**Дані, які буде отримувати програма з інтернету і працювати з ними**

Програма буде працювати з JSON та html файлами. Нижче наведені характеристики їх і їхній компонентів.

Всі **HTML**-сторінки структуровані як ієрархія компонентів (які у HTML визначаються як теги (“tags”)). Великий компонент може включати значне число менших – наприклад, таблиця має багато менших елементів: рядки та колонки. Є багато типів тегів, що виконують різні функції – деякі створюють клітини, інші – таблиці, зображення чи лінки. Теги також можуть мати додаткові властивості (тобто, бути унікальними ідентифікаторами), і можуть належати до груп під назвою «класи», що дає змогу визначати та вилучати індивідуальні елементи у документі. Пошук відповідних елементів у такий спосіб та вилучення їхнього змісту – це ключ до написання програми-скрейпера.

Перегляд елементів веб-сторінки: все можна розкласти на «клітини», що входять до складу інших «клітин».

Для того, щоб вилучати інформацію з веб-сторінок, вам треба дещо знати про різні типи елементів, які можуть бути в HTML документі. Наприклад, елемент table обрамляє всю таблицю, у якій є елементи tr (table row, рядок таблиці) для рядків, а ті, в свою чергу, містять елемент td (table data, дані таблиці) для кожної клітини. Найпоширеніший тип елемента - це div, який може означати буквально будь-який блок контенту.

**Складові JSON файлу**

Рядок: послідовність з нуля або великих символів Unicode. Рядки розділяються символами подвійних лапок і підтримують синтаксис зворотною косою рискою.

Літерали: будь-яке із значень True або False.

Масив: впорядкований список з нуля або більше значень, кожен з яких може бути будь-якого типу. Масиви використовують квадратні дужки з комами.

Об'єкт: невпорядкований набір пар ім'я/значення, де імена (також звані ключами) є рядками. Оскільки об'єкти призначені для подання асоціативних масивів, рекомендується (хоча і не обов'язково), щоб кожен ключ був унікальним всередині нього. Об'єкти поділяють фігурними дужками і використовують коми для поділу кожної пари, у той час як всередині кожної з них символ двокрапки відокремлює ключ або ім'я від його значення.

Null: Пусте значення, використовує слово null. Обмежені пробіли дозволені і можуть бути виставлені навколо або між синтаксичними елементами (значення і пунктуації, але не в межах значення рядка). Для цієї мети тільки чотири спеціальних символів вважаються пробілами: пробіл, горизонтальна вкладка, переклад рядка і коса риска. Зокрема, мітка байтового порядку не повинна генеруватися відповідної реалізацією (хоча вона може бути прийнята при синтаксичному аналізі JSON). Запит у форматі JSON не надає синтаксису для коментарів.