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: **Моніторингова система новинних повідомлень в Інтернет**

Студент

групи КП-73 Защик Михаїл Олександрович

(підпис)

Викладач

к.т.н, доцент кафедри

СПіСКС **Петрашенко А.В.**

(підпис)

Київ – 2020

АНОТАЦІЯ

Даний курсовий проект створений для аналізу популярних новин за певний проміжок часу. Даний документ призначений для звітності процесу розробки програмного забезпечення зазначеної вище системи. У документі викладена актуальність та проблематика аналізу великого обсягу даних, аналіз використаного інструментарію (опис мови програмування, порівняння використаних бібліотек та бази даних), описана структура бази даних (з графічною частиною), опис програмного забезпечення (загальний, опис модулів та основних алгоритмів роботи), аналіз функціонування засобів масштабування, та опис результатів аналізу всієї системи (з графічною частиною та фрагментами коду).

Як результат даного проекту стала інтерактивна аналітика досліджуваних даних (можна переглянути у Додатку А у вигляді графічної частини).

ЗМІСТ

ВСТУП

Новини - це середовище, яке весь час змінюється. Нові статті публікуються, старі статті прочитуються та забуваються. Аналіз інформації, пов’язаної з такими публікаціями може виявити багато цікавого та корисного, допомогти передбачити прийдешні події.

Також аналіз новин може передбачити різні реакції людей на ті чи інші обставини, полегшити обробку статистичних даних або допомогти при їх візуалізації.

Даний проект створюється в навчальних цілях в рамках дисципліни «Бази даних». Метою розробки даного курсового проекту виконання комплексного завдання щодо створення програмного забезпечення – моніторингу та аналізу новин.

Даний курсовий проект призначений для аналізу популярних новин, а також виявлення закономірностей в появі тих чи інших даних.

АНАЛІЗ ІНСТРУМЕНТАРІЮ ДЛЯ ВИКОНАННЯ КУРСОВОГО ПРОЕКТУ

Використана **мова програмування Python 3.8**[(1)]**.** Для даної курсової роботи було обрано мову програмування Python, тому що вона забезпечує ефективне виконання завдань системної інтеграції. Системна інтеграція в інженерії — поєднання компонентів підсистем в єдину систему та забезпечення роботи окремих підсистем як єдиної системи. системна інтеграція є процесом об'єднання різних обчислювальних систем і програмних застосунків фізично або функціонально[(2)].

Python відрізняється двома особливостями, корисними для системної інтеграції і, які не підтримуються багатьма іншими мовами. Одна - типи даних, явно розроблені для зв'язування (interfacing); інша - качина типізація в поєднанні з малим, але гнучким набором стандартних інтерфейсів.

Python портований і працює майже на всіх відомих платформах —від КПК до мейнфреймів.

Серед основних переваг мови Python можна назвати такі:

- Чистий синтаксис (для виділення блоків слід використовувати відступи);
- переносність програм (що властиве більшості інтерпретованих мов);
- стандартний дисрибутив має велику кількість корисних модулів (включно з модулем для розробки графічного інтерфейсу);
- можливість використання Python в діалоговому режимі (дуже корисне для експериментування та розв'язання простих задач);

- стандартний дистрибутив має просте, але разом із тим досить потужне середовище розробки, яке зветься IDLE і яке написане на мові Python;
- зручний для розв'язання математичних проблем (має засоби роботи з компклесними числами може оперувати з цілими числами довільної величини, у діалоговому режимі може використовуватися як потужний калькулятор);

відкритий код (можливість редагувати його іншими користувачами).

Використано базу даних **MongoDB**[3]. Це найпопулярніша серед нереляційних СКБД, документо-орієнтована система керування базами даних із відкритим вихідним 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 даної СКБД обгрунтовується тим що вона забезпечує можливість отримання неприведених до норм даних та подальшу роботу із ними, просто кодується і управляється, швидко працює на множині машин не обмежена експлуатаційними технологіями.

Бібліотеки

- **matplotlib** — це бібліотека Python 2D, яка представляє числові дані у різноманітних форматах та інтерактивних середовищах на різних платформах. Також ця бібліотека - це математичне розширення NumPy. Він надає об'єктно-орієнтований API для вбудови ділянок у додатки, що використовують набір інструментів для загального графічного інтерфейсу, таких як Tkinter, wxPython, Qt або GTK+. Matplotlib може

використовуватися в скриптах Python, оболонках Python та IPython, серверах веб-додатків та чотирьох графічних наборів інструментів для користувацького інтерфейсу. Бібліотека sciPy використовує matplotlib[4].

- **nltk** — це пакет бібліотек і програм для символьної і статистичної обробки природної мови, написаних на мові програмування Python. Містить графічні уявлення і приклади даних. Супроводжується великої документацією, включаючи книгу з поясненням основних концепцій, що стоять за тими завданнями обробки природної мови, які можна виконувати за допомогою даного пакету.

СТРУКТУРА БАЗИ ДАНИХ

База даних товарів була створена самостійно і є вміщує в собі необхідні для аналізу новин. Вона має одну таблицю marks з усіма даними про оцінку. Складається з наступних полів:

- _id – ідентифікатор запису, унікальне поле у базі даних, первинний ключ. Тип даних - ObjectId;
- source – об’єкт джерела новин
    - id.- Унікальний ідентифікатор джерела Тип даних - String;
    - name – назва джерела новин
- author – ім’я автора. Тип даних - String;
- title – Назва статті. Тип даних - String;
- description– Опис статті. Тип даних - String;
- url– посилання на статтю номер класу. Тип даних - String;
- urlToImage– посилання на картинку асоційовану з статтею. Тип даних - String;
- publishedAt– дата публікації статті оцінки. Тип даних - Datatime;
- content – частина контенту статті . Тип даних - String;

ОПИС ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

ЗАГАЛЬНА СТРУКТУРА ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

Структура даного програмного забезпечення складається з 7 файлів, кожен з яких відповідає за важливу функцію даної системи.

- **save.py –** Зберігає базу даних у файл збереження
- **restore.py –** Відновлює базу даних з файлу збереження
- **load.py –** Завантажує з API новини в базу дпних, попередньо провівши валідацію на наявність дублікатів новин
- **clear.py –** Очищає базу даних, попередньо записавши дані у файл збереження
- **word_count.py –** Аналізує дані про найбільш вживані слова з датафрейму та будує гістограму вживаності цих слів у заголовках на описах статей
- **total_amount.py –** Аналізує частоту публікацій новин кожної години доби та будує кругову діаграму
- **word_occurrence.py -** Аналізує кількість входжень слова кожен день та будує лінійну діаграму

ОПИС МОДУЛІВ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

У програмі можна виділити кілька основних логічних модулів, що мають певну самостійність і обмінюються один з одним даними. У програмі є 4 основні модулі: модуль завантаження даних з валідацією, модулі збереження/відновлення даних, модулі аналітики та роботи з графіками. Деякі з них містять вкладені модулі. Взаємодія між модулями відбувається за допомогою виклику методів, але у більшості випадків дані для роботи отримуються безпосередньо з бази даних.

- **Модуль завантаження даних** містить функції, що виконують запис у базу даних з API.
- **Модуль збереження/відновлення –** відповідає за збереження даних, очистку БД та можливість відновлення даних.
- **Модулі аналітики та роботи з графіками** взаємодіє з згенерованими даними аналізуючи, впорядковуючи та візуалізуючи дані.

ОПИС ОСНОВНИХ АЛГОРИТМІВ РОБОТИ

- **Алгоритм побудови графіків** було реалізовано за допомогою потужного математичного пакету matplotlib. Найчастіше це перший пакет, пов'язаний з візуалізацією у Python. Спочатку задаємо назву графіку, потім встановлюємо панель управління, задаємо за які атрибути відповідають X та Y, підписуємо виведені дані та виводимо графік. Всю внутрішню реалізацію містять у собі функції математичного пакету matplotlib.

АНАЛІЗ ФУНКЦІОНУВАННЯ ЗАСОБІВ МАСШТАБУВАННЯ

Масштабування[6] - можливість збільшити продуктивність проекту за мінімальний час шляхом додавання ресурсів. Зазвичай масштабування на увазі не переписування коду, а або додавання серверів, або нарощування ресурсів існуючого. З цього типу виділяють вертикальну та горизонтальну масштабування. Вертикальне - це коли додають більше оперативної пам’яті, дисків і т.д. на вже існуючий сервер, а горизонтальне - це коли ставлять більше серверів в дата-центри, і сервери там взаємодіють.

В базі даних MongoDB, як і в будь-якій NoSQL базі, не виникає проблем з горизонтальним масштабуванням[7], що і необхідно було у даному курсовому проекті для створення засобів резервування та відновлення даних, призначених для оперативного пакетного збереження фрагментів усієї бази даних з можливістю її відновлення з урахуванням необхідності підключення додаткового комп’ютера як елемента горизонтального масштабування (тобто додаткового сервера).

ОПИС РЕЗУЛЬТАТІВ АНАЛІЗУ ПРЕДМЕТНОЇ ГАЛУЗІ

У додатку А представлено 5 графіків демонстрації результатів аналізу американських новин за останні 15 днів.

На рис. 1 представлено фрагмент даних, що зберігаються в базі.

На рис. 2 представлено гістограму найбільш вживаних слів за останні 15 днів.

На рис. 3 представлено кругову діаграму з відсотками статтей та часом їх публікації.

На рис. 4 представлено лінійну діаграму зміни вживаності слова впродовж 15 днів. Наведено 3 моделі для слів coronavirus, pandemic, Trump.

ВИСНОВОК

Була успішно опрацьована відповідна технічна література для успішного написання курсового проекту (перелік літератури наведено нижче). В ході розробки даного програмного забезпечення були додані засоби валідації та фільтрації даних з достатньо великою швидкістю. Були додані засоби аналізу даних та графічне представлення результатів за допомогою існуючих математичних пакетів мови Python. Програмне забезпечення було успішно протестовано на декількох комп'ютерах, швидкість роботи бази даних достатньо велика. Також було реалізовано реплікації бази даних.

В ході виконання даного курсового проекту було досягнуто поставленої мети: було набуто практичних навичок розробки сучасного програмного забезпечення, що взаємодіє з базами даних, а також були здобуті навички оформлення відповідного текстового, програмного та ілюстративного матеріалу у формі проектної документації.

ЛІТЕРАТУРА

1. Python – Вікіпедія [Електронний ресурс]:

   https://uk.wikipedia.org/wiki/Python;

2. Почему Python так хорош в научных вычислениях [Електронний ресурс]. – 2018. – Режим доступу до ресурсу: https://habr.com/post/349482/;

3. MongoDB – Вікіпедія [Електронний ресурс]:

   https://uk.wikipedia.org/wiki/MongoDB:

4. Matplotlib [Електронний ресурс]. – Режим доступу до ресурсу: https://en.wikipedia.org/wiki/Matplotlib;

5. NumPy [Електронний ресурс]. – Режим доступу до ресурсу: https://en.wikipedia.org/wiki/NumPy;

6. Горизонтальное масштабирование: когда и как? [Електронний ресурс]. – Режим доступу до ресурсу: https://habr.com/company/oleg-bunin/blog/319526;

7. 5 причин використовувати Монго - [Електронний ресурс]: http://echo.lviv.ua/dev/9693

ДОДАТКИ

А. ГРАФІЧНІ МАТЕРІАЛИ

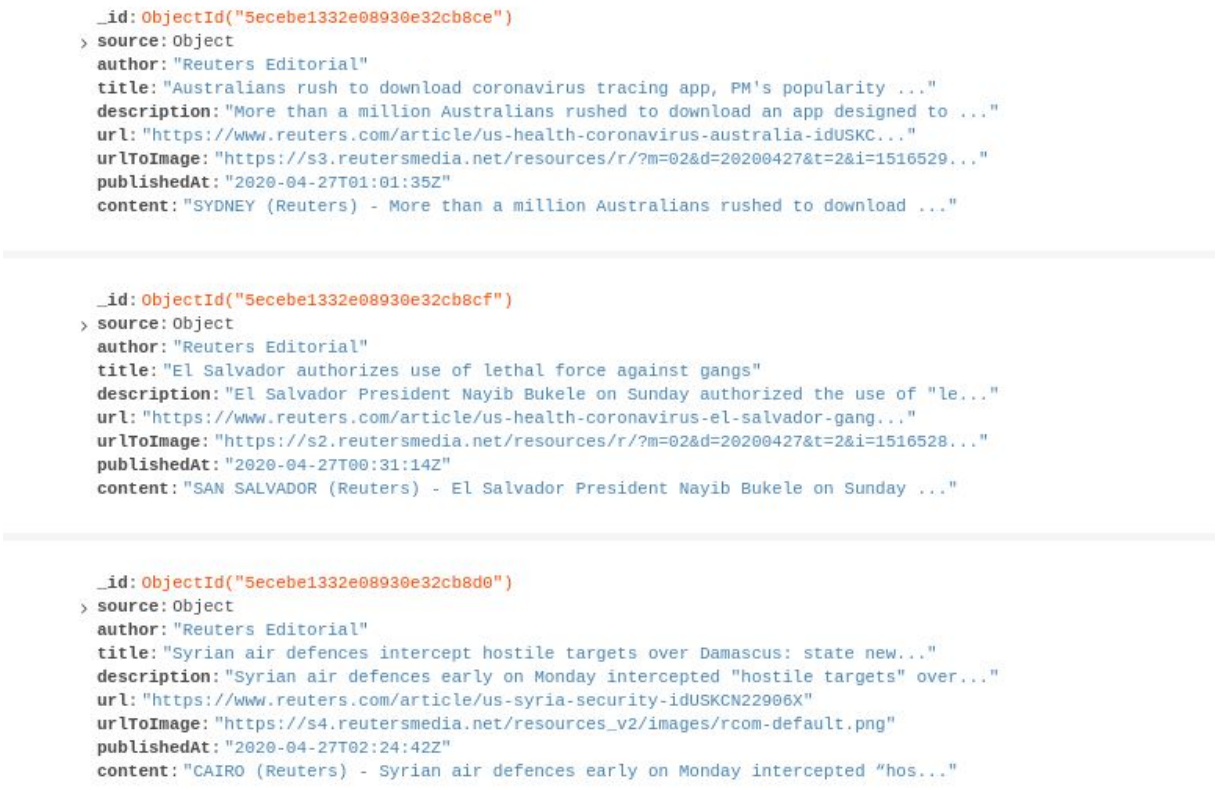

Рис. 1 – Дані в БД

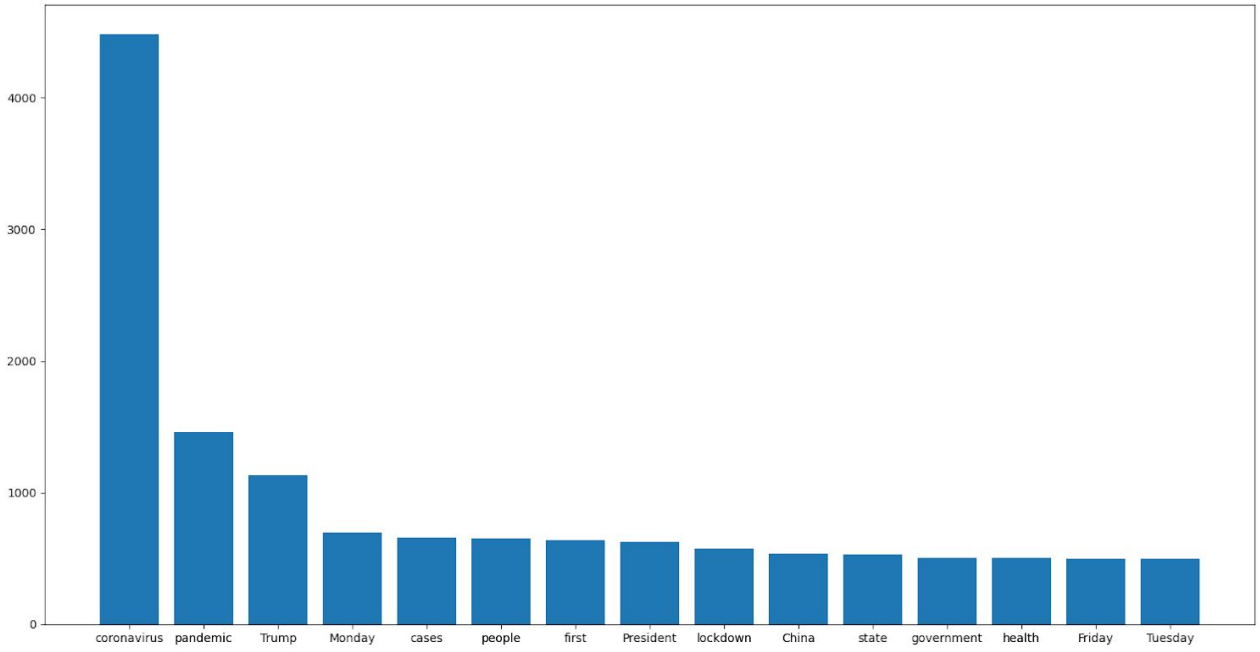

Рис. 2 – Гістограма найпопулярніших слів

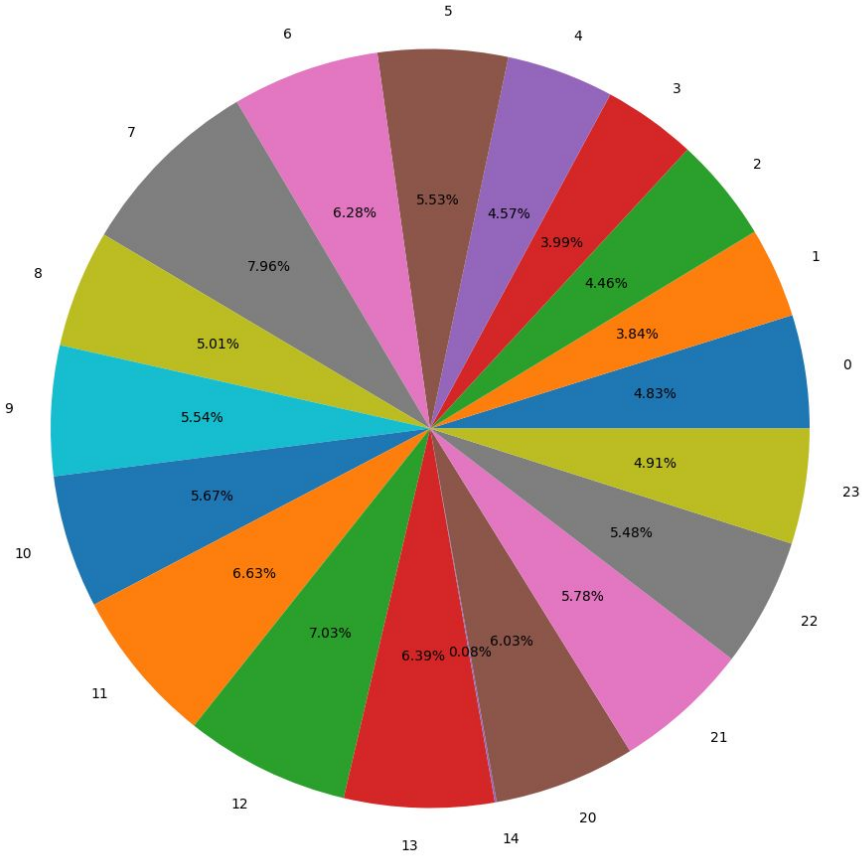

Рис. 3 – Кругова діаграма з відсотками статтей та часом їх публікації

Рис. 4.1 – Лінійна діаграма популярності слова coronavirus

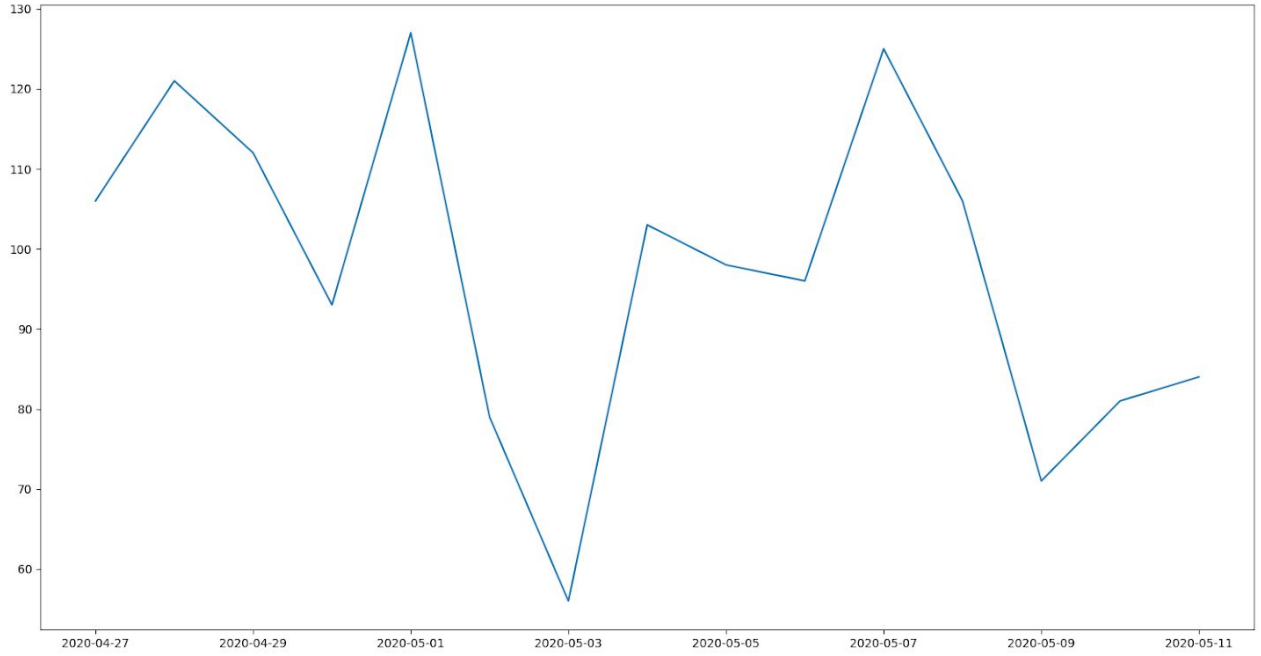

Рис. 4.2 – Лінійна діаграма популярності слова pandemic

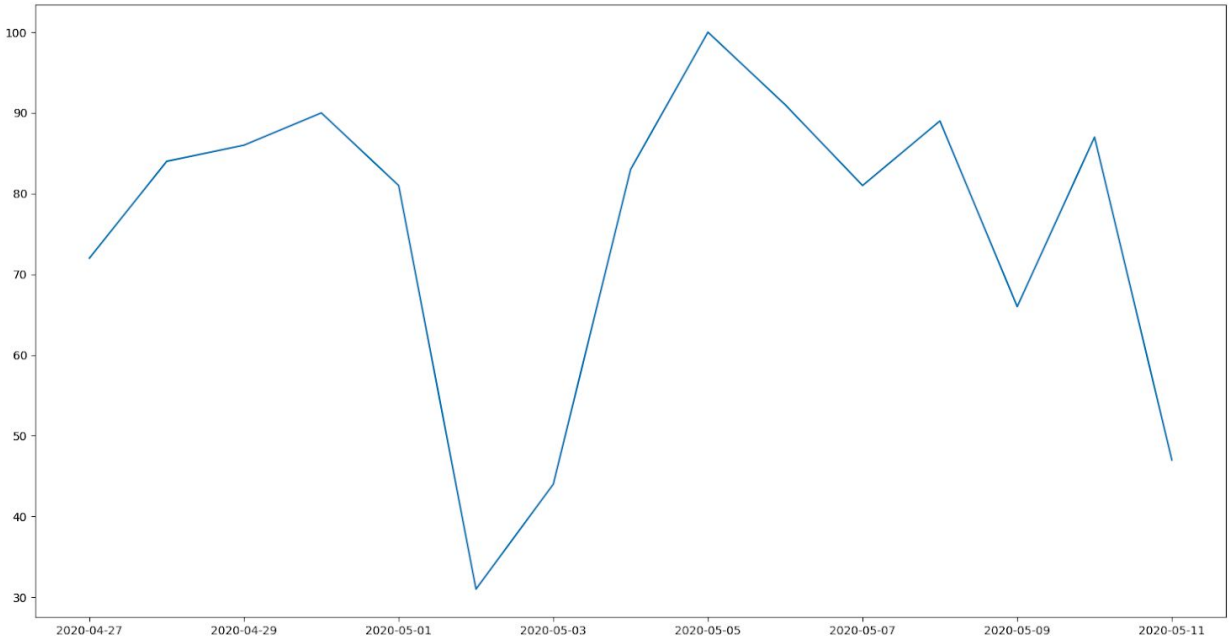

Рис. 4.3 – Лінійна діаграма популярності слова Trump