVe, FastText, BERT, and ELMo. The ultimate goal is to employ these models for 2 applications: section finding in academic papers and classifying notes within folders.

Motivation

With the improvement of natural language processing (NLP), word embedding has seen unneglectable importance due to its ability to represent syntax-level information. Later developments of word embedding models further improves the efficiency and effectiveness of NLP. Based on this fact, we see the aforementioned 2 applications for respective reasons: First, we have been tired of searching concepts, sentences, etc. among a ton of papers, so we decide to turn to NLP to get rid of the tedious work. Second, as time goes on, the notes taken usually appears to be messy, and it’s desirable to find a way to categorize them.

Introduction

- **Applications**

The potential applications of these models include rapidly identifying section-starting keywords in vast academic papers and categorizing diverse notes or text content within specific folders.

- **Methodology**

  Our approach encompasses the selection of suitable word embedding models, data collection and preprocessing, model implementation and training, performance evaluation, fine-tuning (if required), and thorough documentation of the results.

- **Potential Strategies for Performance Enhancement**

  We are considering several strategies to enhance performance and efficiency, including the integration of multiple models, implementation of active learning, incorporation of multilingual support, and real-time application development.

Method

Finding Embedding Model

- Voyage-large-2-instruct
- text-embedding-3-small
- E5-large-v2

https://huggingface.co/spaces/mteb/leaderboard

Model Evaluations

- Wang, B., Wang, A., Chen, F., Wang, Y., & Kuo, C. C. J. (2019). Evaluating word embedding models: Methods and experimental results. *APSIPA transactions on signal and information processing*, *8*, e19.
- Extrinsic evaluator
  - Uses word embedding as input
  - Measures changes in performance
- Intrinsic evaluator
  - Tests independence b/t representation and specific NLP tasks

- Measures syntactic or semantic relationships among words

Model Selection & System Design

- Data preparation

  Hugging Face datasets

  - Spector

    https://huggingface.co/datasets/embedding-data/SPECTER

  - simple-wiki

    https://huggingface.co/datasets/embedding-data/simple-wiki

  - WikiAnswers

    https://huggingface.co/datasets/embedding-data/WikiAnswers

- Model training
- Other designs
  - Use large language model (LLM) for encoding text
- Analyze result

Expected Result

Our biggest goal is to develop applications that rapidly identify sections from a large amount of papers or categorize diverse notes or text content. In addition to this, the results of this project promise not only to advance our technology but also pave the way for future potential breakthroughs in our interaction with information. When constructing the project, we will learn a lot about embedding models and related technologies at the same time.

Resources

https://huggingface.co/blog/getting-started-with-embeddings

https://huggingface.co/blog/how-to-train-sentence-transformers