ThesisSense: Using Large Language Models to extract meaningful information from theses

Margit Antal

Department of Mathematics–Informatics
Sapientia Hungarian University of Transylvania
Târgu-Mureș, Romania

manyi@ms.sapientia.ro

In recent years, significant advancements in natural language processing have led to the development of large language models, exemplified by OpenAI's ChatGPT model. These models offer the capability to effectively extract diverse forms of information from textual documents, encompassing elements such as keywords and emotional tones. Leveraging the OpenAI API, textual data can be transformed into numerical representations, which, upon being stored within vector databases, facilitate the execution of semantic searches across a corpus of documents, thereby enabling interactions with our own textual materials.

This presentation elucidates a systematic approach for the extraction of valuable insights from academic theses. By deconstructing each document into discrete segments, the employed large language model is harnessed to derive keywords from the abstract sections. Subsequently, these keywords are transformed into a word cloud, affording a comprehensive snapshot of the theses encompassing a designated cohort. Furthermore, embeddings generated via OpenAI's technology are crafted for each individual document and then systematically organized within a localized vector database.

Upon projecting these embeddings onto a two-dimensional plane, discernible patterns emerge, visually portraying clusters of related documents. Documents expounding similar subject matters manifest as proximate points within this visual representation. Notably, the incorporation of semantic search functionalities within this document pool becomes feasible, thus enabling nuanced information retrieval from the aggregated body of documents.

Keywords: large language models, vector databases, embeddings, semantic search, information extraction

References

[1] Z. Liu, M. Sun (2023). Representation Learning and NLP. In: Liu, Z., Lin, Y., Sun, M. (eds) *Representation Learning for Natural Language Processing.* Springer, Singapore, 2023.

[2] A. Cohan, S. Feldman, I. Beltagy, D. Downey, D.S. Weld. SPECTER: Document-level Representation Learning using Citation-informed Transformers. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2270–2282, July 5 - 10, 2020.