

Bachelor Thesis Proposal:

LLM-based Text Clustering: Comparing Semantic Embeddings and Traditional Text Representations

Short Abstract

Text clustering aims to automatically organize a collection of documents or text samples into groups according to their similarity. Traditional approaches typically represent text using sparse features such as TF-IDF, while recent language models provide dense semantic representations that can capture more contextual information.

In this thesis, the student will investigate how different text representations affect the quality and characteristics of text clustering. Traditional representations such as TF-IDF will be compared with modern semantic embeddings produced by pretrained language models, such as BERT or Sentence-BERT. Different clustering algorithms can then be applied to these representations, and the resulting clusters will be evaluated using both quantitative metrics and qualitative analysis.

The project provides a practical introduction to modern NLP, language representations, and unsupervised machine learning.

Objectives

The main objectives of the thesis are:

- Implement traditional text representations such as TF-IDF.
- Generate semantic text embeddings using pretrained language models.
- Apply clustering methods such as K-Means or hierarchical clustering.
- Compare the resulting clusters using appropriate clustering evaluation metrics.
- Analyze how different representations influence the discovered semantic structure.
- Visualize and qualitatively inspect the resulting clusters.

The thesis should answer questions such as:

- Do semantic embeddings produce more meaningful text clusters than traditional representations?
- How does the choice of representation affect clustering quality?
- Which types of textual similarity are captured by different representations?

Requirements

- Basic knowledge of Python.
- Basic knowledge of machine learning and data analysis.
- Familiarity with NumPy, pandas, and scikit-learn.
- Basic understanding of NLP is helpful but not required.
- Experience with Hugging Face Transformers or Sentence-Transformers is helpful.
- No prior experience with training large language models is required.

Possible Tools

Python, NumPy, pandas, scikit-learn, Hugging Face Transformers, Sentence-Transformers, and standard visualization libraries.

References

1. Aggarwal, C. C., & Zhai, C. (2012). *A Survey of Text Clustering Algorithms*. In **Mining Text Data**, Springer, pp. 77–128.
2. Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. In **Proceedings of EMNLP-IJCNLP 2019**, pp. 3982–3992.
3. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In **Proceedings of NAACL-HLT 2019**, pp. 4171–4186.

If you are interested in this proposal, please send a short email introducing yourself to Jian Lan (lan@dbis.fh-lmu.de) as soon as possible before 10.10.2026.