**TOPIC SUMMARY**

|  |
| --- |
| Research Topic: Analyze product memory data on Amazon using clustering algorithms. |
| Instructor: Nguyen Ho Duy Tri |
| Working time: from 01/12/2024 to 19/12/2024 |
| Short Description:  This project focuses on implementing the K-means and DBScan clustering algorithms using PySpark to analyze memory product data, which is crawled directly from Amazon. The goal is to utilize the distributed computing power of Apache Spark for efficient data processing. The data undergoes several preprocessing steps, including MinMaxScaler for normalization and StringIndexer and OneHotEncoder to convert categorical features into numerical values, preparing the data for clustering. After preprocessing, the dataset is split into training and testing sets. The clustering algorithms are manually implemented from scratch without relying on machine learning libraries such as MLlib or Scikit-learn. Once the clustering is done, the results are analyzed to uncover patterns and groupings within the memory products. This approach demonstrates how Spark’s parallel processing capabilities can be leveraged to process large-scale data efficiently. The project highlights the practical aspects of data preprocessing, custom algorithm implementation, and the evaluation of results in a distributed environment, showcasing the potential of Spark for scalable machine learning solutions. |