SOFIA UNIVERSITY “St. KLIMENT OHRIDSKI”
FACULTY OF MATHEMATICS AND INFORMATICS

FUNDAMENTALS OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

FINAL PROJECT

Project topic:

Classifying main topics from a collection of documents

Student:

Rangel Plachkov 0MI3400855

February, 2025

Task statement

The goal of this project is to analyze and classify documents from the 20 Newsgroups dataset into their respective topics. The project involves preprocessing the text, transforming it into numerical features using TF-IDF, building and evaluating multiple machine learning models, and performing a baseline experiment with a simple feedforward neural network.

Algorithms used

LinearSVC – Linear Support Vector Classifier for high-dimensional text.
Logistic Regression (OvR) – One-vs-Rest strategy for multi-class classification.
Multinomial Naive Bayes – Probabilistic model for text with TF-IDF features.
Feedforward Neural Network – Simple PyTorch MLP baseline using TF-IDF input.

Implementation description

The implementation starts by loading the 20 Newsgroups dataset and splitting it into training and test sets. Text preprocessing includes removing stop words and converting documents into numerical features using TF-IDF. Pipelines are used to combine vectorization with classifiers for easy training and evaluation. Models including LinearSVC, Logistic Regression (OvR), and Multinomial Naive Bayes are trained and compared. A simple feedforward neural network is implemented in PyTorch as a baseline. The top keywords per topic are extracted using TF-IDF scores to analyze and interpret the topics.

References

https://www.youtube.com/watch?v=ATK6fm3cYfI