УДК 004.8, 374

АНАЛІЗ ЕФЕКТИВНОСТІ РОБОТИ СИСТЕМ КЛАСТЕРИЗАЦІЇ

ANALYSIS OF THE EFFICIENCY OF CLUSTERIZATION SYSTEMS

Ігор Коваль

Луцький національний технічний університет, вул. Львівська, 75, Луцьк, 43000, Україна

Abstract. *Analysis of the performance of clustering systems includes the selection of appropriate metrics, comparison of different algorithms, and assessment of the stability of the results. This allows you to draw reasonable conclusions about the quality of clustering for a specific data set and task.*

Кластеризація дозволяє розділити текст на групи за тими чи іншими ознаками, ознаки можуть бути задані вручну або виведені в процесі машинного навчання моделі. Алгоритм K-Means є одним з найпопулярніших алгоритмів кластеризації. На першому етапі алгоритм ініціалізує центри кластерів випадковим чином. Ітерація. На другому етапі алгоритм ітеративно розподіляє точки набору даних між кластерами, в яких вони знаходяться найближче до центрів кластерів. На рисунку 1 зображена перша стадія досліду, обробка тексту шляхом вилучення з нього так званих «stopwords».

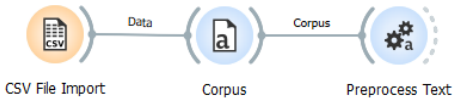

Рисунок 1

На рисунку 2 зображена друга стадія, це дозволить представити документ як вектор чисел, та застосувати методи машинного навчання, які працюють з числовими даними.

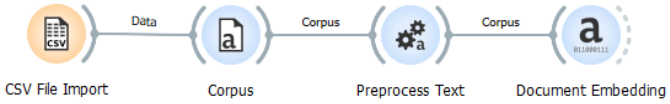

Рисунок 2

На рисунку 3 зображена кінцева модель нашого дослідження відносно часу кластеризації.

Рисунок 3

Результати вимірювання затраченого часу на кластеризацію текстових даних (далі TD) за допомогою аглоритму K-Means: 1600 TD – 5.5 секунд; 10000 TD – 32.3 секунд; 50000 TD – 198 секунд. Результати вимірювання затраченого часу на проходження усіх етапів моделі: 1600 TD – 27.2 секунд; 10000 TD – 126.2 секунд; 50000 TD – 393.4 секунд. Було досліджено поведінку в часі програмної системи ODM за допомогою алгоритму K-Means.

Бібліографія

1. How to use PIL. Режим доступу: https://pillow.readthedocs.io/en/stable/ (дата звернення: 15.04.2024)
2. How to use OpenCv. Режим доступу: https://pypi.org/project/opencv-python/ (дата звернення: 15.04.2024)