Project Proposal: Enhancing SPLADE with Large Vocabularies in Neural Information Retrieval

Background and Objectives

Information retrieval (IR) focuses on developing models that can swiftly find a set of relevant documents in response to a user's query (e.g., Google searches). In the current landscape, the most effective IR models are neural network-based. Among these, SPLADE stands out for its efficiency and effectiveness. SPLADE operates on sparse lexical decomposition, where each document or query is represented in a vector space with components corresponding to terms from a pre-defined vocabulary (i.e., word parts). However, its performance can significantly diminish when handling large vocabularies. This internship aims to extend SPLADE's capabilities and explore whether its performance improves with larger vocabularies.

What you will learn

- Backpropagation in Neural Networks
- How to develop CUDA Kernels (in Python)
- Experiments in Information Retrieval
- How to write a paper

Key Goals

1. **Develop Efficient GPU Kernels (CUDA)**: The intern will design and implement optimized CUDA kernels to address computational bottlenecks in SPLADE. This effort aims to enable the model to process large vocabularies more swiftly and with reduced computational overhead.
2. **Parameter Initialization for Term Space Projection**: Optimal parameter initialization is essential for effectively projecting representations into the term space. The intern will develop methods to initialize these parameters efficiently, thereby enhancing SPLADE's capacity to manage a broad spectrum of terms without losing precision.
3. **Model Experimentation with Standard IR Collections**: The improved SPLADE model will undergo rigorous testing using standard IR collections such as MS MARCO (Microsoft Machine Reading Comprehension) and BEIR (Benchmarking IR). This evaluation will determine the model's effectiveness and efficiency in real-world scenarios, offering valuable insights into its performance with various datasets.

Duration and Location

- The supervisor is Benjamin Piwowarski benjamin.piwowarski@cnrs.fr