Опис проблеми, вирішенню якої буде присвячено цикл домашніх завдань.

Завдання, якому буде присвячено цикл домашніх завдань, полягає у тому, щоб організувати пошук відео TED конференцій на запити українською мовою (найпростішу пошукову систему).

Для цього мені потрібно буде розробити html-сторінку, де користувач вводитиме ключові слова для пошуку (наприклад, «глобальне потепління» або «стрес») та результатом пошуку будуть посилання на знайдені відео у Youtube та, за можливості, й самі відео.

Відбір доречних відео буде проводитися за допомогою субтитрів українською мовою, отримані через youtube-transcripts-api. Пошук інформації відбуватиметься за принципом, описаним у зареферованих дописах до домашнього завдання №0, тобто проводитиметься розрахунок частоти появи слова у субтитрах кожного відео та у множині відібраних відео. Остаточна формула дасть можливість оцінити доречність кожного відео відповідно до запиту, посортувати відібрані відео, а потім представити їх користувачу.

Для організації цього пошуку мені потрібно буде отримати посилання на відео TED конференцій з українськими субтитрами. Оскільки існує лише неофіційний TED API, у якого більшість функцій платні, я буду використовувати youtube-transcripts-api, який дозволить мені робити необмежену кількість запитів у день та отримувати практично всю необхідну мені інформацію.

Єдина проблема полягає у тому, що для отримання субтитрів мені потрібно надавати id на відео у youtube, яке зазначається в кінці його посилання. Тому я буду використовувати id надані з цього датасету(https://www.kaggle.com/goweiting/ted-talks-transcript#ted_metadata_youtube.csv), який містить майже 2,500 відео з TED-конференцій.

Тобто усі дані, які мені потрібні - посилання на відео та субтитри українською мовою, які я отримую у вигляді json-файлу.

Сформувавши дані, можна приступати до їх пошуку. Незважаючи на те, що пошукові системи генерують частоту появи слів у документах систематично залежно від вводу користувача, у моєму проекті я планую підготувати пораховані частоти слів у субтитрах до того, як організовувати пошукову систему. Наскільки мені відомо, тут й потрібно буде використовувати структури даних для їх зберігання.

Проте мені також потрібно буде провести нормалізацію даних, тобто слова, у яких спільний корінь, але різне зацінченння, не мають зберігатися окремо (наприклад, «тепло», «теплом», «теплим»), та інші процедури, потрібні для організації пошуку.

Можливості, які надає API:

- отримати перелік доступних мов субтитрів
- отримати субтитри потрібною мовою
- Якщо потрібної мови немає у списку, є функція перекладу
- Отримавши субтитри, можна доступитися за атрибутами до відео id, код мови, чи можна перекласти ці субтитри та якою мовою

Всі ці функції повертають json-файл, проте його ще потрібно почистити, адже субтитри розділені по частинах, а мені потрібний чистий файл з кодом відео та субтитрами. Поточний репозиторій містить такі модулі:

- приклад роботи з csv-файлом («csv_reader.py»)
- Читаючи csv-файл, записує json-файли субтитрів кожного відео(«data_collector.py»)
- Виділяє потрібну інформацію («data_cleaner.py»)
- Папка data - 4 приклади кінцевого json-файлу
- Папка examples - приклад використання можливостей youtube-transcripts-api («example_with_api.py») та файл «file1.json», який записує зазначений модуль.

Складові частини вимоги на систему

- Спонсор проекту: Олександра Гутор
- Бізнес потреба: Популяризація відео TED з українським перекладом (проте цей проект швидше для того, щоб отримати навички роботи зі структкрами даних).
- Бізнес вимоги: Забезпечує пошук TED відео з українськими субтитрами
- Питання та обмеження:

Граничний термін: 19 травня: хоча краще завершити проект 12 травня, а потім проводити подальші тестування або, за можливості, написати додатковий функціонал.
Також є обмеження по кількості відео, адже не усі мають українські субтитри.