

RULE-BASED NLP APPROACHES FOR ARCHITECTURAL MONUMENTS DOCUMENTS' EXTRACTION

ДОПОВІДАЧ: КИРИЛІН ЄГОР СЕРГІЙОВИЧ

КЕРІВНИК: СМІШ ОЛЕГ РУСЛАНОВИЧ

Мета роботи

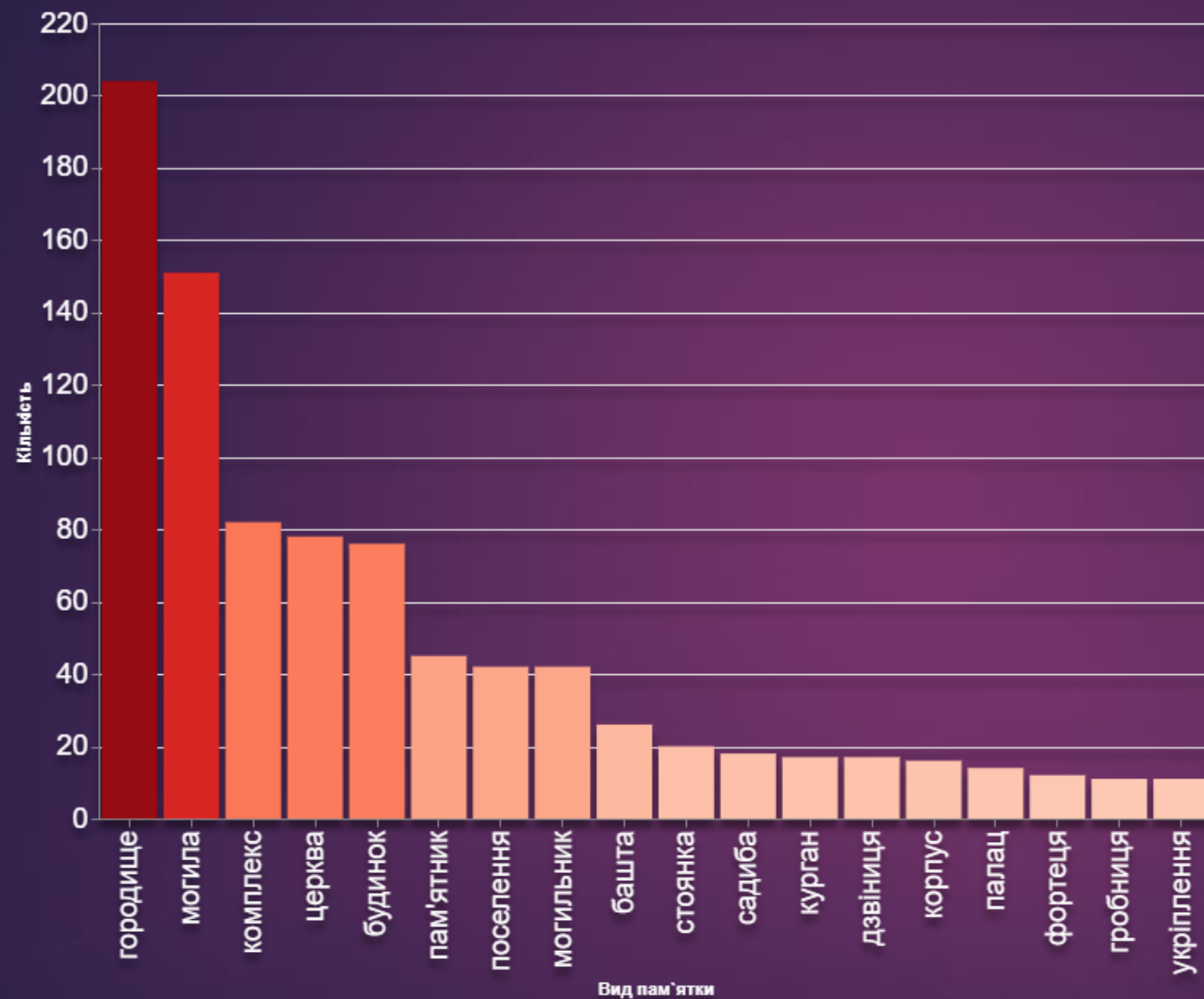
- ▶ Аналіз методів обробки природної мови
- ▶ Дослідження українськомовного набору даних з реєстру пам'яток культури України
- ▶ Розробка парсера для роботи з пам'ятками культури

Завдання роботи

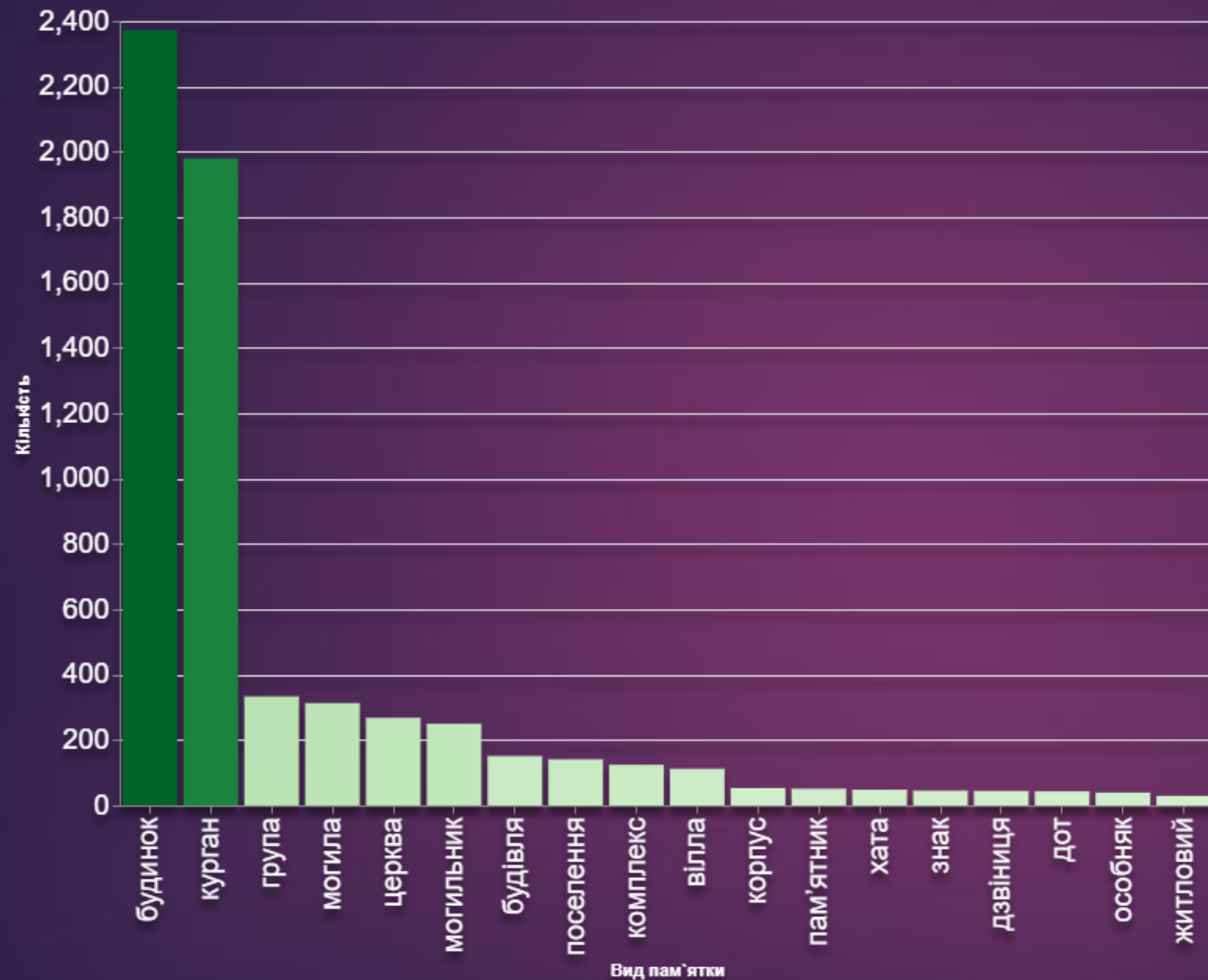
- ▶ Створити парсер для обробки інформації про пам'ятки культури на основі проаналізованих даних з реєстру

Методи обробки природної мови

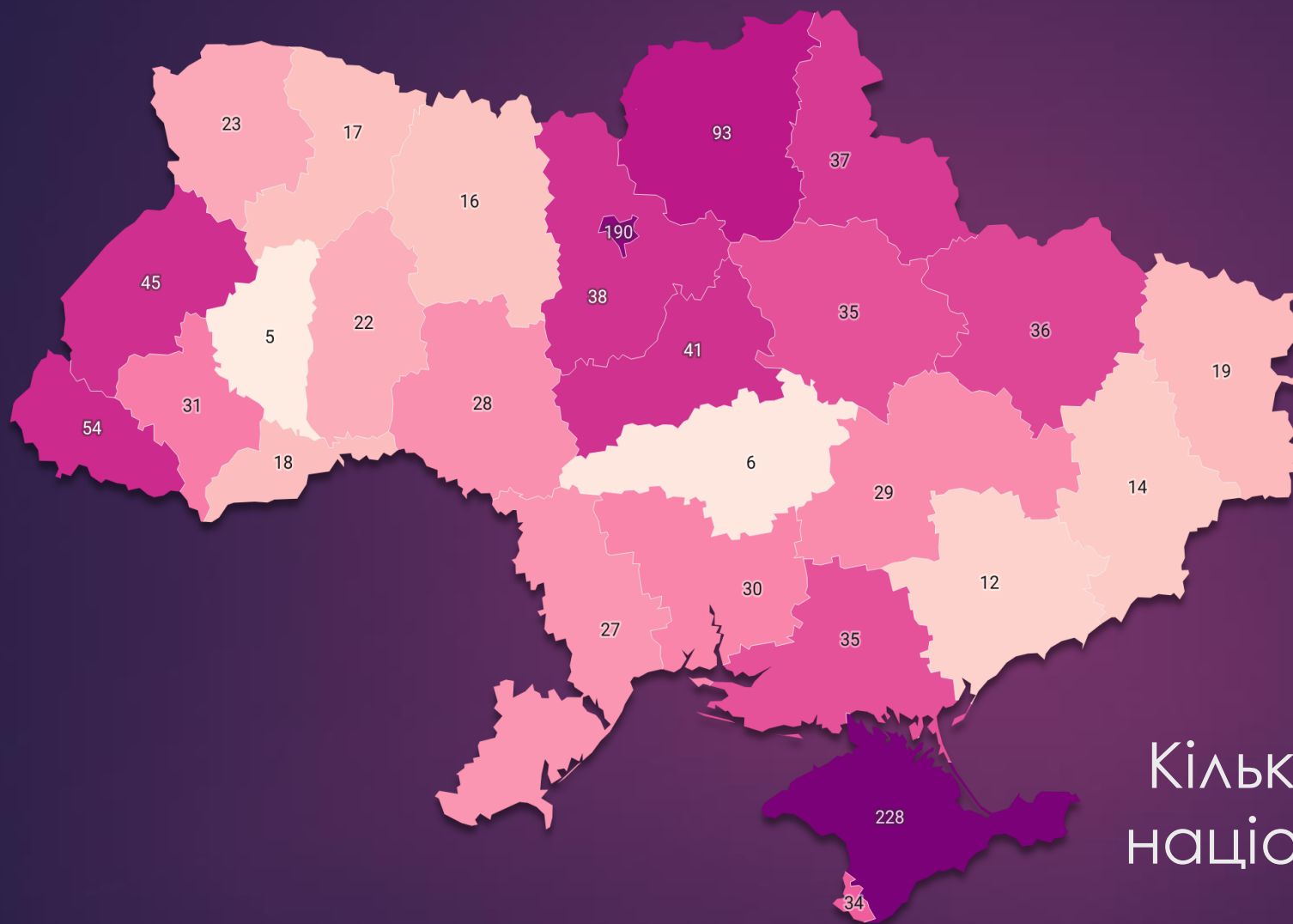
- ▶ Токенізація
- ▶ Лематизація
- ▶ Розмічування частин мови
- ▶ Видобування інформації



Кількісний аналіз типів
пам'яток
національного
значення



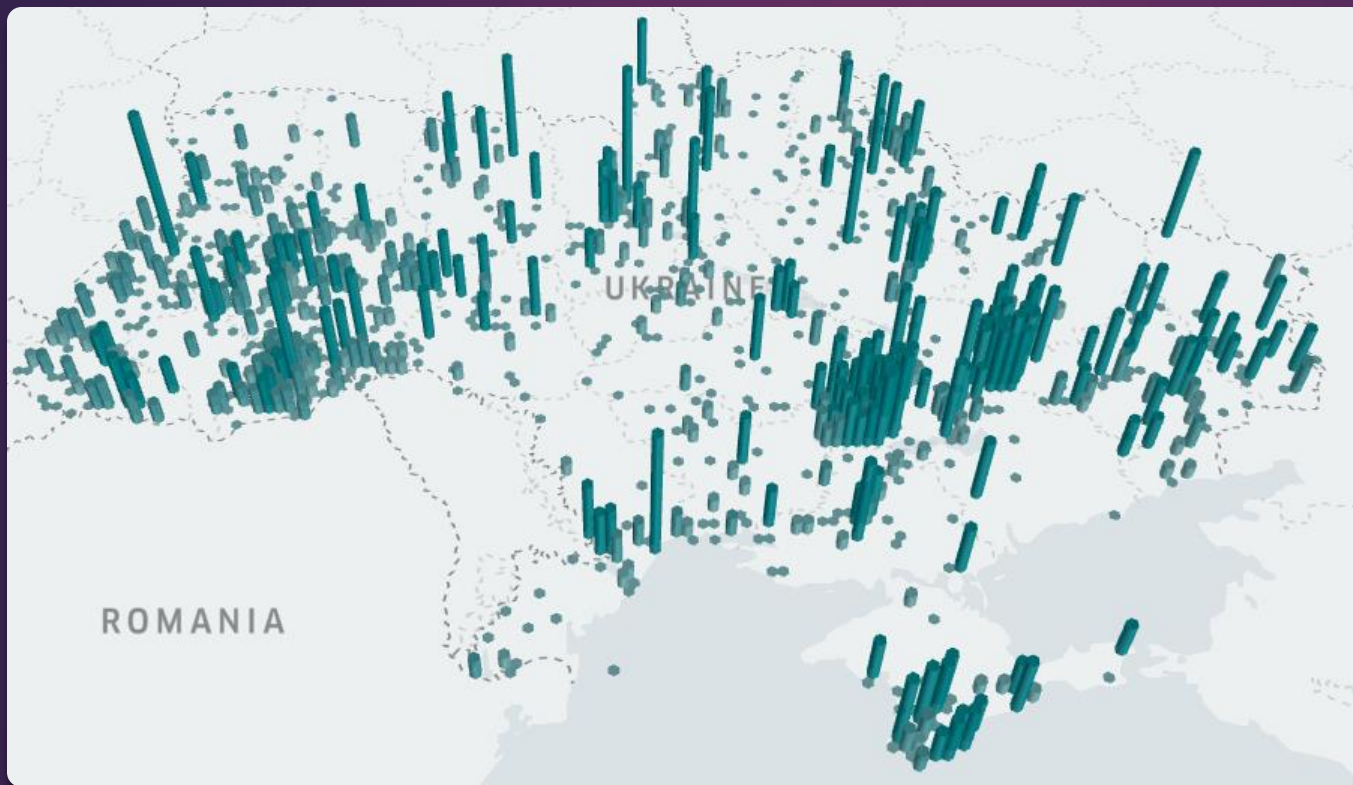
Кількісний аналіз типів
пам'яток місцевого
значення



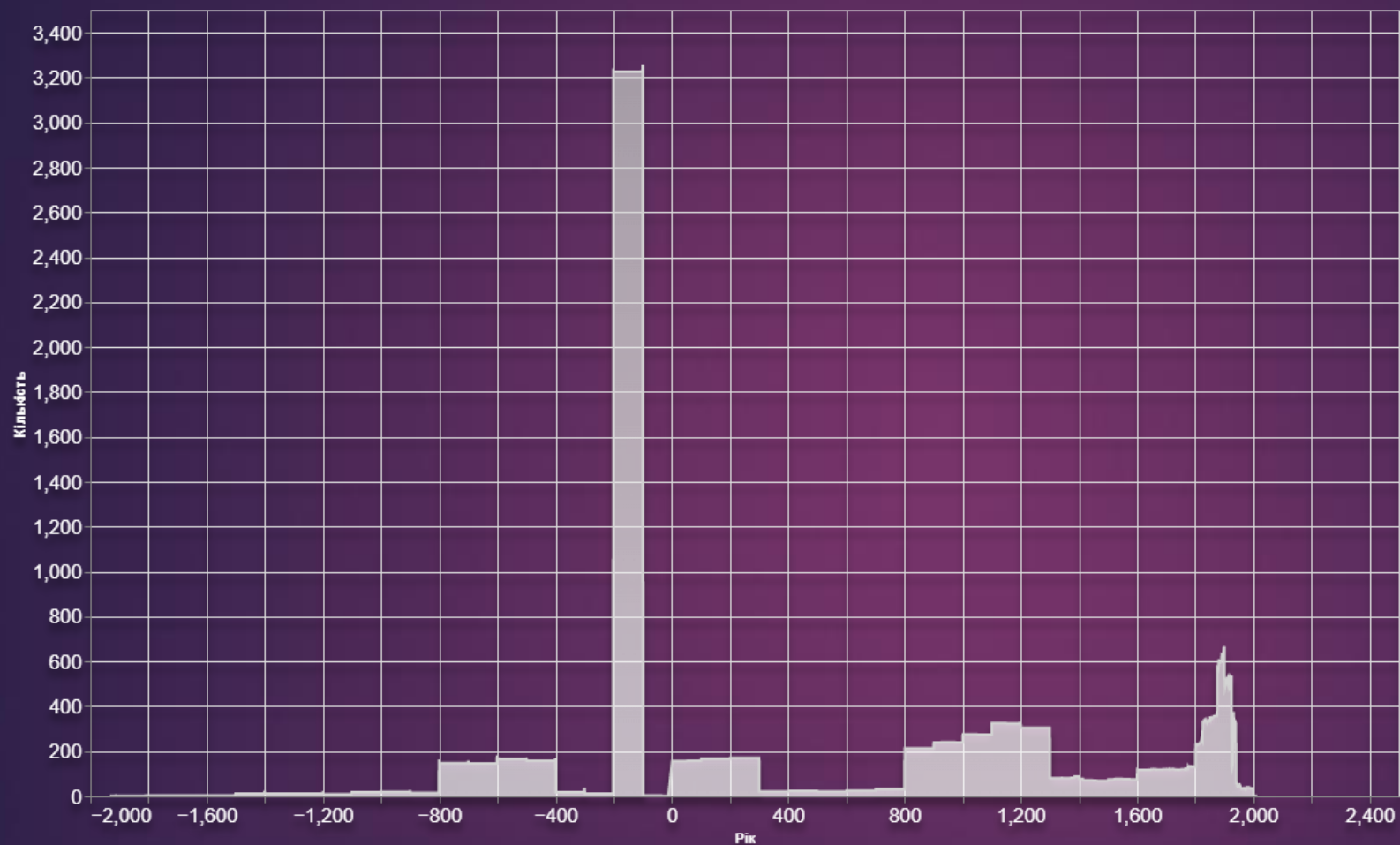
Кількість пам'яток культури
національного значення за
регіоном України



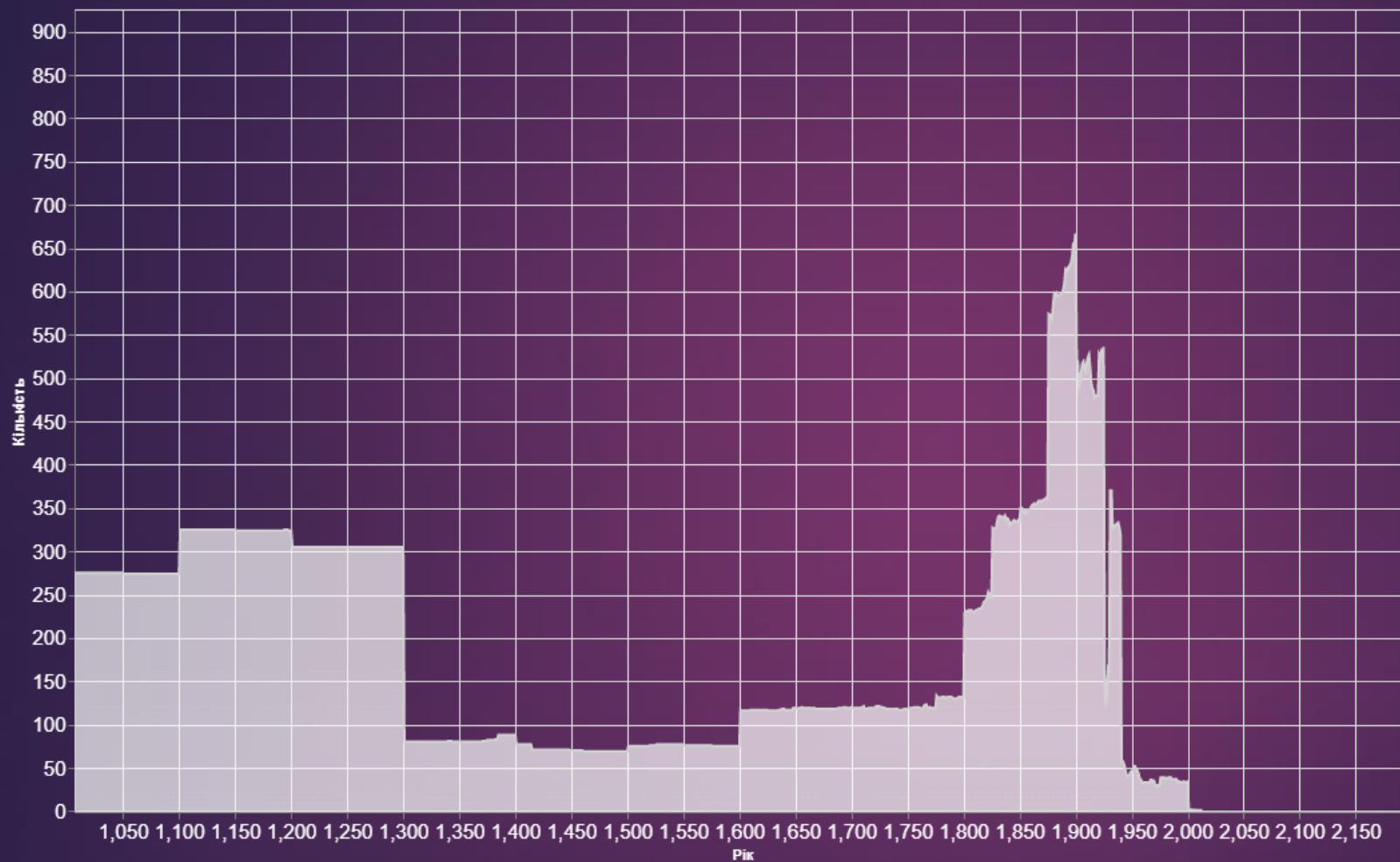
Кількість пам'яток культури місцевого значення за регіоном України



Інтерактивна мапа
географічного
розподілу пам'яток
культури за їхніми
населеними пунктами



Хронологічний
аналіз пам'яток
на найвизначніші
роки



Хронологічний
аналіз пам'яток на
найвизначніші роки
останнього
тисячоліття

Розробка парсера

- ▶ Збереження інформації у форматі JSON
- ▶ Головна перевага – уніфікований формат даних
- ▶ Неможливість порожніх значень – забезпечення цілісності даних
- ▶ Методи для кожного атрибуту

Результат роботи парсера

```
{  
  "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"  
}
```

ВИСНОВКИ

- ▶ Розглянуто технології обробки природної мови
- ▶ Здійснено збір інформації з реєстру
- ▶ Проведено кількісний, географічний та хронологічний аналізи пам'яток
- ▶ Розроблено парсер для уніфікації, обробки та зберігання інформації про пам'ятки

Дякую за увагу!