УДК 004

НЕКЕРОВАНЕ НАВЧАННЯ ДЛЯ ГРУПУВАННЯ ДЕФЕКТІВ У ЛОГ-ФАЙЛАХ: ОЦІНКА ЯКОСТІ КЛАСТЕРИЗАЦІЇ ЗА ДОПОМОГОЮ КОЕФІЦІЄНТА СИЛУЕТУ ТА МЕТОДУ ГОЛОВНИХ КОМПОНЕНТ

Каяфюк А. Г. *– гр. ДФКН-24, аспірант, kaiafiuk.a@knutd.edu.ua*
Стаценко Д. В. *– к.т.н., доц., kems@knutd.edu.ua*
Київський національний університет технологій та дизайну

Мета роботи: дослідити вплив методів попередньої обробки та векторизації лог-файлів на швидкість навчання та точність моделей машинного навчання.

Матеріали та методи: у дослідженні використано датасет HDFS_v3_TraceBench, що містить 276 лог-файлів і понад 370 000 трасувань, частина з яких включає повідомлення про помилки. За допомогою регулярних виразів виділено повідомлення про помилки та сформовано pandas.DataFrame. Для векторизації тексту використано метод TF-IDF (Term Frequency-Inverse Document Frequency), реалізований через TfidfVectorizer зі стандартної бібліотеки Scikit-learn.

Формула для розрахунку TF

$$TF(t,\ d)\ =\ \frac{\text{Кількість разів, які термін } t \text{ присутній у документі } d}{\text{Загальна кількість термінів у документі } d}$$

Формула для розрахунку IDF

$$\mathrm{IDF(t,\ D)}\ =\ \frac{\text{Загальна кількість документів у корпусі D}}{\text{Кількість документів, що включає термін t}}$$

Формула для розрахунку TF-IDF

$$\mathrm{TF-IDF(t,\ d,\ D)}\ =\ \mathrm{TF(t,\ d)}\ \times\ \mathrm{IDF(t,\ D)}$$

Оскільки метою було виявлення кластерної структури, застосовано некероване навчання – зокрема, Gaussian Mixture Model (GMM). Оптимальну кількість кластерів визначено за коефіцієнтом силуету. Для візуалізації результатів після кластеризації використано PCA (2 компоненти) та matplotlib для оцінки якості розділення.

Висновки. Застосування TF-IDF, PCA та Gaussian Mixture дозволило ефективно кластеризувати повідомлення про помилки в лог-файлах. Коефіцієнт силуету вказав на оптимальну кількість кластерів – п'ять, що відповідає типовим збоям у HDFS: Data Disk Failure, Heartbeats and Re-Replication, Cluster Rebalancing, Data Integrity, Metadata Disk Failure.

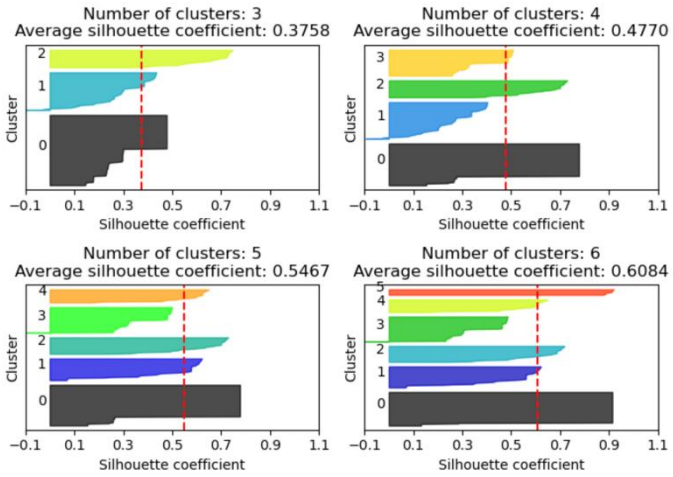

Рисунок 1 – Аналіз коефіцієнта силуету на основі кількості кластерів

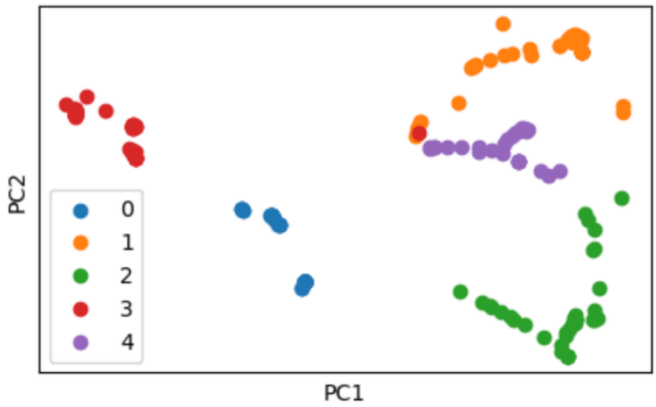

Рисунок 2 – Візуалізація кластерів отриманих в результаті застосування PCA

Візуалізація у двовимірному просторі після PCA підтвердила чітке розділення кластерів. Отримані результати свідчать про практичну цінність підходу – його можна використовувати для автоматичного групування помилок, виявлення нових типів збоїв і порівняння з дефектами, виявленими вручну. Некероване навчання спрощує аналіз логів, зменшує потребу в ручному маркуванні та покращує пошук причин помилок.

Література

1. Jingwen Zhou, Zhenbang Chen, Ji Wang, Zibin Zheng, and Michael R. Lyu. TraceBench: An Open Data Set for Trace-oriented Monitoring, in Proceedings of the 6th IEEE International Conference on Cloud Computing Technology and Science (CloudCom), 2014.

2. Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (3rd ed.). O'Reilly Media.

3. HDFS Architecture Guide. https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html