Опис обсягу даних, які накопичені для подальшої обробки.

Усі необхідні дані зберігаються на pythonanywhere.com - майже 2000 файлів, проте два файли, як приклад, розміщені у папці example_data. Вони також були попередньо нормалізовані, тобто з кожного json-файлу вийнята вся доречна інформація та стемізовано кожне слово.

Результати обчислювальних експериментів.

На основі дослідження було зроблено деякі висновки. Перш за все, я дослідила такі деталі, як неточне сортування українських символів та їх правильна нормалізація. Пошук українською мовою забезпечує те, що відео обов'язково буде мати українські субтитри, що значно полегшує пошук, якщо користувачу потрібні відео саме з українськими субтитрами. Також я переконалася щодо працездатності формули bm-25, адже пошук видає доречні результати на різноманітних тестових випадках.

Оскільки під час розробки я проводила пошук по нормалізованих та ненормалізованих даних, то можу порівняти результати. Перш за все, нормалізація порівняно тривала операція, тому це перша причина, чому краще нормалізовувати дані заздалегідь, проте якщо працювати з нормалізованими субтитрами та нормалізованими вхідними даними, які надає користувач, то залишається лише суттєва інформація слова(корінь) і тому вища ймовірність знайти доречні відео. Головною перевагою пошуку не по назві конференції, а по субтитрах, є те, що досліджується його зміст, що сприяє більш точним результатам.

Також варто зазначити роль структур даних для зберігання результатів дослідження. Оскільки кожне знайдене відео має ще кількість переглядів, опис і т.ін., то для того, щоб інформація до кожного відео зберігалася разом та послідовно, я використала двозв'язний список. Оскільки я запланувала максимальну кількість відео - 7, то я могла б зберігати посилання на початок кожного списку в масиві, проте не завжди можна знайти саме 7 відео. Тому для зберігання посилань на двозв'язні списки я використовую простий список. Однозв'язний список використовується для зберігання tf, id відео та ін. під час самого пошуку. Таке зберігання даних значно полегшує доступ до даних під час обрахунків у формулі, адже дані зберігаються на визначених позиціях та в цілому є в компактному вигляді.

Також спочатку я хотіла зберігати допоміжні дані(tf, df, idf) для кожного слова в окремому файлі та потім доступатися до нього під час пошуку, проте я відмовилася від цього рішення через те, що моя програма містить відносно не багато файлів. Тому я можу рахувати tf і df під час виконання програми, хоча це займе трішки більше часу, проте менше пам’яті. Також під час додавання до даних нових відео потрібно буде заново записувати цей файл, що не є дуже ефективно. Тому структури даних є водночас ефективним та зручним у використанні варіантом у цьому проекті.

Отже, головним результатом експериментів є те, що реалізація проектів українською мовою можлива. Також мені вдалося переконатися щодо ефективності використання структур даних для збереження даних та ефективного доступу до них. Знаючи, що у цьому відео вже є українські субтитри, цей проект буде корисним для тих, хто вивчає англійську мову, адже можна одразу побачити переклад нових слів, прослухати їх вимову та зменшити швидкість виконання відео. Також перегляд конференцій іноземною мовою може бути не лише корисною, а й цікавою справою, адже користувач отримає відео на бажану тему.