Auto-Tagging of Text Using Large Language Models (LLM)

Aleksandr Vashchenko
Email: a.vashchenko@innopolis.university

Grigoriy Nesterov
Email: g.nesterov@innopolis.university

1. Project idea

The project aims to develop an automated tagging system for textual data using advanced Large Language Models (LLMs). The auto-tagging system will be capable of assigning relevant tags to datasets, facilitating improved content organization, searchability, and information retrieval. This system will be used in InnoDataHub project.

2. Method/Technique

2.1. Data Collection and Preparation

- **Data Collection**: Gather metadata (title, subtitle, description, and tags) from various datasets.
- **Preprocessing**: Clean and normalize the text data to prepare it for model training

2.2. Model Selection

- **LLM Utilization**: Use state-of-the-art LLMs to generate tags.
- **Fine-Tuning**: Fine-tune the chosen model with the annotated dataset to learn effective tagging patterns.

2.3. Evaluation

- **Metrics:** Evaluate the accuracy of the tagging using precision, recall, and F1-score.

2.4. Deployment and Integration

- **API Service:** Deploy the auto-tagging system as an API service for easy integration with other applications.

3. Dataset Explanation and Accessible Link

Our dataset will be consist of metadata of datasets from Kaggle in JSON format.

4. Timeline with Individual Contributions

- **05.07-09.07:** Data collection and preprocessing (Aleksandr Vashchenko)
- **10.07-16.07:** Model selection and fine-tuning (Aleksandr Vashchenko & Grigoriy Nesterov)
- **17.07-20.07:** Evaluation and optimization (Aleksandr Vashchenko & Grigoriy Nesterov)
- **21.07-23.07:** Deployment and integration (Grigoriy Nesterov)

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, D. Amodei (2020). Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165.

[2] H. Kopka and P. W. Daly, A Guide to LATEX, 3rd ed., Harlow, England: Addison-Wesley, 1999.