Topic Modeling on Academic Papers Using Latent Dirichlet Allocation (LDA)

Topic modeling has always intrigued me in natural language processing (NLP), and for my project, I decided to explore **topic modeling on academic papers**. This project will involve identifying key themes and topics discussed in a collection of academic research papers using the Latent Dirichlet Allocation (LDA) technique. By uncovering the main topics from large datasets, I hope to create a system that can efficiently categorize papers and provide insight into research trends.

The project’s objective is to help researchers quickly understand the topics within large collections of academic papers and group them into meaningful categories. This can offer value to anyone attempting to stay up to date on key themes in fast-growing research areas. By automating the process of extracting topics, I will gain a deeper understanding of NLP techniques and improve my ability to interpret and present textual data.

**Challenges**

One challenge I anticipate is selecting the appropriate number of topics that will provide material insights without oversimplifying the content. Preprocessing the data, such as dropping stop words and selecting an appropriate stemming or lemmatization method, will also be critical for successful topic modeling.

**Tools & Methods**

So far, my concrete steps are as follows:

1. I will use JupyterLab as my development environment.
2. Programming language will be Python.
3. Project will rely on LDA for topic modeling, this will require the use of libraries like Gensim, NLTK, and spaCy for preprocessing.
4. I will use pyLDAvis and matplotlib to visualize the topics.

**Data**

For the data, I am considering publicly available collections of academic papers from arXiv or PubMed, since they offer a variety of research articles. I will also consider data from Kaggle related to academic publications.

I look forward exploring the field of topic modeling and applying it to academic papers to make research literature more accessible.