

12. Мовні моделі

Метою є навчитись використовувати обчислювальні ресурси та натреновані мовні моделі для розв’язання задач аналізу та генерації текстів.

Завдання:

- I. Перетворити у векторні представлення (embedding) текстові повідомлення датасету з ресурсу Kaggle [1], використовуючи натреновану мовну модель з ресурсу SentenceTransformers [2] та/або Hugging Face Model Hub [3];
- II. Поставити завдання великій мовній моделі LLM (Large Language Model) згенерувати 10 текстових послідовностей різної тематики та складності.

При розв’язанні першої задачі після безпосередньо векторизації текстів на основі навченої моделі, проводиться кластеризація векторних представлень текстових повідомлень за тематикою. Подальша візуалізація результатів кластеризації дозволяє наочно робити висновки щодо якості векторизації: тексти подібної тематики мають мати подібне векторне представлення та знаходитися в одному кластері.

Для розв’язання другої задачі використовуються LLM, яка не потребує налаштувань і після завантаження бібліотеки моделей обробки природної мови можна задавати запити з генерації текстів.

Виконання роботи має відбуватись в середовищі Colab [4], в якому підготовано шаблон розв’язання поставлених задач: *Шаблон роботи_12_LLM.ipynb* [5]. Зазначений шаблон містить інструкції та програмний код, необхідний для завантаження моделей та бібліотек навчання для виконання обох задач.

ЛІТЕРАТУРА ТА РЕСУРСИ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

1. Kaggle. URL: <https://www.kaggle.com/>
2. SentenceTransformers. URL: <https://sbert.net/index.html>
3. Models - Hugging Face. URL:
https://huggingface.co/models?pipeline_tag=sentence-similarity&sort=trending
4. Colab. URL: <https://colab.research.google.com/?hl=uk>
5. Шаблон_роботи_12_LLM.ipynb. URL:
<https://colab.research.google.com/drive/1DIKw4MV3OH4WixbC5yjO0lcQ1iVp1kZu?usp=sharing>