Guided Projects Artificial Intelligence & Machine Learning

Guided Projects: Unsupervised Learning

PLSA: Text Document Clustering

PLSA or Probabilistic Latent Semantic Analysis is a technique used to **model information under a probabilistic framework**. It is a statistical technique for the analysis of two-mode and co-occurrence data. **PLSA characterizes** each word in a document as a sample from a mixture model, where **mixture components** are conditionally independent **multinomial distributions**. Its main goal is to model cooccurrence information under a probabilistic framework in order to discover the underlying **semantic structure** of the data.

Question:

Perform topic modelling using the **20 Newsgroup dataset** (the dataset is also available in sklearn datasets sub-module). Perform the required data cleaning steps using NLP and then model the topics

1. Using Latent Dirichlet Allocation (LDA).
2. Using Probabilistic Latent Semantic Analysis (PLSA)

Dataset Link: https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html