**Міністерство освіти і науки України**

**Національний технічний університет України «КПІ» імені Ігоря Сікорського**

**Кафедра обчислювальної техніки ФІОТ**

**ЗВІТ**

**з лабораторної роботи №5**

**з навчальної дисципліни «Вступ до технології Data Science»**

**Тема:**

**РЕАЛІЗАЦІЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ**

**(MACHINE LEARNING (ML))**

**Виконав:**

Студент 3 курсу кафедри ІПІ ФІОТ,

Навчальної групи ІП-11

Лошак В.І.

**Перевірив:**

Професор кафедри ОТ ФІОТ

Писарчук О.О.

**Київ 2023**

**І. Мета роботи:**

Виявити дослідити та узагальнити особливості аналізу даних з використанням методів

та технологій машинного навчання (Machine Learning (ML))

**ІІ. Завдання:**

**Завдання ІІІ рівня складності 9 балів:** реалізувати на вибір ТРИ з п’яти

сформованих груп технічних вимог.

**Група технічних вимог\_1:**

*Реалізувати кластеризацію вхідних даних, отриманих Вами у ході виконання Дз\_1,*

*модельних та (або) реальних – на власний вибір. Методи Machine Learning з переліку: k-means (k-середніх); Support Vector Machine (машина опорних векторів); k-nearest neighbors (найближчих сусідів); ієрархічна кластеризація – для кластеризації обраних даних обрати самостійно. Провести аналіз та пояснення отриманих результатів, сформувати висновки.*

**Група технічних вимог\_2:**

*Реалізувати кластеризацію за кольоровою ознакою об’єктів на самостійно обраному*

*цифровому зображенні. Методи Machine Learning з переліку: k-means (k-середніх); Support Vector Machine (машина опорних векторів); k-nearest neighbors (найближчих сусідів) – для кластеризації обраного зображення обрати самостійно.*

*За необхідності провести покращення якості зображення: зміна кольору; підвищення*

*контрасту; фільтрація, тощо. Етапи покращення якості та кластерізації повинні забезпечувати виділення геометричних або кольорових ознак обраного на цифровому зображенні об’єкту для його подальшої ідентифікації. Провести аналіз отриманих результатів, сформувати висновки.*

**Група технічних вимог\_3:**

*Підрахувати кількість об’єктів на обраному цифровому зображенні. Об’єкти, що*

*підлягають обрахунку обрати самостійно. Зміст етапів попередньої обробки зображень (корекція кольору, фільтрація, векторизація, кластеризація) має буди результатом R&D процесів, що конкретизується обраним зображенням і об’єктами для підрахунку. Провести аналіз отриманих результатів, сформувати висновки.*

**ІІІ. Результати виконання лабораторної роботи.**

1. ***Реалізувати кластеризацію вхідних даних***

В лабораторній здійснюється обробка даних з електричного мікроскопу наведених за посиланням: <https://www.kaggle.com/competitions/data-science-bowl-2018/data>

Функції для зчитування Id та власне самих зображень.

A screen shot of a computer program

Description automatically generated

Рис. 1 – Функції для читання вхідних даних в заданому форматі(rgb, hsv)

Для успішного виконання поставленого завдання необхідно кластеризувати дані за найактивнішим кольором що присутній у зображенні. Для кластеризації обрано алгоритм K-means що є простим і добре пристосований до задач такого типу. Щоб знайти оптимальні параметри цього алгоритму використовуємо метод Ліктя.

A group of graphs showing the size of a line

Description automatically generated with medium confidenceВибрано декілька зразків даних і знайдено оптимальний параметр K для них використовуючи вище наведений алгоритм.

Рис. 2— Оптимальні значення параметра кластеризації.

A computer screen shot of text

Description automatically generated

Рис. 3— Функція що використовує оптимальний параметр кластеризаці за замовчуванням. Використовується для екстракції домінантного кольору зображення. Повертає гістограму топ кольорів у зображенні .

Для отримання залишкових даних про всі зображення у вигляді зображення таких як ширина та висота використовуємо:

A screen shot of a computer code

Description automatically generated

Рис. 4— Функція для завантаження даних в програму(все крім самого зображення)

A screen shot of a computer code

Description automatically generated

Рис. 5— Функція для побудови графіку «ліктя».

Щоб визначити який параметр кластеризації використовувати для безпосередньо кластеризації зображень в датасеті по групах побудовано графік.

A graph with a blue line

Description automatically generated

Рис. 6— Доцільно використовувати параметр 3 щоб отримати три групи зображень

A screenshot of a computer

Description automatically generated

Рис. 7— Датафрейм з інформацією про зображення

Для кожного кластеру виведено нну кількість зображень що належать цьому кластеру на екран.

A collage of images of dots

Description automatically generated

Рис. 8—Кластер 1.

A collage of cells with numbers and images

Description automatically generated

Рис. 9—Кластер 2.

A collage of images of different shapes

Description automatically generated

Рис. 10—Кластер 3

Отже як видно з результату кластеризація пройшла успішно. Присутній дизбаланс в класах але загалом зображення розподілені за домінуючим кольором зображення.

1. ***Реалізувати кластеризацію за кольоровою ознакою об’єктів на самостійно обраному цифровому зображенні.***

Для аналізу використано перше зображення з другого кластеру в датасеті.

A screenshot of a computer screen

Description automatically generated

Рис. 11— Зображення для кластеризації кольорів.

Щоб краще розуміти що саме відбувається зображено всі пікселі присутні в зображенні на 3D графіку шкалами якого є відповідно R G та B показники зображення

A computer screen shot of a program code

Description automatically generated

Рис. 12— Функція для побудови графіку пікселів за кольорами.

A graph with a purple line

Description automatically generated with medium confidence

Рис. 13— перед застосуванням кластеризації пікселі зображення мають вигляд.

Для того щоб обрати оптимальний параметр кластеризації знову використано лікоть.

A graph with a line

Description automatically generated

Рис. 14— оптимальний параметр для кластеризації 3

A screenshot of a graph

Description automatically generated

Рис. 15 —Пікселі зображення забарвлені кольором свого кластера.

A screenshot of a computer screen

Description automatically generated

Рис. 16 — Реконструйоване зображення в зжатому просторі кольорів.

Залежно від кількості кластерів зображення буде використовувати різну кількість базових кольорів. В ході цієї лабораторної роботи використано алгоритмом підбору параметрів K-means для зменшення втрат під час стиснення зображення.

1. ***Підрахувати кількість об’єктів на обраному цифровому зображенні. Об’єкти, що підлягають обрахунку обрати самостійно.***

Об’єктом підрахунку є клітини що відображені на індивідуальному зображення. Для обробки використано те саме зображення що і в попередньому завданні. Для підрахунку обрано використовувати алгоритм K-means наступним чином. До dataframe використаного в попередньому завданні для кластеризації за кольорами додано також координати кожного пікселя в зображенні. Після цього локалізовано всі пікселі зображення що не належать до заднього плану в окремий датафрейм. На цьому датафреймі що містить лише пікселі клітин проведено пошук за відстанню. Таким чином знайдено кластери для пікселів клітин базуючись на їх близькості. Після підбору оптимального параметра кластеризації цей параметр за значенням дорівнює кількості клітин що видно на фото.

A purple square with white text

Description automatically generated

Рис. 17 — три базові кольори кластеризованого зображення.

A screenshot of a computer program

Description automatically generated

Рис. 18 — dataframe з доданою інформацією про положення пікселя.

A screenshot of a computer

Description automatically generated

Рис. 19 — Пікселі що є частиною клітини.

A graph of a number of clusters

Description automatically generated

Рис. 20 — Пошук оптимального параметра кластеризації з використанням ліктя. Оптимальним є значення параметра 30

A screenshot of a computer program

Description automatically generated

Рис. 20 — Комбінація dataframe без та з кластеризацією за відстанню в один. Присвоєння пікселям фона окремого кластеру.

A screenshot of a computer screen

Description automatically generated

Рис. 21 — Візуалізація кластерів за відстанню у вигляді зображення з рандомних кольорів для кожного окремо взятого кластера

**IV. Висновки.**

Результати лабораторної роботи показують потенціал і багатофункціональність застовування базового алгоритму K-means для досить складних процесів обробки зображень. Як для кластиризації зображень, так і для виділення features конкретно взятого зображення. Також його можна використовувати для побудови алгоритмів компресії, для стилізації зображень і в безлічі інших галузей.

Виконав: студент ФІОТ Лошак В.І. ІП-11