**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**

з дисципліни “Бази даних”

спеціальність 121 – Програмна інженерія

на тему: Система збору та аналізу даних про любительскі статті

**Студент**

**групи** ​КП-72 Кузнєцов Ілля Іванович \_\_\_\_\_\_\_\_\_

**Викладач**

**к.т.н, доцент кафедри** Петрашенко А.В.\_\_\_\_\_\_\_\_\_

**СПіСКС**

Захищено з оцінкою \_\_\_\_\_\_\_\_\_\_

Київ – 2020

## Зміст

[Зміст 1](#_Toc44045466)

[Вступ 2](#_Toc44045467)

[Аналіз інструментарію для виконання курсового проекту 3](#_Toc44045468)

[Структура бази даних 3](#_Toc44045469)

[Збір даних 4](#_Toc44045470)

[Аналіз даних 4](#_Toc44045471)

[Резервне копіювання та відновлення 4](#_Toc44045472)

[Висновки 5](#_Toc44045473)

## Вступ

Метою розробки даного курсового проекту є набуття виконавцем (студентом) практичних навичок розробки сучасного програмного забезпечення, що взаємодіє з постреляційними базами даних, а також здобуття навичок оформлення відповідного текстового, програмного та ілюстративного матеріалу у формі проектної документації. У результаті виконання курсового проекту студент повинен вміти розробляти програмне забезпечення для постреляційних баз даних, володіти основами використання СУБД, а також інструментальними засобами аналізу великих обсягів даних.

Темою даного курсового проекту є створення системи для збору та аналізу даних про любительскі статті. У документі викладені аналіз використаного інструментарію (опис мови програмування, використаних бібліотек та СКБД), описана структура бази даних, опис розробленого програмного забезпечення, та опис результатів виконання завдання.

## Аналіз інструментарію для виконання курсового проекту

В даному курсовому проекті використовується Redis. Це база даних типу ключ-значення. Вона не дозволяє зберігати складні схеми даних, але підходить для обраної задачі, бо використана схема даних дуже проста, кожен пакет зберігає невелику кількість даних, а підтримка множин як структури даних дозволяє реалізувати швидкий пошук саме за необхідними ключами.

Мова програмування Python3 була обрана через наявність великої кількості релевантних бібліотек для роботи з БД, обробки, та візуалізації даних, а також через простоту розробки

Ключові використані бібліотеки:

* os, json - для обробки даних
* redis - клієнт для Mongo
* requests, bs4 – для збору реальних даних

## 

## Структура бази даних

Дані про любительські статті зберігаються у множині структур даних, які визначають всю необхідну інформацію, яка корисна для даної задачі, про статтю.

**Структура даних:**

**- title – назва статті, виступає у ролі ключа в заданій таблиці, оскільки в даному контексті вона унікальна.**

**- rating – поле значення, яке вказує на різницю кількостей позитивної та негативної оцінок читачами. У даній задачі потребується для обрахування середньої оцінки серед n-ої кількості статей**

## 

## Збір даних

Збір даних виконується за допомогою фреймворків **requests, bs4**. Завдяки цим фреймворкам можливо збирати реальні дані про статті, які були опубліковані. Наразі збір даних виконуються з веб-сайту http://www.scp-wiki.net. Оскільки дані реальні та майже не вичерпні, можна зробити висновок, що виконана робота буде мати справедливі та корисні результати. Також задача буде можливою для вирішення при її ускладнені

## Аналіз даних

За допомогою бібліотеки redis був проведений аналіз вхідних даних. А саме у роботі було використано метод пошук середьного арифметичного значення послідовності, що дало змогу знайти середню оцінку статей серед вибірки. Також було визначено найбільш оцінену читачами сторінку. Все це може буде використано для оцінки тегів до яких ці статті прив’язані. Це не входить у початкову задачу але може бути поставлено на меті, якщо задачу ускладнювати.

## 

## Резервне копіювання та відновлення

Mongo не надає вбудованих механізмів резервного копіювання, але має можливість запису стану бази даних у файл та відновлення з цього файлу. На основі цього релізована проста система яка дозволяє робити резервну копію в будь-який момент та відновлюватись до будь-якої зі зроблених копій.

## Висновки

В процесі виконання даного курсового проекту було отримано практичні навички обробки великих масивів даних за допомогою мови програмування Python 3 та СКБД Redis.

Були проаналізовані NoSQL бази даних та обрана така, що найкраще відповідає даній задачі: Redis.

Для забезпечення горизонтального масштабування були використані вбудовані засоби Redis. Дані були слабко зв’язані між собою, що сильно спростило задачу.

Система дозволяє на основі зібраних даних аналізувати поток пакетів між адресами: Які адреси найбільш активно передавали пакети та з якими самими іншими адресами відбувалась передача.

Поставлена мета була виконана частково: було набуто практичних навичок розробки програмного забезпечення, що взаємодіє з NoSQL базами даних, а також були здобуті навички оформлення відповідного текстового, програмного та ілюстративного матеріалу у формі проектної документації. Складний статистичний аналіз даних не проводиться, але демонструється потенційна можливість такого аналізу