**Міністерство освіти і науки України**

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

**Кафедра обчислювальної техніки ФІОТ**

**ЗВІТ**

**з лабораторної роботи №1**

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

**Тема:**

**ПІДГОТОВКА ТА АНАЛІЗ ДАНИХ ДЛЯ СТАТИСТИЧНОГО НАВЧАННЯ**

**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:**

Виявити дослідити та узагальнити особливості застосування методів статистичного

навчання для задач визначення статистичних характеристик вхідного потоку даних з використанням спеціалізованих пакетів мови програмування Python.

**ІІ. Завдання:**

*Лабораторія провідної ІТ-компанії реалізує масштабний проект розробки універсальної платформи з обробки Big Data масиву статистичних даних поточного спостереження для виявлення закономірностей і прогнозування розвитку контрольованого процесу.*

*Платформа передбачає розташування back-end компоненти на власному хмарному сервері з наданням повноважень користувачам заздалегідь адаптованого front-end функціоналу універсальної платформи.*

*Замовниками ресурсів платформи є: державні та комерційні компанії валютного*

*трейдінгу для прогнозування динаміки зміни курсу валют та ціни інших товарів; метеорологічні служби для прогнозування параметрів метеоумов; департаменти охорони здоров’я для прогнозування зміни показників епідеміологічних ситуацій тощо.*

*Вам, як Data Science Engineer поставлено завдання.*

**Завдання ІІІ рівня – максимально 9 балів.**

1. Провести парсинг самостійно обраного сайту. Вміст даних, що підлягають парсингу – обрати самостійно.

2. Результати парсингу зберегти у файлі. Тип файлу обрати самостійно.

3. Оцінити динаміку тренду реальних даних.

4. Здійснити визначення статистичних характеристик результатів парсингу.

5. Синтезувати та веріфікувати модель даних, аналогічних за трендом і статистичними

характеристиками реальним даним, які є результатом парсингу.

6. Провести аналіз отриманих результатів.

**ІІІ. Результати виконання лабораторної роботи.**

**3.1. Провести парсинг самостійно обраного сайту. Вміст даних, що підлягають парсингу – обрати самостійно.**

Для цієї лабораторної роботи я обрав дані про курс криптовалюти bitcoin з сайту: <https://coinmarketcap.com/currencies/bitcoin/historical-data/>. Дані на цьому сайті презентуються в хронологічному порядку у вигляді таблиці що доступна до завантаження. Сайт також надає функціонал для вибору періоду часу на якому знаходяться необхідні дані. Для даної лабораторної роботи було обрано період з

14.09.2022 по 13.09.2023 сумарною тривалістю 365 днів.

Для парсингу відповідних даних було використано бібліотеку selenium в мові python. Для взаємодії з веб застосунком та вибору даних для завантаження було використано набір CSS селекторів. Для взаємодіє з браузером було використано двигун chromium.

Більшість даних не потребують подальшої обробки.

|  |  |
| --- | --- |
| Назва колонки | Опис |
| timeOpen | Мітка часу відкриття періоду історичних даних про Bitcoin |
| timeClose | Мітка часу закриття періоду історичних даних про Bitcoin |
| timeHigh | Мітка часу досягнення найвищої ціни Bitcoin за період |
| timeLow | Мітка часу досягнення найнижчої ціни Bitcoin за період |
| open | Ціна відкриття Bitcoin у початку періоду |
| high | Найвища ціна Bitcoin за вказаний період |
| low | Найнижча ціна Bitcoin за вказаний період |
| close | Ціна закриття Bitcoin в кінці вказаного періоду |
| volume | Обсяг торгів Bitcoin за вказаний період |
| marketCap | Капіталізація ринку Bitcoin за вказаний період |
| timestamp | Мітка часу, що надає додаткову інформацію про час |

Табл. 1 — Опис стовпців даних

A screen shot of a computer program

Description automatically generated

Рис. 1 — скрипт для парсингу даних з вебстайу.

**3.2. Результати парсингу зберегти у файлі. Тип файлу обрати самостійно.**

Після програмного натискання кнопки завантаження дані автоматично зберігаються в файлі Bitcoin\_9\_13\_2022-9\_13\_2023\_historical\_data\_coinmarketcap.csv, що дуже зручно оскільки цей формат підтримує інтеграцію з різноманітними бібліотеками обробки даних.

A screenshot of a computer

Description automatically generated

Рис. 2 — Збережений файл csv

**3.3. Оцінити динаміку тренду реальних даних.**

Для утворення часового ряду використано колонки *timeOpen* та *open* , хоча в цілях даної лабораторної можна використовувати будь-які з колонок що містять відповідно часові та значимі дані. Для візуалізації даних використано бібліотеку matplotlib.

A graph with blue lines

Description automatically generated

Рис. 3 — візуалізація часового ряду

Як видно з рисунку, ціна Bitcoin зростає протягом всього часу. Волатильність висока. Сезонності не виявлено( потребує подальшого дослідження), ряд являє собою часову надмірність(non-stationary time series). Загалом тренд зростаючий хоча кінець даних вказує на спадання вартості.

**3.4. Здійснити визначення статистичних характеристик результатів парсингу.**

Для знаходження статистичних характеристик використаємо бібліотеку Pandas. На наступному скріншоті наведені обчислені дисперсія, середнє арифметичне, медіана, середнє квадратичне та інші характеристики.

A blue rectangle with black lines

Description automatically generated

Рис. 4 — візуалізація часового ряду

**3.5.** **Синтезувати та веріфікувати модель даних, аналогічних за трендом і статистичними характеристиками реальним даним, які є результатом парсингу.**

Для моделювання даних використано LOESS(locally estimated scatterplot smoothing). Щоб підібрати оптимальний параметр захоплення даних(відсоток даних що використовується для побудови локальної регресії) необхідно розбити дані на тренувальні та тестові.

A screenshot of a graph

Description automatically generated

Рис. 5 — візуалізація розбитих даних

A screenshot of a computer program

Description automatically generated

Рис. 6 — Для заповнення прогалин в даних використано лінійну інтерполяцію

A blue line graph with text

Description automatically generated

Рис. 7 — Пошук оптимального значення параметра *frac* для моделювання трендової складової

A graph showing the price of bitcoin

Description automatically generated

Рис. 8 — Візуалізація трендової складової

Віднявши трендову складову від реальних даних отримано Значення випадкової складової яка в свою чергу має параметричну надмірність.

A graph showing a blue line

Description automatically generated

Рис. 9 — візуалізація випадкових залишків.

A graph of a distribution of residuals

Description automatically generated

Рис. 10 — візуалізація розподілу залишків.

На основі візуалізацій сформовано гіпотезу про нормальний розподіл даних. Перевірено використовуючи тест Шапіро-Уілкса.

A screenshot of a computer program

Description automatically generated

Рис. 11 — візуалізація розподілу залишків.

На результат тесту впливають аномалії що присутні в даних. Очищення аномалій:

A graph of a wave

Description automatically generated with medium confidence

Рис. 12 —Дані до очищення

**A blue line graph with white text

Description automatically generated**

Рис. 13 —Дані після очищення

A graph of a number of numbers and a number of numbers

Description automatically generated with medium confidence

Рис. 14 —Повторна перевірка гіпотези про нормальний розподіл.

Дані мають великий ексцес. Це означає що для моделювання оптимально використати Т розподіл(Стюдента). При відображенні розподілу даних згенерованих цією моделлю видно що хвости розподілу занадто видовжені і не відповідають реальним даним. Щоб позбутися цієї неточності використано обрізку розподілу на значеннях що відповідають кінцям діапазону на якому визначений розподіл реальних даних.

A graph of a tower

Description automatically generated with medium confidence

Рис. 15 —Розподіл згенерованих залишків та відображення залишків у вигляді часового ряду з параметричною надмірністю.

Використовуючи моделі тренду та залишків створено комбіновану модель часового ряду.

A screenshot of a graph

Description automatically generated

Рис. 16 — Згенерована модель у порівнянні з реальними даними.

**3.6. Провести аналіз отриманих результатів.**

В результаті отримано модель що добре відображає реальні дані та може бути використаною для для моделювання та дослідження трендів у вже існуючих даних. Модель даних може бути використана для генерації даних аналогічних за трендом але відмінних за згенерованими залишками. Результати від лагодження та тестування довели працездатність розробленого коду. Верифікація функціоналу програмного коду, порівняння отриманих результатів з технічними умовами завдання на лабораторну роботу доводять, що усі завдання виконані у повному обсязі.

**IV. Висновки.**

У ході цієї роботи, я практикував парсинг сайтів для отримання потрібних даних, форматування даних та їх підготовку до статистичного навчання та оцінки. Також ми побудували модель даних, визначили статистичні параметри за допомогою спеціальних пакетів мови python. Код що не потрапив в цей звіт доступний для перегляду в файлі .ipynb і не був включений у звіт з міркувань лаконічності.

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