RPTU
Department of Computer Science
Prof. Dr. Sophie Fellenz

Exploring Topic Embeddings

For: SS 2023
Supervisor: Mayank Nagda

Project: Exploring Topic Embeddings
It is widely recognized that word embeddings offer a comprehensive depiction of a word within the embedding space, enabling the derivation of semantically significant linguistic associations through arithmetic operations on word embeddings [2]. A classic instance of this is the expression "King" - "Man" + "Woman" ≈ "Queen". In this project, we aim to investigate analogous linguistic associations, but for topic embeddings [1], and to establish connections between topic embeddings and word embeddings.

Requirements:

(a) Python (3+)

(b) PyTorch

(c) Deep Learning

(d) Basic NLP knowledge

Test Task: Document Classification
We have access to the 20 Newsgroups dataset (available here), which comprises ∼18K documents divided into 20 classes. The aim of this task is to perform text classification on the dataset by leveraging Neural Networks and PyTorch. While model performance is not the primary evaluation metric, the completeness of the project/pipeline will be assessed.

Note: Do not submit Jupyter Notebooks as the primary submission (you may use them for analysis). The objective is to evaluate your ability to handle substantial Python projects.

Consider the following essential points while completing the task:

(a) Learning from internet sources is permitted, but direct copying is prohibited.

(b) Use the bag-of-words (BoW) representation for the documents.

(c) Avoid using pre-trained models or model modules obtained from providers such as Huggingface.

(d) Organize your code, logs, plots, analysis, and other materials correctly.

(e) Please ensure that you include a comprehensive Readme file that thoroughly documents your work. Additionally, please include a Jupyter Notebook that contains your output analysis.

(f) Please upload your code to your personal GitHub repository and include a link to it in a .txt file.

References

[1] Adji B Dieng, Francisco JR Ruiz, and David M Blei. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020.

[2] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.