Міністерство освіти і науки України

Державний університет «Житомирська політехніка»

Факультет інформаційно-комп'ютерних технологій

Кафедра інженерії програмного забезпечення

**КУРСОВА РОБОТА**

з дисципліни: «Мови інтелектуального аналізу даних»

на тему **«Алгоритм автоматичного визначення кольору зображення»**

Студентів 1-го курсу, групи ІПЗм-19-1  
напряму підготовки 121 «Інженерія програмного забезпечення»

Бойко Таїси Олегівни

(прізвище, ім’я та по-батькові)

Керівник ст. викладач кафедри

ІПЗ Марчук Г.В.

Національна шкала\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Кількість балів:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Оцінка: ECTS \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Члени комісії \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Морозов А.В.

(підпис) (прізвище та ініціали)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Марчук Г.В. .

(підпис) (прізвище та ініціали)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Плечистий Д.Д.

(підпис) (прізвище та ініціали)

Житомир – 2020

ЖИТОМИРСЬКИЙ ДЕРЖАВНИЙ ТЕХНОЛОГІЧНИЙ УНІВЕРСИТЕТ

Факультет інформаційно-комп'ютерних технологій

Кафедра інженерії програмного забезпечення

Напрям підготовки 6.050103 «Програмна інженерія»

«ЗАТВЕРДЖУЮ»

Зав. кафедри інженерії програмного забезпечення

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

«\_\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2020 р.

**ЗАВДАННЯ**

на курсовий проект (роботу)

Студентки 1-го курсу, групи ІПЗм-19-1 Бойко Таїси Олегівни

Тема роботи: Алгоритм роботи автоматичного визначення кольору зобреження.

Строк подання студентом «\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_ 2020 р.

Вихідні дані роботи (зазначається предмет і об’єкт дослідження): об’єктом дослідження є прогнозування оцінки кредитоспроможності фізичних осіб; предметом дослідження є можливості застосування фреймворку ML.NET для прогнозування оцінки кредитоспроможності фізичних осіб.

Консультанти з випускної кваліфікаційної роботи із зазначенням розділів, що їх стосуються

|  |  |  |  |
| --- | --- | --- | --- |
| Розділ | Консультант | Підпис, дата | |
| Завдання видав | Завдання прийняв |
| 1 | Марчук Г.В. |  |  |
| 2 | Марчук Г.В. |  |  |
| 3 | Марчук Г.В. |  |  |

Керівник: Марчук Г.В. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(підпис)

**Календарний план**

|  |  |  |  |
| --- | --- | --- | --- |
| №  з/п | Назва етапів курсового проекту (роботи) | Термін виконання етапів роботи | Примітка |
| 1 | Постановка задачі. | 02.03.2020 | виконано |
| 2 | Аналіз проблематики, методів та засобів вирішення задачі. | 03.03.2020 - 20.03.2020 | виконано |
| 3 | Формуванння технічного завдання | 21.03.2020 - 30.03.2020 | виконано |
| 4 | Опрацювання літературних джерел | 1.04.2020 - 30.04.2020 | виконано |
| 5 | Вибір методів інтелектуального аналізу даних | 1.04.2020 - 24.04.2020 | виконано |
| 6 | Написання програмного коду - прогнозування | 25.04.2020 - 24.05.2020 | виконано |
| 7 | Аналіз результатів | 25.05.2020 - 27.05.2020 | виконано |
| 8 | Написання пояснювальної записки | 28.05.2020 - 30.05.2020 | виконано |
| 9 | Захист роботи | 10.06.2020 | виконано |

Студент: Бойко Т.О. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(підпис)

Керівник: Марчук Г.В. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(підпис)

РЕФЕРАТ

Завданням на курсову роботу було прогнозування оцінки кредитоспроможності фізичних осіб з використанням можливостей ML.NET.

Зміст

[РЕФЕРАТ 2](#_Toc52628445)

[ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ 4](#_Toc52628446)

[ВСТУП 5](#_Toc52628447)

[1. АНАЛІЗ ПРОБЛЕМАТИКИ, МЕТОДІВ ТА ЗАСОБІВ ВИРІШЕННЯ ЗАДАЧІ 7](#_Toc52628448)

[1.1 Аналіз задачі, засобів та методів її вирішення 7](#_Toc52628449)

[1.2 Методи і алгоритми розв’язання задачі визначення кольору зображення 7](#_Toc52628450)

[2. ПРАКТИЧНА РЕАЛІЗАЦІЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ 12](#_Toc52628451)

[2.1 Характеристика джерела даних для проведення аналізу 12](#_Toc52628452)

[3. РОЗДІЛ 3 13](#_Toc52628453)

[3.1 Взаємодія компонентів системи 13](#_Toc52628454)

[ВИСНОВОК 14](#_Toc52628455)

[ПЕРЕЛІК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ 15](#_Toc52628456)

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

AI (англ. Artificial Intelligence) – штучний інтелект.

ML (англ. Machine Learning) – машинне навчання.

MART (англ. Multiple Additive Regression Trees) – множинні аддиктивні регресійні дерева.

LBFGS (англ. Limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm) – aлгоритм Бройдена – Флетчера – Голдфарба – Шенно.

GDM (англ. Gradient boosting machine) – метод зростання градієнта.

SVM (англ. Support vector machine) – метод опорних векторів.

SDCA (англ. Stochastic Dual Coordinate Ascent) – стохастичний метод подвійного координатного сходження.

SGD (англ. Stochastic Gradient Descent) – метод стохастичного градієнта.

# ВСТУП

З стрімким розвитком інформаційних технологій кількість користувачів та їх даних стрімко збільшується щодня. В умовах цієї тенденції, виникла гостра необхідність змінювати підходи до зберігання та обробки даних, для того щоб впоратись з їх кількістю. В побутовому житті кожного з нас, на підприємствах та в державних установах невпинно зростає потік інформації, важливої для роботи. В результаті розвитку соціальних мереж та різноманітних сервісів для мультимедійних даних безперервно ростуть потреби в засобах, що надають можливість ефективно працювати з цими даними. Окрім збереження та обробки інформації, гостро стало питання аналізу даних користувачів. В результаті для органів державної влади й управління, телекомунікаційних та Інтернет-компаній, банків, підприємств роздрібної торгівлі, енергетики, житлово-комунальному господарстві накопичена інформація стає особливо важливим ресурсом, від ефективності управління яким суттєво залежить діяльності підприємств та компаній.

Для роботи з такою кількістю інформації постійно створюються нові апаратні та програмні засоби, які здатних оперативно обробляти великі обсяги даних і дозволяють робити процес обробки більш економічно вигідним. Саме тому, такі умови сучасного світу інформаційних технологій створили напрямок в інженерії програмного забезпечення, що носить назву великі дані.

Значення терміну великі дані має інтуєтивно зрозумілий характер. Великі дані (англ. Big Data) це набори структурованої або неструктурованої інформації які за розмірами настільки великі, що традиційні способи та підходи їх управління не можуть бути застосовані до них[]. Інколи, великими даними називають феноменальне прискорення нагромадження даних та їх ускладнення.

Галузь освіти не є виключенням серед інших галузей людської діяльності, що продукують велику кількість даних. Оцінки студентів, навчальні програми, методи оцінювання досягнень учнів та дані створені іншими процеси діяльності різних навчальних закладів. Вся ця інформація накопичується щоденно. Саме, важливо мати інстументи аналізу цих даних.

Виходячи з наданої інформації, можна встановити завдання до курсової роботи.

*Мета* курсової роботи. Проектування та аналіз даних про успішність вступників до вищих навчальних закладів на території України.

З мети можна сформувати наступні *завдання* до курсової роботи:

* Теоретичний аналіз проектування та реалізації OLAP-систем
* Вибір фактів та вимірів для збереження
* Проектування сховища даних
* Запонення сховища даних
* Вибір методів інтелектуального аналізу даних
* Реалізація звітності для сховища даних

*Предметом* дослідження є засоби аналізу та розробки сховища данних, проектування сховища даних.

*Об’єктом* дослідження вступники до ВНЗ України та їх успішність.

# АНАЛІЗ ПРОБЛЕМАТИКИ, МЕТОДІВ ТА ЗАСОБІВ ВИРІШЕННЯ ЗАДАЧІ

## Аналіз задачі, засобів та методів її вирішення

Інтелектуальний аналіз даних - це обробка інформації та виявлення в ній моделей і тенденцій, які допомагають приймати рішення. Необхідність інтелектуального аналізу даних виникла в кінці XX століття в результаті повсюдного поширення інформаційних технологій, що дозволяють детально протоколювати процеси бізнесу і виробництва. Великі обсяги даних, широту і різноманітність інформації привели до вибухового зростання популярності методів інтелектуального аналізу даних. Одним з результатів розвитку методів інтелектуального аналізу є можливість використання раніше зібраних даних для прийняття кращих рішень, наприклад, рішення видачі кредиту.

Метою курсової роботи є розробка додатку, який визначатиме основний колір наданого зображення. Результатом аналізу зображення повинне бути значення переважного кольору (або список основних кольорів) в одному з форматів: HEX або RGB. Дана задача є типовою задачею роботи комп’ютерного зору.

Для реалізації поставлено завдання, основними етапами є:

- обробка зображення;

- інтелектуальний аналіз даних;

- визначення результату.

## Методи і алгоритми розв’язання задачі визначення кольору зображення

Задача визначення кольору зображення є задачею кластерного аналізу.

Кластерний аналіз - задача розбиття заданої вибірки об'єктів (ситуацій) на підмножини, які називаються кластерами, так, щоб кожен кластер складався з схожих об'єктів, а об'єкти різних кластерів істотно відрізнялися. Завдання кластеризації відноситься до статистичної обробки, а також до широкого класу завдань навчання без вчителя.

Кластерний аналіз — це багатовимірна статистична процедура, яка виконує збір даних, що містять інформацію про вибірку об'єктів і потім упорядковує об'єкти в порівняно однорідні групи — кластери (Q-кластеризація, або Q-техніка, власне кластерний аналіз). Основна мета кластерного аналізу — знаходження груп схожих об'єктів у вибірці. Спектр застосувань кластерного аналізу дуже широкий: його використовують в археології, антропології, медицині, психології, хімії, біології, державному управлінні, філології, маркетингу, соціології та інших дисциплінах.

Розглянемо формальне визначення кластеризації:

Нехай - множина об'єктів, — множина номерів (імен, міток) кластерів. Задано функцію відстані між об'єктами . Є кінцева вибірка об'єктів . Потрібно розбити вибірку на непересічні підмножини, що називаються кластерами, так, щоб кожен кластер складався з об'єктів, близьких по метриці , а об'єкти різних кластерів істотно відрізнялися. При цьому кожному об'єкту приписується номер кластеру .

Алгоритм кластеризації — це функція , яка будь-якому об'єкту ставить у відповідність номер кластера . Множина в деяких випадках відома заздалегідь, проте частіше ставиться завдання визначити оптимальне число кластерів, з погляду деякого критерію якості кластеризації.

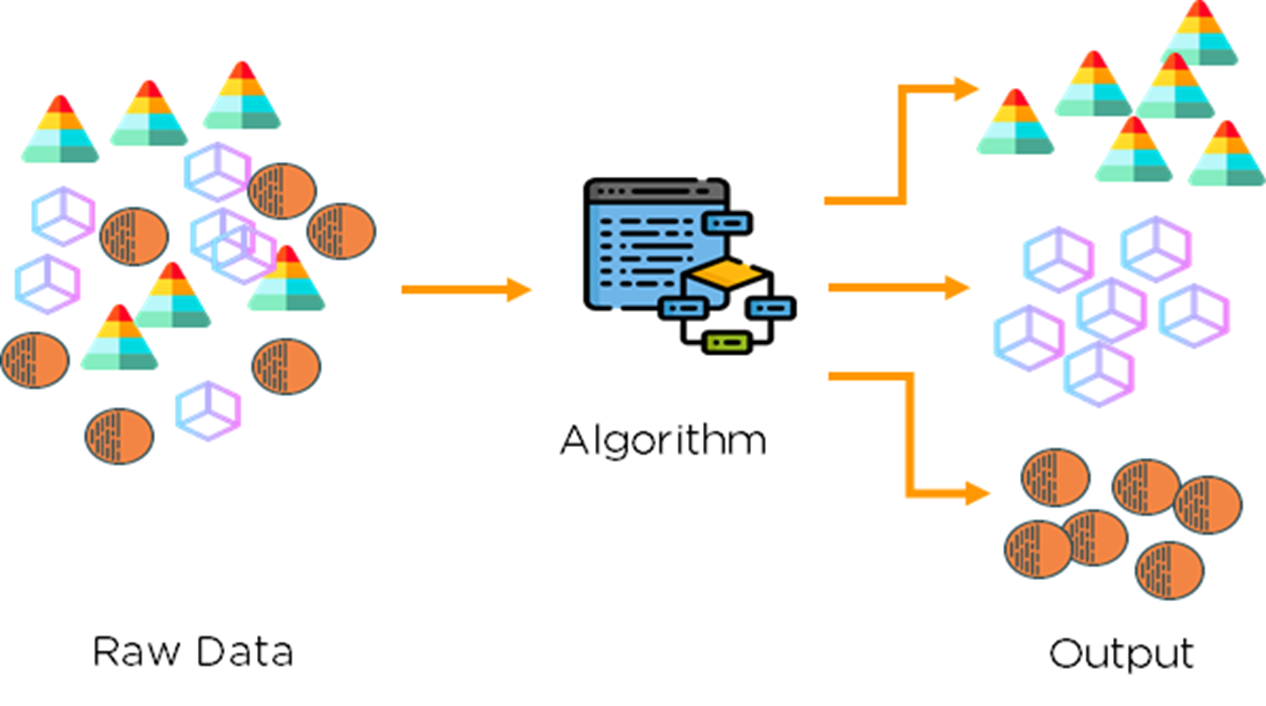


Рисунок 1.1 – Загальний принцип роботи алгоритмів кластеризації

Задачі кластеризації можуть вирішуватись різними алгоритмами. Власне задача визначення кольору розв’язується за допомогою методу k-середніх.

Кластеризація методом k-середніх - популярний метод кластеризації, — впорядкування множини об'єктів в порівняно однорідні групи. Винайдений в 1950-х роках математиком Гуґо Штайнгаузом і майже одночасно Стюартом Ллойдом. Особливу популярність отримав після виходу роботи МакКвіна.

Головні переваги методу k-середніх — його простота та швидкість виконання. Метод k-середніх більш зручний для кластеризації великої кількості спостережень, ніж метод ієрархічного кластерного аналізу (у якому дендограми стають перевантаженими і втрачають наочність).

Одним із недоліків простого методу є порушення умови зв'язності елементів одного кластера, тому розвиваються різні модифікації методу, а також його нечіткі аналоги (англ. fuzzy k-means methods), у яких на першій стадії алгоритму допускається приналежність одного елемента множини до декількох кластерів (із різним ступенем приналежності).

Незважаючи на очевидні переваги методу, він має суттєві недоліки:

1. Результат класифікації сильно залежить від випадкових початкових позицій кластерних центрів
2. Алгоритм чутливий до викидів, які можуть викривлювати середнє
3. Кількість кластерів повинна бути заздалегідь визначена дослідником

Метод k-середніх є доволі простим і прозорим, тому успішно використовується у різноманітних сферах — маркетингових сегментаціях, геостатистиці, астрономії, сільському господарстві тощо.

Мета методу — розділити n спостережень на k кластерів, так щоб кожне спостереження належало до кластера з найближчим до нього середнім значенням. Метод базується на мінімізації суми квадратів відстаней між кожним спостереженням та центром його кластера, тобто функції

де - метрика, - і-ий об'єкт даних, а - центр кластера, якому на j-ій ітерації приписаний елемент .

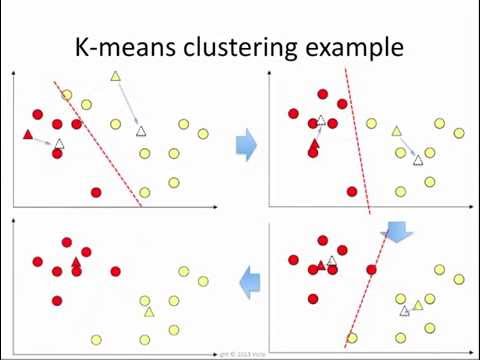


Рисунок 1.2 – Приклад виконання методу k-середніх

Опис алгоритму k-середніх

Маємо масив спостережень (об'єктів), кожен з яких має певні значення по ряду ознак. Відповідно до цих значень об'єкт розташовується у багатовимірному просторі.

1. Дослідник визначає кількість кластерів, що необхідно утворити
2. Випадковим чином обирається k спостережень, які на цьому кроці вважаються центрами кластерів
3. Кожне спостереження «приписується» до одного з n кластерів — того, відстань до якого найкоротша
4. Розраховується новий центр кожного кластера як елемент, ознаки якого розраховуються як середнє арифметичне ознак об'єктів, що входять у цей кластер
5. Відбувається така кількість ітерацій (повторюються кроки 3-4), поки кластерні центри стануть стійкими (тобто при кожній ітерації в кожному кластері опинятимуться одні й ті самі об'єкти), дисперсія всередині кластера буде мінімізована, а між кластерами — максимізована

Вибір кількості кластерів відбувається на основі дослідницької гіпотези. Якщо її немає, то рекомендують створити 2 кластери, далі 3,4,5, порівнюючи отримані результати. На рисунку 1.2 зображено приклад вирішення задачі методом k-середніх.

Сучасні зображення містять велику кількість кольорів та відтінків. Перераховування кожного кольору в зображенні може сформувати великий список даних, які потребують фільтрації. Для цієї фільтрації необхідно визначити основні кольори способом вирахування відмінності кольорів.

Формула кольорної відмінності (англ. Color difference) або колірна відстань (відстань між кольорами) - математичне уявлення, що дозволяє чисельно висловити відмінність між двома кольорами в колориметрії. Поширені визначення колірного відмінності зазвичай використовують формулу обчислення відстані в евклідовому просторі, проте варто зауважити що при цьому не кожне колірне простір є евклідовим з суворою математичної точки зору. Міжнародний комітет CIE (фр. Commission Internationale de l'Eclairage) задає визначення колірної різниці через метрику

ВИСНОВКИ ДО РОЗДІЛУ 1

У першому розділі курсової роботи було розглянуто теоретичні засади виконання задачі визначення кольору зображення. Було детально описано задачу кластеризації, її алгоритми та метод розв’язання задач кластеризації методом k-середніх.

# ПРАКТИЧНА РЕАЛІЗАЦІЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

## Характеристика джерела даних для проведення аналізу

В якості джерела даних для проведення аналізу використовуються користувацькі зображення.

ВИСНОВКИ ДО РОЗДІЛУ 2

У другому розділі курсової роботи було розроблено

# РОЗДІЛ 3

## Взаємодія компонентів системи

ВИСНОВКИ ДО РОЗДІЛУ 3

У третьому розділі курсової роботи було створено

# ВИСНОВОК

В ході виконання курсової роботи була спроектована

# ПЕРЕЛІК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Zimmer, Walter (1995), "Relationships between Design Patterns"
2. Software Design Patterns [Електронний ресурс] – Режим доступу: https://www.geeksforgeeks.org/software-design-patterns/

ДОДАТКИ