Міністерство освіти і науки України
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО–МОГИЛЯНСЬКА АКАДЕМІЯ»
Кафедра мультимедійних систем факультету інформатики

RULE-BASED NLP APPROACHES FOR ARCHITECTURAL MONUMENTS DOCUMENTS' EXTRACTION

Текстова частина до курсової роботи
за спеціальністю «Комп’ютерні науки» 122

Керівник курсової роботи
ас. Смиш О.Р.

_____
(підпис)
“____” _____ 2023 р.

Виконав студент КН–3
Кирилін Є.С.
“____” _____ 2023 р.

Київ 2023

Міністерство освіти і науки України

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»

Кафедра мультимедійних систем факультету інформатики

ЗАТВЕРДЖУЮ

Зав.кафедри мультимедійних систем,

Доцент., к. ф.-м. н. О.П. Жежерун

_____(підпис)

“____”_____2022 р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

на курсову роботу

студенту 3-го року БП КН факультету інформатики

Кириліну Єгору Сергійовичу

ТЕМА: Rule–based NLP approaches for architectural monuments documents' extraction

Зміст ТЧ до курсової роботи:
Анотація
Вступ
1. Обробка природної мови
2. Аналіз даних з реєстру
3. Географічний аналіз пам’яток
4. Хронологічний аналіз пам’яток
5. Розробка методів для парсера
Висновки
Список використаних джерел
Додатки

Дата видачі “___” _____ 2022 р. Керівник _____

(підпис)

Завдання отримав _____

(підпис)

Тема:

_____

Календарний план виконання роботи:

№ п/п	Назва етапу курсової роботи	Термін виконання етапу	Примітка
1	Отримання завдання на курсову роботу	16.11.2022 – 18.11.2022	
2	Огляд технічної літератури за темою роботи	23.11.2022 – 13.12.2022	
3	Збір та підготовка даних з реєстру для аналізу	15.12.2022 – 18.01.2023	
4	Проведення кількісного, географічного та хронологічного аналізів пам’яток	21.01.2023 – 25.02.2023	
5	Розробка методів для парсера	02.03.2023 – 06.04.2023	
6	Написання пояснювальної роботи	08.04.2023 – 01.05.2023	
7	Створення слайдів для доповіді та написання доповіді	18.04.2023 – 02.05.2023	
8	Аналіз отриманих результатів з керівником	21.04.2023 – 22.04.2023	
9	Корегування роботи згідно з результатами перевірки наукового керівника	25.04.2023 – 03.05.2023	
10	Остаточне оформлення пояснювальної роботи та слайдів	04.05.2023 – 14.05.2023	
11	Захист курсової роботи	22.05.2023	

Зміст

Анотація

Курсову роботу присвячено застосуванню методів обробки природної мови з метою розробки парсера для нерухомої спадщини України за допомогою створених програм аналізу даних з державного реєстру. У роботі досліджено статистичні характеристики пам’яток культури та сформовано інтерактивну карту з географічним розподілом цих об’єктів.

Вступ

У сучасному світі зростає обсяг даних, який збирається та зберігається в електронному форматі, але підхід до цих задач не завжди є коректним, що ускладнює подальшу обробку та аналіз, а також збільшує ризик втрати інформації. В Україні існує державний реєстр нерухомої спадщини, представлений Міністерством культури та інформаційної політики, який містить інформацію про тисячі пам'яток культури. Проте, проблема полягає в тому, що репрезентовані дані не є уніфікованими, про це свідчить відсутність унікального формату для низки атрибутів у наборах даних.

Метою роботи є детальний аналіз методів обробки природної мови для використання їх для дослідження та роботи з україномовним набором даних з реєстру пам'яток культури України. Досліджено географічний та хронологічний розподіли пам'яток, а також способи написання датувань у реєстрі. У процесі роботи виявлено, що існує значна невиправдана варіативність у написанні дат та інших атрибутів пам'яток, що створює складнощі у подальшій обробці та аналізі. Також, опрацьовано найпоширеніші типи пам'яток культури в реєстрі. Кінцевим продуктом є парсер, створений для розв’язання проблеми різноманіття в написанні даних, з використанням методів обробки природної української мови, що дає змогу єдиним та стандартизованим способом вводити інформацію в реєстр, що полегшує подальший аналіз та обробку і забезпечує точність даних.

Для створення застосунку використано мову програмування Python, оскільки вона має бібліотеки для аналізу та візуалізації даних, обробки геоданих та інших задач.

Завдання роботи полягає в розробці парсера для пам’яток культури на основі проаналізованих даних з реєстру.

Перший розділ присвячено обробці природної мови та огляду основних методів, які використано в роботі.

У другому розділі здійснено збір та підготовку даних для подальшого аналізу, а також проведено кількісний аналіз лематизованих форм назв типів пам’яток.

У третьому розділі роботи проведено географічний аналіз пам'яток, розроблено алгоритм знаходження їхніх координат та сформовано інтерактивну карту візуалізації розташування пам'яток на території України.

Четвертий розділ присвячено аналізу датувань. Розроблено та оцінено ефективність методу уніфікації датувань, а також знайдено найбільш визначні роки.

У п’ятому розділі описано методи, використані для парсингу, розроблено відповідний парсер та оцінено його ефективність.

1. Обробка природної мови

Мова – це одна з найважливіших та невіддільних частин людського існування, за допомогою якої ми вільно виражаємо думки, почуття, емоції та ідеї, розуміємо інших людей, передаємо та отримуємо нові знання. Саме ту мову, яка використовується для комунікації між людьми називають природною. Проте, природна мова не є ефективним засобом комунікації з комп’ютером через свою складну структуру. Тому, для досягнення адекватної взаємодії з комп’ютером розроблено спеціальні алгоритми та методи, які б забезпечували розпізнавання та відтворення живої мови.

NLP (Natural Language Processing) – це галузь комп'ютерних наук, яка застосовується для аналізу та синтезу людської мови комп'ютерними системами, основна ідея якої є написання програм, які здатні розуміти, аналізувати, та взаємодіяти з людьми за допомогою мови. Обробка природної мови уможливлює комп’ютерам розуміти не тільки слова в тексті, але й його контекст. Це дає змогу автоматично розпізнавати, аналізувати настрій і емоції, чим користуються при розроблюванні чат-ботів та інших інтерактивних інтерфейсів, які можуть «розмовляти» з користувачами. NLP для української мови, у порівнянні з англійською, ще є в зародковому стані, оскільки українська використовується для вмісту сайтів в дев’яносто вісім разів рідше, ніж англійська [1]. Незважаючи на це, вона активно розвивається з урахуванням збільшення обсягу українськомовного контенту на просторах Інтернету. Наразі обробка природної української мови дає змогу працювати для таких основних методів, як лематизація, токенізація, розмічування частин мови тощо.

Лематизація – процес пошуку «леми», тобто словникової форми слова. Лематизація тексту передбачає використання словника та виконання морфологічного аналізу слів, щоб видалити лише флективні закінчення та повернути базову або словникову форму слова [2].

Токенізація, або сегментація – це процес поділу фраз і абзаців на менші частини, щоб моделі обробки природної мови могли їх визначати та використовувати. Токени – це найменші можливі одиниці обробленого тексту [3].

Part-of-speech tagging, або PoS-tagging – алгоритм присвоєння міток для кожного слова в тексті відповідно до його частини мови. Однак, проблема з використанням цього методу в українській мові полягає в складнощах з визначенням частин мови для деяких слів, які можуть мати декілька різних значень та граматичних форм. Наприклад, слово «три» може бути як і числівником, так і наказовим способом другої особи однини дієслова «терти», в залежності від контексту використання.

Видобування інформації (Information Extraction) – це процес автоматичного визначення та виділення значущих фрагментів з неструктурованого або малоструктурованого тексту [4].

2. Аналіз даних з реєстру

2.1. Збір та підготовка даних для аналізу

Для отримання інформації використано сайт Міністерства культури та інформаційної політики України. Вебсайт має розділ «Культурна спадщина», що охоплює різноманітні підрозділи, які охоплюють нормативно-правові аспекти охорони культурної спадщини. Підрозділ «Нерухома культурна спадщина» містить Державний реєстр нерухомих пам'яток України. Цей реєстр поділено на дві частини: "Реєстр пам'яток місцевого значення" та "Реєстр пам'яток національного значення" [5]. Для роботи використано обидва, оскільки вони містять інформацію про пам’ятки культури всіх двадцяти п’яти областей України, а також міста Києва та міста Севастополя. Однак, складність полягала в тому, що усі файли в реєстрі були у форматах DOCX(Document Extended) або PDF(Portable Document Format), а не CSV(Comma-separated Values), який зазвичай використовують для обробки та аналізу датасетів. Для формування набору даних переглянуто кожен файл, що відповідає за окремий регіон та складено CSV-файли для кожного з них, після чого об’єднано. Попередньо кожен файл перевірено на наявність невидимих знаків розриву рядка та абзацу для коректного перенесення даних в датасет. Деякі файли з реєстру мали PDF-формат, їх конвертовано у DOCX-формат для подальшого редагування та перетворення у CSV-формат.

2.2. Модель UDPipe

UDPipe – сучасний інструмент обробки природних мов, що забезпечує ефективний та точний спосіб виконання токенізації, лематизації, розмічування частин мови, морфологічного аналізу та аналізу залежностей [6].

Однією з ключових переваг UDPipe є використання Universal Dependencies, що є міжлінгвістичною структурою для анотування синтаксичних залежностей. Цей фреймворк надає стандартний набір рекомендацій щодо анотацій, які можна

застосувати до будь-якої доступної мови. Завдяки тому, що UDPipe навчено на Universal Dependencies для більш ніж шістдесяти мов, його можна використовувати для різноманітних завдань обробки природної мови [7].

Щоб перевірити роботу UDPipe, можна скористатися онлайн-сервісом, що надано на його вебсайті, який дає змогу копіювати текст для обробки або завантажити файл з текстом. UDPipe також надає API (Application Programming Interface), яке можна під'єднати до власного проєкту. Для використання API необхідно відправити запит за допомогою команди cURL – утиліти командного рядка, що застосовується для взаємодії з URL-адресами і зазвичай використовується для отримання вмісту вебсторінок [8]. Запит містить параметри, які вказують на модель мови, що використовуватиметься для аналізу тексту, а також інші параметри для налаштування обробки тексту. Результати аналізу повертаються у форматі CoNLL-U (Computational Natural Language Learning), який містить інформацію про розбиття тексту на токени, морфологічні теґи та залежності між словами. У запиті передаються два файли – вхідний файл з даними для обробки та вихідний файл для збереження результату обробки за допомогою UDPipe.

Приклад запиту до вебсервісу UDPipe, де «test.txt» – вхідний файл з даними, а «out.txt» – файл для виведення результату обробки даних:

"curl -F model=ukrainian-iu-ud-2.10-220711 -F data=@test.txt -F tokenizer= -F tagger= -F parser= http://lindat.mff.cuni.cz/services/udpipe/api/process > out.txt"

2.3. Лематизація та кількісний аналіз типів пам'яток

Наступний етап дослідження – використання UDPipe для обробки тексту, що початково містив назви пам'яток архітектури. Проте, ці дані були непідхожими для поставленої задачі, тому перед основним аналізом використано інший алгоритм для фільтрації непотрібних частин мови з опрацьованого тексту, таких як сполучник, прикметник, прийменник тощо. Залишивши лише root-слова,

тобто ті, від яких йде залежність до інших слів у реченні, отримано назви типів пам’яток архітектури.

Після попереднього етапу отримано файл з лематизованими формами назв типів пам’яток, наприклад «церква», «могила», «будинок» тощо. Наступним кроком є підрахунок кількості входжень кожного типу в датасеті. Для досягнення цього використано новий алгоритм, за допомогою якого утворено словник, у який додано леми та їхні кількісні показники. За словником утворено відсортований список пар «тип пам’ятки – число їхніх екземплярів», і записано у новий файл.

2.4. Бібліотека статистичної візуалізації Vega–Altair

Бібліотека Altair – це потужний інструмент візуалізації даних, який дає змогу створювати різноманітні графіки, діаграми та візуалізації. Altair пропонує декларативний підхід до візуалізації даних, що уможливлює визначення елементів графіків та їхніх взаємозв'язків [9].

Altair підтримує такі типи графіків, як лінійні діаграми, стовпчикові діаграми, кругові діаграми, точкові графіки, гістограми, графіки розсіювання тощо.

Одним з головних переваг Altair є те, що вона інтегрується з розповсюдженими інструментами для аналізу даних, такими як Pandas та Numpy. Це дає змогу обробляти та візуалізувати великі обсяги даних без необхідності застосування додаткових інструментів.

Крім того, в Altair є змога створення інтерактивних графіків.

2.5. Візуалізація результатів кількісного аналізу лем

Розглянуто візуалізацію результатів кількісного аналізу лем. З метою зручності та кращого сприйняття результатів, вирішено зробити дві окремі візуалізації: одну для пам’яток національного значення, другу – для пам’яток місцевого значення. Використовуючи бібліотеку Altair, створено графік, який

демонструє кількість пам’яток різних типів. Кожен стовпчик відповідає одному типу пам’ятки, а його висота відображає кількість згадувань цього типу в наборі даних. Колір стовпчиків залежить від кількості згадувань – чим більше входжень, тим темніший колір (див. рисунок 2.5.1, рисунок 2.5.2).

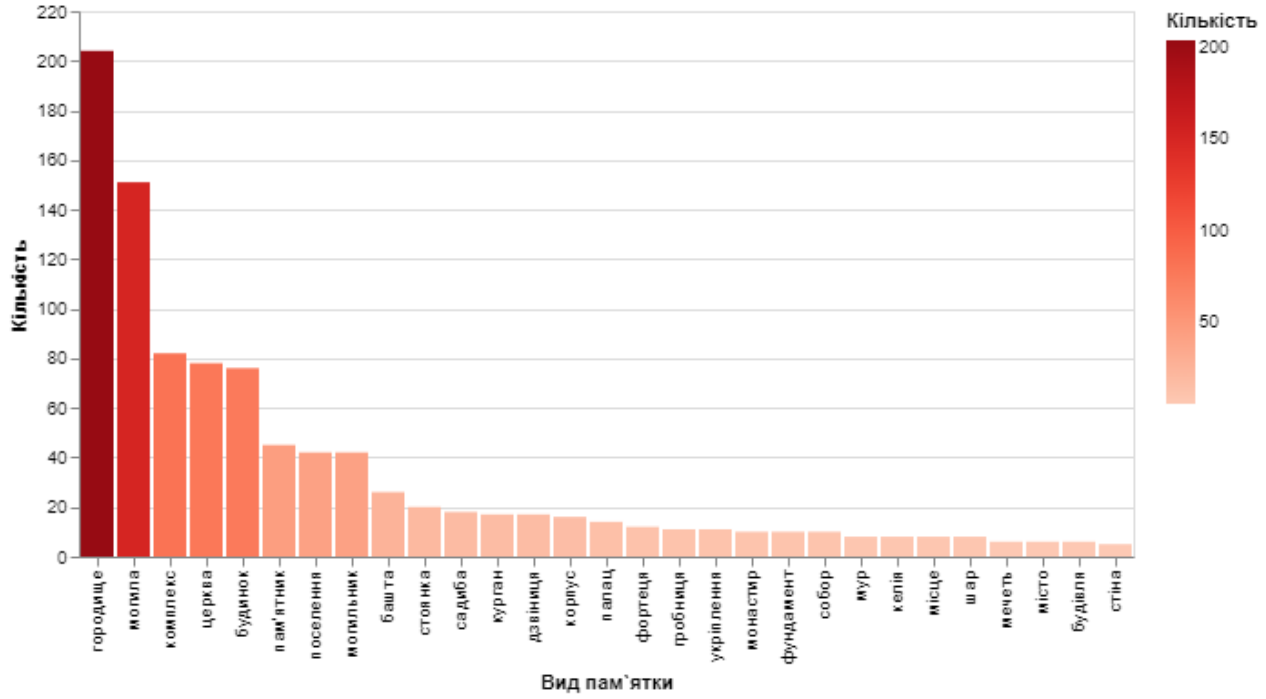

Рисунок 2.5.1 – Візуалізація результатів кількісного аналізу типів пам’яток національного значення

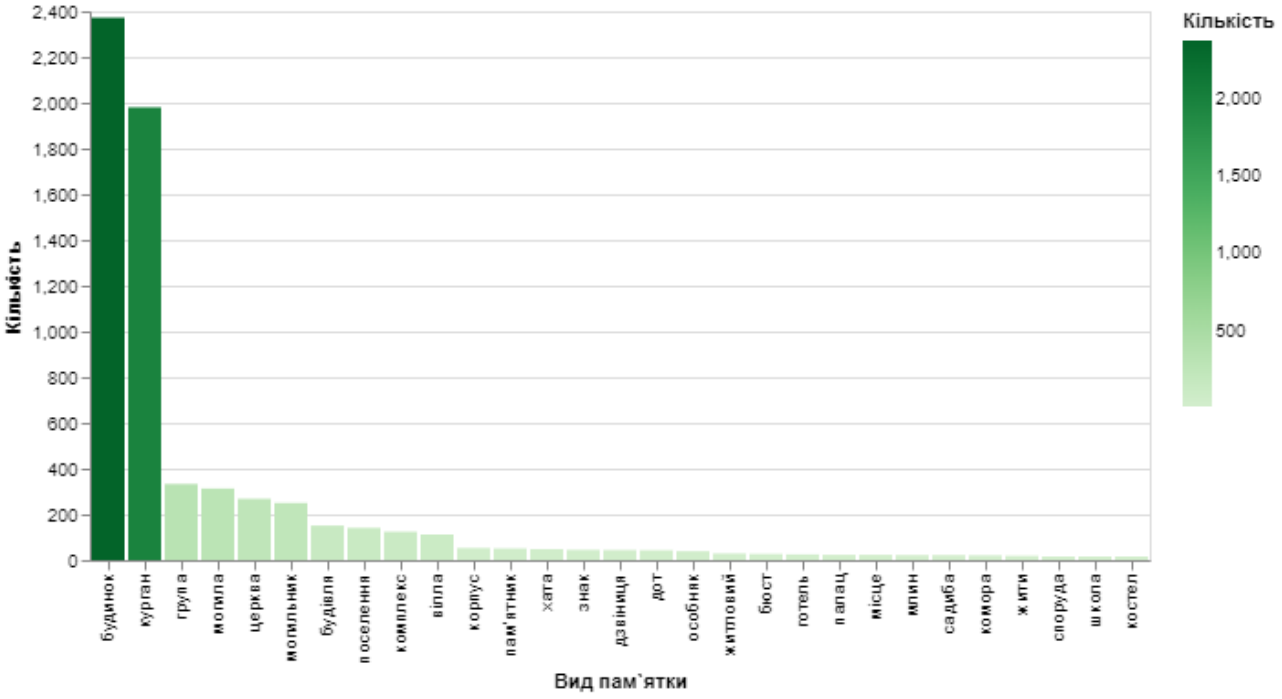

Рисунок 2.5.2 – Візуалізація результатів кількісного аналізу типів пам’яток місцевого значення

З графіків видно, що деякі леми є поширеними серед обох категорій пам’яток, наприклад «будинок», «могила» та «церква». Ці леми можна вважати більш репрезентативними для аналізу, оскільки вони є типовими для обох груп пам’яток. Таким чином, саме ці та деякі інші найбільш поширені леми використано для швидкого вибору типу пам’ятки в парсері.

3. Географічний аналіз пам’яток

3.1. Аналіз географічного розподілу пам’яток

Географічний аналіз є важливим етапом у дослідженні культурної спадщини. Для первинної візуалізації використано хороплетні, або фонові карти, які показують кількість пам’яток у кожному регіоні, використовуючи кольори та їхні відтінки. Створено дві окремі мапи для пам’яток місцевого та національного значення для порівняння розподілу на території країни (див. рисунок 3.1.1, рисунок 3.1.2). Мапи для географічного аналізу представлено за допомогою Datawrapper – онлайн-інструменту для візуалізації даних [10].

Рисунок 3.1.1 – Фонова картограма кількості пам’яток національного значення

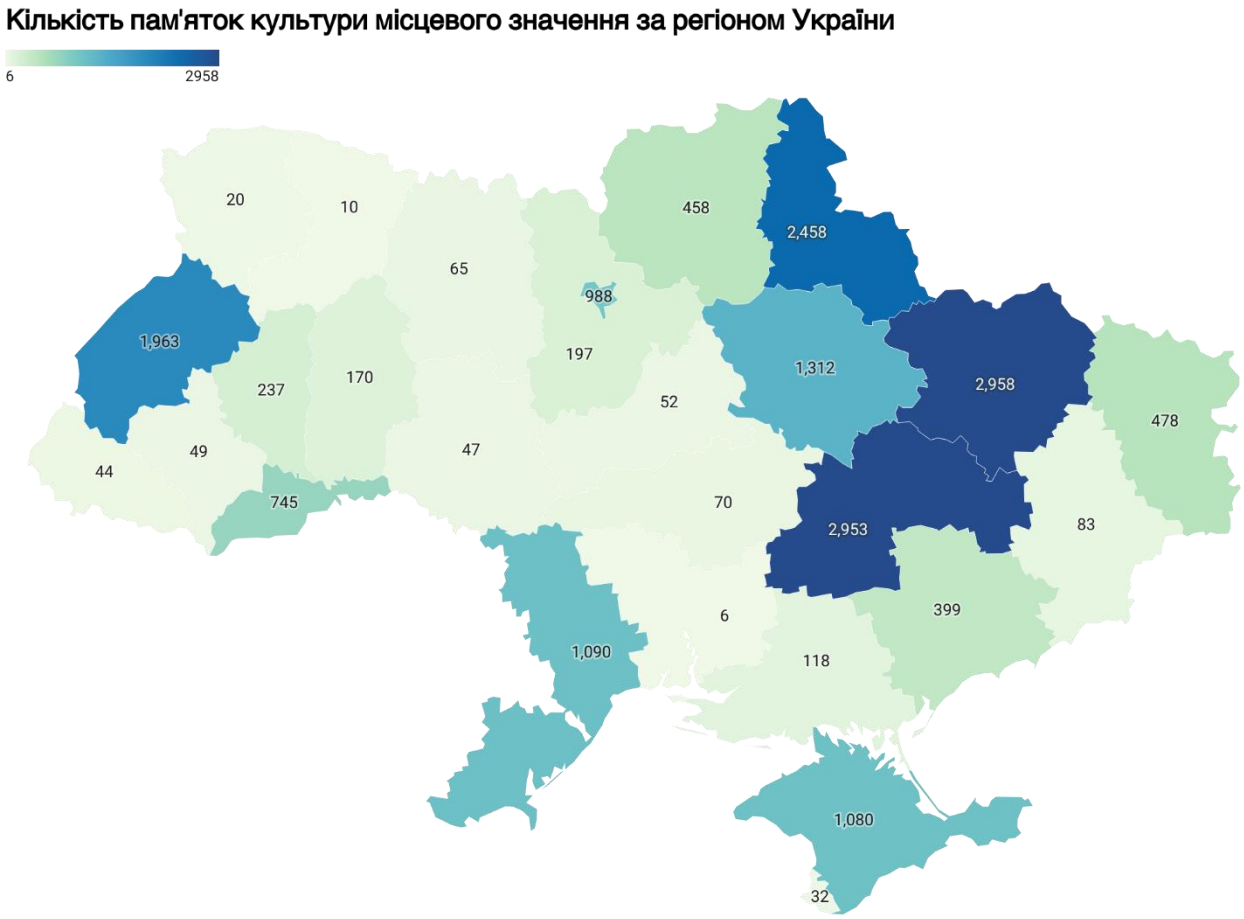

Рисунок 3.1.2 – Фонова картограма кількості пам’яток місцевого значення

На картограмах видно різницю в розподілі пам’яток між регіонами України. Для пам’яток національного значення, найбільші скупчення об’єктів зосереджено в Києві, на Кримському півострові та в Чернігівській області. З іншого боку, для пам’яток місцевого значення найбільша кількість пам’яток зосереджена в Дніпропетровській та Харківській областях. У розділі 3.3. результати формування інтерактивної мапи порівняно з цими мапами, щоб проаналізувати різницю в представленні даних.

3.2. Розробка алгоритму знаходження координат

Для подальшого аналізу та візуалізації географічного розташування пам’яток культури необхідно мати точні координати, проте точні локації левової частки об’єктів не знайдено автоматично, тому вирішено використати

координати їхніх населених пунктів, як найближчий еквівалент для подальшого формування інтерактивної мапи.

Створено алгоритм, що шукає координати населених пунктів, де розміщені пам’ятки та прив’язує їх до пам’яток. Для цього алгоритму імпортується бібліотека geopy, що визначає координати міст, країн та орієнтирів по всьому світу. Ця бібліотека містить Nominatim – інструмент від OpenStreetMap для геокодування, тобто перетворення адрес на координати [11]. Для кожного рядка вхідного файла з назвами населених пунктів використано вбудовану функцію геокодера, яка повертає об’єкт з координатами відповідного населеного пункту. Якщо координати знайдені, то їх записано у два нові стовпці з широтою та довготою і рядок дописано до вихідного файлу.

Щоб забезпечити можливість коректного пошуку координат, назви населених пунктів перекладено англійською мовою, оскільки геокодер не працює з українською. Незважаючи на це, знайдено координати лише для частини пам’яток, але використання автоматизованого алгоритму є ефективнішим та швидшим варіантом, ніж пошук координат вручну для більш ніж десяти тисяч об’єктів.

3.3. Формування інтерактивної мапи

Для формування інтерактивної мапи використано kepler.gl – інструмент, призначений для створення інтерактивних карт, що використовується для візуалізації різноманітних геопросторових даних. За допомогою kepler.gl можна створювати картографічні візуалізації, які надають інформацію про геопросторові риси, такі як розташування, територіальний розподіл, відстані тощо. Kepler.gl надає користувачам можливість налаштовування та візуалізації даних на картах. Інструмент надає багато функцій, таких як можливість вибирати різні типи карти, налаштування кольорової палітри та шкали, вибирати способи відображення даних на карті, визначати рівень деталей тощо [12].

Для створення карти використано датасет, що містить назви пам’яток та координати їхніх населених пунктів. Візуалізація даних на карті використовує метод hexbin. Hexbin – це метод групування точок на шестикутниках з метою візуалізації щільності точок. Цей метод демонструє географічний розподіл пам’яток та їхню концентрацію в певних областях. Мапу створено з можливістю інтерактивної навігації та фільтрації даних, що дало змогу досліджувати різні аспекти географічного розподілу об’єктів.

Під час створення датасету з координатами населених пунктів, що використано для формування карти, виникла проблема з неправильно визначеними координатами низки населених пунктів, що спричинило помилкове відображення об’єктів на мапі за межами України, тому помилкові координати відфільтровано та видалено з датасету (див. рисунок 2.8.1).

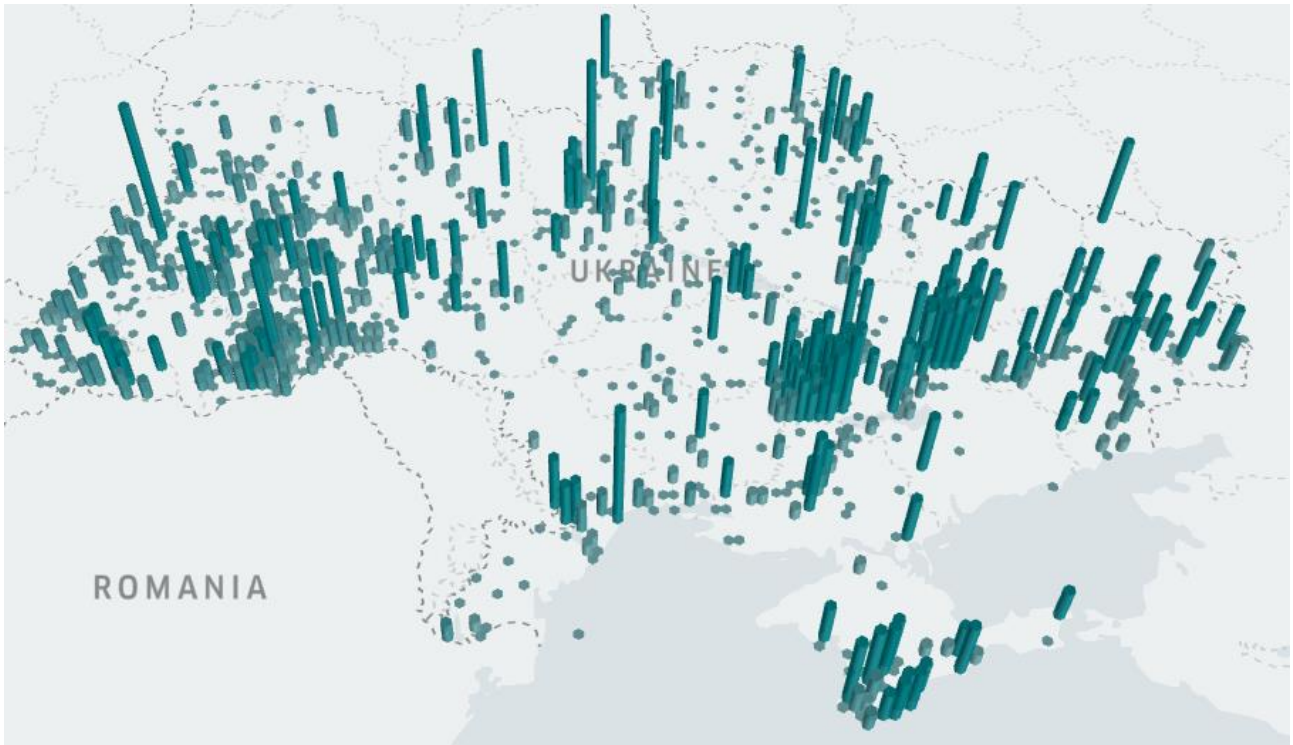

Рисунок 2.8.1 – Інтерактивна мапа географічного розподілу пам’яток культури за їхніми населеними пунктами

При порівнянні інтерактивної мапи зі створеними раніше фоновими картограмами, помітно, що втрачено чимало координат на північному сході

країни. Це свідчить про те, що створення набору даних з використанням геокодера є неточним методом і вимагає додаткової обробки даних для отримання точних результатів.

4. Хронологічний аналіз пам’яток

4.1. Дослідження способів написання датувань у реєстрі

Досліджено способи написання датувань у сформованому датасеті пам’яток культури. Оскільки хронологічні атрибути об’єктів мають безліч форматів виникає необхідність уніфікувати датування, адже варіативність датувань ускладнює подальший аналіз даних пам’яток за часом їхнього виникнення.

Датасет має різні формати датувань, такі як один рік, діапазон років, декади, одне століття або діапазон століть, а також початок, середина, кінець, перша або друга половина століття тощо. Століття записано як арабськими, так і римськими цифрами, тому для алгоритму уніфікації розроблено метод, який переводитиме римські цифри в арабські.

Для уніфікації датувань у датасеті обрано формат «один рік» або «діапазон років». Такі формати є зручними для аналізу та порівняння даних.

4.2. Написання методу для уніфікації датувань

Уніфікація датувань є одним із ключових кроків, який дасть змогу отримати точні результати під час дослідження. Для цього написано метод, з використанням двадцяти п’яти регулярних виразів під усі вищезазначені формати дат. Регулярні вирази – це шаблон, що використовується для пошуку та заміни тексту відповідно до певних правил, оскільки датасет містить безліч форматів, використано регулярні вирази для пошуку відповідності кожному рядку текстового файлу та переведення датування в правильний формат «один рік» або «діапазон років».

Написано дві додаткові функції, перша з яких призначена для перетворення століття, записаного арабським числом, у діапазон років, а друга функція перетворює римські числа у відповідні арабські.

4.3. Оцінка ефективності написаного алгоритму

Для оцінки точності написаного методу розроблено алгоритм обрахунку ефективності. При оцінці ефективності алгоритму уніфікації використано спеціальну змінну, що виконує функцію лічильника усіх датувань, що знайшли відповідник серед регулярних виразів.

Усі датування зібрано в один текстовий файл для аналізу. Код методу шукає відповідність у регулярних виразах кожному рядку текстового файлу і, якщо знаходить, то форматує, збільшує лічильник та додає до вихідного файлу, якщо ні – програма виводить значення цього датування в консоль для подальшого аналізу та виводить кількість входжень цього датування. Використання лічильника уможливило визначення частоти датувань, які пройшли перевірку, та оцінювання ефективності алгоритму.

Приклад виводу у консоль датування, що не пройшло перевірку:

приблизно II – I тисячоліття
6

За результатами оцінки встановлено, що метод ефективно спрацював на 87% датувань у досліджуваному датасеті.

4.4. Обрахунок визначних років

Розроблено алгоритм, що використовує уніфіковані значення датувань з отриманого раніше файлу та підраховує кількість пам’яток, що створено в кожному році.

Створений метод зберігає результати у словник, де ключами є роки, а значеннями – кількість пам’яток, побудованих у цьому році. Для цієї задачі використано Counter – структуру даних, що підраховує кількість входжень кожного елементу в датасеті.

Якщо датування містить один рік, то кількість пам’яток збільшиться винятково для лічильника цього року, але якщо датування подано як діапазон років, то алгоритм враховуватиме кількість пам’яток для кожного року в цьому діапазоні.

Отриманий файл із кількістю пам’яток у кожному році використано для візуалізації результатів.

4.5. Візуалізація отриманих результатів підрахунку значущих років

Графічне зображення кількості пам’яток, побудованих у різний період часу допомагає з’ясувати, які роки були найбільш визначними для спорудження пам’яток.

Для візуалізації результатів, що отримано раніше, використано бібліотеку Altair. За допомогою цієї бібліотеки побудовано інтерактивний графік, який відображає залежність кількості об’єктів від року.

Графік має вигляд затемненої ділянки, де затемнення залежить від кількості пам’яток, що були споруджені в окремий рік. Якщо рік на графіку є від’ємним – це означає, що це є роком до нашої ери.

На графіку (Рисунок 3.5.1) видно, що найпопулярнішими роками для побудови є роки 2–го століття до нашої ери, а за останнє тисячоріччя (Рисунок 3.5.2) найбільше пам’яток побудовано наприкінці 19 століття.

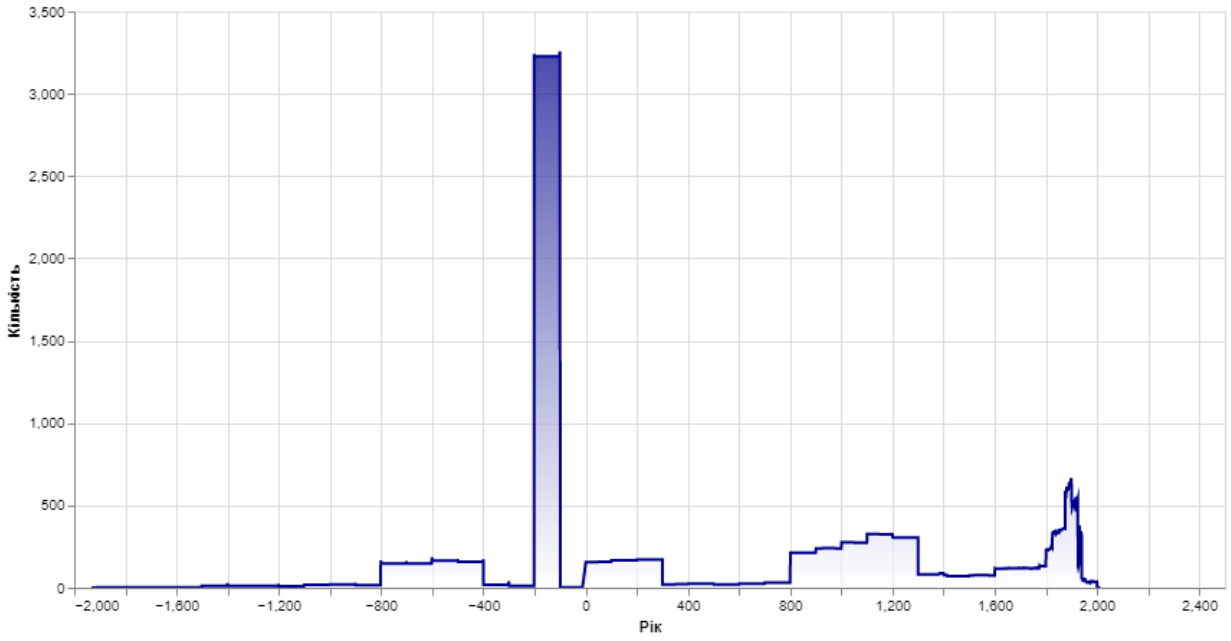

Рисунок 3.5.1 – Графік кількості пам’яток для всіх досліджених років

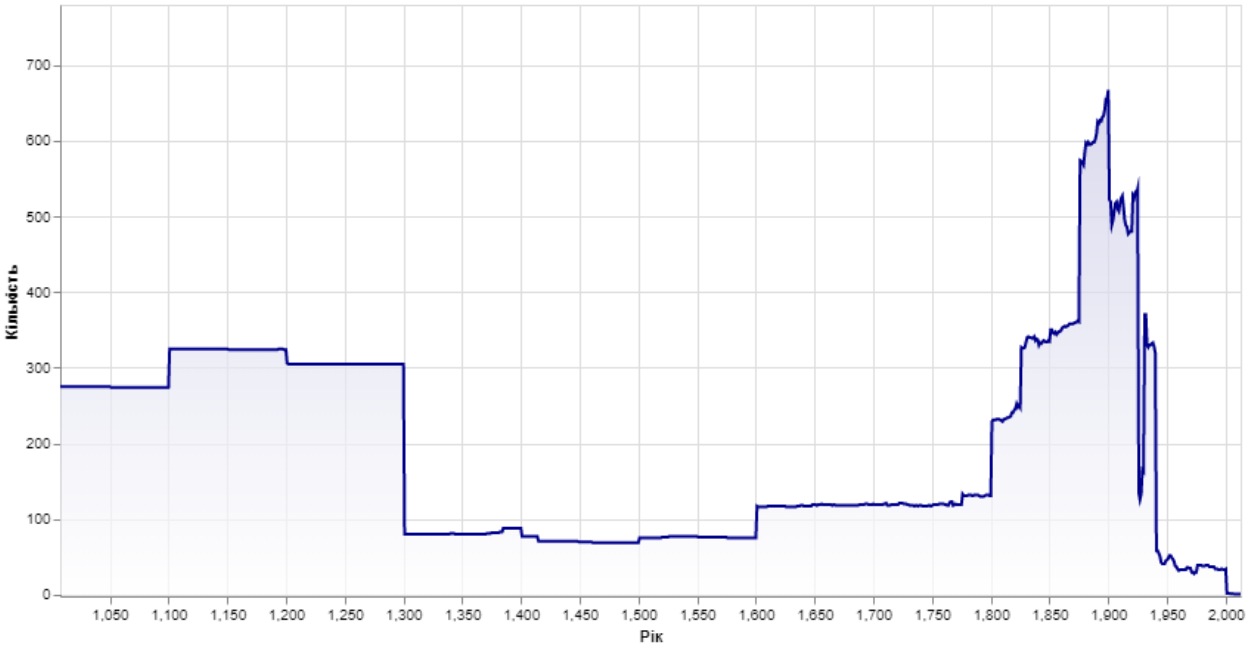

Рисунок 3.5.2 – Графік кількості пам’яток для останнього тисячоліття

5. Розробка методів для парсера

П’ятий розділ дослідження присвячено написанню парсера для пам’яток культури, який уможливлює збереження інформації про них у форматі JSON(JavaScript Object Notation), що спрощує майбутнє розширення або оновлення реєстру. Завдяки цьому рішенню оптимізовано зберігання інформації та забезпечено цілісність даних.

Парсер розроблено на основі ретельного аналізу структури та характеристик різних записів у датасеті, його головна перевага – уніфікований формат даних. Гарантовано, що всі дані представлено в однаковому форматі, що забезпечує точність та надійність результатів, а також створення даних без порожніх значень, що полегшує подальший аналіз. Таким чином, зникає необхідність спеціально підготовлювати дані для дослідження, що дає змогу використовувати інформацію для потрібного аналізу після отримання.

5.1. Вибір типу та назви пам’ятки

При запуску парсера, спочатку потрібно обрати тип пам’ятки серед десятки найпоширеніших, які визначено на основі проведення кількісного аналізу лематизованих форм типів пам’яток (див. розділи 2.3 та 2.5), або надано можливість ввести власний тип пам’ятки. Обрано такі стандартні типи пам’яток: будинок, курган, могила, могильник, церква, будівля, городище, комплекс, пам’ятник та поселення.

Після вибору типу, запропоновано ввести назву пам’ятки цього типу.

5.2. Вибір типу населеного пункту та розташування

Надано можливість обрати тип населеного пункту з таких варіантів, застосованих у реєстрі: місто, село, смт, селище, мис та урочище, далі

запропоновано ввести назву населеного пункту. Якщо вибір не входить у список, можна обрати опцію «інше» та ввести одночасно тип і назву населеного пункту.

В Україні запроваджено сім типів вулиць: власне вулиця, проспект, бульвар, набережна, тупик, провулок та узвіз – усі ці варіанти доступні для вибору розташування пам’ятки. Якщо локація пам’ятки не передана точною адресою, доступна опція «інше» для самостійного введення інформації.

5.3. Вибір виду пам’ятки

Надано можливість обрати вид пам’ятки. Реєстр містить широкий спектр пам’яток, що можуть бути класифіковані таким чином: пам’ятки історії, архітектури, монументального мистецтва, археології, науки, техніки, містобудування та ландшафту.

Є можливість обрати будь-який набір з цих видів, обираючи їхні номери, та завершити вибір, ввівши «0», тоді здійсниться перехід до наступного блоку.

5.4. Вибір охоронних номерів

Після вибору виду пам’ятки необхідно внести її охоронний номер. Для пам’яток національного та місцевого значення формати охоронних номерів різні, проте піддаються перевірці за допомогою регулярних виразів.

Зокрема, охоронний номер національного значення може бути шестизначним або семизначним числом, окрім цього може містити велику літеру через дефіс, або одну чи дві цифри через знак дробу, або комбінацію двох останніх варіантів.

Для пам’яток місцевого значення номер містить від однієї до шести цифр, а також може містити велику та маленьку літери через дефіс, або, окрім них, одну або дві цифри через знак дробу.

Запитано про тип пам’ятки – національного або місцевого значення, після чого перевірено введений номер на відповідність будь-якому з форматів за

допомогою регулярних виразів. Якщо введений номер не відповідає жодному із зазначених форматів – виведено відповідне повідомлення та повторно запитано номер.

5.5. Вибір установи, дати та номеру постанови про взяття під охорону

Після вибору охоронного номеру, введено дату та номер постанови про взяття пам’ятки під охорону. Перевірено правильність введених даних, зокрема валідність дати, яка повинна мати рік більший за 1990-ий та менший за 2024-ий та цифровий вміст номера. Запитано про установу, що видала відповідний наказ або постанову. У реєстрі варіантів є п’ять установ: Кабінет Міністрів України, який видавав постанови про охорону пам’яток національного значення, а також Міністерство культури України, Міністерство культури та інформаційної політики, Міністерство культури, молоді та спорту і Міністерство культури та туризму.

5.6. Вибір датування

Для визначення датування пам’ятки використано результати попереднього дослідження (див. розділ 4). Перевірку введеної дати на валідність проведено за допомогою регулярних виразів. Якщо введене датування не відповідає жодному формату, запропоновано ввести діапазон років, який використовуватиметься як датування.

Запитано, чи є це датуванням до нашої ери. Якщо так, початковий та кінцевий роки змінено місцями. Якщо ні – перевірено кінцевий рік діапазону, чи є він більшим за 2023, і у разі схвальної відповіді – кінцевому року надається значення 2023.

5.7. Збереження результатів парсингу

Після виконання роботи парсера, надано можливість записати та зберегти результат у файл формату JSON. Для цього запропоновано ввести назву файлу, в який збережено результати парсингу. Це дає змогу зберігати інформацію про пам’ятки культури для подальшого використання її для власних цілей.

5.8. Оцінка ефективності парсера

Після розробки парсера обраховано оцінку ефективності. Розглянуто ефективність алгоритму уніфікації датувань та перевірки охоронних номерів.

Згідно з результатами оцінки алгоритму уніфікації датувань, ефективність становила 87% (див. розділ 4.3). Більшість дат, зазначених у початковому датасеті, успішно переведено до єдиного формату.

Ідентичним способом оцінено ефективність алгоритму перевірки охоронних номерів, згідно з чим алгоритм показав ефективність у 99%. Майже всі охоронні номери, з–поміж десяти тисяч у реєстрі, успішно перевірено на відповідність формату охоронних номерів пам’яток національного та місцевого значення за допомогою регулярних виразів. Залишок в 1% може вказувати на помилки при заповненні реєстру, оскільки непідхожі до форматів номери не відповідають жодному зі встановлених форматів.

5.9. Приклад використання парсера

Для демонстрації роботи парсера випадковим чином обрано пам’ятку культури національного значення – городище «Золотий мис» у селі Золота Балка Херсонської області. Введено атрибути пам’ятки через консоль програми, обрано відповідні значення зі списку або введено текстові дані.

Спочатку обрано тип пам’ятки – «Городище», та введено назву «Золотий мис». Далі, вказано тип населеного пункту – село, та вписано назву Широка Балка. У третьому пункті вказано датування пам’ятки – «IX століття», та підтверджено, що це датування є часом до нашої ери. Програма перейшла до

наступного блоку після введення датування і не запросила введення діапазону років – це означає, що серед регулярних виразів знайшовся відповідник до введеного значення датування, його опрацьовано та переформатовано на уніфікований формат діапазону років. Оскільки точний тип розташування невідомий, обрано «Інше» та введено «с. Широка Балка». Обрано вид пам’ятки – «пам’ятка історії» та вказано, що це є пам’яткою національного значення для перевірки формату охоронного номеру – «210016-Н». Оскільки програма продовжила своє виконання після введення охоронного номеру, це означає, що введений номер є валідним та відповідає формату. Вказано номер та дату постанову Кабінету Міністрів України – «929» та «10.10.2012».

Завдяки парсеру опрацьовано інформацію про тип і найменування пам'ятки, її населений пункт, датування, розташування, вид, охоронний номер та постанову про взяття під охорону.

Результати виконання програми збережено у форматі JSON (див. додаток 1), що уможливлює подальшу зручну обробку даних.

Висновки

Отже, у роботі розглянуто технологію обробки природної мови, що застосовується для аналізу та синтезу людської мови комп'ютерними системами. Досліджено доступні для української мови методи обробки природної мови, що дають змогу аналізувати тексти українською. Однак, українська мова в цьому напрямі є на ранньому етапі розвитку, оскільки обсяг українськомовного контенту в мережі Інтернет є низьким [1].

Здійснено збір інформації з реєстру в окремий датасет. Проведено лематизацію та кількісний аналіз типів пам'яток та встановлено, які типи є найпоширенішими.

Проведено географічний аналіз пам'яток, що охоплює створення фонових мап географічного розподілу пам'яток та формування інтерактивної карти за населеними пунктами.

Під час дослідження атрибутів пам'яток, виявлено різноманіття форматів та способів написання датувань, що ускладнювало проведення подальшого аналізу, тому розроблено метод, що уніфіковує датування до єдиного формату.

Результатом роботи є парсер для пам'яток культури, що забезпечує ефективну обробку та зберігання інформації про них у форматі JSON. Для кожного з атрибутів пам'яток розроблено методи для введення та перевірки інформації, що забезпечує цілісність даних. Обраховано оцінку ефективності парсера та встановлено, що 87% датувань реєстру спрощено та уніфіковано, і 99% охоронних номерів пройшло перевірку за допомогою регулярних виразів.

У реєстрі пам'яток культури виявлено неточності, що ускладнювали обробку та аналіз даних – різні формати та варіанти написання атрибутів, а також порожні поля. Однак, розроблений парсер забезпечує уніфікованість та неможливість порожніх значень, що спрощує майбутнє розширення або оновлення реєстру пам'яток культури та забезпечує ефективну роботу з даними.

Список використаної літератури

1. Petrosyan A. Common languages used for web content 2023, by share of websites / Ani Petrosyan – 2023.

2. Pykes K. Stemming and Lemmatization in Python / Kurtis Pykes. – 2023.

3. Webster J. TOKENIZATION AS THE INITIAL PHASE IN NLP / J. Webster, K. Chunyu // City Polytechnic of Hong Kong. – 1992.

4. Kurama V. Information Extraction / Vihar Kurama. – 2021.

5. Державний реєстр нерухомих пам'яток України [Електронний ресурс] – Режим доступу до ресурсу: https://mkip.gov.ua/content/derzhavniy–reestr–neruhomih–pamyatok–ukraini.html.

6. UDPipe [Електронний ресурс] – Режим доступу до ресурсу: https://lindat.mff.cuni.cz/services/udpipe/run.php.

7. Straka M. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing / M. Straka, J. Hajic, J. Strakova.

8. Saladas J. What is cURL and how does it relate to APIs? / Johanna Saladas. – 2021.

9. Vega-Altair: Declarative Visualization in Python [Електронний ресурс] – Режим доступу до ресурсу: https://altair–viz.github.io/index.html.

10. Datawrapper [Електронний ресурс] – Режим доступу до ресурсу: https://www.datawrapper.de/.

11. GeoPy’s documentation [Електронний ресурс] – Режим доступу до ресурсу: https://geopy.readthedocs.io/en/stable/#.

12. Kepler.gl [Електронний ресурс] – Режим доступу до ресурсу: https://docs.kepler.gl/docs/user–guides.

Додаток А

(обов’язковий)

Результати парсингу інформації про пам’ятку культури


```
{
 "monument_type": "Городище",
 "name": "Золотий мис",
 "location": "с. Широка Балка",
 "dating": "900–801",
 "location_description": "с. Широка Балка",
 "type": "Пам'ятка історії",
 "protection_number": "210016-Н",
 "decision": "Постанова Кабінету Міністрів України від 10.10.2012 №929"
}
```

Додаток Б

(обов’язковий)

Перелік прийнятих скорочень

NLP – Natural Language Processing – Обробка природної мови

PoS-tagging – Part-of-Speech tagging – Розмічування частин мови

DOCX – Document Extended

PDF – Portable Document Format

CSV – Comma-separated Values

UD – Universal Dependencies – Універсальні залежності

API – Application Programming Interface – Прикладний програмний інтерфейс

CoNLL-U – Computational Natural Language Learning

JSON – JavaScript Object Notation