МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
ІМЕНІ ІГОРЯ СІКОРСЬКОГО”

Факультет прикладної математики
Кафедра програмного забезпечення комп’ютерних систем

КУРСОВИЙ ПРОЕКТ
ТЕХНІЧНЕ ЗАВДАННЯ

з дисципліни “Бази даних”

спеціальність 121 – Програмна інженерія

на тему: “Моніторингова система рейтингу фільмів у рейтингових сервісах”

Студентка
групи КП-81 **Мозгова Катерина Олегівна** _____
(підпис)
(ПІБ)

Викладач
к.т.н, доцент кафедри
СПіСКС **Петрашенко А.В.** _____
(підпис)

Київ – 2021

Назва: моніторингова система рейтингу фільмів у рейтингових сервісах.

Галузь застосування розробки: використання даних для оцінки популярності фільмів.

Дата початку проекту: 18.03.2021

Дата закінчення проекту: 21.06.2021

Мета розробки: збір, фільтрація та аналіз рейтингів фільмів різних жанрів з метою визначення їх популярності серед глядачів. Результати можуть бути використані власниками онлайн-кінотеатрів для підвищення продажу реклами шляхом підбору найбільш популярних фільмів чи покращення рейтингової підбірки в цілому.

1. **Вимоги до програмного забезпечення**
    - **Підсистема попередньої обробки даних містить у собі:**
        - Засоби генерації даних: має бути реалізовано ПЗ для збору інформації про рейтинги кінострічок в різних онлайн-кінотеатрів.
    - **Засоби фільтрації та валідації даних:**
        - Доповнення ПЗ з попереднього пункту функціоналом фільтрації та перевірки даних на коректність, відсіювання дублікацій.
    - **База даних:**
        - MongoDB
    - **Засоби реплікації даних:**
        - оскільки для використання у даній розробці була обрана нереляційна база даних MongoDB, то як засіб реплікації використовуватимемо реплісети (Replica Sets).
    - **Засоби масштабування:**
        - шардинг (sharding), який використовує MongoDb для обробки великої кількості даних.
    - **Засоби аналізу даних:**
        - NumPy – бібліотека для роботи із великими масивами даних.

    - Matpotlib – бібліотека для візуалізації даних у вигляді 2D і 3D графіків.
    - Pandas – бібліотека, яка використовуватиметься як надбудова до NumPy для структуризації роботи із масивами даних.
- **Задачі аналізу даних:**
    - Структурувати усі отримані дані з різних веб-ресурсів про фільми: назва, режисер,оцінка на ресурсах.
    - Здійснити валідацію даних та прибрати зайву інформацію із усіх масивів даних.
    - Об’єднати інформацію та заповнити пусті місця у даних.
    - Провести кореляцію по всіх даних.
    - У місцях, де найбільша кореляція, здійснити більш детальний аналіз.
    - Знайти зв'язок між режисером та середнім рейтингом його фільмів.
- **Засоби резервування та відновлення даних:**
    - Передбачені при використанні Replica Sets у MongoDb.

2. Вимоги до бази даних

Генерування даних має відбуватись відповідно до вимог масштабування: а саме, має бути проілюстровано, що на двох комп’ютерах ефективність обробки та аналізу даних підвищується. Підтвердити ці положення результатами дослідження: таблицями, графіками, діаграмами.

3. Обґрунтування вибору СКБД

Була обрана найпопулярніша серед нереляційних СКБД - MongoDb. Це документо-орієнтована система керування базами даних із відкритим кодом, яка не потребує опису схеми таблиць. MongoDB займає нішу між швидкими і масштабованими системами, що оперують даними у форматі ключ/значення, і реляційними СКБД, функціональними і зручними у формуванні запитів.

Вибір нереляційної СКБД обгрунтовується наявністю великої кількості ненормалізованих даних, які необхідно обробити швидко. Вона забезпечує можливість отримання неприведених до норм даних та подальшу роботу із ними. Формування додаткових таблиць при появі додаткової інформації у екземплярі (реляційні СКБД) було б недоцільним.

4. Вимоги до інтерфейсу користувача

Інтерфейс користувача має бути мінімалістичним саме тому він буде консольним. Задачею інтерфейсу користувача є налаштування засобів та підсистем, запуск/завершення їх роботи, генерація звітної інформації (графіків, діаграм тощо) у вигляді збережених файлів- зображень. Звітна інформація стосується візуалізації роботи засобів аналізу даних.

5. Вибір засобів розробки

Мова програмування: Python 3.7

Бібліотеки:

- *scrapy:* генерація даних
- *sklearn*: (бібліотека алгоритмів машинного навчання) класифікація досліджених даних;
- *pandas*: обробка та аналіз даних, використовується для первинної обробки даних;
- *numpy*: робота із великими масивами даних
- *matplotlib*: графічне представлення даних

6. Етапи розробки

№	Назва етапів розроблення	Термін виконання
1	Підготовка технічного завдання на затвердження	18.03.2021
2	Аналіз постановки задачі	23.03.2021

3	Розробка засобів генерації даних.	27.03.2021
4	Додавання засобів фільтрації та валідації даних.	09.04.2021
5	Реалізація збереження, реплікації та масштабування інформації розробленої моніторинг-системи.	13.04.2021
6	Підключення та розробка засобів обробки даних	20.04.2021
7	Налаштування резервування та відновлення даних	27.04.2021
8	Тестування програми	10.05.2021
9	Аналіз результатів. Підготовка матеріалів курсового проекту та оформлення пояснювальної записки	15.05.2021
10	Захист курсової роботи	22.06.2021