Кафедра комп'ютерних наук

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

на тему *Метод структурування тексту оголошень для об'єктів нерухомості засобами NLP*

Галузь знань *12 – Інформаційні технології*
Шифр і назва галузі знань

Спеціальність *122 – Комп'ютерні науки*
Шифр і назва спеціальності

Освітня програма *Комп'ютерні науки*
Назва освітньої програми

Виконав: *студент 2 курсу, група КНм-22-1* _____ *Д.Т. Карлечук*
Курс, група виконавця | Підпис | Ініціали, прізвище

Керівник: *к.т.н., доцент кафедри КН* _____ *Р.О. Багрій*
Науковий ступінь, посада | Підпис | Ініціали, прізвище

Нормоконтроль: *к.т.н., доцент кафедри КН* _____ *Р.О. Багрій*
Науковий ступінь, посада | Підпис | Ініціали, прізвище

До захисту допускаю:
Зав. кафедри КН, д.т.н., професор _____ *О.В. Бармак*
Підпис | Ініціали, прізвище

8 грудня 2023 р.

Хмельницький 2023

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Факультет *інформаційних технологій*
Кафедра *комп'ютерних наук*
Освітній ступінь *магістр*
Галузь знань *12 – Інформаційні технології*
Спеціальність *122 – Комп'ютерні науки*

ЗАТВЕРДЖУЮ
Завідувач кафедри *комп'ютерних наук*

_____
(підпис)
д.т.н., професор О.В. Бармак
«*01*» *вересня* *2023* року

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ МАГІСТРА

1. Тема кваліфікаційної роботи магістра: *«Метод структурування тексту оголошень для об'єктів нерухомості засобами NLP»*

2. Завдання видано студенту *Карлечуку Дмитру Траяновичу*
(прізвище, ім'я, по батькові)

3. Керівник роботи *доцент кафедри КН Багрій Руслан Олександрович*
(прізвище, ім'я, по батькові)

4. Затверджені наказом університету від «*15*» *серпня* *2023* р. № 30

5. Зміст пояснювальної записки (перелік задач) та вихідні дані:

Мета кваліфікаційної роботи магістра – розробити метод структурування тексту оголошень для об'єктів нерухомості для подальшої аналітики та візуалізації. Для досягнення поставленої мети було визначено наступні задачі: провести аналіз методів обробки природної мови; провести аналіз методу розпізнавання іменованих сутностей; провести аналіз методологій та підходів з ідентифікації та класифікації об'єктів у тексті; розробити метод структурування тексту оголошень; провести експериментальне тестування реалізованого сервісу за еталонними наборами даних;

Реферат

Кваліфікаційна робота магістра присвячена розробленню методу структурування тексту оголошень (тобто виявлення ключових слів) для об'єктів нерухомості для подальшої аналітики та візуалізації.

Актуальність теми. У процесі виконання магістерської роботи було розроблено метод структурування текстів оголошень у сфері нерухомості, використовуючи засоби обробки природної мови (NLP).

Структуровані дані являють собою високоорганізовану, фактичну та точну інформацію [1]. Вони мають чітко визначений та організований формат, який повинен відповідати стійкій схемі або моделі даних, яка визначає тип та розташування цих елементів у самому наборі даних. Відмінною рисою структурованих даних є їхня передбачуваність та послідовність. Зазвичай такі дані мають форму букв і цифр, які добре вписувати в рядки та стовпці таблиць, саме тому їх так часто подають у табличній формі аналогічній файлам Excel або документам Google.

Неструктуровані дані являють собою неорганізований набір інформації [2]. Вони не мають певної структури, представлені у всьому різноманітті форм і містять величезну кількість інформації, що робить їх ідеальним кандидатом для категоризації контенту та інтелектуального аналізу даних. До прикладів неструктурованих даних можна віднести зображення, текстові документи, відео та аудіо файли.

Процес перетворення неструктурованих даних у структуровані дані відіграє ключову роль у багатьох сферах, особливо у сфері обробки природної мови (NLP) – галузі штучного інтелекту, де застосовуються передові методи для трансформації та обробки тексту [3]. Структурування тексту має вирішальне значення, оскільки воно спрощує комплексну й розпливчасту інформацію до короткої та конкретної, цим самим збільшує її ясність і доступність, а також слугує основою для машинного навчання.

Використання алгоритмів з обробки природної мови у сфері нерухомості надасть можливість вилучити та категоризувати ключові відомості про сам об’єкт

нерухомості. Метод структурування текстів оголошень дозволить автоматично ідентифікувати та виділяти важливі сутності в неструктурованому тексті, що дозволить мати конкретну, визначену інформацію. Оскільки на сьогоднішній стан сфера нерухомості постійно розвивається в інформаційно-технологічному просторі – інтеграція NLP для структурування тексту оголошень є не лише розумним кроком, а й важливим внеском для подальшого розвитку цієї галузі.

Мета і задачі роботи. Метою даної магістерської роботи є розробка методу структурування тексту оголошень для об'єктів нерухомості для подальшої аналітики та візуалізації. Для виконання поставленої мети було визначено наступні завдання:

- провести аналіз методів обробки природної мови;
- провести аналіз методу розпізнавання іменованих сутностей;
- провести аналіз методологій та підходів з ідентифікації та класифікації об’єктів у тексті;
- розробити метод структурування тексту оголошень;
- провести експериментальне тестування реалізованого сервісу за еталонними наборами даних;

Об’єкт дослідження – процес визначення ключових слів для структурування текстової інформації

Предмет дослідження – моделі, методи та технології обробки текстових даних для структурування текстової інформації.

Наукова новизна одержаних результатів. В результаті проведеної роботи були отримані наступні результати:

- вдосконалено метод розпізнавання іменованих сутностей (NER) для задачі категоризації об'єктів у текстах оголошень. Це досягається за рахунок використання архітектурного підходу EEAP для ідентифікація та класифікація сутностей.
- розроблено метод структурування тексту оголошень для об'єктів нерухомості за допомогою методу розпізнавання іменованих сутностей. Це дозволило визначити послідовності слів, які представляють іменовані об'єкти в тексті оголошення.

Практичне значення одержаних результатів. У результаті проведених досліджень – розроблений метод допомагає виявити та вилучити важливі дані про об’єкт нерухомості із неструктурованого тексту оголошення, розподілити їх по категоріях та зберегти їх у базу даних. Практична значимість цих результатів полягає в їхньому потенціалі для оптимізації маркетингу нерухомості, поліпшення взаємодії з користувачами, прийняття важливих інвестиційних рішень на основі структурованих даних та забезпечення конкурентної переваги. Ці переваги можуть змінити те, як індустрія нерухомості взаємодіє з клієнтами та інвесторами в епоху цифрових технологій.

Апробація результатів кваліфікаційної роботи магістра та публікації.

Основні наукові й практичні результати кваліфікаційної роботи магістра доповідались на XV Всеукраїнській науково-практичній конференції «Актуальні проблеми комп’ютерних наук АПКН-2023» (17-18 листопада 2023 року) у доповіді на тему «Метод структурування тексту оголошень для об'єктів нерухомості засобами NLP» за темою роботи автором виконано наукову публікацію: Карлечук Д.Т., Багрій Р.О., Скрипник Т.К., Тищенко О.О. Метод структурування тексту оголошень для об'єктів нерухомості засобами NLP. Збірник наукових праць за матеріалами XV Всеукраїнської науково-практичної конференції «Актуальні проблеми комп’ютерних наук АПКН-2023». - Хмельницький, 2023. - С. 111-115.

Структура та обсяг роботи. Кваліфікаційна робота магістра складається із завдання, реферату, змісту, переліку скорочень, вступу, 4 розділів, висновків, переліку посилань із 41 найменувань та 5 додатків. Загальний обсяг кваліфікаційної роботи магістра становить 105 сторінок, з них 84 сторінок основного тексту та 21 сторінок додатків. У роботі наведено 38 рисунків та 4 таблиці.

Ключові слова: НЛП, оголошення, структуровані дані, нерухомість, метод, забудова, неструктуровані дані.

Зміст

Перелік скорочень 4
Вступ 5
РОЗДІЛ 1 Аналіз методів структурування текстів оголошень засобами NLP 8
1.1 Опис предметної області 8
1.2 Обробка текстової інформації 10
1.3 Аналіз методу обробки природної мови з розпізнавання іменованих сутностей 13
1.4 Аналіз існуючих підходів розпізнавання іменованих сутностей 16
1.5 Постановка задачі 22
Висновки до розділу 1 23
РОЗДІЛ 2 Аналіз механізму та алгоритму роботи по розпізнаванню іменованих сутностей 24
2.1 Основні принципи керування та обробки текстової інформації 24
2.2 Попередня обробка текстової інформації 27
2.2.1 Токенізація текстових даних 27
2.2.2 Усунення шуму в текстових даних 30
2.2.3 Лематизація та стемінг в текстових даних 31
2.2.4 Порівняння підходів для попередньої обробки текстових даних 33
2.3 Ідентифікація та класифікація сутностей з тексту оголошень 34
2.4 Оцінки ефективності та якості розпізнавання 38
Висновки до розділу 2 41
РОЗДІЛ 3 Проектування інформаційної системи структурування текстів оголошень для об’єктів нерухомості засобами NLP 43
3.1 Вимоги до розроблених програмних засобів 43
3.2 Обґрунтування вибору засобів розробки програмного застосунку з розпізнавання сутностей 44
3.3 Архітектура та процеси бібліотеки SpaCy 44
3.4 Компоненти модулів програмного застосунку розпізнавання сутностей засобами NLP 50

Висновки до розділу 3 53
РОЗДІЛ 4 Дослідження методу розпізнавання сутностей для об’єктів нерухомості засобами NLP 54
4.1 Характеристика датасету та набору даних для тренування моделі з розпізнавання сутностей 54
4.2 Налаштування та тренування моделі для розпізнавання іменованих сутностей. 61
4.3 Дослідження ефективності методу розпізнавання іменованих сутностей 65
Висновки до розділу 4 73
Загальні висновки 74
Перелік посилань 76
ДОДАТКИ

Перелік скорочень

Скорочення, термін, позначення	Пояснення
НЛП	Нейролінгвістичне програмування
РІС	Розпізнавання іменованих сутностей
NLP	Neuro-linguistic programming
NER	Named-entity recognition
БД	База даних
ІТ	Інформаційні технології
ПЗ	Програмне забезпечення
ПП	Програмний продукт
СКБД	Система керування базами даних
ОН	Об’єкти нерухомості
ШІ	Штучний інтелект
EEAP	Embed Encode Attend Predict

Вступ

Кваліфікаційна робота магістра присвячена розробленню методу структурування тексту оголошень (тобто виявлення ключових слів) для об'єктів нерухомості для подальшої аналітики та візуалізації.

Актуальність теми. У процесі виконання магістерської роботи було розроблено метод структурування текстів оголошень у сфері нерухомості, використовуючи засоби обробки природної мови (NLP).

Структуровані дані являють собою високоорганізовану, фактичну та точну інформацію [1]. Вони мають чітко визначений та організований формат, який повинен відповідати стійкій схемі або моделі даних, яка визначає тип та розташування цих елементів у самому наборі даних. Відмінною рисою структурованих даних є їхня передбачуваність та послідовність. Зазвичай такі дані мають форму букв і цифр, які добре вписувати в рядки та стовпці таблиць, саме тому їх так часто подають у табличній формі аналогічній файлам Excel або документам Google.

Неструктуровані дані являють собою неорганізований набір інформації [2]. Вони не мають певної структури, представлені у всьому різноманітті форм і містять величезну кількість інформації, що робить їх ідеальним кандидатом для категоризації контенту та інтелектуального аналізу даних. До прикладів неструктурованих даних можна віднести зображення, текстові документи, відео та аудіо файли.

Процес перетворення неструктурованих даних у структуровані дані відіграє ключову роль у багатьох сферах, особливо у сфері обробки природної мови (NLP) – галузі штучного інтелекту, де застосовуються передові методи для трансформації та обробки тексту [3]. Структурування тексту має вирішальне значення, оскільки воно спрощує комплексну й розпливчасту інформацію до короткої та конкретної, цим самим збільшує її ясність і доступність, а також слугує основою для машинного навчання.

Використання алгоритмів з обробки природної мови у сфері нерухомості надасть можливість вилучити та категоризувати ключові відомості про сам об’єкт нерухомості. Метод структурування текстів оголошень дозволить автоматично ідентифікувати та виділяти важливі сутності в неструктурованому тексті, що дозволить мати конкретну, визначену інформацію. Оскільки на сьогоднішній стан сфера нерухомості постійно розвивається в інформаційно-технологічному просторі – інтеграція NLP для структурування тексту оголошень є не лише розумним кроком, а й важливим внеском для подальшого розвитку цієї галузі.

Мета і задачі роботи. Метою даної магістерської роботи є розробка методу структурування тексту оголошень для об'єктів нерухомості для подальшої аналітики та візуалізації. Для виконання поставленої мети було визначено наступні завдання:

- провести аналіз методів обробки природної мови;
- провести аналіз методу розпізнавання іменованих сутностей;
- провести аналіз методологій та підходів з ідентифікації та класифікації об’єктів у тексті;
- розробити метод структурування тексту оголошень;
- провести експериментальне тестування реалізованого сервісу за еталонними наборами даних;

Об’єкт дослідження – процес визначення ключових слів для структурування текстової інформації

Предмет дослідження – моделі, методи та технології обробки текстових даних для структурування текстової інформації.

Наукова новизна одержаних результатів. В результаті проведеної роботи були отримані наступні результати:

- вдосконалено метод розпізнавання іменованих сутностей (NER) для задачі категоризації об'єктів у текстах оголошень. Це досягається за рахунок використання архітектурного підходу EEAP для ідентифікація та класифікація сутностей.
- розроблено метод структурування тексту оголошень для об'єктів нерухомості за допомогою методу розпізнавання іменованих сутностей. Це

дозволило визначити послідовності слів, які представляють іменовані об'єкти в тексті оголошення.

Практичне значення одержаних результатів. У результаті проведених досліджень – розроблений метод допомагає виявити та вилучити важливі дані про об’єкт нерухомості із неструктурованого тексту оголошення, розподілити їх по категоріях та зберегти їх у базу даних. Практична значимість цих результатів полягає в їхньому потенціалі для оптимізації маркетингу нерухомості, поліпшення взаємодії з користувачами, прийняття важливих інвестиційних рішень на основі структурованих даних та забезпечення конкурентної переваги. Ці переваги можуть змінити те, як індустрія нерухомості взаємодіє з клієнтами та інвесторами в епоху цифрових технологій.

Апробація результатів кваліфікаційної роботи магістра та публікації.

Основні наукові й практичні результати кваліфікаційної роботи магістра доповідались на XV Всеукраїнській науково-практичній конференції «Актуальні проблеми комп’ютерних наук АПКН-2023» (17-18 листопада 2023 року) у доповіді на тему «Метод структурування тексту оголошень для об'єктів нерухомості засобами NLP» за темою роботи автором виконано наукову публікацію: Карлечук Д.Т., Багрій Р.О., Скрипник Т.К., Тищенко О.О. Метод структурування тексту оголошень для об'єктів нерухомості засобами NLP. Збірник наукових праць за матеріалами XV Всеукраїнської науково-практичної конференції «Актуальні проблеми комп’ютерних наук АПКН-2023». - Хмельницький, 2023. - С. 111-115.

Структура та обсяг роботи. Кваліфікаційна робота магістра складається із завдання, реферату, змісту, переліку скорочень, вступу, 4 розділів, висновків, переліку посилань із 41 найменувань та 5 додатків. Загальний обсяг кваліфікаційної роботи магістра становить 105 сторінок, з них 84 сторінок основного тексту та 21 сторінок додатків. У роботі наведено 38 рисунків та 4 таблиці.

Ключові слова: НЛП, NER, оголошення, структуровані дані, нерухомість, метод, забудова, неструктуровані дані.

РОЗДІЛ 1 Аналіз методів структурування текстів оголошень засобами NLP

1.1 Опис предметної області

Оголошення – це короткі текстові повідомлення, котрі містять різну за призначенням інформацію, зазвичай приватного рекламного характеру [4]. Зміст будь-якого оголошення варіюється залежно від його типу, але яке б це оголошення не було воно загалом повинно містити коротке повідомлення, яке може бути пропозицією або іншою відповідною інформацією. Наприклад, оголошення про надання послуг, може містити такі реквізити [4]:

- назва документа;
- заголовок;
- основний текст;
- адреса автора оголошення.

В свою чергу – оголошення, яке містить інформацію про певну подію, буде містити зовсім інші реквізити [4]:

- назва виду документа (оголошення);
- текст, що містить дату;
- місце проведення заходу;
- хто є організатором;
- зміст події та умови;

Зі свого боку – оголошення, у якому наведена інформація про купівлю/продаж об’єкта нерухомості матиме свої дані релевантні для цієї галузі. Наприклад, до таких даних можна віднести: назва виду документа (оголошення про купівлю/продаж ділянки), ціна маєтку, адреса розташування, кількість кімнат, стан нерухомості й інші.

Як бачимо, інформація яка подається в оголошеннях існує в багатьох різних формах, розмірах, кількості та якості, але яка би то інформація не була, вона може бути представлена у вигляді структурованих і неструктурованих даних.

Інформація, яку можна побачити в оголошенні про об’єкт нерухомості являється прикладом неструктурованих даних, оскільки у ньому вся інформація не має визначеної структури, включає весь об'єм текстових даних та розкинута по різним частинам оголошення, як це зображено нижче на рисунках 1.1-1.2.

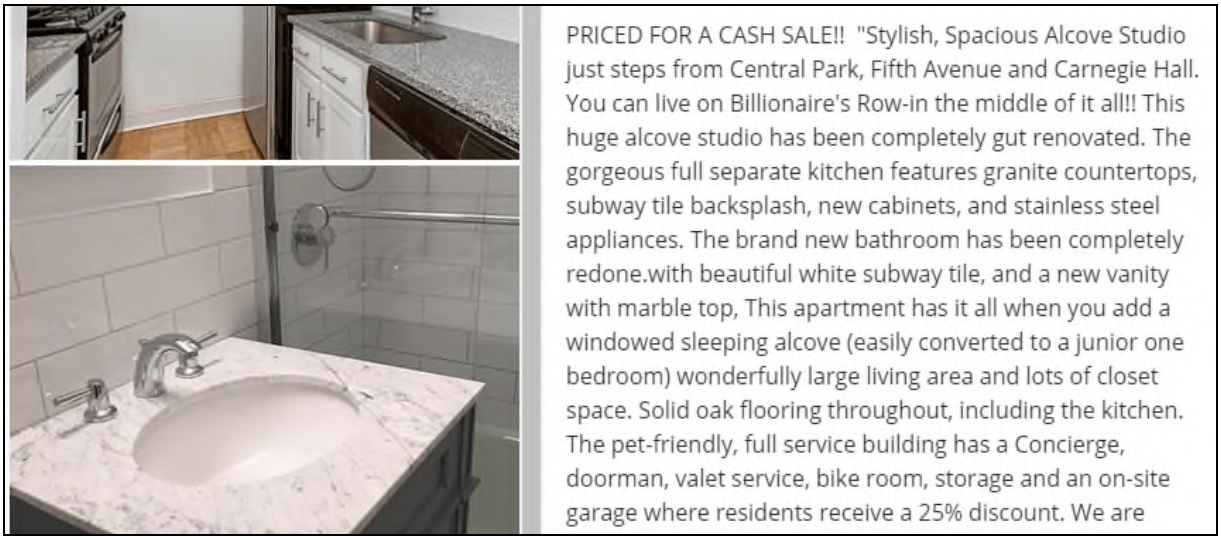


Рисунок 1.1 – Приклад оголошення з продажу нерухомості

Рисунок 1.2 – Ще один приклад подання інформації про нерухомість

Структуровані дані – навпаки, являють собою організовані, визначені, конкретні та відформатовані дані, що спрощує їх пошук та аналіз. Конкретний приклад структурованих даних зображений нижче на рисунку 1.3.

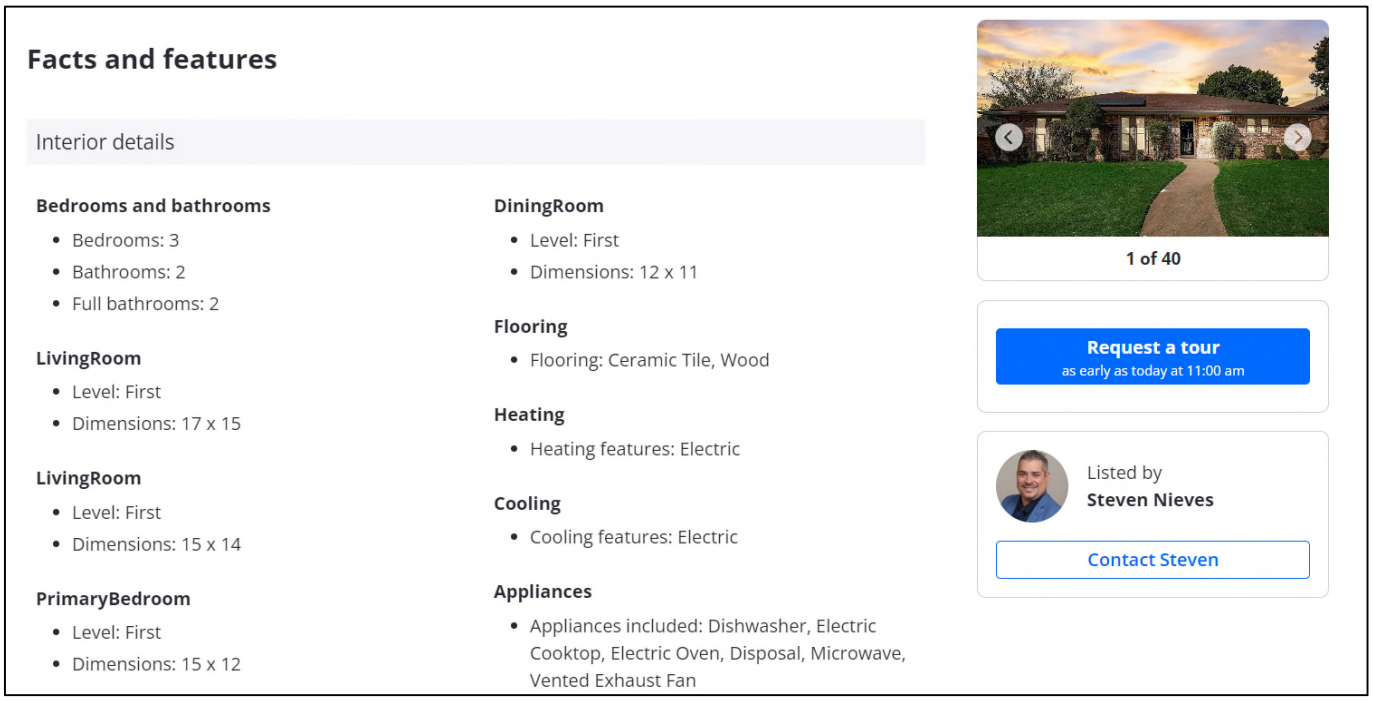


Рисунок 1.3 – Приклад структурованих даних із нерухомості

Сам процес отримання структурованих даних (таких, як кількість кімнат, стан маєтку, функціональні можливості приміщення) з неструктурованих називається структуруванням тексту. Щоб отримати необхідну інформацію з тексту потрібно використовувати різні методи, прийоми та цілі підходи для вирішення такого завдання, але на сучасний стан уже існує багато інструментів, засобів та утиліт з обробки текстової інформації, які спрощують роботу в цьому напрямку.

1.2 Обробка текстової інформації

Обробка текстової інформації – це важливий процес у сучасному інформаційному суспільстві, який включає в себе різні методи й техніки для аналізу, інтерпретації, екстракції і використання текстових даних. Оскільки обсяг текстових даних постійно продовжує зростати, необхідно використовувати досконалі та

ефективні методи з роботою над текстовою інформацією. На теперішній стан існує величезна область обробки природної мови, яка має різноманітний запас можливостей по роботі з текстовими даними.

Обробка природної мови (Natural Language Processing) – це область штучного інтелекту (ШІ), яка базується на тому, щоб надати комп'ютерам здатності читати та розуміти, як текстові дані так і розмовні слова [5]. НЛП не лише дозволяє комп'ютерам читати текстову інформацію, але й розуміти настрої в повідомленні, інтерпретувати їх та згенерувати новий текст на основі заданого. Вперше термін НЛП було представлено Річардом Бендлером та Джоном Гріндером у 1970 році [6] і розшифровувалося, як нейролінгвістичне програмування, що використовується в психотерапії. Разом вони вивчали основні моделі спілкування та поведінки, які основуються на терапевтичних методах та методах комунікації.

Учені намагалися розпізнати закономірності в мисленні та діях успішних особистостей, щоб використати їх для навчання інших людей, та змогли опублікувати при цьому серію видатних досліджень [7]. Ці дослідження заклали основу нейролінгвістичного програмування та ознаменували його ранню спрямованість на розуміння і моделювання людської поведінки, та розпізнавання шаблонів спілкування. Трохи згодом, з появою та розвитком інформаційних технологій, термін НЛП, як нейролінгвістичне програмування (Neuro-Linguistic Programming), розвинувся в обробку природної мови (Natural Language Processing) і фактично являє собою інтелектуальний аналіз тексту – спрямований на вирішення великого спектру завдань з обробки текстової інформації [8].

Обробка природної мови охоплює широкий спектр задач, які можна розділити на дві категорії [9], зображені на рисунку 1.4. З рисунку видно, що існує два основних типи завдань, заснованих на НЛП:

- аналіз текстової інформації;
- генерація текстових даних;

Рисунок 1.4 – Завдання засновані на НЛП

До першої категорії відноситься аналіз тексту. Аналіз тексту – це процес отримання необхідної інформації з неструктурованих текстових даних [10]. Він передбачає використання обчислювальних методів та алгоритмів обробки природної мови для перетворення великих обсягів текстової інформації в структуровані дані, з якими легко взаємодіяти. Текстовий аналіз допомагає організаціям та дослідникам розібратися в текстових даних, перетворюючи їх у формат, який можна аналізувати, візуалізувати та використовувати для обґрунтування бізнес-стратегій, наукових досліджень та інших додатків.

До другої категорії відноситься генерація тексту. Генерація тексту – це автоматизований процес створення людського текстового контенту з використанням комп'ютерних алгоритмів та методів обробки природної мови [11]. Він передбачає створення зв'язків між словами, у такий спосіб створюючи відповідний текст, який може включати що завгодно: відповіді на питання, переклад, моделювання інформації, тощо. Створення тексту відіграє вирішальну роль у наданні машинам

можливості спілкуватися з людьми більш природним та контекстно-орієнтованим способом.

Як видно з рисунка 1.1 – обробка природної мови охоплює широкий спектр завдань у сучасній обробці даних. Ці завдання варіюються від аналізу до генерації тексту, але разом спрямовані на те, щоб дозволити комп’ютерам розуміти, інтерпретувати й генерувати людську мову. Не зважаючи на таку велику різноманітність завдань, у сфері НЛП вирізняється одна задача, яка слугує фундаментальною основою цієї області та робить значний внесок у її ефективність та універсальність. Це задача є не що іншим, як розпізнаванням об'єктів у тексті (Named Entity Recognition), скорочено називається NER. Вона є важливою для різних додатків та інших завдань обробки природної мови та одночасно слугує ідеальним кандидатом у структуруванні текстів оголошень.

1.3 Аналіз методу обробки природної мови з розпізнавання іменованих сутностей

Розпізнавання іменованих сутностей (NER) – це метод обробки природної мови (NLP), що використовується для ідентифікації та класифікації іменованих об'єктів у неструктурованому тексті за заздалегідь визначеними категоріями [12]. Іменовані об'єкти – це слова або фрази, які представляють певні сутності, такі як місце розташування, організації, дати, імена людей, числові значення або інші категорії з унікальними назвами, тощо. NER необхідний для розпізнавання конкретних фрагментів інформації в текстовому наборі даних, і їх класифікації за встановленими категоріями.

У 1997 році фраза "Розпізнавання іменованих сутностей" була вперше використана Чинчором і Робінсоном на 6-й конференції з розпізнавання повідомлень (MUC) для ідентифікації висловлювань [13]. Від свого моменту появи, термін «Розпізнавання іменованих сутностей» значно змінився і еволюціонував, та на сьогоднішній день має широкий спектр застосування з різних галузей, ось деякі з них:

- NER у соціальних мережах *«Weakly labeled data augmentation for social media named entity recognition»* [14]. У дослідженні обговорюється використання методу розпізнавання іменованих сутностей у контексті соціальних мереж та користувацькому контенті відомому, як блогінг. Тут ілюструються реальні проблеми, з якими стикається NER. Контент соціальних мереж, наприклад твіти, часто буває коротким, неформальним та містить багато непотрібної інформації, що створює проблеми для точного розпізнавання об'єктів. Запропонований метод NER, який використовує *передавальне навчання (transfer learning)* [15], бере знання з вихідного набору даних для підвищення продуктивності роботи в цільовому наборі даних. У такий спосіб, метод NER використовується в цьому контексті для покращення розуміння та обробки текстових даних із соціальних мереж, де контент часто короткий, неформальний та складний для розуміння. Запропонований підхід спрямований на підвищення ефективності обробки контенту користувача на платформах соціальних мереж.

- NER у біомедичній галузі *«Proximity-based document representation for named entity retrieval»* [16]. Згідно цього дослідження, у медичній сфері NER відіграє життєво важливу роль, отримуючи ключову інформацію з клінічних записів пацієнтів, забезпечуючи цим вхідні дані для наступних задач, таких, як формування статусу клінічного затвердження, реєстрація дозволу на використання клінічних інструментів та ідентифікація конфіденційних даних. Значимість цього дослідження полягає в нестачі зручних у використанні інструментів NER, адаптованих для клінічної області. Результати дослідження вказують на можливість впровадження передових алгоритмів NER у клінічні робочі процеси НЛП, задовольняючи гостру потребу у цій галузі.

- Системи запитань та відповідей «*An approach for extracting exact answers to question answering (QA) system for English sentences*» [17]. Дослідження підкреслює важливість розпізнавання іменованих об'єктів для систем контролю якості, забезпечуючи цим точні та інформативні відповіді. Системи контролю якості покладаються на здатність розуміти зміст питань та шукати відповідну інформацію у великих текстових словниках. NER допомагає в цьому процесі, оскільки визначає

конкретні об'єкти в тексті, дозволяючи системам контролю якості отримувати найбільш релевантні та контекстуально точні відповіді. Важливість NER у таких системах допомагає фільтрувати та отримувати потрібну інформацію з великого набору даних, підвищуючи ефективність у наданні цінних відповідей на запити реальних справжніх користувачів.

- Машинний переклад [18]. У досліджені «*Transliteration normalization for information extraction and machine translation*», транслітерація призводить до відмінностей у тому, як імена, місця чи об'єкти представляються у різних алфавітах чи мовах. Запропонована система NER допомагає розпізнавати та нормалізувати ці різноманіття, спрощуючи вилучення інформації та роботу з нею.

Використання NER підвищує точність вилучення інформації, гарантуючи, що правильні об'єкти ідентифікуються незалежно від алфавіту чи мови із збереженням контексту при перекладі тексту. Запропонований метод гарантує, що контекст іменованих об'єктів зберігається під час транслітерації, оскільки це допомагає отримувати точні та значущі результати. Без NER ризик втрати контексту або неправильного перекладу іменованих об'єктів у транслітерованому тексті значно вищий.

Як можна побачити застосування NER виходить за рамки своєї основної ролі – ідентифікації сутностей, та зустрічається у різних сферах інформаційних технологій, таких як пошук інформації, машинний переклад та багато інших. У зв'язку з тим, що системи NER аналізують та класифікують іменовані об'єкти в текстових даних, вони формують основу для широкого спектру додатків, забезпечуючи ясність, контекст та ефективність при роботі з різними джерелами інформації.

Оскільки, застосування методу з розпізнавання сутностей можна застосувати в різних сферах обробки текстової інформації, він слугує ідеальним варіантом для ідентифікації даних із рекламних оголошень по об’єктах нерухомості, адже основна суть це розпізнавання даних, ідентифікація, та класифікація не важливо з яких предметних областей та категорій.

1.4 Аналіз існуючих підходів розпізнавання іменованих сутностей

Розпізнавання іменованих сутностей має декілька методологій та підходів, які використовуються для вирішення такого завдання як ідентифікація об’єктів у тексті. Кожен з цих підходів має свої унікальні переваги та сфери застосування. Найбільш поширеними з них є:

- підхід, що ґрунтується на правилах;
- контрольований підхід;
- неконтрольований підхід;

Підхід, що ґрунтується на правилах:

Підхід, що ґрунтується на правилах, включає набір правил, ретельно розроблених фахівцями, та опирається на заздалегідь визначені певні шаблони, які часто формулюються за допомогою регулярних виразів, граматичних правил або словників [20]. Ці правила точні й засновані на мовних, синтаксично-лексичних та предметно-специфічних знаннях. Такі правила обов'язково розробляються з урахуванням специфіки предметної галузі, оскільки кожна предметна область має більш точне та ефективне визначення того чи іншого слова.

Підходи, що ґрунтуються на правилах, не є надто узагальненими, оскільки створені правила обмежені лише конкретною областю, тому системи, що використовують цей підхід зазвичай не виходять за межі своєї галузі. Крім того, ці системи вимагають втручання людини для розробки правил у поєднанні з вищими навичками програмування. Нижче на рисунку 1.5 наведено короткий огляд роботи такого підходу [20]:

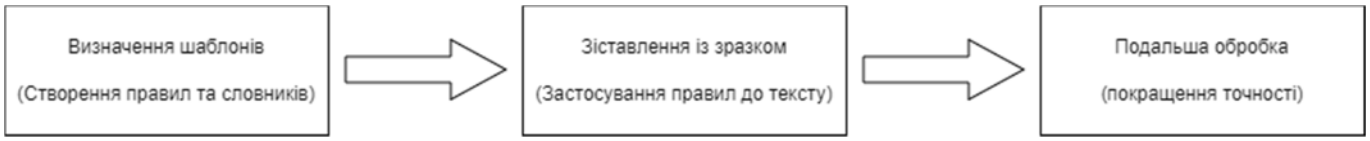


Рисунок 1.5 – Робота систем заснованих на словниках

1. Визначення шаблону. Системи NER, що базуються на правилах, починаються з визначення шаблонів та правил. Ці правила можуть бути як простими, наприклад, такими як регулярні вирази для зіставлення адреси електронної пошти, так і більш складними, що враховують лінгвістичний контекст слова.

2. Зіставлення зі зразком: Створені правила далі застосовуються до вхідного тексту. Коли правило відповідає певному сегменту тексту, воно позначається як розпізнана сутність.

3. Подальша обробка: Пройшовши етап зіставлення слів із зразком далі відбувається покращення початкової ідентифікації іменованих об'єктів. Цей процес уточнення спрямований на підвищення точності та послідовності об'єктів, що успішно були розпізнаними.

Підхід заснований на правилах може використовувати регулярні вирази для порівняння довгих складних рядків зі слів за допомогою використання базових простих або складних шаблонів. Їх можна використовувати для виявлення закономірностей і збігів елементів у тексті. *«Boosting drug named entity recognition using an aggregate classifier»* [21] детально описує застосування підходу на правилах, використовуючи регулярні вирази. У роботі запропонований підхід дозволив розпізнати назви ліків у тих ситуаціях, коли дані для навчання, пов’язані з медичними препаратами, недостатні або відсутні взагалі. Тут використовується покращений механізм розпізнавання, який може поєднувати прогнози від різних систем розпізнавання сутностей. Крім того, схожість та подібність рядків, у назвах медичних препаратів, знайдених у базі даних ліків, зросла за допомогою поєднання регулярних виразів із методом ідентифікації об’єктів.

Підхід, що ґрунтується на правилах, також можна використовувати з іншими методами у вигляді двоетапного процесу. Будь-який підхід, заснований на правилах, використовує початковий етап, у якому відбувається отримання даних, а наступний етап можна використати для класифікації сутностей за допомогою контрольованого, напівконтрольованого або неконтрольованого алгоритму, як було представлено у дослідженні «*A hybrid rule-based pattern extractor and named entity recognition*» [22].

Щоби безперешкодно розпізнати сутності на основі невеликого словнику даних та ідентифікувати вже існуючі об'єкти в наборі записів про аварії на дорогах пропонується підхід на основі правил та методу напівконтрольованого розпізнавання іменованих сутностей. Тут використовуються Стенфордський тегувальник частин мовлення (POS) та аналізатор граматики [21]. Екстрактор використовує мовний тегер для пошуку необхідної сутності та вилучення шаблону пов’язаною із цією сутністю. Щоб класифікувати об'єкти за певними категоріями, вилучення шаблонів далі передається до напівконтрольованого методу NER. Після цього оцінка точної відповідності використовується для оцінки продуктивності двох різних об'єктів таких, як дата та місце.

Існує кілька інших систем NER, що ґрунтуються на правилах, наприклад як в цьому дослідженні з вилучення інформації написаною різними мовами [23]. Розроблена система LaSIE ідентифікує необхідні сутності, не зважаючи на те, що вони написані 4-ма різними мовами. Створені словники а також інструменти текстових анотацій (токенізатор, лематизатор та тегер), допомагають класифікувати текст на кілька мов, таких як німецька, англійська, іспанська та італійська.

Хоча розглянуті джерела вдало класифікують та ідентифікують сутності, вони засновані на правилах та ґрунтуються на мовних, синтаксично-лексичних та предметно-орієнтованих знаннях, і через це мають певний основний недолік. Такі правила застосовуються лише до однієї предметної області, тому вони не є універсальними.

NER з використанням некерованого навчання:

Підхід до некерованого навчання, більш відомого як навчання без учителя, використовуються для даних, що не мають наперед заданого маркування чи помічених анотацій. Навчання без вчителя – це ціла парадигма машинного навчання, в якій моделі вивчають шаблони та структури на основі немаркованих даних [24]. На відміну від контрольованого навчання, для навчання не потрібні анотовані приклади. Неконтрольовані методи NER досліджують внутрішні структури тексту виявлення іменованих об'єктів, не покладаючись на попередньо

створені набори даних. Поширеними підходами до навчання без вчителя є асоціація та кластеризація. У соціальних мережах для вирішення проблеми недостатньої кількості інформації та занадто коротких твітів дослідження «*Two-stage NER for tweets with clustering*» [25] пропонує унікальну техніку, яка використовує алгоритм подібності твітів шляхом двоетапного розпізнавання іменованих сутностей для кількох схожих повідомлень. Зокрема, тут попередньо маркується кожен твіт за допомогою послідовного маркувальника, побудованого спочатку на моделі лінійних умовних випадкових полів. Після цього повідомлення діляться на кластери (групуються), щоб зібрати разом твіти, які мають схожий, подібний контент. Зрештою, для кожного кластера точно налаштовується мітка кожному повідомленні, використовуючи покращену модель умовного випадкового поля з алгоритмом іменованих сутностей, що враховує дані рівня кластера, тобто мітки поточного слова та сусідніх із ним слів у всіх повідомленнях у поточному кластері.

У біомедичній сфері, «*Study of Named Entity Recognition methods in biomedical field*» [26] розроблений метод для ідентифікації сутностей ліків легко адаптується до різних семантичних категорій та текстових жанрів, оскільки він не залежить від створених вручну екземплярів медичних препаратів та правил анотованих сутностей. Розпізнавання сутностей ґрунтується на термінології, поверхневому синтаксичному розумінні та статистиці даних, а не на словниках. Згідно з експериментальними результатами, цей метод дає результати на двох добре відомих наборах даних клінічних записів із різними жанрами, включаючи біомедичні дані та різні зв'язані типи об'єктів.

Оцінка методів навчання без вчителя складніша, ніж його аналог з учителем, оскільки для навчання моделі потрібні лише немарковані вхідні дані. Основна причина цього полягає в тому, що результати методів навчання без вчителя не мають міток та наперед визначених анотацій, які були б присутніми в створеному вручну наборі анотованих даних. Через відсутність міток для навчання моделі, немає і результативних міток присутня в якості навчання. Через це в результаті неможливо правильно оцінити алгоритм без учителя, використовуючи традиційну

парадигму оцінок порівняння з використанням таких показників продуктивності, як точність або F-показника [24].

NER з використанням керованого навчання (з учителем):

Навчання з учителем – це ще парадигма машинного навчання, в якій алгоритм вчиться робити прогнози чи рішення на основі наперед анотованих даних для навчання [27]. Під час навчання з учителем алгоритму надається набір даних, де кожен приклад складається з вхідних даних та відповідного правильного результату, або конкретної цілі. Мета алгоритму полягає в тому, щоб вивчити зіставлення вхідних даних та вихідних даних, аби дозволити йому робити точні прогнози на основі нових, наперед не тренованих даних. Термін «контрольоване навчання» стосується того факту, що алгоритм навчається під керівництвом «вчителя», який дає правильні відповіді під час навчання, що дозволяє йому узагальнювати та робити прогнози чи класифікації для майбутніх немаркованих даних.

Використовуючи підхід з учителем, для правильного функціонування систем NER, вирішальне значення має вибір правильного алгоритму навчання. Найбільш популярними алгоритмами в навчанні з учителем для розробки систем NER є наступні [27]:

- прихована марковська модель;
- умовне випадкове поле;
- метод опорних векторів;

У біомедичній галузі в дослідженні «*Medical named entity recognition based on dilated convolutional neural network*» [28] розроблено систему розпізнавання іменованих сутностей на основі прихованої марковської моделі. Система розпізнавання біомедичних іменованих об'єктів ідентифікує спеціальні явища, такі як абревіатури, та каскадні іменовані медичні дані. У цій статі інтегровано орфографічну, морфологічну та семантичну частини мови в метод розпізнавання сутностей, який працює на основі прихованої марковської моделі для розширення розпізнавання біомедичних іменованих об'єктів. Розроблений метод правильно зміг ідентифікувати всі подані на вхід речення, які містять інформацію пов’язаною з

біохімічними явищами. Обмеженням такого методу є лише те, що деякі складні конструкції в реченні, такі як "і/або", не можуть бути ефективно оброблені створеним методом. Не зважаючи на це, експеримент показує, що система працює помітно краще, ніж на попередніх підходах таких, як навчання без вчителя.

У наступному дослідженні, пов’язаному з категоризацією та ідентифікацією сутностей, на маратхській мові «*Named Entity Recognition using Conditional Random Fields*» [29], було розроблена статистична система NER, яка використовує умовні випадкові поля для ідентифікації та класифікації елементів. Алгоритм, який працює з попередньо створеним набором даних, що містить 27177 елементів, використовує умовні випадкові поля де іменовані об'єкти в системі виявляються та розподіляються за категоріями. Статистичні методи успішно ідентифікують і класифікують дані цією морфологічно складною мовою, але даних для тренування було замало, тому додаткова інформація врази підвищила би точність розпізнавання.

У наступному дослідженні «*Named entity correction in neural machine translation using the attention alignment map*» [30], запропоновано двофазний метод NER, заснований на методі опорних векторів, який включає етап визначення меж анотованих даних і етап семантичної категоризації. Даний підхід вирішує проблему незбалансованого розподілу класів, крім того, будучи заснованим на ієрархічній семантичній класифікації, покращує продуктивність та складність обчислення алгоритму розпізнавання.

Як бачимо, існує 3 основних підходи для побудови системи розпізнавання сутностей, кожен із них успішно використовується для вирішення завдань по ідентифікації об’єктів.

Підхід, що ґрунтується на правилах, спирається на заздалегідь визначені шаблони та лінгвістичні правила, що робить його легко інтерпретованим та легким у настроюванні, але обмежується тим наскільки складними є правила, тобто можливості системи обмежені рівнем складності та деталізації використовуваних нею правил.

Навчання з учителем включає в себе навчання моделі на анотованих даних для прогнозування сутностей, що дозволяє їй узагальнювати наперед нетреновані

дані на основі вивчених шаблонів. Цей підхід забезпечує високу точність, але вимагає значної кількості розмічених навчальних даних.

Навчання без вчителя не спирається на розмічено-анотовані дані, а натомість виявляє закономірності та об’єктивно існуючий зв’язок у тексті. Він більш гнучкіший, але може мати меншу точність через відсутність наперед правильно анотованих даних.

Згідно інформації з різних джерел та досліджень, розробка методу розпізнавання іменованих сутностей для структурування текстів оголошень із нерухомості буде використовувати контрольований підхід через наступні ознаки:

- контрольований підхід використовує анотовані дані в якості навчання моделі, використовуючи великий обсяг даних. Оскільки сфера оголошень з нерухомості має величезний набір готової інформації, з неї можна легко створити значний набір даних із наперед позначеними об'єктами;
- оголошення з нерухомості часто містять унікальні терміни та вислови, специфічні для цієї галузі. Контрольований підхід дозволить адаптувати модель NER для ефективного розпізнавання об'єктів, конкретних для цієї предметної області. За рахунок такого підходу можна точно налаштувати модель з урахуванням тонкощів опису властивостей, гарантуючи її оптимальну, та правильну роботу;
- на відміну від решти підходів, контрольований підхід, за час навчання, забезпечує високу точність та якість розпізнавання об'єктів. Враховуючи важливість правильної ідентифікації та структурування інформації про нерухомість, контрольований підхід є найкращим кандидатом для забезпечення надійності та точності класифікації даних.
- натреновану модель набагато легше масштабувати, та коректувати, маючи набір готових результатів, які показують повинна працювати модель.

1.5 Постановка задачі

Метою даної магістерської роботи є розробка методу структурування тексту оголошень для об'єктів нерухомості для подальшої аналітики та візуалізації. Тобто,

маючи текст оголошення по об’єкту нерухомості необхідно з нього виділити важливу інформацію таку, як кількість кімнат, ціна, місце розташування забудови і т.д. та зберегти її в базу даних. Для виконання поставленої мети було визначено наступні завдання:

- провести аналіз методів обробки природної мови;
- провести аналіз методу розпізнавання іменованих сутностей;
- провести аналіз методологій та підходів з ідентифікації та класифікації об’єктів у тексті;
- розробити метод структурування тексту оголошень;
- провести експериментальне тестування реалізованого сервісу за еталонними наборами даних;

Висновки до розділу 1

У першому розділі було проведено аналіз методів структурування текстів оголошень засобами NLP та здійснено детальний огляд предметної області з обробки природної мови у вигляді структурування текстів оголошень за допомогою засобів по обробці текстової інформації. Також було розглянуто основні підходи та стратегії по ідентифікації і класифікації сутностей у неструктурованому тексті і здійснено аналіз методу з розпізнавання іменованих сутностей.

Проведено дослідження існуючих публікацій та статей із використанням методу обробки природної мови по розпізнаванню іменованих сутностей для знаходження та категоризації даних у неструктурованому тексті, та описано їхні переваги й недоліки.

Визначено різноманітні методології з розпізнавання іменованих сутностей, включаючи контрольовані, неконтрольовані та засновані на правилах підходи, де більш детально було зосереджено увагу на контрольованому підході (навчанні з учителем) через його відмінні характеристики та помітні переваги над іншими підходами в рамках даної задачі.

РОЗДІЛ 2 Аналіз механізму та алгоритму роботи по розпізнаванню іменованих сутностей

2.1 Основні принципи керування та обробки текстової інформації

Галузь з обробки природної мови та її метод з розпізнавання іменованих сутностей, слугує важливою складовою між явно складною, неструктурованою природою людської мови та структурованими даними. На відміну від комп'ютерних мов, які засновані на строгих правилах та синтаксисі, людська мова має безліч нюансів і може сильно відрізнятися залежно від мовлення, ситуації та контексту. У цьому багатогранному мовному контексті завдання правильної ідентифікації та розпізнавання сутностей набуває важливого значення.

Якщо розглянути сферу нерухомості, сектор, який має свої унікальні нюанси, контекстно-залежну лексику, характерні тонкощі мови, та свій словниковий запас, то можна зробити висновок, що така галузь є досить багатогранною. Термінологія значень об’єктів нерухомості може сильно варіюється від одного маєтку до іншого, та містити специфічні терміни починаючи від земельної ділянки та закінчуючи багатоповерховим будинком. Оскільки така галузь вміщує в собі значні лексикони, то в ній обов’язково повинно використовувати добре навчені методи, які точно відображають тонкощі мови, пов'язаної з нерухомістю. Зважаючи на такі відмінності, ідентифікація та класифікація сутностей повинна мати здатність розшифровувати контекстуальні тонкощі та відповідним чином маркувати об'єкти, гарантуючи, що вони відповідають мовним особливостями цієї конкретної галузі. Саме тому метод розпізнавання сутностей, якраз і є призначений для вмілого подолання цих складнощів та тонкощів.

Метод розпізнавання іменованих сутностей працює наступним чином, зображеним на рисунку 1.2. NER-система приймає вхідне речення, розбиваючи його на слова, а на виході дає набір міток та відповідні їм дані, згадані у вхідному тексті. На рисунку 2.1 наглядно показаний випадок, як повинна працювати система NER на

прикладі правильної ідентифікації 3-ох сутностей у реченні, пов’язаних з об’єктом нерухомості [12].

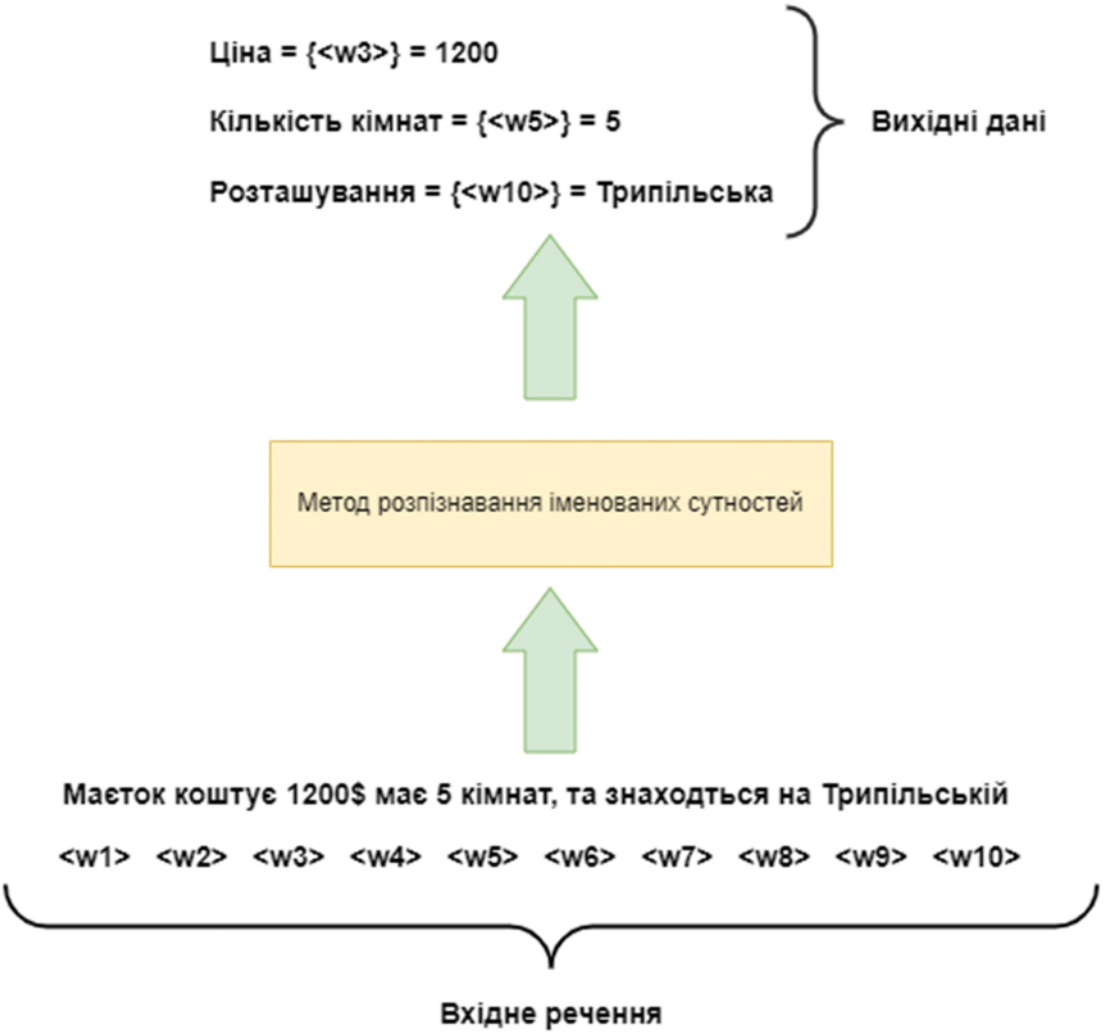


Рисунок 2.1 – Схема роботи методу розпізнавання сутностей в оголошенні

Як видно з рисунка вище, NER відіграє ключову роль у вилученні та класифікації сутностей із неструктурованих текстових даних, які присутні в оголошеннях. Система дає можливість розрізняти та маркувати важливу інформацію, тим самим, покращуючи розуміння тексту. Однак вся суть розпізнавання іменованих сутностей полягає не в тому, як гарно представлені результати на виході, а в алгоритмічних процесах, що лежать в основі методу, що дозволяють досягати високої точності в розпізнаванні об'єктів. Нижче на рисунку 2.2 наведено механізм, що лежить в основі обробки даних методу NER [19].

Рисунок 2.2 – Алгоритм роботи методу розпізнавання сутностей

Згідно рисунку, для того, щоб отримати розмічені дані, відбувається процес, роботу якого можна розділити на кілька ключових етапів:

1. Попередня обробка тексту. Перший крок включає підготовку текстових даних для аналізу. Зазвичай це включає такі завдання, як токенізація, видалення шумів, лематизація та стемінг.

2. Ідентифікація та класифікація сутності. Після попередньої обробки тексту алгоритми NER сканують його, щоб визначити послідовності слів, які відповідають об'єктам. Тому після попередньої обробки даних розробляється алгоритм розпізнавання сутностей. Алгоритм може автономно аналізувати текст і призначати певні теги, мітки, або категорії на основі вмісту текстових даних, наприклад, вказавши мітку «1200» як ціна а «Трипільська» як вулиця розташування об’єкта нерухомості.

3. Оцінка якості розпізнавання. За результатом розпізнавання, можна оцінити якість та точність розпізнавання інформації за рахунок перевірки натренованої моделі через метрики оцінок.

2.2 Попередня обробка текстової інформації

2.2.1 Токенізація текстових даних

Токенізація, або інша назва сегментація тексту – це процес розбиття цілого тексту на окремі одиниці, зазвичай слова, фрази, символи або під слова, залежно від конкретного варіанту використання [31]. Ці одиниці називаються "токенами". Токенізація є фундаментальним кроком в обробці природної мови (NLP) і служить для попередньої обробки неструктурованих текстових даних. Процес токенізації залежить від конкретної мови та конкретних вимог завдання. У своїй найбільш поширеній формі токенізація передбачає поділ тексту на слова за рахунок знаходження пробілів (пустих символів) та розділових знаків, які слугують як межі між словами. Однак більш складніші варіації токенізації можуть також обробляти складові слова, вирази з кількох слів, або навіть під-слів, більш відомі як морфеми або символи.

Токенізація служить початковим кроком у будь-якому робочому процесі обробки природної мови (NLP) і істотно впливає на всі наступні операції у конвеєрі. Токенізатор розбиває неструктурований текстовий вміст та вирази на окремі елементи, кожен з яких представляє окрему частину інформації. Отримані токени використовують в якості векторів, що представляють текстовий документ. Таке перетворення тексту створює структурований формат даних, необхідний для машинного навчання.

Основний підхід токенізації передбачає сегментування тексту окремі слова. в багатомовних та складних лінгвістичних контекстах, токенізація опирається на словники або лексикони для забезпечення значущої та точної сегментації тексту. Ну сучасний стан НЛП вирішило цю проблему за рахунок сегментації на словники, зберігаючи лінгвістичні значення слів. Отже, навіть коли модель НЛП зустрічає незнайоме слово, вона може визначити конкретне слово, за рахунок морфем. Метод, який ілюструє такий підхід токенізації з урахуванням вище сказаного зветься WordPiece [32] зображений на рисунку 2.3.

Рисунок 2.3 – Сегментація речення методом WordPiece

У методі токенізації WordPiece текст спочатку попередньо розбивається на слова, зазвичай шляхом визначення розділових знаків і пробілів, як символи, по яким розбивається текст. Після цього кожне слово розбивається на більш дрібні одиниці за рахунок стратегії найбільшого збігу для більш конкретної токенізації попередньо розбитих на кусочки слів. Він ітеративно вибирає найдовший відповідний префікс із тексту, що залишився, відповідно до словника моделі. Цей підхід зветься MaxMatch і працює як квадратичну складність щодо довжини вхідного слова [32].

Іншим альтернативним варіантом від компанії Google є алгоритм LinMaxMatch, який працює із чітко лінійно-залежним часом токенізації, що робить його ефективним для сегментації тексту на елементарні одиниці, такі як слова та під-слова. [33]. Основна ідея LinMaxMatch полягає в структуруванні словнику токенів у вигляді дерева. Кожне ребро в дереві є символом, а шлях від кореня дерева до певного вузла в дереві відповідає префіксу токена в словнику. Ця деревоподібна структура дозволяє ефективно ідентифікувати токени у вхідному тексті шляхом обходу дерева символом за символом. У цьому алгоритмі токени словника спочатку структуруються як дерева. Кожний край дерева є символом, а шлях від кореня до певного вузла в дереві є префіксом токена в словнику. На рисунку 2.4 показано край дерева та вузли між ними. Цей деревоподібний підхід дозволяє ефективно

ідентифікувати словниковий токен, який відповідає вхідному тексту, рухаючись символ за символом по черзі в циклі.

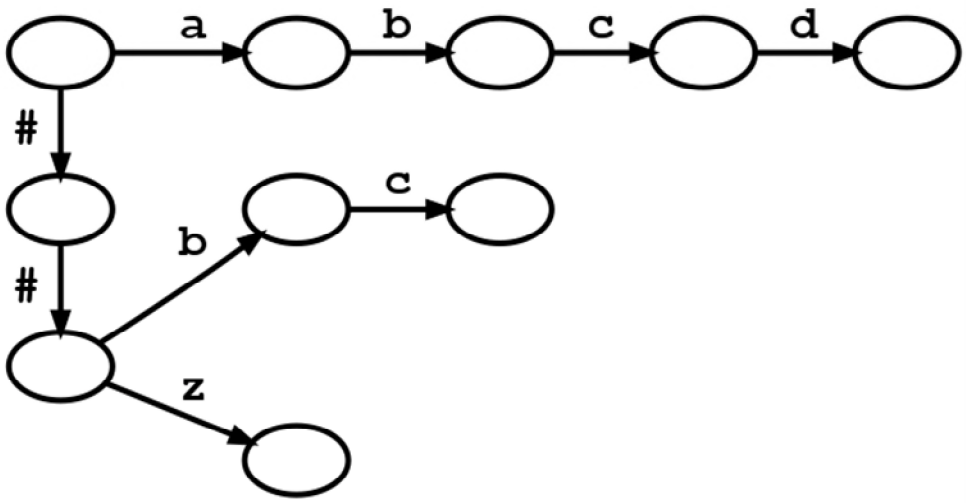


Рисунок 2.4 – Деревоподібний підхід токенізації слів

Процедура виконання алгоритму WordPiece працює, як описано нижче:

1. Необхідно взяти існуючі, або створити власноруч готовий великий набір даних, іншими назвами датасет або корпус.
2. Встановити необхідну розмірність словника, в якому будуть задані слова.
3. Розділити вхідні слова на окремі символи.
4. Отримані символи використати для ініціалізації словника.
5. Створити лексичну модель за допомогою створеного словника ініціалізованого на попередньому кроці.
6. Циклічно повторювати 5-й крок, доки словник під-слів не досягне визначеного розміру, вказаному на кроці 2.

Ще одним підходом токенізації є алгоритм створений розробниками бібліотеки Spacy [34]. Цей алгоритм використовує методи, що ґрунтуються на правилах, та комбінацію евристики словників запасів, специфічних для конкретної мови, що слугує основним поділом тексту на токени. Токенізація залежить від мови з урахуванням нюансів, специфічних для цієї мови. На рисунку 2.5 зображено

приклад алгоритму розподілу речення на окремі токени, на ньому детально демонструється кожен етап, через який проходить вхідне речення для детальної попередньої обробки.

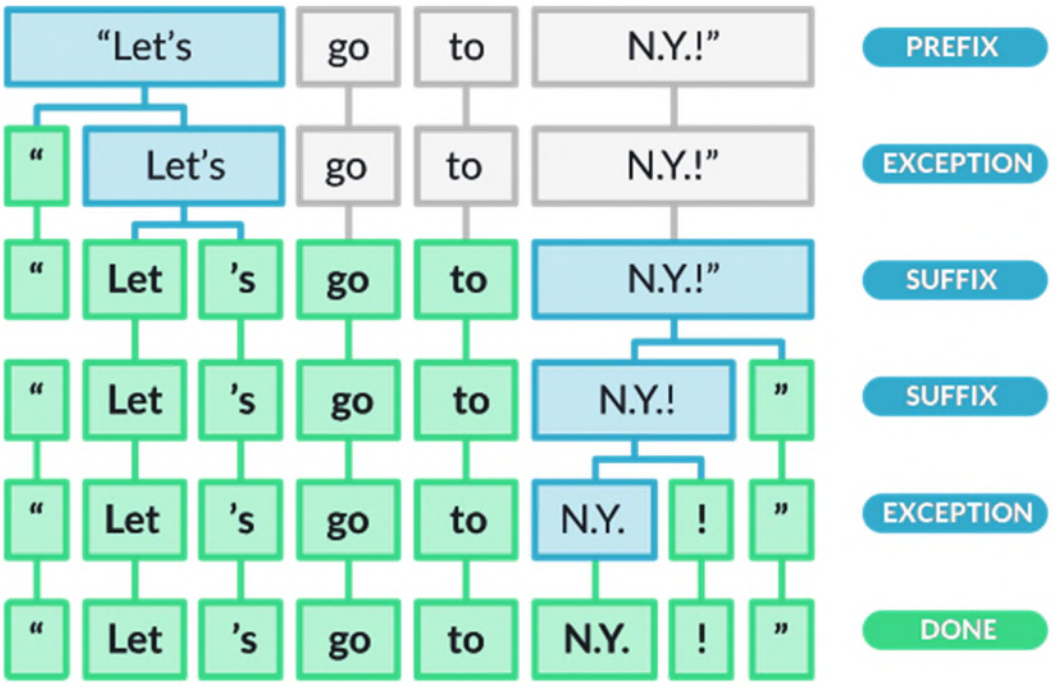


Рисунок 2.5 – Підхід токенізації слів від бібліотеки Spacy

На початку речення розподіляється на слова простим знаходженням пробілів між словами, потім токенізатор обробляє текст зліва направо. Для кожного слова він виконує дві перевірки, чи слово співпадає з винятком в правилах і чи можна в слові розділити префікс, суфікс та інфікс. У такий спосіб, SpaCy може розділяти складні вкладені лексеми, такі як комбінації скорочень із кількома розділовими знаками. Хоча правила пунктуації зазвичай є загальними, наперед задані винятки токенізатора сильно залежать від особливостей конкретної мови, через це кожна мова має свій власний підклас, із сильно запрограмованими правилами.

2.2.2 Усунення шуму в текстових даних

Видалення текстового шуму – це ще один важливий етап у попередній обробці даних в алгоритмі обробки природної мови, метою якої є очищення

текстової інформації шляхом усунення непотрібних або небажаних слів, які не потрібні для машинного навчання та які лише погіршуватимуть результат [35]. Взагалі слово «Шум» у текстових даних означає стоп-слова (такі як «до», «в», «на»), розділові знаки, спеціальні символи та інші елементи не потрібних в аналізі тексту.

У завданні з розпізнавання сутностей по оголошеннях з нерухомості ці слова по суті не мають семантичного значення, що робить їх менш актуальними для аналізу. Тому потрібно відфільтрувати інформацію низького рівня, дозволяючи більш чітко зосередитися на суттєвому змісті. Видалення стоп-слів не надає шкідливого впливу на модель.

Процес усунення стоп-слів із поданої текстової інформації можна описати наступним алгоритмом:

1. Ідентифікація шуму та створення списку топ слів, зазвичай до них входять розділові знаки, спеціальні символи та інші не потрібні слова.
2. Токенізація цільового документа, за рахунок розподілу окремих слів із речення.
3. Зчитування стоп-слова із створеного списку стоп-слів.
4. Порівняння стоп-слова з цільовим нетокенізованим текстом за рахунок ітеративного циклу.
5. Якщо збіг знайдено, слово видаляється з масиву токенізованих слів, порівняння триває до завершення довжини масиву.

2.2.3 Лематизація та стемінг в текстових даних

Лематизація – це процес, перетворення слів у їх базову або кореневу форму, гарантуючи, що слова з однаковим значенням групуються разом [36]. Цей процес зазвичай включає звернення до словника, проведення морфологічного аналізу, видалення закінчень і відновлення слів у формі словникової леми. Лематизація включає групування слів із загальним коренем, але з різними суфіксами та префіксами, щоб ці слова можна було проаналізувати як єдине ціле. Основна мета лематизації полягає в отриманні базової форми слова.

Варто зазначити, що якихось важливих алгоритмів в лематизації не існує і зазвичай використовують простий підхід базового пошук у словнику. Такий підхід ефективний для обробки простих відмінюваних форм слів, хоча в більш складних випадках, наприклад, у мовах з розширеними та складними словами, може знадобитися використання системи, що базується на правилах. Ці правила можуть бути створені вручну або отримані за допомогою автоматичного навчання з корпусу, анотованого для цієї мети.

Стемінг – це метод нормалізації тексту, що використовується в обробці природної мови (НЛП), метою якого є зведення слів до їхньої кореневої або базової форми [37]. Цей процес включає видалення афіксів, які являють собою префікси, або суфікси, щоб отримати основну, значущу частину слова. Мета стемінгу полягає в тому, щоб згрупувати різні форми або варіанти слова разом, розглядаючи їх як єдине ціле, що має загальне семантичне значення. Для досягнення цієї мети використовуються різні алгоритми стемінгу, кожен з яких має власний набір правил і методів [37]:

- Портер Стеммер: Алгоритм використовує ряд правил для зведення слів до їхньої кореневої форми. Правила розбиті на різні етапи, кожен із яких виконується послідовно до слова.
- Ловінс Стеммер: Цей алгоритм призначений лише для англійської мови. Він застосовує набір правил і використовує великий словниковий запас слів для ухвалення обґрунтованих рішень.
- Хаск Стеммер: Цей алгоритм використовує комбінацію правил та пошуку за словником. Він відомий своєю високою точністю, але може бути не таким швидким, як деякі інші алгоритми.

Як бачимо суть алгоритмів зводиться в будь якому випадку до використання словника і на даний момент є безліч готових бібліотек та засобів, які використовують це для обробки тексту.

Лематизація та стемінг дає нормалізацію сегментів текстових даних. На рисунку 2.6, показано відмінність між 2-ма підходами за допомогою англійських слів, та наведено порівняння цих двох підходів на прикладі англійського слова

«Improve». Лематизація враховує слова, враховує граматичні закінчення та афікси, і намагається знайти слова в словнику, у той час, як стемінг просто відкидає афікси, не завжди враховуючи граматичні правила.

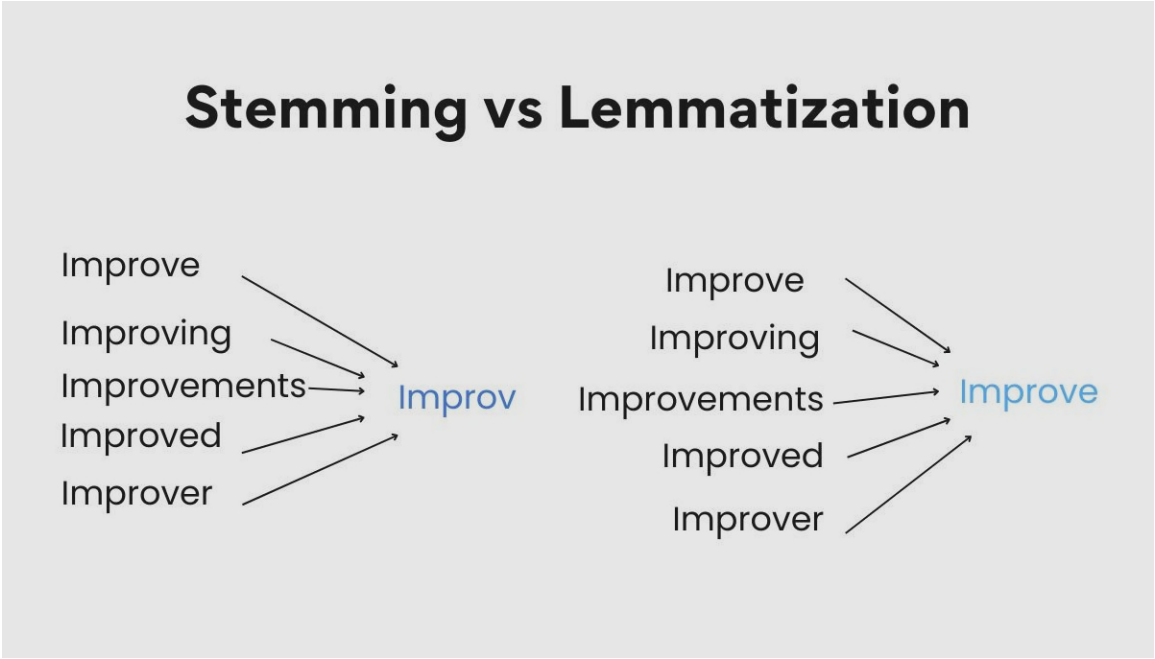


Рисунок 2.6 – Приклад лематизації та стемінгу на слові «improve»

2.2.4 Порівняння підходів для попередньої обробки текстових даних

WordPiece – це алгоритм токенізації тексту для мовних моделей на основі нейронних мереж. Однак WordPiece має деякі обмеження, через які він не так добре справляється з завданнями як токенізатор SpaCy через наступні причини:

- не завжди правильно обробляє розділові знаки;
- часто створює дуже детальні токенізації слів, що може ускладнити інтерпретацію результатів завдань НЛП з розпізнавання сутностей, оскільки результат розпізнавання може видати не правильні межі слів;
- не сильно підходить для мов з багатою морфологією, та специфічних слів притаманних конкретній галузі;

З іншого боку, алгоритм для токенізації слів від SpaCy є токенізатором який вже зразу є налаштованим та заточеним під задачі пов’язанні з галуззю обробки природної мови, через те що сама бібліотека SpaCy налаштована під NLP.

LinMaxMatch – це алгоритм сегментації слів, що ґрунтується на ідеї максимізації контексту слів у реченні. Незважаючи на широке використання в НЛП, цей алгоритм досі вимагає великих обчислень [38], а також має наступні недоліки:

- не враховує регістр, що може утруднити обробку власних назв, таких як назва вулиці, міста розташування нерухомостей та інших подібних випадків;
- не справляється зі скороченнями, такими як «ні» та «я»;
- не обробляє розділові знаки, такі як коми і крапки;

Згідно цих обмежень, токенізатор SpaCy – це потужний і універсальний інструмент, який можна використовувати для вирішення різних завдань в сфері NLP. Він точніший і надійніший, ніж WordPiece, LinMaxMatch або й word2vec, і має ширший набір функцій. Також алгоритм токенізації SpaCy заснований на правилах і дозволяє налаштовувати себе під різні параметри з можливість конфігурації під різні області, що дозволяє адаптувати його до різних мов та галузей. В результаті це найкращий вибір для більшості додатків заснованих на завданнях пов’язаних з NLP.

2.3 Ідентифікація та класифікація сутностей з тексту оголошень

Як вже було не раз наголошено розпізнавання сутностей є процесом вилучення інформації з речень, абзаців, текстових даних або інших форм неструктурованої інформації. Після попередньої обробки цих даних, відбувається механізм розпізнавання, тобто ідентифікації та класифікації сутностей (таких, як ціна, кількість кімнат, стан маєтку і т.д.), які подаються як частина набору навчальних даних. Для того, щоб розпізнавати таку інформацію з неструктурованого тексту використовують інструменти вилучення об'єктів за допомогою глибокого навчання, тим самим створюючи модель розпізнавання сутностей, яку потім можна застосувати до нових наперед невідомих текстових даних.

На рисунку 2.7 показано спрощено алгоритм. що робить натренована модель, коли на вхід подається інформація з оголошення про продаж нерухомості. Маючи речення з оголошення, модель розпізнавання використовує покрокові внутрішні механізми та етапи для ідентифікації сутностей, цим самим на виході дає правильно розпізнані результати у вигляді міток та слів, які ці мітки позначають.

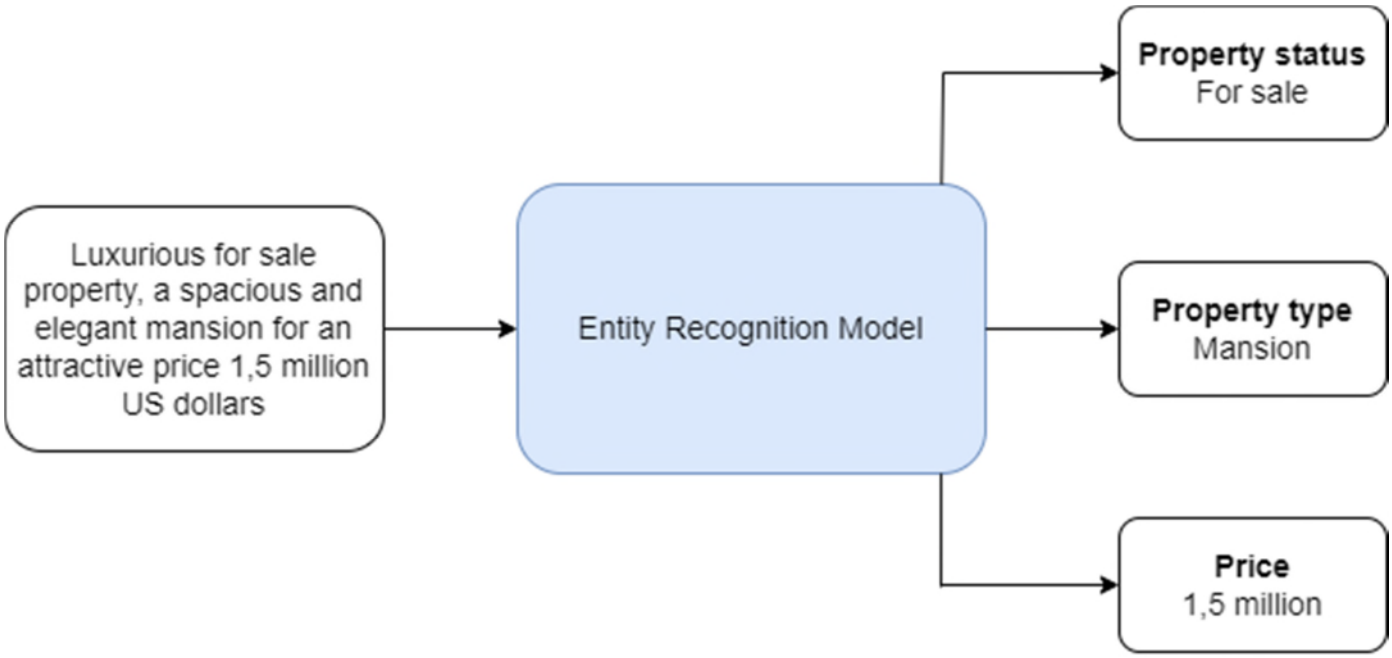


Рисунок 2.7 – Приклад розпізнавання сутностей через навчену модель

Сам алгоритм ідентифікації та класифікації сутностей в задачі розпізнавання іменованих сутностей має назву «Embed, Encode, Attend, Predict» і стосується концептуальної основи Метью Хоннібала з глибокого навчання по обробці природної мови [39]. Алгоритм складається із 4-ох фаз:

1. Embed – це процес перетворення тексту в числові вектори слів. Таке перетворення збирає семантичну інформацію із вхідного тексту, і з ним набагато простіше працювати моделі. Ці вектори генеруються таким чином, що слова зі схожим значенням або контекстом стають близько один до одного в цьому числовому просторі. Таким чином, цей крок допомагає комп'ютеру детальніше вловити нюанси слів, їх значення та відносини між ними. Така карта векторів допомагає комп'ютеру глибше зрозуміти значення та контекст слів та відношення між ними. Ці вектори слів часто генеруються за допомогою таких методів, як

Word2Vec, GloVe, або навіть більш складних моделей, таких Word Embeddings та BERT. Процес перетворення робить текстові дані більш керованими та допомагає представити семантику слів у математичній формі. Даний етап зображений на рисунку 2.8

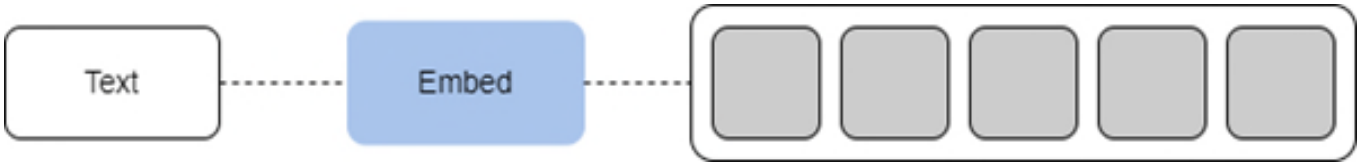


Рисунок 2.8 – Перетворення тексту в числові вектори

2. Encode – Це процес кодування послідовності числових векторів слів у матрицю з використанням таких методів як рекурентні нейронні мережі (RNN) або мереж з довгою короткочасною пам’яттю (LSTM). Більш прозоре розуміння цього, що відбувається зображено на рисунку 2.9. На цьому етапі алгоритм прагне створити розширене уявлення вхідних текстових даних, яке може послідовно фіксувати відносини між словами та токенами.

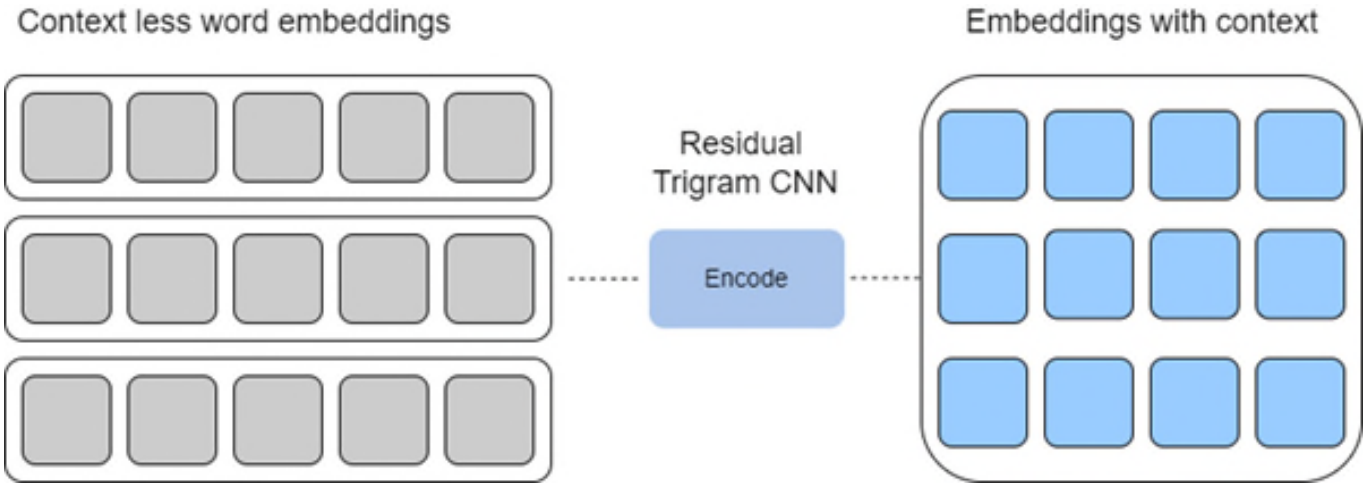


Рисунок 2.9 – Кодування векторів в матрицю векторів

3. Attend – вихідні дані матриці на попередньому кроці зводяться до одного єдиного вектора, який буде далі переданий у нейронну мережа прямого поширення для здійснення прогнозування. Цей крок має той самий ефект, як і механізм уваги.

Механізми уваги використовуються для зважування важливості різних частин вхідного тексту. Це допомагає моделі зосередитися на найбільш релевантній інформації під час прогнозування. На рисунку 2.10 зображено процес перетворення.

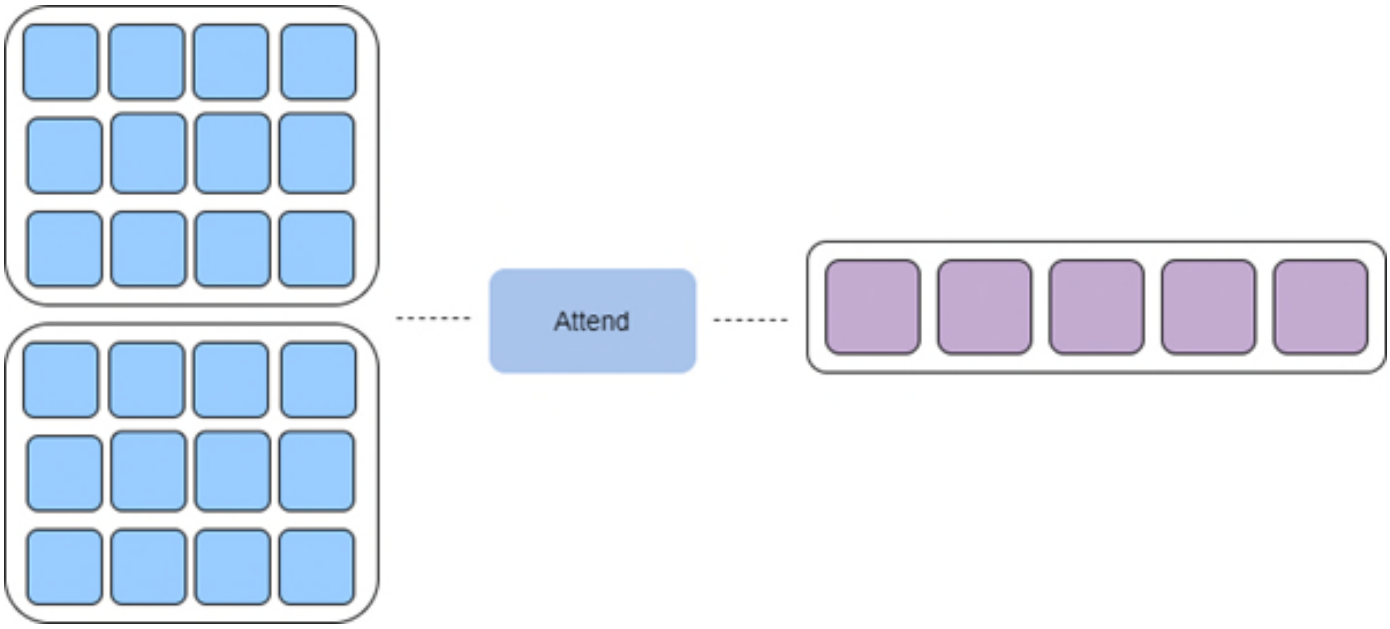


Рисунок 2.10 – Механізм роботи етапу Attend

4. Predict – Останнім кроком моделі є прогнозування на основі вхідного вектору з попереднього кроку. На цьому етапі цей вектор передається багатошаровому перцептрону для виведення ідентифікатора мітки об'єкта. Саме тут відбувається генерація бажаного результату у вигляді мітки. Характер прогнозу варіюється залежно від проблеми та задачі, яка використовується для розпізнавання сутностей. На рисунку 2.11 зображено останній результат передбачення мітки «ID».

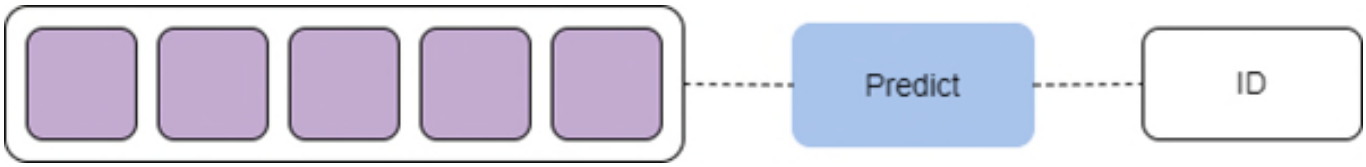


Рисунок 2.11 – Механізм роботи етапу Predict

Як можна побачити даний алгоритм розпізнавання іменованих сутностей є досить складним, оскільки використовує декілька нейронних мереж для вирішення

завдання. Цей алгоритм якраз і пропонує структурований підхід до розуміння того, як моделі глибокого навчання опрацьовують та інтерпретують дані природної мови. Використання такого алгоритму допоможе вирішити завдання по структуризації тексту оголошень

2.4 Оцінки ефективності та якості розпізнавання

Оцінка ефективності методу машинного навчання є фундаментальним етапом у його розробці. Навіть, якщо модель вдалося натренувати – вона все рівно може показати як правильні результати, при її оцінці на основі точності, так і неправильні результати. У багатьох випадках точність служить основним показником для оцінки продуктивності моделі, проте бувають ситуації, коли вона не відповідає дійсності. Тому існують наступні показники, за якими порівнюють ефективність роботи натренованих нейронних моделей:

Оцінка F1: цей показник являє собою середнє гармонійне значення між точністю та повнотою вибірок. Оцінка F1 вимірює як точність класифікатора (правильні класифікації), так і його здатність уникати пропуск важливих екземплярів. Оцінка F1 потрапляє в діапазон [0, 1], причому вищі значення означають кращу продуктивність моделі.

Площа під кривою (AUC): найпоширеніший показник оцінки, що використовується в основному в задачах двійкової класифікації. Цей показник відповідає ймовірності того, що класифікатор поставить випадково обране позитивне значення набагато вище, ніж випадково обране негативне значення.

Логарифмічна втрата: цей показник працює таким чином, де намагається «покарати» модель за неправильну класифікацію значень, та ефективно підходить для сценаріїв класифікації з кількома класами. При використанні цієї метрики, класифікатор обов’язково повинен присвоїти кожному класу певну ймовірність класифікації.

Точність класифікації: ця метрика розраховує відношення правильно прогнозованих результатів із всього ряду вхідних прогнозованих значень (як правильних так і не правильних результатів).

Повнота класифікації: цей показник вираховує відношення правильно прогнозованих результатів згідно всієї кількості зразків, які повинні були б бути правильно розпізнанні в самому ідеальному варіанті.

Середня абсолютна помилка (MSE): це важливий показник, який обчислює середнє значення між прогнозованими та початковими значеннями, показуючи при цьому ступінь відхилення прогнозованих значень від правильних значень. Однак дана метрика не вказує напрямок та природу помилки.

Середньоквадратична помилка (MSE): цей показник зводить у квадрат різницю між прогнозованими та фактичними значеннями. MSE має перевагу в обчисленні розрахунку градієнта, завдяки чому спрощується оптимізація моделі. Беручи різницю помилки в квадрат, MSE вдало цим підкреслює велику розбіжність в прогнозованих значеннях, дозволяючи моделі приділити першочергову увагу виправленню істотних помилок.

Як видно існує багато метрик для оцінювання методу розпізнавання сутностей, оскільки оцінка є необхідною складовою для його правильної роботи і ретельного прогресу. Серед великої кількості метрик для оцінювання, було обрано ті, які найчастіше використовувалися на конференціях MUC та CONLL по розпізнаванню сутностей [40]. До таких оцінок відноситься точність розпізнавання, повнота і середнє гармонійне значення між ними (f-міра). Обидві конференції використовують ці метрики для оцінювання роботи моделі, проте мають деякі відмінності щодо їх ефективності:

1. Оцінювання часткової відповідності згідно конференції MUC.

Згідно конференції MUC, ефективність результату розпізнавання сутностей базується на двох критеріях:

- здатність розпізнати правильний тип сутності(Entity Type).
- здатність знайти правильні границі сутності (Entity Text).

Entity Type вважається розпізнаним, якщо послідовності текстовим даним присвоєно правильний тип, незалежно від границь. Entity Text вважається розпізнаним, якщо границі текстових даних визначено правильно, незалежно від того чи тип на цю сутність розпізнано правильно.

Для обох критеріїв (Entity Type та Entity Text) присвоюється три показники:

- correct answers – кількість правильно розпізнаних сутностей.
- actual answers – кількість правильних (correct answers) і неправильних результатів розпізнавання (incorrect answers).
- possible answers – загальна кількість усього можливих розпізнаних сутностей

Згідно цих показників, підсумковий результат – це f-оцінка, тобто середнє гармонійне значення точності й повноти розпізнавання, що розраховується з урахуванням обох критеріїв (Entity Type, Entity Text). В даному випадку, точність – частка правильно класифікованих сутностей (correct answers) серед усіх розпізнаних сутностей (actual answers), ідентифікованих класифікатором. Обчислюється за формулою 1.1.

$$precision = \frac{\text{correct answers}}{\text{actual answers}} \tag{1.1}$$

Повнота – частка загального числа правильно класифікованих сутностей (correct answers), які було знайдено серед усіх можливих сутностей, які мали б бути правильними (possible answers). Обчислюється за формулою 1.2.

$$recall = \frac{\text{correct answers}}{\text{possible answers}} \tag{1.2}$$

Маючи результати precision і recall, гармонійне середнє буде розраховуватися за наступною формулою:

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (1.3)$$

2. Оцінювання точної відповідності, згідно конференції CONLL.

Згідно конференцій CONLL, для визначення правильності роботи з розпізнавання сутностей, використовують точний протокол оцінювання. Системи розпізнавання сутностей також порівнюються на основі f-оцінки, водночас, окрема іменована сутність буде вважатися правильною лише у тому випадку, якщо одночасно її Entity Type розпізнаний правильно і якщо границі текстових даних також визначено правильно.

Отже, за допомогою таки оцінок можна буде правильно оцінити роботу натренованої моделі.

Висновки до розділу 2

У цьому розділі проведено детальний аналіз механізму та алгоритму роботи по розпізнаванню іменованих сутностей де було розглянуто основні принципи керування та обробки текстової інформації необхідних при категоризації та ідентифікації даних. Детально було виділено увагу попередній обробці даних, що є необхідним кроком перед початком тренування моделі. Серед великого вибору алгоритмів по попередній обробці даних було вирішено використовувати алгоритм токенізації від SpaCy, оскільки він вже зразу йде налаштованим під обробку природної мови, легко конфігурується, та може налаштовуватися під специфічні предметні області, що ідеально підходить для оголошень з продажу нерухомостей.

Крім того було вирішено розробити метод розпізнавання іменованих сутностей (NER) для задачі категоризації об'єктів у текстах оголошень за рахунок використання архітектурного підходу EEAP для ідентифікація та класифікації сутностей, оскільки такий підхід використовує в собі поєднання декількох нейронних мереж, які краще дозволяють вловити зв’язки та суть між словами, що в свою чергу повинно дати більш точніші результати розпізнавання слів.

Насамкінець було розглянуто метрики, та оцінки ефективності які використовуються для оцінювання якості результатів роботи моделі з розпізнавання сутностей. Згідно конференції MUC та CONLL, для порівняння прогнозів розпізнавання, було обрано використовувати такі показники ефективності роботи моделі як точність та повнота розпізнавання сутностей, з яких легко обчислити середнє гармонійне значення (f-міру).

РОЗДІЛ 3 Проектування інформаційної системи структурування текстів оголошень для об’єктів нерухомості засобами NLP

3.1 Вимоги до розроблених програмних засобів

При проектуванні та оцінці методу структурування текстів оголошень для об’єктів нерухомості з використанням засобів НЛП, необхідно правильно розробити спеціалізоване програмне забезпечення для оптимізації процесів роботи із системою. Це в свою чергу повинно дати можливість користувачам вводити бажане оголошення в якому описуватиметься інформація про нерухомість, обробляти введені дані та показувати результат розпізнавання сутностей. Для досягнення цієї мети програмне забезпечення має давати змогу виконувати наступні функції:

1. *Побудова моделі структурування тексту:*

Розробити алгоритми структурування текстових даних, пов'язаних із оголошеннями про нерухомість.

Використати можливості NER із обробки природної мови для ідентифікації ти категоризації відповідної інформації відомостей про об'єкт нерухомості.

2. *Інтерфейс для взаємодії:*

Використати компоненти графічного інтерфейсу для зручності користування інтерфейсом.

Забезпечити візуально інтуїтивний вигляд, що дозволить користувачам безперешкодно взаємодіяти із системою.

3. *Можливість введення тексту:*

Надати механізм введення неструктурованих текстових даних, таких як оголошення про нерухомість.

Реалізувати функції, які дозволяють легко редагувати, оновлювати чи вводити нову текстову інформацію.

4. *Представлення розпізнаних сутностей:*

Розробити функціонал для відображення структурованої інформації витягнутої з тексту.

Відобразити розпізнану інформацію яку легко можна побачити у чіткій та зручній формі.

За допомогою таких функціональних вимог, створене програмне забезпечення дозволить користувачам ефективно розпізнавати текстові дані про нерухомість, що сприятиме результативності та практичності розробленого методу.

3.2 Обґрунтування вибору засобів розробки програмного застосунку з розпізнавання сутностей

Перед розробкою програмного застосунку, було вирішено використовувати мову програмування Python, через його унікальні переваги і можливості для структурування тексту за допомогою методів NLP. Надійна екосистема Python для обробки природної мови, а також широкі можливості використання машинного навчання та штучного інтелекту [41] у поєднанні з великим вибором бібліотек і застосуванням в сфері обробки природної мови роблять Python найкращим вибором для запропонованого методу структурування тексту оголошень з продажу нерухомості. Крім того, із наявних бібліотек доступних на мові Python найдоцільніше використовувати бібліотеку SpaCy через її ідеальний баланс між продуктивністю, широким функціоналом, простотою в розумінні і широкою інфраструктурою для завдань по розпізнаванню іменованих сутностей (NER). Можливість SpaCy використовувати попередньо мінімально навчені моделі розпізнавання іменованих об'єктів з великим функціоналом по NLP роблять мову Python ідеальним вибором для розробки точного та надійного методу з розпізнавання іменованих сутностей.

3.3 Архітектура та процеси бібліотеки SpaCy

Архітектура бібліотеки SpaCy передбачає використання різноманітних структур даних для ефективного управління та обробки лінгвістичних анотацій. Центральними структурами даних у SpaCy є компоненти «*Language*», «*Vocab*»,

«*Doc*», «*Token*» і «*Span*», вони зображені на рисунку 3.1 і являються важливими елементами архітектури, які надають механізми для керування словниковим запасом та налаштуваннями, специфічними для мови на якій проводиться розпізнавання.

Рисунок 3.1 – Схема Архітектури бібліотеки Spacy

1. *Клас Language (Мова)*

- Language – це основний компонент, який керує всім конвеєром обробки текстових даних. Він поєднує різні компоненти, такі як об’єкт «Vocab», токенізацію,

тегування частин мови, синтаксичний аналіз та багато іншого, а також відповідає за ініціалізацію та координацію конвеєра обробки. Цей клас надає користувачам зручний інтерфейс для доступу до лінгвістичних анотацій та налаштування мовних параметрів.

- Під час ініціалізації конвеєра SpaCy компонент Language створює об'єкт Vocab, забезпечуючи доступність словника для подальшої обробки тексту, крім того користувачі можуть створювати екземпляри цього класу з різними конфігураціями для вирішення більш конкретних та специфічних мовних завдань. Наприклад, SpaCy надає попередньо навчені моделі для різних мов, кожна з яких має власний об'єкт Language, адаптований конкретно під цю мову.

2. *Об'єкт Vocab (Словник)*

- Об’єкт Vocab представляє собою словник, який містить набір унікальних слів, з якими система зіткнулася при обробці текстових документів, або про які їй відомо. Цей об’єкт відповідає за керування словником конвеєра SpaCy та включає в себе як токени слів (конкретні екземпляри слів у вхідному реченні), так і лексеми (абстрактні уявлення слів без контексту).

- Словник містить в собі детальний опис слів, таку як числові вектори слова, частоту появи слів та іншу споріднену інформацію. Це допомагає ефективно працювати із окремими словами цілого вхідного речення, а це сприяє вирішенню багатьох різних завдань НЛП.

- Об'єкт Vocab є спільним для всіх екземплярів конвеєра SpaCy, тому кілька документів будуть завжди використовувати один і той же словниковий запас, підвищуючи цим ефективність використання пам'яті та сталу послідовність подання слів.

- Словник створюється та оновлюється динамічно у міру появи нових слів під час обробки документів. Це особливо корисно для обробки слів, які виходять за межі словникового запасу, що трапляються в реальних текстах.

3. *Об'єкт Doc (Документ):*

- Об'єкт Doc – це головний компонент, що представляє оброблений документ. Він інкапсулює інформацію про весь наданий текст, включаючи його окремі токени, анотації, залежності, іменовані об'єкти та багато іншого.
- Документ створюється шляхом обробки неструктурованого тексту через конвеєр обробки SpaCy, який включає токенізацію, тегування, аналіз залежностей та інші завдання з попередньої обробки, які можна додати вручну.
- Даний об’єкт основною служить точкою входу що дає доступ до різних лінгвістичних анотацій та полегшує перегляд і маніпулювання лінгвістичними елементами документа.

4. *Об'єкт Token:*

- Токен собою представляє окреме слово у документі. Кожен токен в об'єкті Doc є екземпляром класу Token.
- Об’єкти токени зберігають інформацію пов’язану із словом, його початкову форму, та іншу споріднену інформацію. Такі об'єкти розташовані в документі у строгій послідовності, що дозволяє користувачам отримувати доступ до них та аналізувати.

5. *Об'єкт Span:*

- Span представляє діапазон токенів у документі. Якщо Token відповідає окремому слову із речення, то Span це декілька токенів (діапазон слів із речення)
- Користувачі створюють об'єкти Span для посилання на певні частини документа на основі індексів токенів. Це корисно для роботи з фразами, пропозиціями чи будь-якою іншою задачею в якій необхідно працювати з певною послідовністю токенів.
- Як і об'єкт Doc, об'єкти Span підтримують лінгвістичні анотації, дозволяючи користувачам отримувати та аналізувати інформацію на рівні екземпляра об’єкта Span.

Важливим нюансом є те що об'єкти Vocab і Language використовують перехресні посилання. Це означає, що кожен об'єкт Token у екземплярі Doc міститиме посилання на об'єкт Vocab, що дозволяє отримати доступ до інформації зі словника. З рисунку 2.12 можна виділити окремо процес який відповідає за обробку

неструктурованого тексту в структурований. Дана діаграма зображена на рисунку 3.2.

Рисунок 3.2 – Діаграма процесу обробки тексту

З даної діаграми чітко видно що відбувається з даними від початку до кінця:

1. *Вхідний текст*: процес починається із зчитування необробленого тексту. Це може бути оголошення, реклама, або документ, який далі передається в мовну модель.

2. *Мовна модель*: необроблений текст обробляється мовною моделлю, яка містить усі компоненти обробки, та інформацію специфічну для мови в якій подано текст. Після цього мовна модель запускає конвеєр обробки.

3. *Конвеєр обробки*. Конвеєр обробки є серією компонентів, які виконують різні завдання над необробленим текстом. Сюди входить токенізація, лематизація, стемінг, розпізнавання іменованих сутностей, тегування, та багато іншого. Кожен компонент конвеєрі обробляє текст послідовно, і вихідні дані одного компонента стають вхідними даними для наступного.

4. *Документ.* Вихідними даними конвеєра обробки є документ. Цей об'єкт містить інформацію про слова з речення, їх лінгвістичні особливості та зв'язки між ними і являється контейнером для обробленого тексту.

5. *Вихідні дані:* Останній крок – отримання вихідних даних. Сюди входить інформація про токени слів, розпізнані іменовані сутності та інші лінгвістичні особливості.

Окремої уваги заслуговує конвеєр обробки, точніше ті внутрішні процеси які він використовує. Дана схема обробки зображена на рисунку 3.3 нижче:

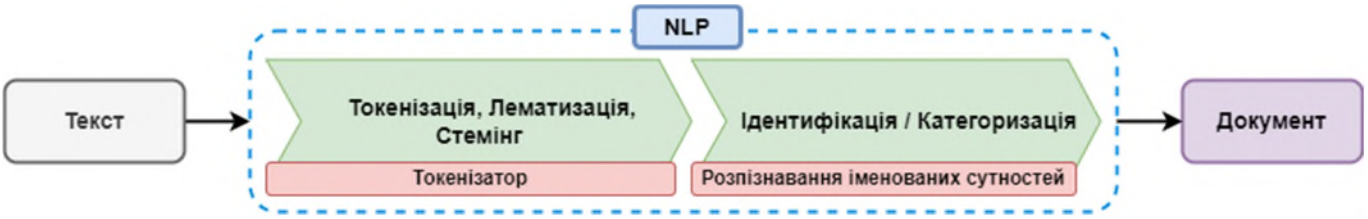


Рисунок 3.3 – Схема процесу конвеєра обробки тексту

Як видно з рисунку. токенізатор запускається перед основним модулем розпізнавання. Крім цього до цих компонентів можна додавати готові інші або свої власні компоненти. Вони можуть містити статистичну модель, навчену модель, підправлені ваги, та інші елементи. На початку текст передається в конвеєр, над яким відбувається маніпуляції по попередній обробці, після цього запускається інший компонент конвеєра з розпізнавання сутностей в обробленому тексті.

3.4 Компоненти модулів програмного застосунку розпізнавання сутностей засобами NLP

Згідно методу розпізнавання іменованих сутностей в структуризації текстів оголошень об’єктів нерухомості засобами обробки природної мови, з використанням можливостей бібліотеки SpaCy, необхідно запровадити такі елементи модулів до інформаційної системи розпізнавання об’єктів:

- модуль попередньої обробки текстових даних для очищення, покращення необроблених текстових даних, гарантуючи, ефективність отримання інформації;
- модуль що відповідає за ідентифікацію та категоризацію необхідних даних, тобто розпізнавання сутностей;
- модуль тренування моделі розпізнавання;
- модуль кінцевого представлення даних розпізнаних об’єктів;
- модуль який відповідає за введення текстових даних;

Таким чином до реалізованих модулів побудовано діаграму їх взаємодії (рисунок 3.4):

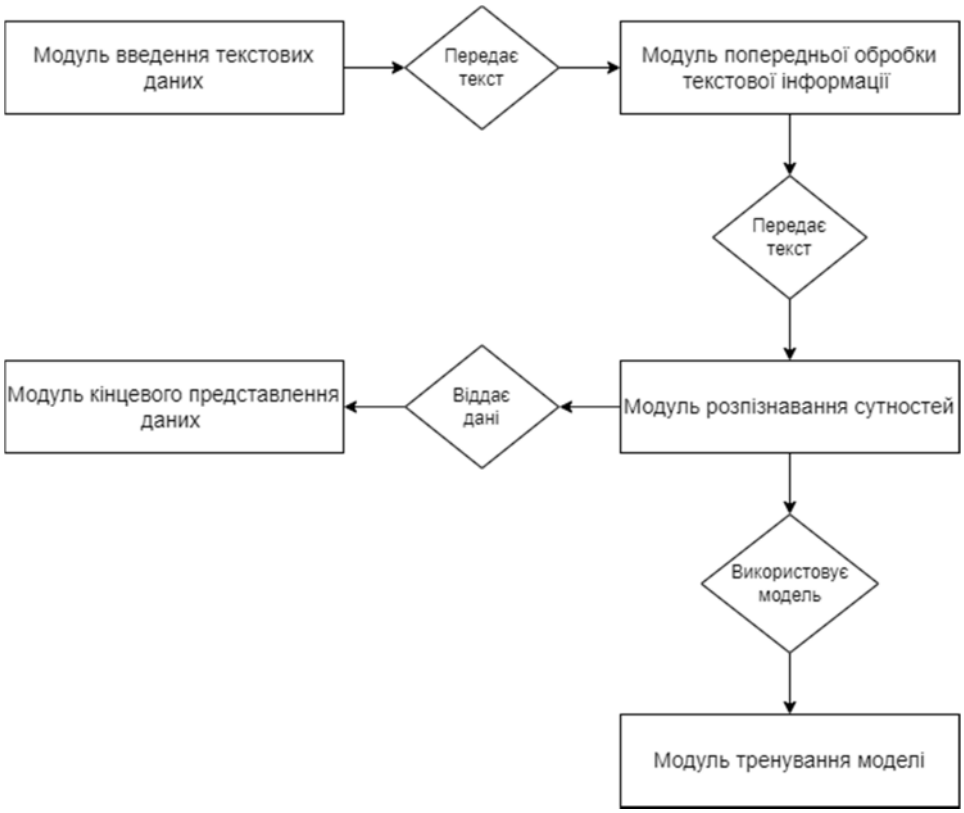


Рисунок 3.4 – ER-діаграма взаємодії модулів програми

Процес роботи програмного застосунку буде відбуватися так як зображено на діаграмі нижче:

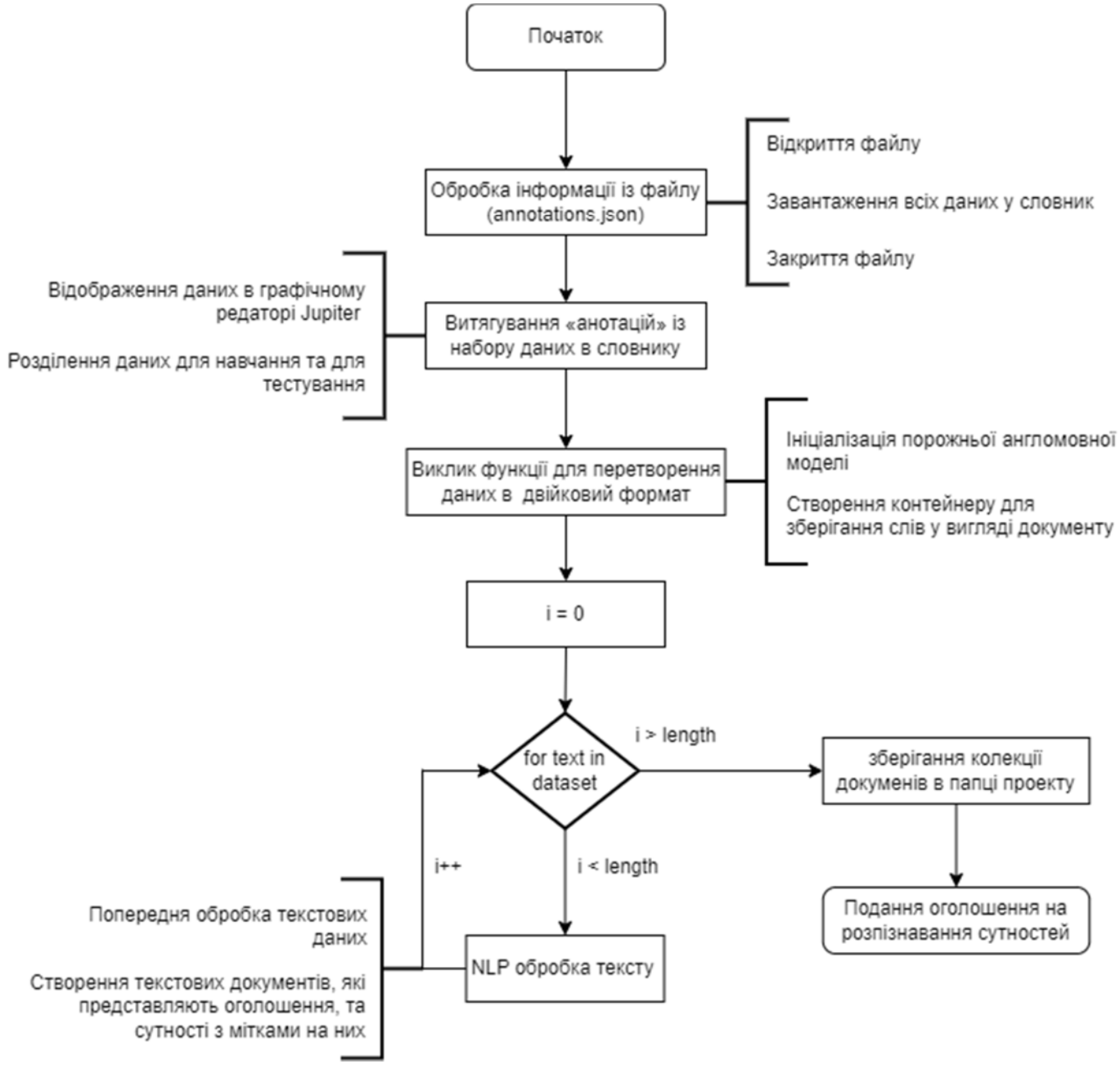


Рисунок 3.5 – Діаграма роботи системи

З даної діаграми видно, що для роботи програмного застосунку необхідно завантажити файл з вхідними оголошеннями, та готовими на них анотаціями, для того щоб натренувати модель розпізнавання сутностей. Файл має розширення «*.json*», та виглядає так, як зображено на рисунку 3.6 нижче:

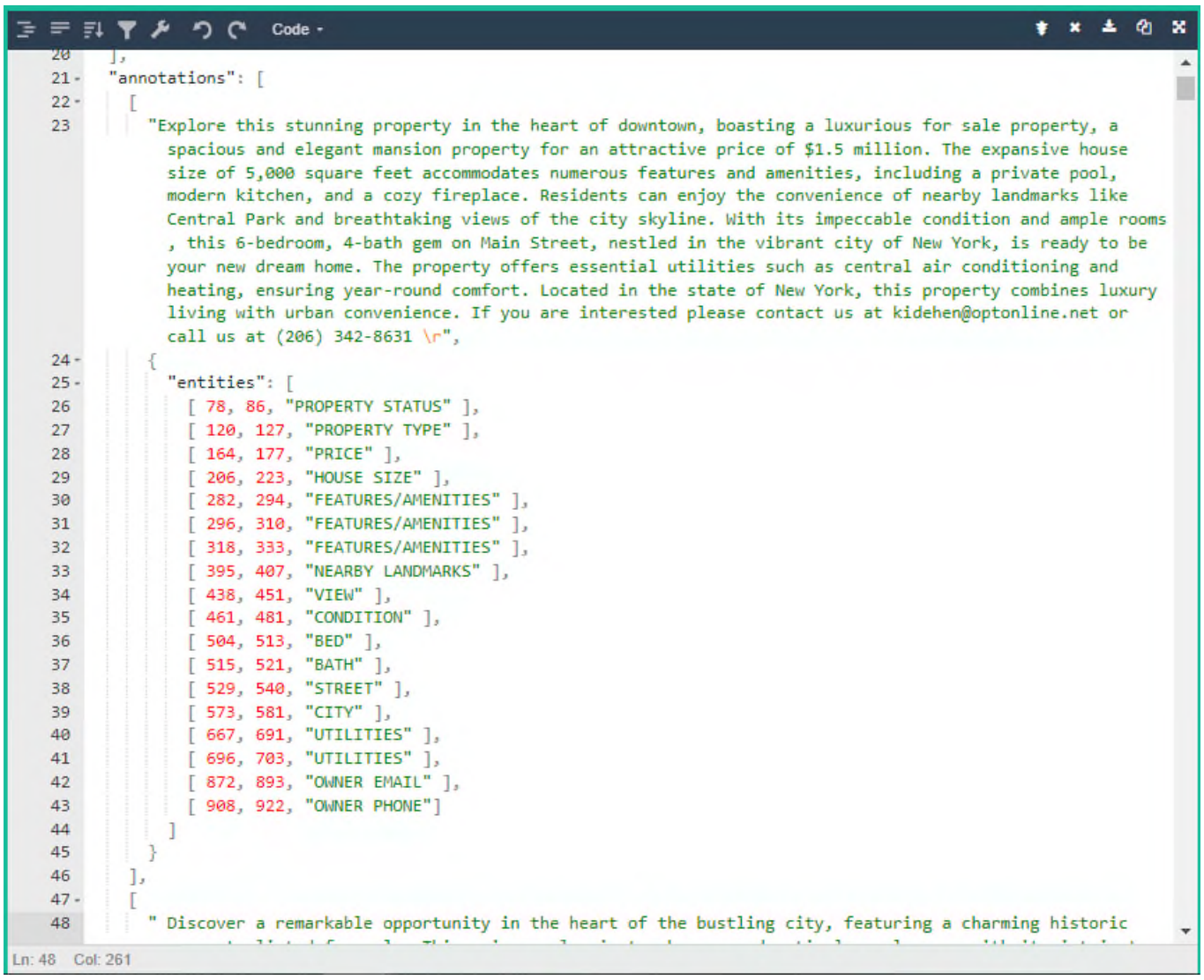


Рисунок 3.6 – Список слів з готовими сутностями

Файл містить набір оголошень та анотацій на них, а також початкові і кінцеві позиції слів з мітками на них які позначають готову сутність. З діаграми видно, витягнувши необхідну інформацію, створюється 2 набори даних для навчання та розпізнавання (train_data та dev_data). Після розбивання вхідного набору даних, формується документ, який буде використовуватися для тренування NER моделі. На діаграмі видно, що спочатку при тренуванні ініціалізується нова мовна порожня модель з англійським конвеєром обробки даних. Конвеєр слугує набором інструкцій, де кожна інструкція проводить модифікацію над даними або витягує з них необхідну інформацію. Далі в циклі створюються текстові документи, що представляють оголошення, та окремо сутності з мітками на них. Кожне оголошення з анотаціями існує, як окремий документ, який потім додається до загальної

кількості, де потім зберігаються як двійковий формат даних у папці проекту. Після цього можна подавати оголошення які система ще не бачила і вони зможуть бути розпізнані.

Висновки до розділу 3

У цьому розділі було розглянуто більш детально архітектуру з компонентами інформаційної моделі та проектування програмної системи з розпізнавання іменованих сутностей для об’єктів нерухомості за допомогою засобів NLP. Для розробки програмного застосунку з розпізнавання іменованих сутностей – було обрано мову програмування Python та застосування бібліотеки SpaCy. Цей вибір обумовлений тим що Python має велику екосистему з обробки природної мови, машинного навчання та штучного інтелекту що позиціонує себе як найбільш оптимальну мову для розробки методу, спрямованого на структурування тексту оголошень у сфері нерухомості по розпізнаванню іменованих сутностей.

Було розглянуто архітектуру компонентів бібліотеки SpaCy та як її використання зможе допомогти в реалізації поставленого завдання.

Також було визначено, та сформовано модулі які використовуватимуться в програмній системі з розпізнавання сутностей. Ці модулі відповідатимуть за попередню обробку текстових даних, ідентифікацію та категоризацію необхідних сутностей, тренування та навчання моделі розпізнавання і представлення даних користувачеві. Запропоноване розбиття програми на модулі дозволить легко та ефективно інкапсулювати дані, а зв’язок між ними дасть можливість правильній взаємодії даних між собою.

РОЗДІЛ 4 Дослідження методу розпізнавання сутностей для об’єктів нерухомості засобами NLP

4.1 Характеристика датасету та набору даних для тренування моделі з розпізнавання сутностей

Розпізнавання іменованих сутностей відіграє вирішальну роль у вилученні структурованої інформації з неструктурованих текстових даних. Для того, щоб могти вилучати і категоризувати вхідну інформацію необхідно якісно та чітко натренувати модель для ідентифікації та класифікації сутностей. Чим більш правильніше і влучніше подати дані для моделі, тим якісніше й точніше буде працювати алгоритм із розпізнавання.

Джерела даних

Навчальні дані, за якими тренувалася модель NER, були взяті із широкого спектру веб-застосунків з продажу нерухомості і включали відомі платформи, такі як Zillow, Realtor, RealEstate та аналогічні веб-застосунки з лістингами нерухомостей. Перелічені он-лайн-джерела разом містять величезну кількість текстової інформації, що належать до великого спектру земельних ділянок, маєтків, та цілих конгломерацій (рисунок 4.1-4.2)

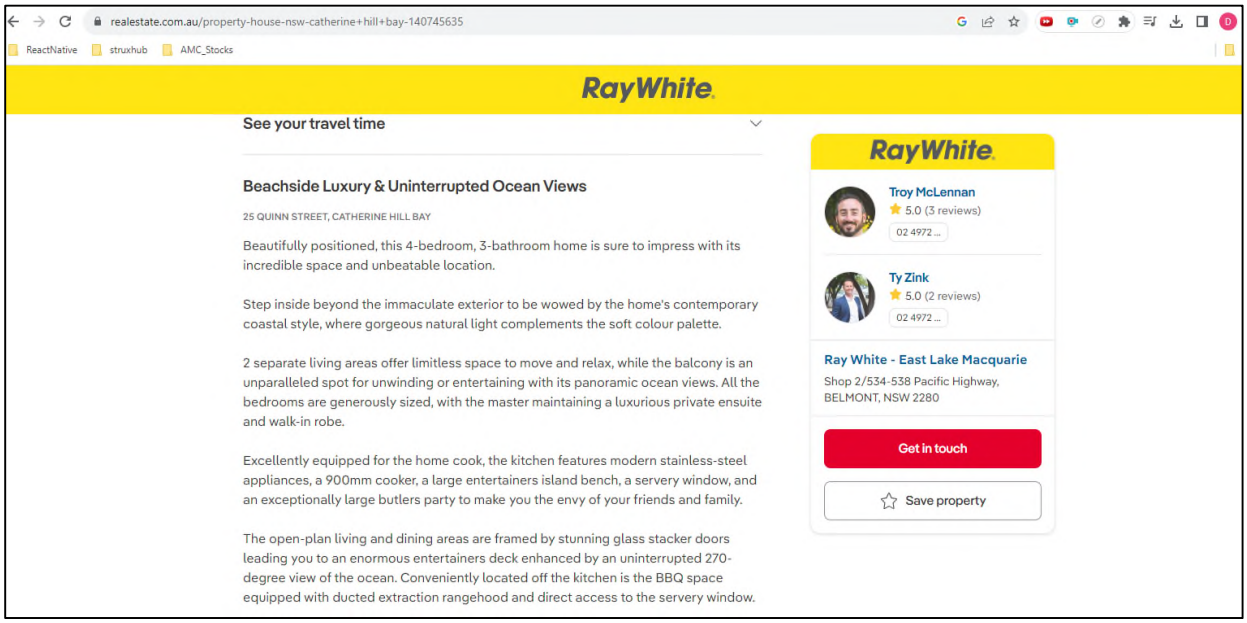


Рисунок 4.1 – Оголошення про об’єкт нерухомості із веб-застосунку RealEstate

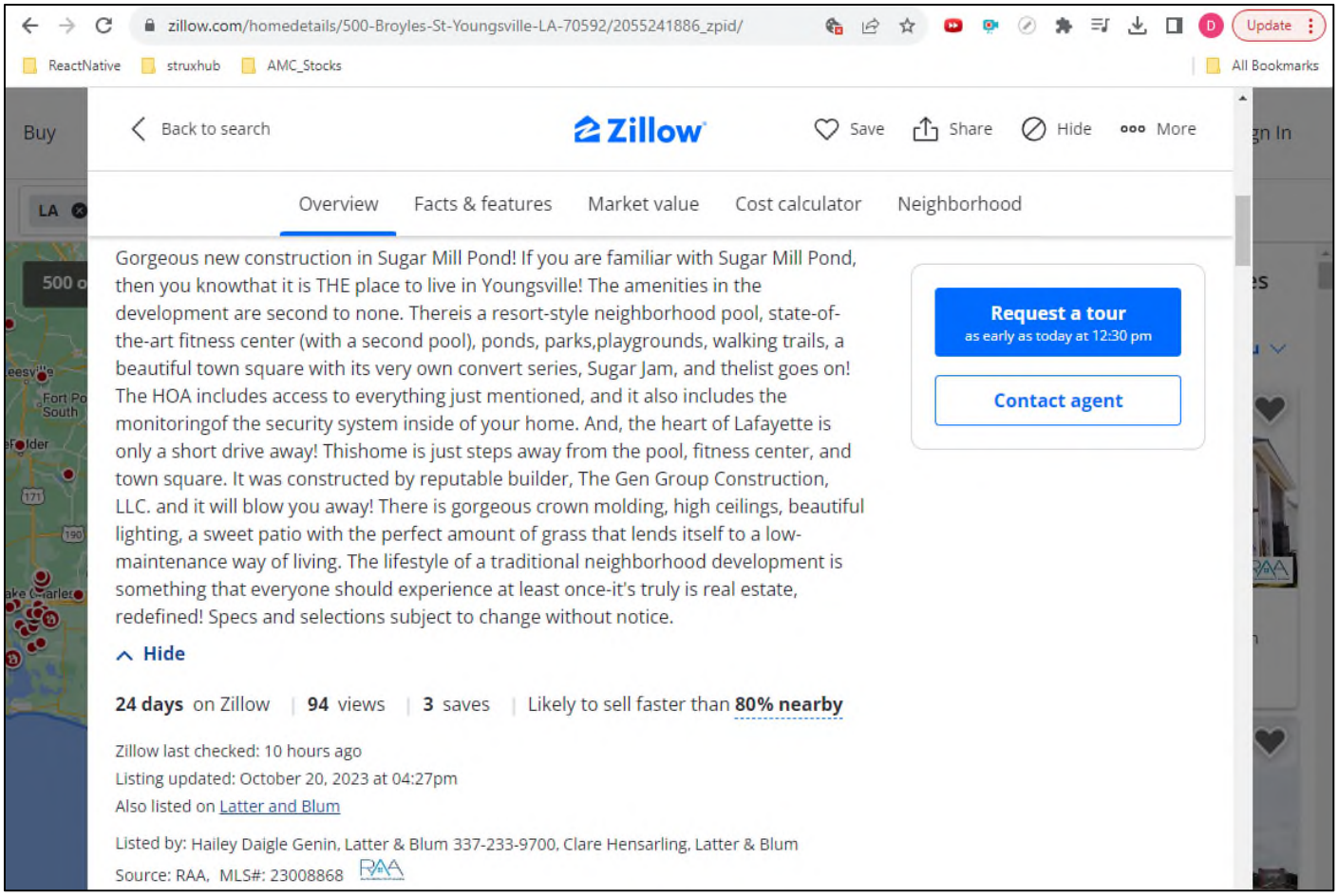


Рисунок 4.2 – Оголошення про об’єкт нерухомості із веб-застосунку Zillow

З рисунків добре видно наскільки детально описується інформація про той чи інший об’єкт нерухомості. Ці оголошення включають різні атрибути, специфічні слова й характеристики властивостей об’єктів, що робить їх ідеальними кандидатами для навчання моделі NER.

Збір та підготовка даних

Процес збору даних проводився систематично, щоб забезпечити високу якість навчання. Весь процес можна поділити на наступні етапи:

1. Вебскрейпінг: Спочатку було здійснено пошук вебсторінок, де якомога докладніше описувалася інформація про нерухомості. Ціль полягала в тому, щоб отримати набір текстів з оголошень про продаж маєтків, які представляли різні типи нерухомостей, включаючи житлові будинки, квартири, комерційні приміщення та багато іншої сумісної з ними різноманітної інформації. Ця різноманітність якраз і мала вирішальне значення для навчання надійної моделі NER.

2. Попередня обробка. Зібрані необроблені текстові дані піддалися попередній обробці для очищення та стандартизації. На цьому етапі були видалено всі непотрібні HTML-теги, спеціальні символи та інші перешкоди в тексті, які було отримано під час збору, що суттєво би завадили процесу навчання моделі. Оброблені текстові дані було збережено у звичайний Word файл із розширенням .docx (рисунок 4.3). Кожний текст з оголошення розміщений на окремій сторінці.

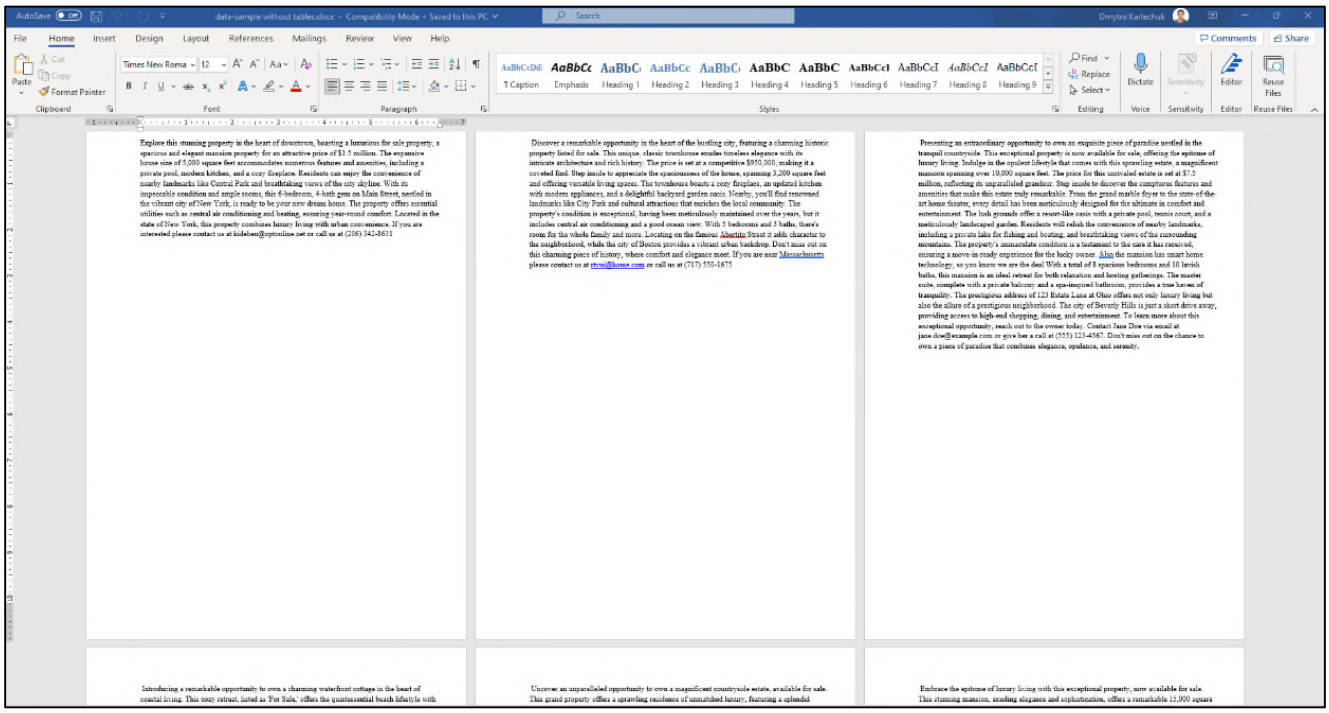


Рисунок 4.3 – Набір даних після попередньої обробки

3. Маркування та створення анотацій. На цьому етапі, який являється ключовим кроком у підготовці даних, було проведено маркування іменованих об'єктів. Для цього, було використано он-лайн-інструмент анотацій NER. Цей інструмент дав змогу вручну робити анотацію необхідних сутностей в тексті. Даний процес зображений на рисунку 4.4. Цей інструмент дозволяє виконувати наступні дії:

– завантаження набору даних;

– створення анотацій на розпізнавання;

– маркування тексту анотаціями;

– завантаження результатів, як файл JSON, для навчання моделі

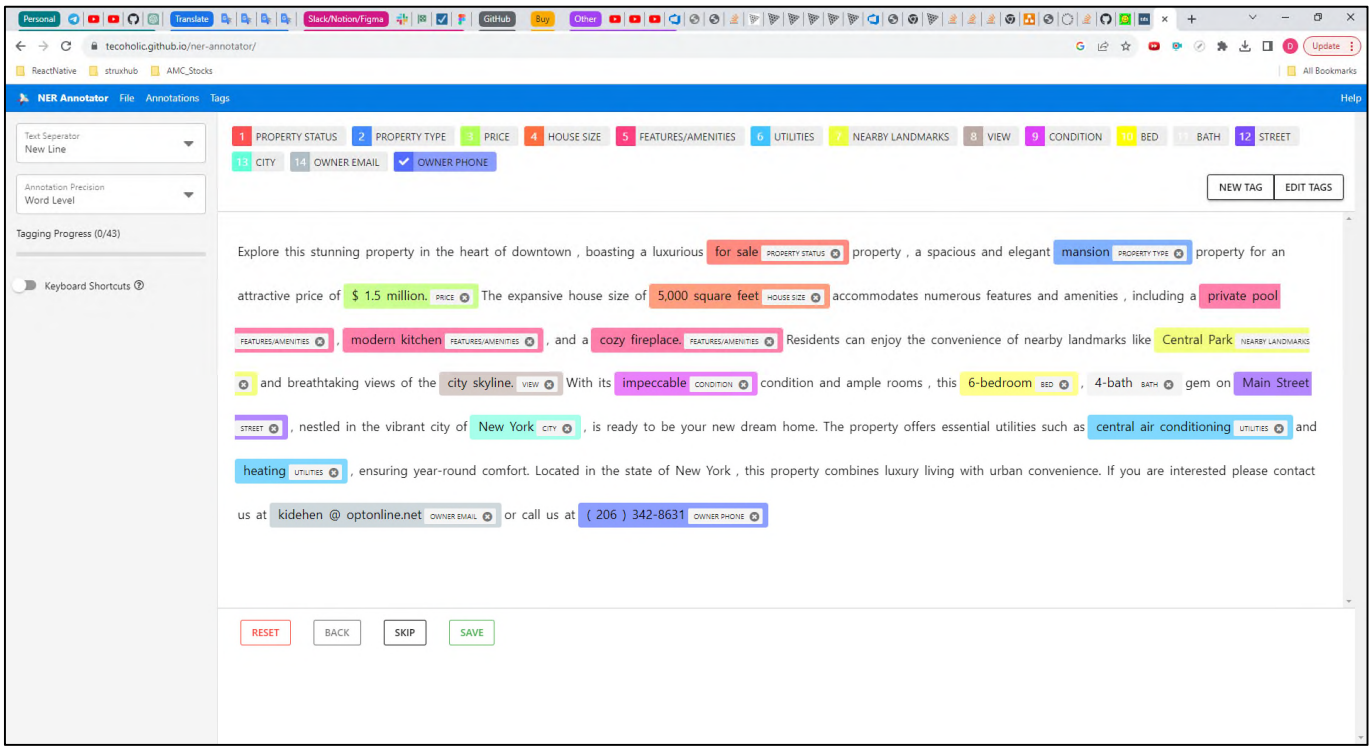


Рисунок 4.4 – Он-лайн інструмент анотацій NER

Сутності, по яким відбувалося анотування даних, було обрано з власних міркувань, попередньо проаналізувавши інформацію з різних вебсторінок. До таких сутностей належать наступні атрибути:

- property status: ця мітка відображає статус нерухомості й показує поточний її стан, а саме «на продаж», «купується», «здається в оренду» чи будь-який інший відповідний стан;
- property type: ідентифікує конкретний тип або категорію нерухомості, наприклад «Будинок на одну сім'ю», «Кондомініум», «Квартира», «Комерційне приміщення» і так далі;
- price: ціна показує грошову вартість нерухомості;
- house size: розмір нерухомості, визначає такий атрибут, як площа у квадратних футах або квадратних метрах, тобто описує площу об'єкта;
- features/amenities: ця категорія охоплює різні особливості та зручності, які містить у собі власність, до прикладів можна віднести «басейн», «сучасна кухня»,

«2-камерний камін» або інші атрибути, що підвищують ціну та привабливість об'єкта;

- utilities: ця категорія належить до комунальних послуг, доступних на території об'єкта нерухомості. Сюди можуть входити такі послуги, як «Центральне кондиціювання» та «Опалення» і тому подібне;
- nearby landmarks: являє собою найближчі пам'ятки, та ідентифікує відомі орієнтири, місця чи споруди поряд з об'єктом нерухомості, які можуть надати контекст його розташування;
- view: ця категорія позначає вид, який можна побачити з власності, наприклад «Вигляд на горизонт міста» або «Неймовірний вигляд на гори»;
- condition: описує загальний стан об'єкта нерухомості, вказуючи, чи він є новим, відремонтованим або будь-яким іншим відповідним станом;
- bed: кількість спальних місць, включає кількість кімнат, які знаходяться у нерухомості;
- bath: кількість ванних кімнат, включає кількість кімнат, які знаходяться у нерухомості;
- street: вказує назву або опис вулиці, на якій розташований об'єкт, та містить необхідну адресну інформацію;
- city: показує місто, у якому розташована власність, надаючи важливу географічну інформацію про місцезнаходження об'єкта;
- owner email: показує контактний адреси електронної пошти власника нерухомості або агента, який займається продажем, що дає змогу зацікавленим сторонам зв’язатися за необхідністю до власника маєтку;
- owner phone: номер телефону власника нерухомості;

Як видно, категорії були обрані у такий спосіб, щоб разом відповідати більш повній інформації про нерухомість.

4. Отримання даних: після призначення анотацій конкретним словам в оголошенні, тим самим, ідентифікуючи сутності, он-лайн-інструмент анотацій NER структурував всю помічену інформацію у форматі JSON для зручності обробки та

сумісності з навчанням моделі NER. На цьому етапі, дані у форматі JSON, представлені у напівструктурованому вигляді.

Представлення даних

Формат JSON містить позначені зразки тексту з готовим результатом, а також початкові та кінцеві позиції об'єктів у тексті. Це зображено нижче на рисунках 4.5

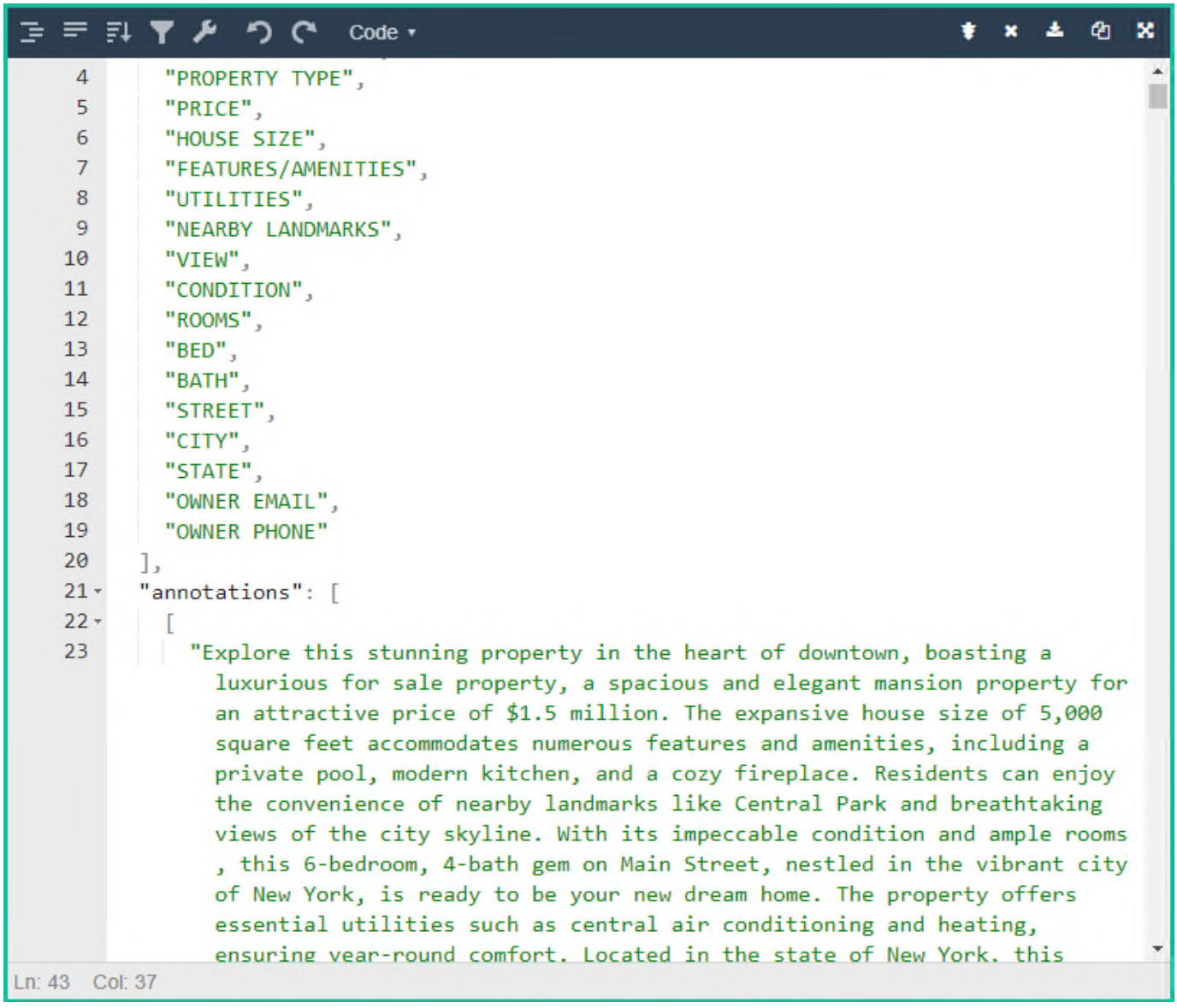


Рисунок 4.5 – Список сутностей у файлі annotations.json

На самому початку файл містить атрибут «classes», що являє собою увесь список сутностей, які можуть зустрітися в тексті. Нижче, розташовується атрибут «annotations», який представляє масив, в якому знаходиться неструктурований текст

оголошення та масив позиції слів у реченні з готовими призначеними їм сутностями. Даний етап зображено на рисунку 4.6 нижче:

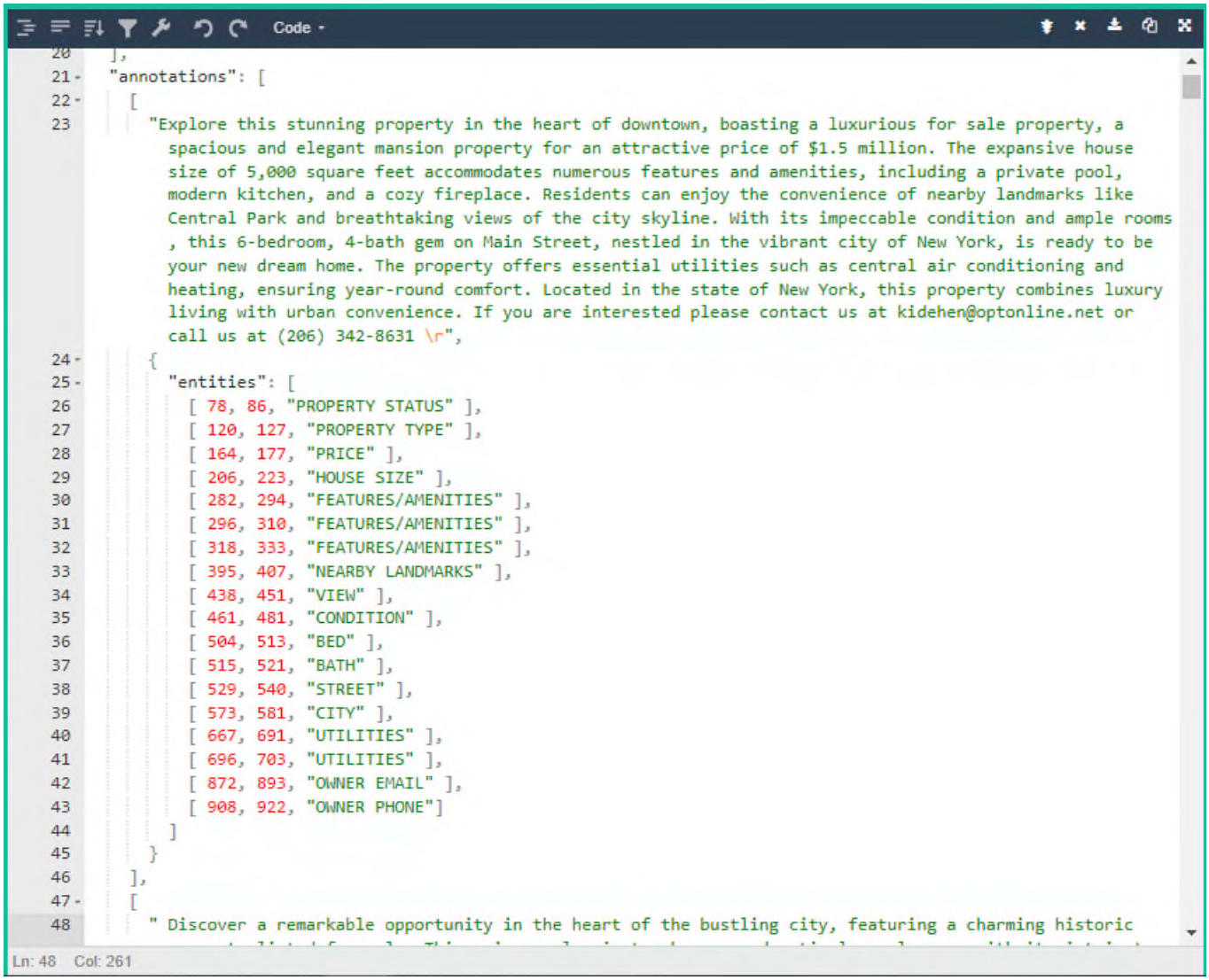


Рисунок 4.6 – Список слів з готовими сутностями

Як видно, процес збирання даних для навчання моделі NER вручну є досить складним. Дані, отримані з вебсторінок нерухомості, були ретельно підготовлені шляхом очищення, попереднього оброблення та ручного анотування іменованих об'єктів із використанням інструменту анотації NER. У таких випадках правильно використовувати готовий набір даних, але для такої предметної області, готового датасету не було. Отримані анотовані дані є необхідною складовою для навчання моделі з розпізнавання і класифікації сутностей із неструктурованого тексту і використовуються для досягнення точних результатів із розпізнавання.

4.2 Налаштування та тренування моделі для розпізнавання іменованих сутностей

Для розробки програмного застосунку з розпізнавання іменованих сутностей для об’єктів нерухомості засобами NLP, потрібно налаштувати робоче середовище, підключити правильний мовний пакет та саму бібліотеку SpaCy. Це виконується за допомогою наступних команд:

```
pip install -U pip setuptools wheel
pip install -U spacy
python -m spacy download en_core_web_sm
```

Ці команди відповідають за налаштування середовища Python з необхідними інструментами для роботи із цією бібліотекою. Крім того, також було обрано мовну модель «*en_core_web_sm*», яка підтягує словник англійських слів, словосполучень, та набір правил з англійської мови, оскільки оголошення із продажу нерухомостей були взяті з англомовних веб-застосунків.

Маючи дані для тренування в двійковому, необхідно створити файл конфігурації, який міститиме необхідні налаштування і параметри для тренування. Щоб не створювати файл вручну, файл конфігурації було ініціалізовано за допомогою наступної команди, вказавши лише необхідні параметри:

```
!python -m spacy init config config.cfg --lang en --pipeline ner --optimize
efficiency --force
```

Ця команда генерує файл config.cfg, який міститиме інформацію, про те, що для тренування буде використовуватися англомовна модель, конвеєр обробки буде використовувати компонент «Розпізнавання іменованих сутностей (NER)», та мусить бути оптимізований під точність, а не на продуктивність (рисунок 4.7). На даному кроці будуть всі необхідні дані для тренування моделі. На вхід необхідно подати датасет у двійковому форматі і файл конфігурації, а на виході буде створена

модель із налаштованими вагами. За допомогою наступної команди відбулося тренування моделі, а результат представлено на рисунку 4.8

```
!python -m spacy init config config.cfg --lang pt --pipeline ner --optimize efficiency --force
[i] Generated config template specific for your use case
- Language: pt
- Pipeline: ner
- Optimize for: efficiency
- Hardware: CPU
- Transformer: None
[+] Auto-filled config with all values
[+] Saved config
config.cfg
You can now add your data and train your pipeline:
python -m spacy train config.cfg --paths.train ./train.spacy --paths.dev ./dev.spacy
```

Рисунок 4.7 – Результат створення файлу конфігурації

```
!python -m spacy train config.cfg --output ./ --paths.train ./test.spacy --paths.dev ./real.spacy
```

```
!python -m spacy train config.cfg --output ./ --paths.train ./test.spacy --paths.dev ./real.spacy
[i] Saving to output directory: .
[i] Using CPU

=========================== Initializing pipeline ===========================
[+] Initialized pipeline

============================= Training pipeline =============================
[i] Pipeline: ['tok2vec', 'ner']
[i] Initial learn rate: 0.001
E    #       LOSS TOK2VEC  LOSS NER  ENTS_F  ENTS_P  ENTS_R  SCORE
---  ------  ------------  --------  ------  ------  ------  ------
  0       0          0.00    174.03    0.00    0.00    0.00    0.00
  6     200       5537.44   9788.41   38.16   51.79   30.21    0.38
 13     400        314.13   1735.42   44.96   50.32   40.62    0.45
 20     600        278.94    997.03   47.16   55.24   41.15    0.47
 26     800        224.10    735.28   48.85   54.49   44.27    0.49
 33    1000        224.50    582.06   49.23   60.15   41.67    0.49
 40    1200        629.28    583.92   47.13   60.66   38.54    0.47
 46    1400        220.08    454.60   48.55   54.55   43.75    0.49
 53    1600        257.34    417.74   52.07   60.27   45.83    0.52
 60    1800        297.13    403.92   48.48   57.97   41.67    0.48
 66    2000        318.26    396.99   50.29   57.33   44.79    0.50
 73    2200        294.21    342.04   50.15   56.95   44.79    0.50
 80    2400        368.51    349.94   48.10   61.29   39.58    0.48
 86    2600        329.14    329.72   47.51   54.36   42.19    0.48
 93    2800        359.04    346.21   49.00   54.09   44.79    0.49
100    3000        391.91    316.73   51.58   57.32   46.88    0.52
106    3200        463.09    339.27   47.48   55.17   41.67    0.47
[+] Saved pipeline to output directory
```

Рисунок 4.8 – Результат тренування моделі

Як видно з рисунку, при такій кількості даних у 700 екземплярів, точність навчання не перевищує 52%. Це свідчить про те, що така кількість даних є замалою, оскільки на кожну сутність може припадати, як декілька простих слів, так і складених слів. Згідно рекомендацій, правильно було б подати датасет з мінімальною кількістю 20000 екземплярів, оскільки є 17 класів, під які можуть бути розпізнані сутності, при чому на 1 клас може бути розпізнано до 100 слів. При правильній роботі необхідно було б на кожен класи подати мінімум 1000 екземплярів, але незважаючи на такий показник, якщо подати оголошення, що містить слова які були вже розгадані системою (а не нові), то точність буде врази більша, оскільки натренована модель уже матиме асоціації цих слів у векторному представленні.

У кінцевому варіанті, на рисунку 4.9 нижче показано як працює модель, подавши речення з оголошення про продаж нерухомості.

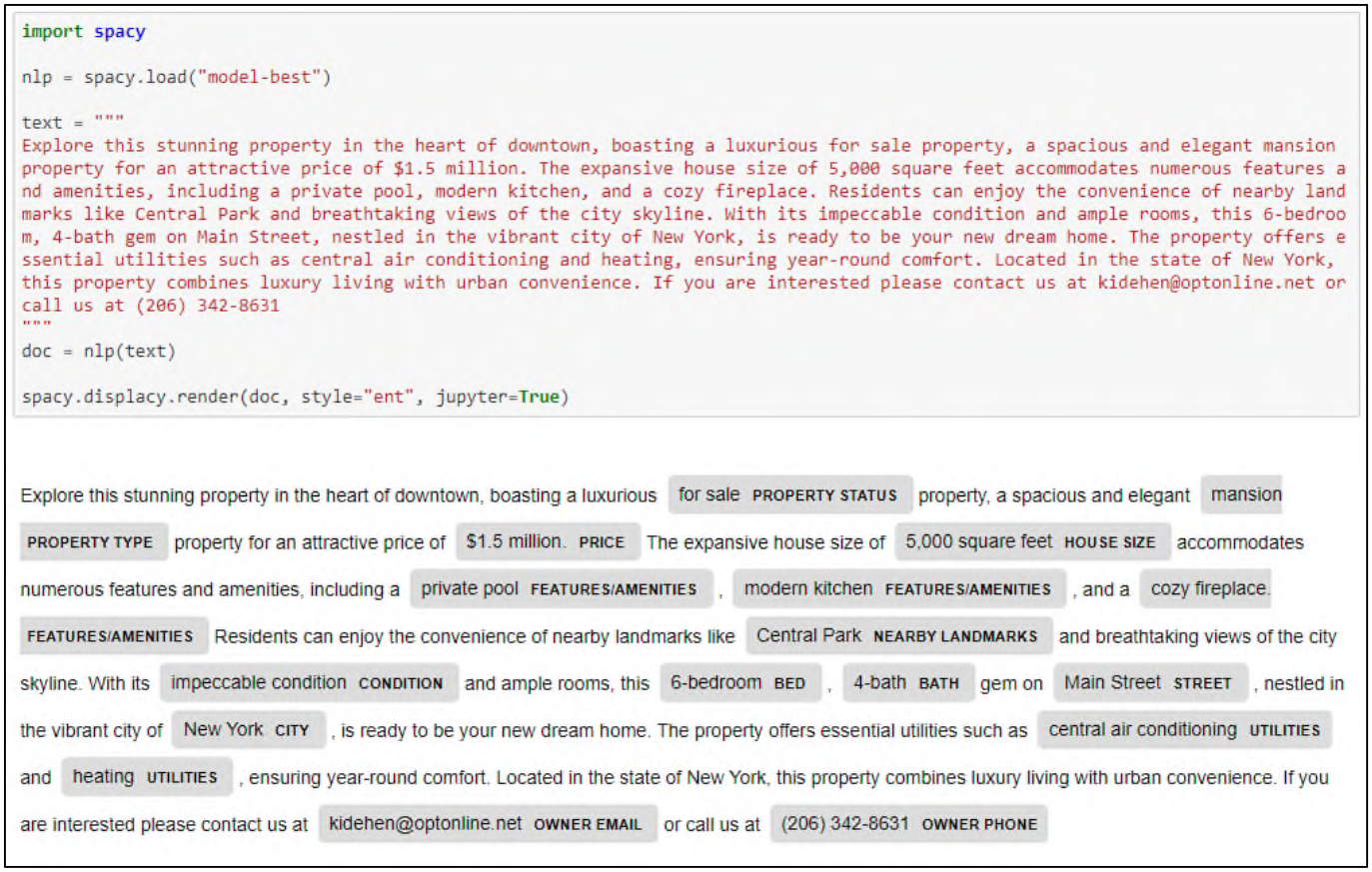


Рисунок 4.9 – Результат тренування моделі

З рисунку можна побачити, як натренована модель обробляє вхідний текст, створюючи документ, який містить розпізнані сутності. Результат розпізнавання представлений завдяки рендерингу модуля для візуалізації іменованих об'єктів у тексті. За допомогою різних налаштувань які притаманні рендерингу, розпізнані сутності по нерухомості можна підсвітити різними кольорами для більш кращого візуального представлення (рисунок 4.10):

Рисунок 4.10 – Розпізнані сутності в більш кращому представленні

4.3 Дослідження ефективності методу розпізнавання іменованих сутностей

Оцінювання методу розпізнавання сутностей є необхідною складовою для його правильної роботи і ретельного прогресу. У завданнях з розпізнавання сутностей, оцінка правильності роботи системи базується та формується на урахуванні того, як результати роботи, які видані системою, оцінювалися б людьми в галузі лінгвістики [40]. Тому, як вже було наголошено перевірка роботи програмного застосунку відбувалося за показниками, які використовувалися на конференціях MUC та CONLL.

Оцінювання проводилося на спеціально підібраному вхідному реченні, яке містить абсолютно всі сутності, за якими повинно проводитися розпізнавання. Вхідне речення подавалося 2 рази на порівняння результатів та якість роботи системи розпізнавання сутностей: перший раз при кількості екземплярів для навчання рівній 135 вибірок, та 2-ий раз при кількості екземплярів для навчання рівній 740. Нижче представлено це вхідне речення:

“Exceptional property for sale – a luxurious townhome featuring 6 bedrooms and 8 baths across 18,000 sq ft in Pennsylvania, Hamilton Street. Price: $18 million. The grand foyer leads to a living area with views of surrounding hills. Gourmet kitchen, private terrace, home theater, fitness center, wine cellar, infinity-edge pool, cabana, tennis court. Nearby Rodeo Drive, Greystone Mansion, downtown. Backup generator, solar panels and water filtration system. Also the property condition is it needs TLC. Contact owner, Lisa Johnson: lisa.johnson@example.com, (555) 123-4567.”

Варто відмітити дане речення, ні разу не зустрічалося, і не було використано для навчання моделі. Воно являє собою оголошення про продаж нерухомості із великою кількістю сутностей, які можуть зустрічатися кілька разів на один клас.

При правильній роботі методу, результат розпізнавання повинен був би бути таким, як це зображено на рисунку 4.11 нижче (тобто, усі сутності з тексту є розпізнаними):

Рисунок 4.11 – Робота методу розпізнавання сутностей

На рисунку вище показано ідеальний результат роботи розпізнавання іменованих сутностей із текстів оголошення про нерухомість. Але подавши такий текст на розпізнавання при початковій вибірці в 135 екземплярів, результат вийшов наступним (рисунок 4.12):

Рисунок 4.12 – Результат розпізнавання сутностей при кількості даних 135

З рисунків візуально можна помітити, де система помилилася, а нижче в таблиці 1.1 описано фрагмент відмінностей між очікуваним результатом і тим що дала система. Повний результат (тобто всі відмінності) наведено в додатках.

Дана таблиця містить правильний результат для сутностей, які мали би бути розпізнані, а також результат, який дала система з поясненнями де система помилилася.

Таблиця 1.1 – короткий зразок результатів розпізнавання сутностей програмного застосунку при вибірці в 135 екземплярів:

Правильний результат	Результат виданий системою	Опис результату
<Entity type="location"> For Sale </Entity >	-	Сутність була повністю пропущена системою
<Entity type="property type"> Townhome </Entity >	-	Сутність була повністю пропущена системою
<Entity type="bed"> 6 bedrooms </Entity >	<Entity type="bed"> 6 bedrooms </Entity >	Сутність правильно розпізнана
<Entity type="bath"> 8 baths </Entity >	<Entity type="bath"> 8 baths </Entity >	Сутність правильно розпізнана
<Entity type="house size"> 18,000 sq </Entity >	<Entity type="house size"> Across 18,000 </Entity >	Система помітила, що є сутність, але неправильно визначила її межі

З таблиці 1.1 можна зробити висновок, що система видала 15 різних помилок, та одночасно з тим правильно розпізнала лише 12 сутностей.

Маючи такий результат, оцінка ефективності була розрахована за наступними показниками:

1. Оцінювання згідно конференції MUC.

Згідно цієї конференції сутність вважається розпізнаною, якщо послідовності текстовим даним присвоєно правильний тип, незалежно від їх границь. Entity Text вважається розпізнаним, якщо границі текстових даних визначено правильно, незалежно від того чи тип на цю сутність розпізнано правильно. Обидва критерії мають 3 показники: кількість правильно розпізнаних сутностей, кількість правильних і неправильних результатів розпізнавання, та загальну кількість усього можливих розпізнаних сутностей

Згідно результатів розпізнавання представлених у таблиці 1.1 – correct answers буде складати 23 (14 Entity Type + 9 Entity Text), actual answers – 34 (17 Entity Type + 17 Entity Text), possible answers – 52 (26 Entity Type + 26 Entity Text).

Підставивши ці значення у формули – влучність буде дорівнювати 67.65%, повнота 44.23%, а гармонійне середнє буде 53.49%.

Зміст такого оцінювання полягає в тому, що воно враховує всі можливі типи помилок, представлені в таблиці 4, або ті, які можуть виникнути лише на одному критерії (або Entity Type, або Entity Text). Оскільки, максимально може бути присутні два критерії, то кожен із них оцінюється в один бал. Як можна побачити, бувають випадки, коли помилки оцінюються в 0 балів через відсутність обох критеріїв як Entity Type, так і Entity Text.

2. Оцінювання точної відповідності, згідно конференції CONLL.

Аналогічно до оцінювання по MUC, тут також розраховується точність та повнота розпізнавання, але окрема іменована сутність буде вважатися правильною лише у тому випадку, якщо її тип розпізнано правильно і одночасно границі текстових даних визначено правильно.

У попередньому прикладі є 26 істинних сутностей, які повинні бути розпізнані, при чому 17 припущень було здійснено системою і лише 6 правильно розпізнаних сутностей. У такий спосіб, точність становитиме 35.30%, повнота – 23.07%, а f-оцінка – 27.90%. Результат представлено нижче на рисунку 4.13

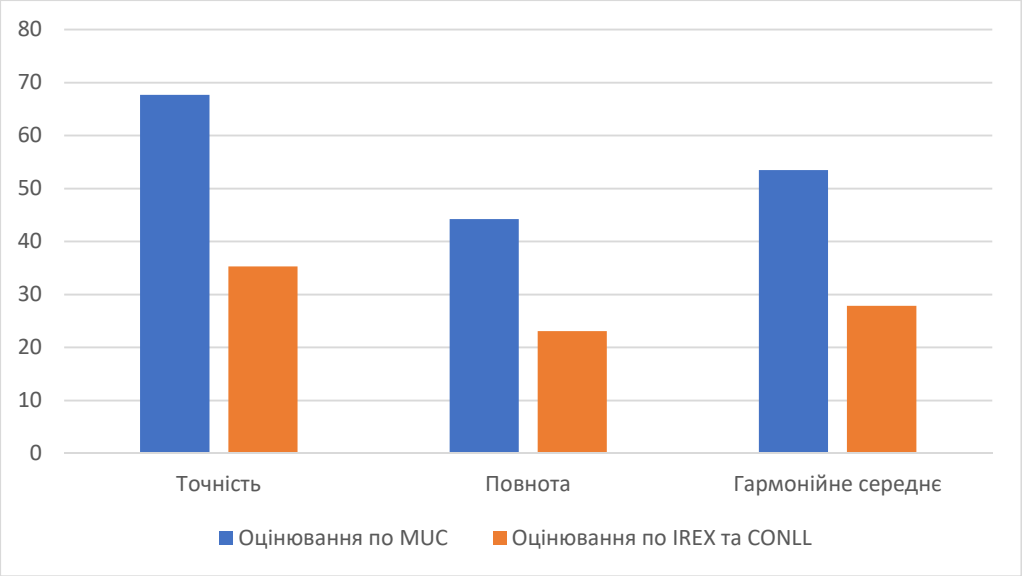


Рисунок 4.13 – Графік порівняння точності, повноти та середнього гармонійного при розпізнаванні сутностей на навчальній вибірці даних рівній у 135 екземплярів

З графіку вище можна зробити такий висновок, що навіть при малій кількості даних для навчання, оцінка згідно конференції MUC може досягти великого відсотку так як розпізнана сутність розраховується як сума правильно ідентифікованої анотації (без границь) та правильно ідентифікованого тексту (без анотації на нього). Але незважаючи на це результат можна покращити, подавши більшу кількість екземплярів для навчання моделі. Збільшивши кількість, та подавши аналогічний текст на розпізнавання при вибірці в 740 екземплярів, результат вийшов наступним (рисунок 4.14):

Рисунок 4.14 – Робота методу розпізнавання сутностей при кількості даних 740

З даного рисунку, можна помітити, що результат уже відрізняється, більш точніше розпізнано екземпляри, але все рівно система має помилки. Нижче в таблиці 1.2, наведеною фрагмент помилок в розпізнаванні, та описано результати видані системою. Повну версію відмінностей описано в додатках.

Таблиця 1.2 – короткий зразок результатів розпізнавання сутностей програмного застосунку при вибірці в 740 екземплярів:

Правильний результат	Результат виданий системою	Опис результату
<Entity type="location"> For Sale </Entity >	-	Сутність була повністю пропущена системою
<Entity type="property type"> Townhome </Entity >	-	Сутність була повністю пропущена системою

Правильний результат	Результат виданий системою	Опис результату
-	<Entity type="nearby landmarks"> Nearby </Entity >	Система висунула гіпотезу про існування сутності, де її зовсім немає
<Entity type="bath"> 8 baths </Entity >	<Entity type="bath"> 8 baths </Entity >	Сутність правильно розпізнана
<Entity type="house size"> 18,000 sq </Entity >	<Entity type="house size"> Across 18,000 </Entity >	Система помітила, що є сутність, але неправильно визначила її межі

Згідно оцінювання по конференції MUC та результатів розпізнавання представлених у таблиці 1.2 – correct answers буде складати 36 (19 Entity Type + 17 Entity Text), actual answers – 50 (25 Entity Type + 25 Entity Text), possible answers – 52 (26 Entity Type + 26 Entity Text). Підставивши ці значення у формули – влучність буде дорівнювати 72%, повнота 69.23%, а гармонійне середнє буде 70.59%.

Згідно оцінювання по конференції CONLL, маємо 26 істинних сутностей, які повинні бути розпізнані, при чому 25 припущень було здійснено системою і лише 12 правильно розпізнаних сутностей. Отже, точність становитиме 48%, повнота – 46.15%, а f-оцінка – 47.06%. Обидва результати представлено на рисунку 4.15.

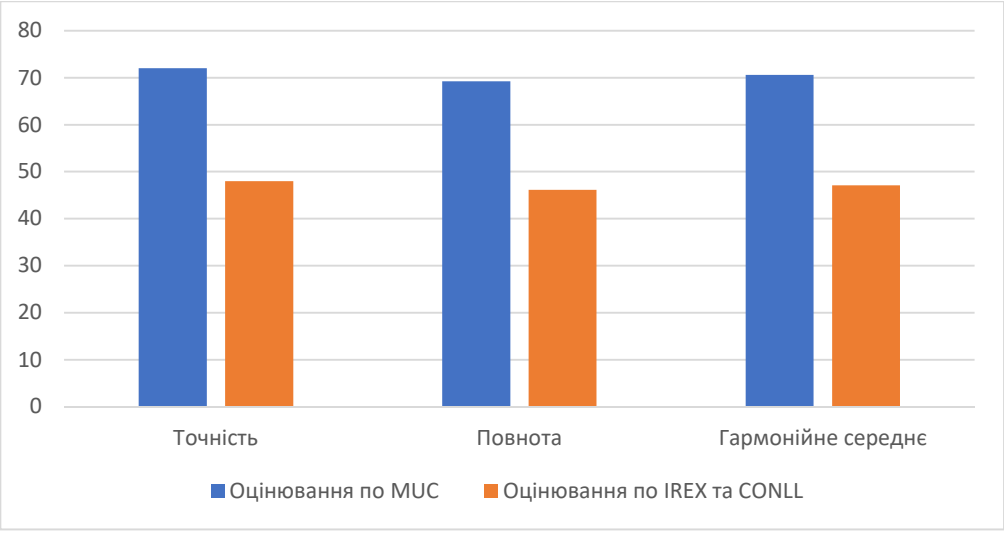


Рисунок 4.15 – Графік порівняння точності, повноти та середнього гармонійного при розпізнаванні сутностей на навчальній вибірці даних рівній в 740 екземплярів

Маючи результати оцінювання при вибірці рівній в 135 та 740 екземплярів, точність, повнота, та влучність покращили своє значення. Результати порівняння представлено нижче на рисунку 4.16.

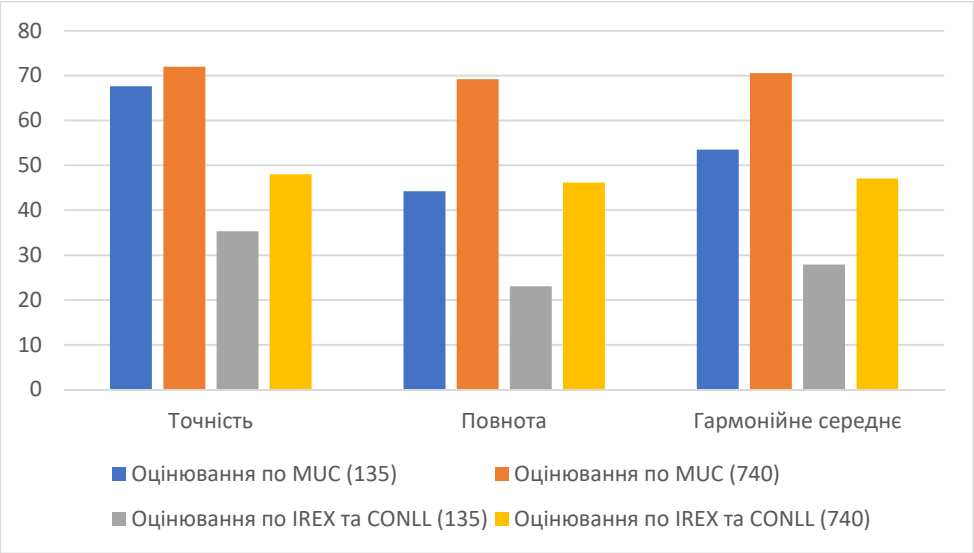


Рисунок 4.16 – Графік порівняння точності, повноти та середнього гармонійного між 135 екземплярами та 740

З графіку вище можна помітити пропорційну залежність зростання ефективності розпізнавання в порівнянні з різною кількістю екземплярів для навчання. При вибірці в 740 екземплярів (згідно оцінювання по MUC) – точність розпізнавання зросла на 4.35%, повнота на 25%, гармонійне середнє на 17.1%. При вибірці в 740 екземплярів (згідно оцінювання по CONLL) – точність розпізнавання зросла на 12.7%, повнота на 23.08%, гармонійне середнє на 19.16%. Порівняння розпізнавання між кількістю екземплярів зображено в таблиці 1.3 та 1.4.

Таблиця 1.3 – Порівняння якості розпізнавання згідно оцінювання по MUC:

	Вибірка в 135 екземплярів	Вибірка в 740 екземплярів	Різниця
Точність	67.65%	72%	4.35%
Повнота	44.23%	69.23%	25%
Гармонійне середнє	53.49%	70.59%	17.1%

Таблиця 1.4 – Порівняння ефективності розпізнавання згідно оцінювання по CONLL:

	Вибірка в 135 екземплярів	Вибірка в 740 екземплярів	Різниця
Точність	35.3	48	12.7%
Повнота	23.07	46.15	23.08%
Гармонійне середнє	27.9	47.06	19.16%

Отже, при оцінюванні ефективності роботи методу розпізнавання сутностей для об’єктів нерухомостей засобами NLP, було здійснено дослідження якості розпізнавання даних за допомогою таких показників як точність, повнота і гармонійне середнє значення між ними. Дослідження було проведено на справжніх оголошеннях, які було взято з веб-застосунків з продажу нерухомості та базувалося на 2-ох критеріях: оцінювання запропоноване на конференції MUC, та оцінювання запропоноване на конференції CONLL.

Згідно результатів дослідження, було встановлено закономірність, що при збільшенні кількості навчальних даних для тренування моделі з розпізнавання сутностей, відбувається приріст ефективності самої ідентифікації та категоризації об’єктів. Проаналізувавши якість розпізнавання на основі 135 тренувальних екземплярів та 740, точність розпізнавання (згідно MUC) зросла від 67.65% до 72%, а повнота від 44.23% до 69.23%. Згідно CONLL точність розпізнавання зросла від 35.3% до 48%, а повнота від 23.07% до 46.15%. Різниця між 2-ма критеріями оцінювання полягає в тому, що в 1-му випадку сутність вважається розпізнаною якщо її тип ідентифіковано правильно (незалежно від того чи границі сутності розпізнано правильно), або якщо самі границі сутності ідентифіковано правильно (незалежно від того чи тип на неї розпізнано правильно). В 2-му випадку сутність вважається розпізнаною, якщо і тип і границі розпізнані правильно, тому можна зустріти таку відмінність у метриках.

Висновки до розділу 4

При дослідженні методу розпізнавання сутностей для об’єктів нерухомості засобами NLP, було детально розглянуто структуру прикладних компонентів, а також роботу методу загалом. Описано особливості формування датасету для навчання та тренування моделі з розпізнавання сутностей, а також показано, як працює модель на реальних даних, та як в кінцевому варіанті виглядає результат розпізнавання.

Також було здійснено оцінювання ефективності роботи методу розпізнавання сутностей для об’єктів нерухомостей засобами NLP, за допомогою таких показників як точність та повнота. Якість розпізнавання порівнювалася на справжніх оголошеннях, які було взято з веб-застосунків з продажу нерухомості. Згідно результатів дослідження, було встановлено закономірність, що при збільшенні кількості навчальних даних для тренування моделі з розпізнавання сутностей, відбувається приріст ефективності самої ідентифікації та категоризації об’єктів. Проаналізувавши якість розпізнавання на основі 135 тренувальних екземплярів та 740, точність розпізнавання (згідно MUC) зросла від 67.65% до 72%, а повнота від 44.23% до 69.23%. Згідно CONLL точність розпізнавання зросла від 35.3% до 48%, а повнота від 23.07% до 46.15%. Різниця між 2-ма критеріями оцінювання полягає в тому, що в 1-му випадку сутність вважається розпізнаною якщо її тип ідентифіковано правильно (незалежно від того чи границі сутності розпізнано правильно), або якщо самі границі сутності ідентифіковано правильно (незалежно від того чи тип на неї розпізнано правильно). В 2-му випадку сутність вважається розпізнаною, якщо і тип і границі розпізнані правильно, тому можна зустріти таку відмінність у метриках.

Загальні висновки

На основі результатів виконаної кваліфікаційної роботи магістра, було розроблено метод структурування тексту оголошень для об'єктів нерухомості засобами NLP. Створений програмний застосунок вміє розпізнавати інформацію пов’язану з нерухомістю, таку як ціна маєтку, розмір площі забудови, особливості будівлі та інші споріднені дані.

Під час виконання роботи, було вирішено наступні завдання:

- проведено аналіз методів з обробки природної мови (NLP). Було детально оглянуто предметну область методів з обробки текстової інформації та розглянуто основний підхід по ідентифікації та класифікації даних у не структурованому тексті, за допомогою методу розпізнавання іменованих сутностей (NER), цим самим підкреслюючи ключову роль NER у перетворенні неструктурованих даних у структуровані.
- здійснено аналіз методу розпізнавання іменованих сутностей, де більш детально було розібрано, як працюють системи NER, а саме їх алгоритм по ідентифікації та класифікації об'єктів. Згідно аналізу для створення методу було використано алгоритм розпізнавання сутностей, який базується на підході з учителем, оскільки він дає точніші результати в порівнянні з іншими підходами.
- розроблено метод структурування тексту оголошень для об'єктів нерухомості, використовуючи засобами NLP. Було розглянуто та показано архітектуру з компонентами інформаційної моделі та проектування програмної системи з розпізнавання іменованих сутностей для об’єктів нерухомості, а також визначено модулі які використовуватимуться в програмному застосунку, оскільки вони відповідатимуть за його правильне функціонування.
- удосконалено метод розпізнавання іменованих сутностей (NER) для задачі категоризації об'єктів у текстах оголошень. Це досягається за рахунок використання архітектурного підходу EEAP для ідентифікація та класифікація сутностей.
- проведено експериментальне тестування реалізованого сервісу на справжніх оголошеннях, які було взято з популярних веб-застосунків по продажу

нерухомості. Згідно результатів тестування, було встановлено закономірність, що при збільшенні кількості навчальних даних для тренування моделі з розпізнавання сутностей, відбувається приріст ефективності самої ідентифікації та категоризації об’єктів. Проаналізувавши якість розпізнавання на основі 740 тренувальних екземплярів, точність розпізнавання (згідно оцінювання по конференції MUC) становить 72%, а згідно оцінювання по конференції CONLL точність розпізнавання становить 48%. Різниця між 2-ма критеріями оцінювання полягає в тому, що в 1-му випадку сутність вважається розпізнаною якщо її тип ідентифіковно правильно (незалежно від того чи границі сутності розпізнано правильно), або якщо самі границі сутності ідентифіковно правильно (незалежно від того чи тип на неї розпізнано правильно). У 2-му випадку сутність вважається розпізнаною якщо і тип і границі розпізнані правильно, тому можна зустріти таку відмінність у метриках.

Розроблений програмний застосунок із розпізнавання іменованих сутностей для об'єктів нерухомості засобами NLP можна покращити за рахунок наступних шляхів удосконалення:

- підвищити точність та ефективність розпізнавання сутностей, за рахунок навчання моделі NER на більшому та різноманітнішому наборі даних оголошень про нерухомості.
- покращити здатність програми розуміти контекст слів за рахунок взаємодії з готовими моделями такими як BERT чи GPT-3, які можуть розв’язувати задачі в рамках NLP. Об’єднане використання моделей дасть змогу більш правильніше розпізнати сутності оскільки ці моделі натреновані на те, щоб розуміти нюанси в описах слів.
- поліпшити інтерфейс користувача, за рахунок продуманого дизайну програмного застосунку.

За виконаною роботою було підготовлено та подано тези на Всеукраїнську конференцію АПКН 2023.

Перелік посилань

1. Структуровані дані. [Електронний ресурс]. – Режим доступу: https://www.altexsoft.com/blog/structured-unstructured-data/
2. Неструктуровані дані. [Електронний ресурс]. – Режим доступу: https://www.researchgate.net/publication/277615592_Unstructured_Data_Analysis-A_Survey
3. Трансформація структурованих даних в неструктуровані. [Електронний ресурс]. – Режим доступу: https://www.graphable.ai/blog/natural-language-is-structured-data/
4. Різновид оголошеннь. [Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Classified_advertising
5. Natural Language Processing (NLP). [Електронний ресурс]. – Режим доступу: https://www.researchgate.net/publication/364302290_A_Brief_Survey_on_Natural_Language_Processing_Based_Text_Generation_and_Evaluation_Techniques
6. T. J.Legg, "What is NLP and what is it used for?," Medical News Today, 27 May 2022. [Електронний ресурс]. – Режим доступу:: https://www.medicalnewstoday.com/articles/what-is-nlp-andwhat-is-it-used-for.
7. Mihaela Amalia Petrovici, "Effective Methods of Learning and Teaching: A Sensory Approach" [Електронний ресурс]. – Режим доступу: https://www.sciencedirect.com/science/article/pii/S1877042813032710
8. From neuro-linguistic programming to natural language processing. [Електронний ресурс]. – Режим доступу: https://inkerrobotics.com/blog/natural-language-processing-vs-neuro-linguistic-programming
9. J. A. Gulla, "New Language Models in NorwAI," NorwAI, [Електронний ресурс]. – Режим доступу: https://www.ntnu.edu/norwai/newlanguage-models-in-norwai.
10. Text analysis in today’s world: [Електронний ресурс]. – Режим доступу: https://aws.amazon.com/what-is/text-analysis/
11. Text generation [Електронний ресурс]. – Режим доступу: https://www.datacamp.com/blog/what-is-text-generation

12. Basra Jehangir, Saravanan Radhakrishnan, Rahul Agarwal, "A survey on Named Entity Recognition": [Електронний ресурс]. – Режим доступу: https://www.sciencedirect.com/science/article/pii/S2949719123000146
13. Chinchor and Robinson, "The 6th Message Understanding Conference (MUC)": [Електронний ресурс]. – Режим доступу: http://ai-center.botik.ru/Airec/images/stories/Articles/Collections/muc-guidelines/MUC_7_NER_markup.pdf
14. Kim J., Kim Y., Kang S. "Weakly labeled data augmentation for social media named entity recognition" Expert Syst. Appl., 209 (2022), Article 118217. [Електронний ресурс]. – Режим доступу: https://www.sciencedirect.com/science/article/abs/pii/S0957417422013707
15. Transfer Learning. [Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Transfer_learning
16. Petkova D., Croft W.B, "Proximity-based document representation for named entity retrieval". Eur. Phys. J. B. [Електронний ресурс]. – Режим доступу: https://www.sciencedirect.com/science/article/pii/S2665963822000793
17. Raju B., Gulfishan2 F.A., Nepal B. An approach for extracting exact answers to question answering (qa) system for english sentences. Procedia Eng., 30 (2012), pp. 1187-1194. [Електронний ресурс]. – Режим доступу: https://www.sciencedirect.com/science/article/pii/S1877705812009897
18. Yuval Marton, Yuval Marton I.Z, "Transliteration normalization for information extraction and machine translation" J. King Saud Univ. Comput. Inf. Sci., 26. [Електронний ресурс]. – Режим доступу: https://www.sciencedirect.com/science/article/pii/S1877705812009897
19. Named Entity Recognition for Sensitive Data Discovery. [Електронний ресурс]. – Режим доступу: https://www.mdpi.com/2076-3417/10/7/2303
20. Rule-based approach. [Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Rule-based_machine_learning

21. Boosting drug named entity recognition using an aggregate classifier. [Електронний ресурс]. – Режим доступу: https://www.sciencedirect.com/science/article/pii/S0933365715000780.
22. Rule-based pattern extractor and named entity recognition: A hybrid approach. [Електронний ресурс]. – Режим доступу: https://ieeexplore.ieee.org/abstract/document/5561392
23. Rule-based NER LAsie system. [Електронний ресурс]. – Режим доступу: https://aclanthology.org/M98-1007.pdf
24. Unsupervised learning. [Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Unsupervised_learning
25. Two-stage NER for tweets with clustering. [Електронний ресурс]. – Режим доступу: https://www.sciencedirect.com/science/article/abs/pii/S0306457312000714
26. Study of Named Entity Recognition methods in biomedical field. [Електронний ресурс]. – Режим доступу: https://www.sciencedirect.com/science/article/pii/S1877050919316813
27. Supervised learning. [Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Supervised_learning
28. Medical named entity recognition based on dilated convolutional neural network. [Електронний ресурс]. – Режим доступу: https://www.sciencedirect.com/science/article/pii/S2667241321000197
29. Named Entity Recognition using Conditional Random Fields. [Електронний ресурс]. – Режим доступу: https://www.sciencedirect.com/science/article/pii/S1877050920308978
30. Named entity correction in neural machine translation using the attention alignment map. [Електронний ресурс]. – Режим доступу: https://www.mdpi.com/2076-3417/11/15/7026
31. Text Segmentation Techniques: A Critical Review. [Електронний ресурс]. – Режим доступу: https://core.ac.uk/download/pdf/159510906.pdf

32. A Fast WordPiece Tokenization System. [Електронний ресурс]. – Режим доступу: https://blog.research.google/2021/12/a-fast-wordpiece-tokenization-system.html
33. Google’s LinMaxMatch approach. [Електронний ресурс]. – Режим доступу: https://analyticsindiamag.com/google-adds-fast-wordpiece-tokenization-to-tensorflow
34. Linguistic tokenization. [Електронний ресурс]. – Режим доступу: https://spacy.io/usage/linguistic-features#tokenization
35. Noise removal [Електронний ресурс]. – Режим доступу: https://www.kdnuggets.com/2019/04/text-preprocessing-nlp-machine-learning.html
36. Lemmatization survey [Електронний ресурс]. – Режим доступу: https://ijrat.org/downloads/Conference_Proceedings/icatest2015/ICATEST-2015127.pdf
37. Stemming survey [Електронний ресурс]. – Режим доступу: https://www.researchgate.net/publication/261439174_A_survey_of_stemming_algorithms_in_information_retrieval
38. Limitations in LinMaxMatch algorithm: https://analyticsindiamag.com/google-adds-fast-wordpiece-tokenization-to-tensorflow/
39. Neural Architectures for Named Entity Recognition. [Електронний ресурс]. – Режим доступу: https://arxiv.org/pdf/1603.01360.pdf
40. D. Nadeau and S. Sekine, “A survey of named entity recognition and classification”, in Lingvisticae Investigationes, 30(1), 2021, pp. 3–26.
41. Python – programming language. [Електронний ресурс]. – Режим доступу: https://www.python.org/doc/essays/blurb/

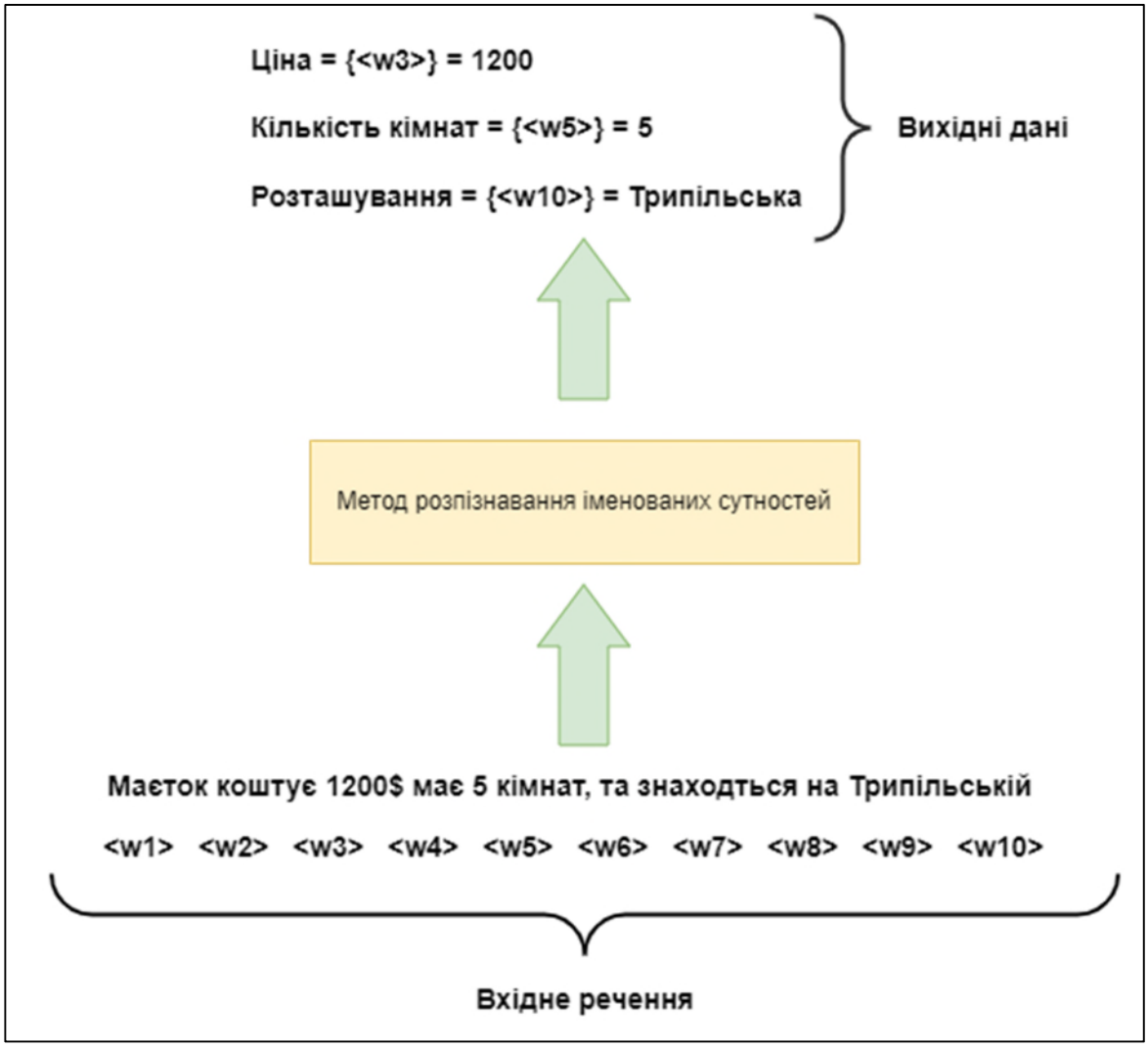
ДОДАТКИ

Додаток А

Алгоритм роботи методу розпізнавання сутностей

Додаток Б

Схема роботи методу розпізнавання сутностей в оголошенні



Додаток В

Таблиця 1.1 – результати розпізнавання сутностей програмного застосунку між при вибірці в 135 екземплярів:

Правильний результат	Результат виданий системою	Опис результату
<Entity type="property status"> For Sale </Entity >	-	Сутність була повністю пропущена системою
-	<Entity type="condition"> Exceptional property </Entity >	Система висунула гіпотезу про існування сутності, де її зовсім не повинно бути
<Entity type="property type"> Townhome </Entity >	-	Сутність була повністю пропущена системою
<Entity type="bed"> 6 bedrooms </Entity >	<Entity type="bed"> 6 bed </Entity >	Система помітила, що є сутність, але неправильно визначила її межі
<Entity type="bath"> 8 baths </Entity >	<Entity type="bath"> 8 baths </Entity >	Сутність правильно розпізнана
<Entity type="house size"> 18,000 sq </Entity >	-	Сутність була повністю пропущена системою
<Entity type="city"> Pennsylvania </Entity >	<Entity type="city"> Pennsylvania </Entity >	Сутність правильно розпізнана
<Entity type="street"> Hamilton Street </Entity >	<Entity type="street"> Street </Entity >	Система помітила, що є сутність, але неправильно визначила її межі
<Entity type="price"> 18 million </Entity >	<Entity type="price"> 18 million </Entity >	Сутність правильно розпізнана
<Entity type=" features/amenities"> grand foyer </Entity >	-	Сутність була повністю пропущена системою
-	<Entity type="property type"> living area </Entity >	Система висунула гіпотезу про існування сутності, де її зовсім не повинно бути
-	<Entity type="view">	Система висунула гіпотезу

	with views </Entity >	про існування сутності, де її зовсім немає
<Entity type="view"> surrounding hills </Entity >	<Entity type="nearby landmarks"> surrounding hills </Entity >	Система помітила сутність, але надала неправильну анотацію
<Entity type=" features/amenities"> Gourmet kitchen </Entity >	-	Сутність була повністю пропущена системою
<Entity type=" features/amenities"> private terrace </Entity >	-	Сутність була повністю пропущена системою
<Entity type=" features/amenities"> home theater </Entity >	<Entity type=" features/amenities"> home theater </Entity >	Сутність правильно розпізнана
<Entity type=" features/amenities"> fitness center </Entity >	-	Сутність була повністю пропущена системою
<Entity type=" features/amenities"> wine cellar </Entity >	-	Сутність була повністю пропущена системою
<Entity type=" features/amenities"> infinity-edge pool </Entity >	<Entity type=" features/amenities"> infinity-edge pool </Entity >	Сутність правильно розпізнана
<Entity type=" features/amenities"> cabana </Entity >	-	Сутність була повністю пропущена системою
<Entity type=" features/amenities"> tennis court </Entity >	<Entity type=" features/amenities"> tennis court </Entity >	Сутність правильно розпізнана
<Entity type="nearby landmarks"> Rodeo Drive </Entity >	-	Сутність була повністю пропущена системою
<Entity type="nearby landmarks">	-	Сутність була повністю пропущена системою

Greystone Mansion </Entity >		
<Entity type="utilities"> Backup generator </Entity >	-	Сутність була повністю пропущена системою
<Entity type="utilities"> solar panels </Entity >	<Entity type="utilities"> solar panels </Entity >	Сутність правильно розпізнана
<Entity type="utilities"> water filtration system </Entity >	<Entity type="utilities"> filtration system </Entity >	Система помітила, що є сутність, але неправильно визначила її межі
<Entity type="condition"> needs TLC </Entity >	<Entity type="condition"> property condition </Entity >	Система помітила, що є сутність, але неправильно визначила її межі
<Entity type="owner email"> lisa.johnson@example.com </Entity >	<Entity type="owner email"> Lisa Johnson: lisa.johnson@example.com </Entity >	Система помітила, що є сутність, але неправильно визначила її межі
<Entity type="owner phone"> (555) 123-4567 </Entity >	<Entity type="owner phone"> 123-4567 </Entity >	Система помітила, що є сутність, але неправильно визначила її межі

Таблиця 1.2 – результати розпізнавання сутностей програмного застосунку при вибірці в 740 екземплярів:

Правильний результат	Результат виданий системою	Опис результату
<Entity type="location"> For Sale </Entity >	-	Сутність була повністю пропущена системою
<Entity type="property type"> Townhome </Entity >	-	Сутність була повністю пропущена системою
<Entity type="bed"> 6 bedrooms </Entity >	<Entity type="bed"> 6 bedrooms </Entity >	Сутність правильно розпізнана
<Entity type="bath"> 8 baths </Entity >	<Entity type="bath"> 8 baths </Entity >	Сутність правильно розпізнана
<Entity type="house size"> 18,000 sq </Entity >	<Entity type="house size"> Across 18,000 </Entity >	Система помітила, що є сутність, але неправильно визначила її межі
-	<Entity type="house size"> Sq ft in </Entity >	Система висунула гіпотезу про існування сутності, де її зовсім немає
<Entity type="city"> Pennsylvania </Entity >	<Entity type="city"> Pennsylvania, Hamilton </Entity >	Система помітила, що є сутність, але неправильно визначила її межі і віднесла одну сутність до іншої
<Entity type="street"> Hamilton Street </Entity >	<Entity type="street"> Street </Entity >	Система помітила, що є сутність, але неправильно визначила її межі
<Entity type="price"> 18 million </Entity >	<Entity type="price"> 18 million </Entity >	Сутність правильно розпізнана
<Entity type=" features/amenities"> grand foyer </Entity >	<Entity type=" features/amenities"> grand foyer </Entity >	Сутність правильно розпізнана
-	<Entity type="nearby landmarks"> living area </Entity >	Система висунула гіпотезу про існування сутності, де її зовсім немає
<Entity type="view">	<Entity type="view">	Система помітила, що є

surrounding hills </Entity >	hills </Entity >	сутність, але неправильно визначила її межі
<Entity type=" features/amenities"> Gourmet kitchen </Entity >	<Entity type="utilities"> Gourmet kitchen </Entity >	Система помітила сутність, але надала неправильну анотацію
<Entity type=" features/amenities"> private terrace </Entity >	<Entity type="utilities "> private terrace </Entity >	Система помітила сутність, але надала неправильну анотацію
<Entity type=" features/amenities"> home theater </Entity >	<Entity type="utilities "> home theater </Entity >	Система помітила сутність, але надала неправильну анотацію
<Entity type=" features/amenities"> fitness center </Entity >	<Entity type=" features/amenities"> fitness center </Entity >	Сутність правильно розпізнана
<Entity type=" features/amenities"> wine cellar </Entity >	<Entity type=" features/amenities"> wine cellar </Entity >	Сутність правильно розпізнана
<Entity type=" features/amenities"> infinity-edge pool </Entity >	<Entity type=" features/amenities"> infinity-edge pool </Entity >	Сутність правильно розпізнана
<Entity type=" features/amenities"> cabana </Entity >	-	Сутність була повністю пропущена системою
<Entity type=" features/amenities"> tennis court </Entity >	<Entity type=" features/amenities"> tennis court </Entity >	Сутність правильно розпізнана
-	<Entity type="nearby landmarks"> Nearby </Entity >	Система висунула гіпотезу про існування сутності, де її зовсім немає
<Entity type="nearby landmarks"> Rodeo Drive </Entity >	-	Сутність була повністю пропущена системою
<Entity type="nearby landmarks">	<Entity type="street"> Rodeo Drive, Greystone	Система помітила, що є сутність, але неправильно

Greystone Mansion </Entity >	Mansion </Entity >	визначила її межі і надала неправильну анотацію
<Entity type="utilities"> Backup generator </Entity >	<Entity type="utilities"> Backup generator </Entity >	Сутність правильно розпізнана
<Entity type="utilities"> solar panels </Entity >	<Entity type="utilities"> solar panels </Entity >	Сутність правильно розпізнана
<Entity type="utilities"> water filtration system </Entity >	-	Сутність була повністю пропущена системою
<Entity type="condition"> needs TLC </Entity >	<Entity type="condition"> the property condition is it needs TLC </Entity >	Система помітила, що є сутність, але неправильно визначила її межі
<Entity type="owner email"> lisa.johnson@example.com </Entity >	<Entity type="owner email"> lisa.johnson@example.com </Entity >	Сутність правильно розпізнана
<Entity type="owner phone"> (555) 123-4567 </Entity >	<Entity type="owner phone"> (555) 123-4567 </Entity >	Сутність правильно розпізнана

Додаток Г

Світлини наукових публікацій, виконаних при роботі над кваліфікаційною роботою магістра

(ксерокопії титульної сторінки, сторінки змісту та всіх сторінок із публікацією)

Перелік наукових публікацій:

1. Карлечук Д.Т., Багрій Р.О., Скрипник Т.К., Тищенко О.О. Метод структурування тексту оголошень для об'єктів нерухомості засобами NLP.Метод оптимальної доставки замовлень у динамічній мережі на базі хмарних технологій. Збірник наукових праць за матеріалами XV Всеукраїнської науково-практичної конференції «Актуальні проблеми комп’ютерних наук АПКН-2023». – Хмельницький, 2023. - С. 111-115.

Міністерство освіти і науки України
Хмельницький національний університет

ЗБІРНИК НАУКОВИХ ПРАЦЬ

за матеріалами XV Всеукраїнської науково-практичної конференції
«Актуальні проблеми комп'ютерних наук АПКН-2023»

17-18 листопада 2023

Хмельницький 2023

Залуцька О.О., Молчанова М.О., Віт Р.В., Мазурець О.В.
Конфігурування нейронної мережі для класифікації емоційної тональності текстової інформації за показниками семантичної зв'язності 102

Запорожець М.В., Молчанова М.О., Скрипник Т.К.
Метод виявлення патологій мозку за зображеннями магнітно-резонансної терапії нейромережевими засобами 108

Карлечук Д.Т., Багрій Р.О., Скрипник Т.К., Тищенко О.О.
Метод структурування тексту оголошень для об'єктів нерухомості засобами NLP111

Карпович В.В., Дрозд А.І., Жуковський П.О., Мельник В.В.
Методи вирішення проблем пропускної здатності дисків для застосунків з інтенсивним обсягом даних 116

Каушан. С.О., Лисенко С.М.
Дослідження інформаційних систем електронного рекрутингу персоналу 118

Качур А.В., Лисенко С.М.
Виклики в розвитку технології віртуальної реальності: оптимізація архітектури VR 121

Качур О.І.
Перспективні напрямки розвитку сучасного антивірусного захисту мереж та роль методів на основі генетичних алгоритмів 124

Кирилюк О.О., Онишко О.Г.
Дослідження використання інструменту Elasticsearch для оптимізації вебдодатків, розроблених з використанням фреймворку Laravel 128

Кльоц Ю.П., Петляк Н.С., Чвалов А.А.
Технології тестування безпеки вебресурсів 130

Коберник Д.С.
Мобільний додаток для читання книг з Google Books: методології програмної інженерії та архітектурні рішення 133

Козакевич В.А., Собко О.В., Тищенко О.О., Вознюк Л.О., Медведчук В.Ю.
Метод автоматизованої генерації текстових повідомлень заданої семантичної спрямованості з використанням лексичних n-грам 136

Козельський О.В.
Методи та засоби створення мультикомп'ютерних систем з подвійною автентифікацією потоків даних в корпоративних мережах 142

УДК 004.4

Карлечук Д.Т., Багрій Р.О., Скрипник Т.К., Тищенко О.О.

Хмельницький національний університет

МЕТОД СТРУКТУРУВАННЯ ТЕКСТУ ОГОЛОШЕНЬ ДЛЯ ОБ'ЄКТІВ НЕРУХОМОСТІ ЗАСОБАМИ NLP

Розглянуто метод розпізнавання іменованних сутностей, який дозволяє ідентифікувати та класифікувати по категоріях розпізнані дані із текстових оголошень у сфері нерухомості, за допомогою підходів обробки природної мови, тим самим забезпечуючи трансформацію неструктурованих даних в структуровані. Також тут описано алгоритм роботи такого методу, внутрішні процеси зв'язані з ним та результат розпізнавання сутностей

A named entity recognition method is proposed that allows one to identify and categorize recognized data from text advertisements in the real estate industry using natural language processing methods, thereby ensuring the transformation of unstructured data into structured ones. The algorithm of this method, the internal processes connected with it and the result of entity recognition are described here

Вступ

Структуровані дані являють собою високоорганізовану, фактичну та точну інформацію [1]. Вони мають чітко визначений та організований формат, який повинен відповідати стійкій схемі або моделі даних, яка визначає тип та розташування цих елементів у самому наборі даних. Неструктуровані дані являють собою неорганізований набір інформації [2]. Вони не мають певної структури, та представлені у всьому різноманітті форм. Одним з прикладів неструктурованих даних, де може міститися інформація про об'єкти нерухомості являється оголошення.

Оголошення – це короткі текстові повідомлення, котрі містять різну за призначенням інформацію, зазвичай приватного рекламного характеру [3]. Зміст будь-якого оголошення варіюється залежно від його типу, та предметної області яку він представляє. Щоб отримати структуровані дані з неструктурованого тексту оголошень необхідно застосувати методи обробки текстової інформації, чим займається окрема галузь під назвою обробка природної мови (Natural Language Processing) [4].

Основна частина

Перетворення неструктурованих даних з оголошення у структуровані здійснюється за допомогою обробки текстової інформації, що включає в себе різні підходи і техніки з використанням текстових даних. Одним з таких підходів є

розпізнавання іменованих сутностей (NER). Це метод обробки природної мови, що використовується для ідентифікації та класифікації іменованих об'єктів у неструктурованому тексті за заздалегідь визначеними категоріями [5].

Метою роботи є розробка методу розпізнавання іменованих сутностей з текстів оголошень у сфері нерухомості. На рисунку 1 представлено цей процес.

Рисунок 1 – Процес роботи розпізнавання сутностей

Згідно рисунку, можна виділити на два ключових етапи:

1. Попередня обробка тексту. Перший крок включає підготовку текстових даних для аналізу. Він включає такі завдання, як токенізація (розбиття речення на слова), видалення шумів (видалення зайвих сполучних слів), лематизація та стемінг (спрощення слова до його основного кореня).

2. Ідентифікація та класифікація сутності. Після попередньої обробки тексту алгоритми NER сканують його, щоб визначити послідовності слів, які відповідають об'єктам. Для цього було використано алгоритм розпізнавання сутностей, який може базуватися на підході з учителем, оскільки він дає точніші результати в порівнянні з іншими підходами [5]. Цей алгоритм може аналізувати текст і призначати теги, мітки, та категорії на основі вмісту текстових даних, наприклад, вказавши мітку «1200» як ціна маєтку (рисунок 2)

З рисунка видно, що на початку NER-система приймає вхідне речення про об'єкт нерухомомті і розбиває його на окремі слова. Тобто, речення: «*Маєток коштує 1200$ має 5 кімнат, та знаходиться на Трипільській*» розбивається на список слів $<w_n>$, де w – означає слово з речення, а n – порядковий номер слова.

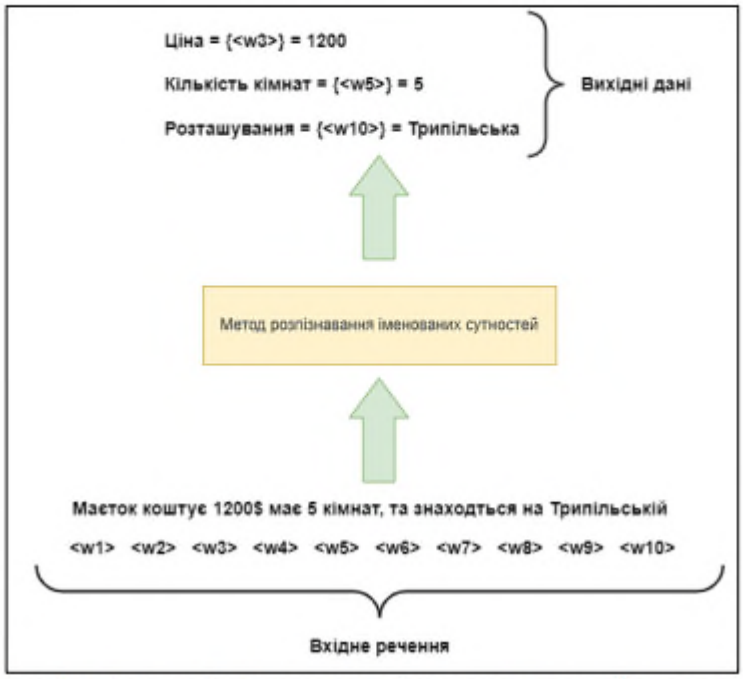


Рисунок 2 – Схема роботи методу розпізнавання сутностей в оголошенні

Отримавши слова з речення до них застосовується метод розпізнавання іменованих сутностей, який всередині являє собою цілий фреймворк ідентифікації та категоризації і має назву «Embed, Encode, Attend, Predict», що стосується концептуальної основи Метью Хоннібала з обробки природної мови [6]. Він зображений на рисунку нижче:

На етапі «*Embed*» кожне слово з речення «*Маєток коштує 1200$ має 5 кімнат, та знаходиться на Трипільській*» перетворюється в числовий вектор слова, тим самим, збираючи семантичну інформацію із вхідного тексту. Це саме відбувається із словами що позначають готові категорії. Таким чином слова зі схожим контекстом стають близько один до одного в числовому просторі.

Після цього відбувається етап «*Encode*» де кодується послідовність числових векторів слів (з попереднього етапу) у матрицю з використанням рекурентних нейронних мереж (RNN). На цьому етапі фіксуються відносини між словами.

Далі етап «*Attend*» – приймає матриці (сформовані на попередньому кроці) та зводить їх до одного єдиного вектора, який буде далі переданий у нейронну мережу прямого поширення для здійснення прогнозування.

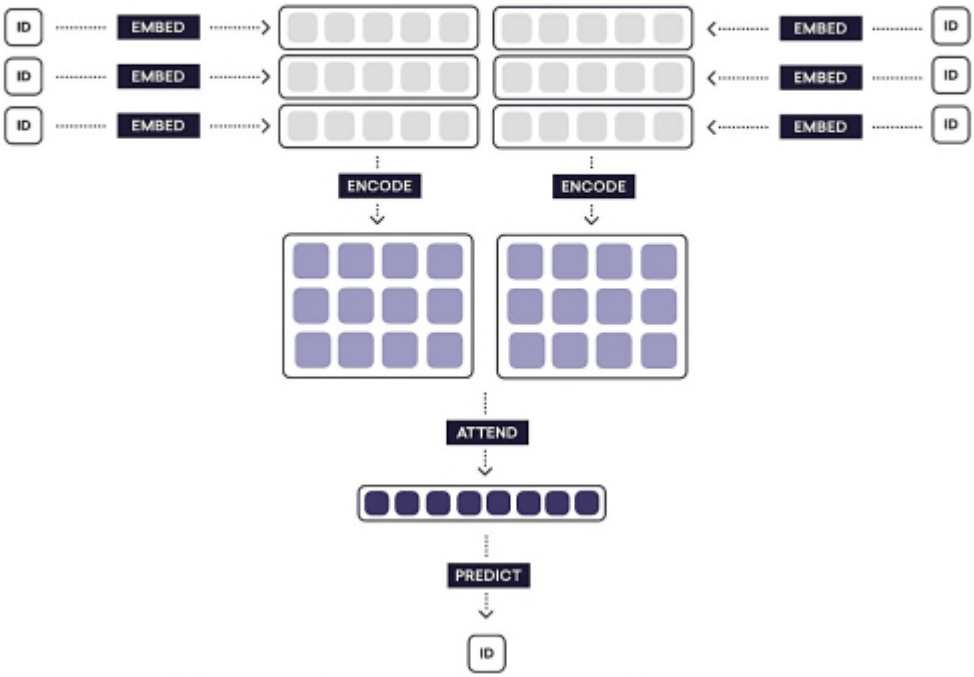


Рисунок 3 – Фреймворк «Embed, Encode, Attend, Predict»

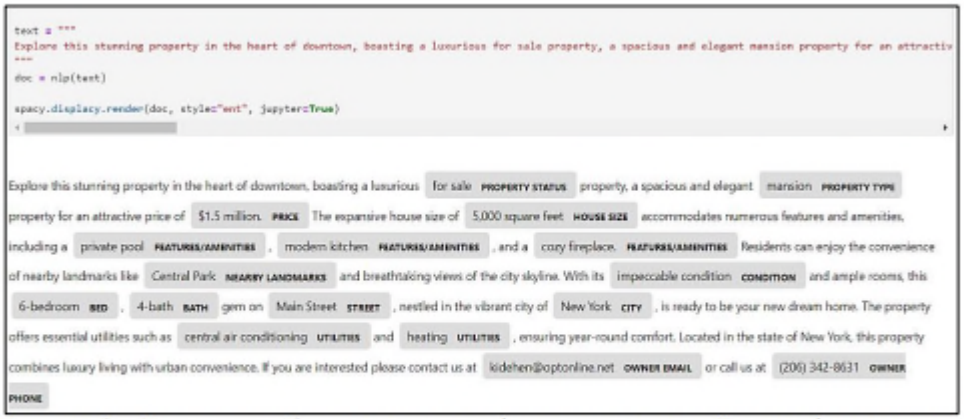


Рисунок 3 – Результат роботи методу з розпізнавання сутностей з текстів оголошень по об'єктах нерухомості

Останній етап «*Predict*» – прогнозує (передає вектор з попереднього кроку) багатошаровому перцептрону для виведення ідентифікатора мітки об'єкта. Саме тут відбувається генерація бажаного результату.

Розібравши роботу розпізнавання іменованих сутностей (NER), важливо відзначити, що практична реалізація NER є значно полегшеною за рахунок уже відомих потужних бібліотек обробки природної мови. Одна з таких комплексних бібліотек, яку було використано в дослідженні має назву SpaCy [7], що забезпечує зручний інтерфейс для NER та значно спрощує процес. На рисунку 3, представлено результат роботи методу розпізнавання сутностей подавши на вхід текст оголошення з нерухомості.

Висновки

Отже, проведено аналіз методів обробки природної мови в структуруванні текстів оголошень для об'єктів нерухомості засобами NLP. Розглянуто та описано основний підхід по ідентифікації та класифікації сутностей у неструктурованому тексті, та аналіз методу з розпізнавання іменованих сутностей. На основі попередньої інформації розроблено та реалізовано метод структурування тексту оголошень. Розроблений метод дозволяє розпізнавати, ідентифікувати, категоризувати та витягувати розпізнану інформацію із звичайних текстів оголошень у сфері нерухомості.

Перелік посилань

1. Структуровані дані. [Електронний ресурс]. – Режим доступу: https://www.altexsoft.com/blog/structured-unstructured-data/
2. Неструктуровані дані. [Електронний ресурс]. – Режим доступу: https://www.researchgate.net/publication/277615592_Unstructured_Data_Analysis-A_Survey
3. Різновид оголошень. [Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Classified_advertising
4. Natural Language Processing (NLP). [Електронний ресурс]. – Режим доступу: https://www.researchgate.net/publication/364302290_A_Brief_Survey_on_Natural_Language_Processing_Based_Text_Generation_and_Evaluation_Techniques
5. Basra Jehangir, Saravanan Radhakrishnan, Rahul Agarwal, "A survey on Named Entity Recognition": [Електронний ресурс]. – Режим доступу: https://www.sciencedirect.com/science/article/pii/S2949719123000146
6. Embed Encode Attend Predict. [Електронний ресурс]. – Режим доступу: https://github.com/explosion/talks/blob/master/2018-04-12_Embed-Encode-Attend-Predict.pdf
7. Spacy a python library for NLP tasks. [Електронний ресурс]. – Режим доступу: https://spacy.io/api/entityrecognizer

DOI 10.31891/2307-5732-2021-301-5-32-35
УДК 004.055:331.103.22

RADIUK PAVLO M.
Khmelnytskyi National University
ORCID ID: 0000-0003-3609-112X
e-mail: radiukpavlo@gmail.com
SKRYPNYK TETIANA K.
Khmelnytskyi National University
ORCID ID: 0000-0002-8531-5348
e-mail: marine_1996@ukr.net
KARLECHUK DMYTRO T.
Khmelnytskyi National University
ORCID ID: 0000-0003-2239-2394
e-mail: dimon4yk.karlechuk@gmail.com

APPLYING MENTAL MODELS TO MAKING CONTROLLED CRITICALLY SAFE DECISIONS IN IT PROJECT MANAGEMENT

Over the past decades, numerous researchers and practitioners have focused on the use case modeling for information technology project management. However, few studies have examined ways to improve the discovery phase of an information technology project through modifying use case diagrams. In this work, a mental model approach is suggested to integrate a visualization stage into the discovery phase by creating a conceptual mental model before constructing the use case diagram. It has been established that designing a mental model by visualizing the end user's needs improves the ability of junior analysts to capture, analyze, and determine the user and functional requirements of information systems. Moreover, junior business analysts can extrapolate from the workflow presented in the conceptual mental model to determine the system's actions to benefit the associated actor, which must appear in the use case diagram. The paper describes the design of user precedents through a conceptual mental model. Altogether, the proposed approach's effectiveness in capturing, analyzing, and determining the requirements of information systems is measured in terms of accuracy, completeness, and simplicity.

Keywords: information technology project, project management, discovery phase, use case diagram, mental model, controlled critically safe decisions, business analysis

П.М. РАДЮК, Т.К. СКРИПНИК, Д.Т. КАРЛЕЧУК
Хмельницький національний університет

ЗАСТОСУВАННЯ МЕНТАЛЬНИХ МОДЕЛЕЙ ДО ПРИЙНЯТТЯ КОНТРОЛЬОВАНИХ КРИТИЧНО-БЕЗПЕКОВИХ РІШЕНЬ В УПРАВЛІННІ ІТ ПРОЄКТАМИ

Упродовж останніх десятиліть багато дослідників та практиків зосереджують свою увагу на моделюванні прецедентів для управління інформаційними технологіями. Втім мало досліджень розглядають шляхи вдосконалення фази виявлення проєкту інформаційних технологій за допомогою діаграм прецедентів. У цій роботі пропонується підхід на основі ментальних моделей до фази виявлення вимог, що полягає в інтегруванні етапу візуалізації для побудови користувацьких діаграм. Структурування ментальної моделі з погляду користувачів покращує здатність бізнес-аналітиків-початківців для захоплення, аналізу та визначення функціональних вимог інформаційних систем. Крім того, аналітики можуть екстраполювати з робочого процесу, поданого в концептуальній ментальній моделі, щоб визначити дії, які необхідно виконати системою, щоб скористатися асоційованим актором, і це повинно з'явитися в діаграмі корпусу користувача. У роботі обговорюється сінтеза структурування ментальної моделі у вигляді концептуальної ментальної моделі до визначення функціональних вимог за допомогою користувацьких діаграм. Запропонований підхід на основі ментальних моделей ґрунтується на незалежних та залежних змінних. Незалежною змінною є кількість замовлень від клієнтів; водночас залежними змінними є загальна якість системних функціональних вимог, якість кожного випадку використання в термінах правильності та повноти, якість посилань із погляду правильності та повноти, надлишкові зв'язки або використання випадків. Застосування концептуальних ментальних моделей може викликати певні недоліки. Наприклад, некоректне впровадження ментальної моделі під час фази виявлення може призвести до проблеми валідування зібраних функціональних вимог. Загалом, ця проблема поділяється на чотири групи: висновок, внутрішня перевірка, дизайн та зовнішня перевірка.

Ключові слова: проєкт інформаційних технологій, управління проєктами, фаза виявлення, діаграма прецедентів, ментальна модель, контрольоване критико-безпекове рішення, бізнес-аналіз.

Problem statement

Within most information technology (IT) projects, discovering requirements for developing information systems (IS) often rely on controlled critically safe decisions made by different stakeholders. As so, practitioners utilize the Use Case Diagrams (UCDs) [1] to define functional system requirements to perform the robust discovery phase. A sample of a UCD is presented in fig. 1.

The UCD has been considered a valuable and easy-to-use tool for representing and specifying various project requirements within the business analysis (BA) community. In general, UCD comprises two components: 1) a graphical diagram to model and represent customer requirements, and 2) use case descriptions to elaborate and specify these requirements in detail.

The presented study aims to facilitate the discovery phase performed by junior BA practitioners. We propose a novel approach based on mental models to improve UC diagrams while identifying, capturing, and specifying functional system requirements.

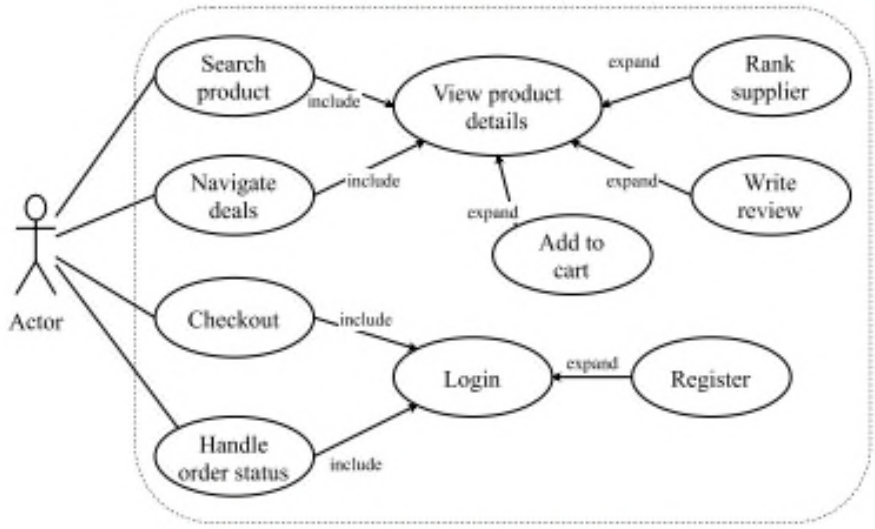


Fig. 1. A use case diagram for an IT project

Analysis of recent research

Over the last decades, numerous studies have aimed to improve the quality of UC model outcomes through comparison-based experiments. For instance, in [2], the authors enhanced UC descriptions by proposing advanced guidelines to write good use cases, while [3] were focused on improving UC descriptions inspection checklists with screen mockups. Another work [4] showed that using predefined templates allows designing a UCD easier to comprehend and read comparing to guidelines without specific details.

Throughout its history, the IT domain has challenged the issue of interpreting and explaining computational results. And to tackle this issue, mental models might be a feasible solution. The term "mental model" originates from the fields of cognitive psychology and human-computer interaction (HCI) [5, 6]. In [7], authors argued that "a mental model represents a person's thought process for how something works, i.e., a person's understanding of the surrounding world." For practical use, mental models require some physical representation, commonly verbal or spatial [8, 9]. For example, an IS can be viewed as a conceptual mental model (CMM) – a diagram expressing the structure of an IS from the user's perspective. The model, in this case, is created in a controlled process according to the preset guidelines.

Another recent study [9] discussed the cognitive significance of visualization for software engineers for whom an IT project perceives as a combination of various tasks; visualization allows covering all business requirements as a whole. Thus, a CMM is a graphical representation of the user's mental activity and behavior within an IS. An example of a CMM is illustrated in fig. 2.

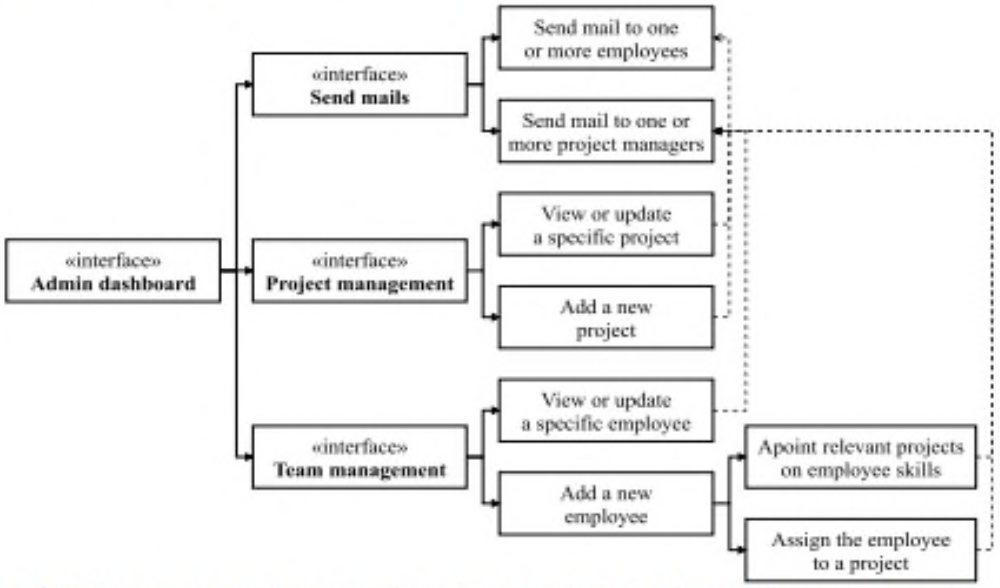


Fig. 2. Definition of user's work processes through a conceptual mental model. Bold rectangles in the scheme depict the user interface, while ordinary rectangles define a screen or a function. User actions in the information system are specified by arrows

Based on the mental model approach, it is proposed to apply the user demands' visualization at the beginning of the discovery phase, which implies formalizing a CMM before designing a UCD. It should be noted that both a CMM and UCD refer to similar components of an IS, such as actors, relations, functionality, etc. Formalizing a mental model in terms of user needs can improve the ability of junior professionals to collect, analyze, and define system requirements for information systems. In particular, junior business analysts can extrapolate from the IS workflow presented in the CMM to determine system's actions to benefit the associated actor, which must appear in the further UCD. Finally, the visualization of user demands through a CMM provides improvement in the accuracy, completeness and simplicity of the system requirements definition by inexperienced practitioners.

Conceptual mental model

According to the literature review and practical observations, those junior professionals responsible for collecting system requirements for the product who "imagine" the future IS's functionality and its workflow in terms of user demands during the discovery phase perform analysis and identification of actors and relationships between them more effectively and better model these relationships under appropriate elements of UCD. Motivated by this insight, an integrative approach is suggested to benefit the discovery phase while providing UC modeling. The proposed approach is aimed to enhance the discovery phase by structuring the "imaginary" process to a formal stage of visualization, which must be performed before creating a UC diagram. Therefore, team members within an IT project must formalize a CMM based on external data collected from key stakeholders before the analysis and definition of IS for a UCD. Fig. 3 depicts a scheme of the proposed approach implemented to an IT project.

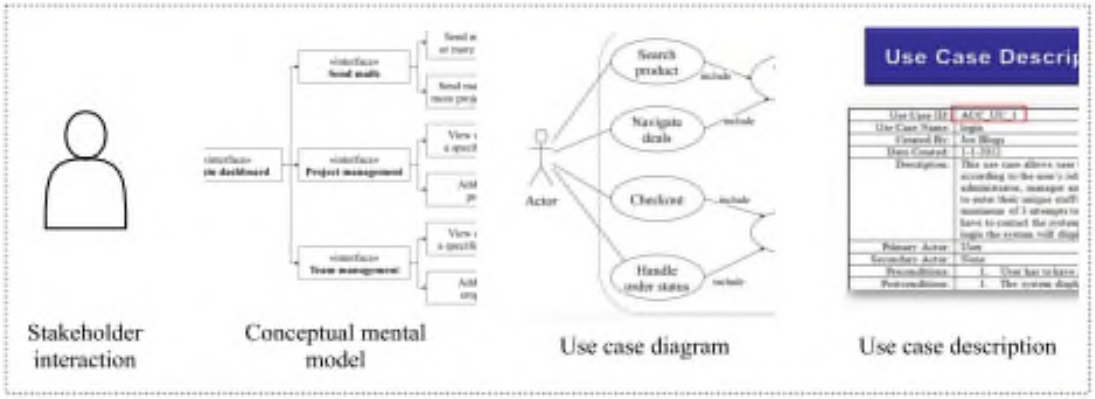


Fig. 3. A scheme of a discovery stage with a conceptual mental model before creating a use case diagram.

Our mental model approach also suggests quality criteria to evaluate the possible outcome of an examined discovery phase. Based on a quality evaluation framework [11], three quality indicators were chosen: accuracy, completeness, and simplicity. The accuracy indicator determines whether the model corresponds to the data modeling methodology and is measured by the number of errors in the created model. Completeness is determined by the number of user system requirements that are not present in the model. Simplicity determines the number of subjects and relationships in the model, i.e., shows how simplified the model is.

The independent and dependent variables of the proposed approach are summarized below:

1. Independent variables: request to perform a task within an IT project.
2. Dependent variables:

a) the estimated quality of the defined user and functional requirements, scored from 0 to 60;

b) the estimated quality of each use case by accuracy and completeness: 0 – absent, 2 – incorrect, 4 – correct;

c) the estimated quality of the defined connections between actors and use cases by accuracy and completeness: 1 point is deducted from the final score if the connection is redundant, 2 points are subtracted from the final score if the use case is redundant.

This section discusses the possible disadvantages of the proposed mental model approach. Incorrect results of the cognitive model's implementation to the discovery phase may manifest due to the validation of the collected requirements. This problem can be divided into four components: conclusion, internal validation, design, and external validation. Conclusion concerns the credibility of the inferences reached within the discovery phase. Internal validation refers to factors that might affect the dependent variables. Construction refers to how mental models are reflected in actual measures. Finally, external validation concerns the generalization of results, both to the study of the population and to other settings.

Conclusion

The present study proposes a mental model approach which implies including a visualization stage before the discovery phase for project assessment in the information technology domain. This approach aims to increase the efficiency of junior business analysts and project managers in collecting, analyzing, and defining custom and functional requirements for information system development. Applying mental models to the discovery stage will enhance the

construction of user case diagrams by accuracy, completeness, and simplicity. Furthermore, future research needs to examine more closely how to transform a set of conceptual mental model elements into a set of use case diagram constructs.

References

1. Faitelson D. UML diagram refinement (focusing on class- and use case diagrams) / D. Faitelson, S. Tyszberowicz // 2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE) : Proceedings, Buenos Aires, 20–28 May 2017. – 2017. – P. 735–745. – DOI : https://doi.org/10.1109/ICSE.2017.73

2. Tiwari S. UCAnalyzer: A tool to analyze use case textual descriptions / S. Tiwari, M. Laddha // 2017 IEEE 25th International Requirements Engineering Conference (RE) : Proceedings, Lisbon, 4–8 September 2017. – 2017. – P. 448–449. – DOI : https://doi.org/10.1109/RE.2017.39

3. Reggio G. DUSM: A method for requirements specification and refinement based on disciplined use cases and screen mockups / G. Reggio, M. Leotta, R. Filippo, D. Clerissi // Journal of Computer Science and Technology. – 2018. – № 33 (5). – P. 918–939. – DOI : https://doi.org/10.1007/s11390-018-1866-8

4. Skjæveland M. G. Practical ontology pattern instantiation, discovery, and maintenance with reasonable ontology templates / M. G. Skjæveland, D. P. Lupp, L. H. Karlsen, H. Forssell // 17th International Semantic Web Conference, Monterey, CA, USA, 8–12 October 2018. – Lecture Notes in Computer Science : Proceedings. – Cham : Springer, 2018. – Vol. 11136, No. 1. – P. 477–494. – DOI : https://doi.org/10.1007/978-3-030-00671-6_28

5. Seinfeld S. User representations in human-computer interaction / S. Seinfeld, T. Feuchtner, A. Maselli, J. Müller // Human–Computer Interaction. – 2020. – № 36 (5-6). – P. 400–438. – DOI :

6. https://doi.org/10.1080/07370024.2020.1724790

7. Oran A. C. A framework for evaluating and improving requirements specifications based on the developers and testers perspective / A. C. Oran, G. Santos, B. Gadelha, T. Conte // Requirements Engineering. – 2021. – № 26 (4). – P. 481–508. – DOI : https://doi.org/10.1007/s00766-021-00352-6

8. Krak Iu. Using visual analytics to develop human and machine - centric models: A review of approaches and proposed information technology / Iu. Krak, O. Barmak, E. Manziuk // Computational Intelligence. – 2020. – P. 1–26. – DOI : https://doi.org/10.1111/coin.12289

9. Barmak O. Web-based information technology for classifying and interpreting early pneumonia based on fine-tuned convolutional neural network / O. Barmak, P. Radiuk // Computer systems and information technologies. – 2021. – № 3 (1). – P. 12–18. – DOI : https://doi.org/10.31891/CSIT-2021-3-2

10. Barmak O. Information technology of visual analysis of X-ray images for interpretation of pneumonia diagnostic results / O. Barmak, P. Radiuk // Herald of Khmelnytskyi National University. Technical sciences. – 2021. – № 295 (2). – P. 52–55. – DOI : https://www.doi.org/10.31891/2307-5732-2021-295-2-52-55

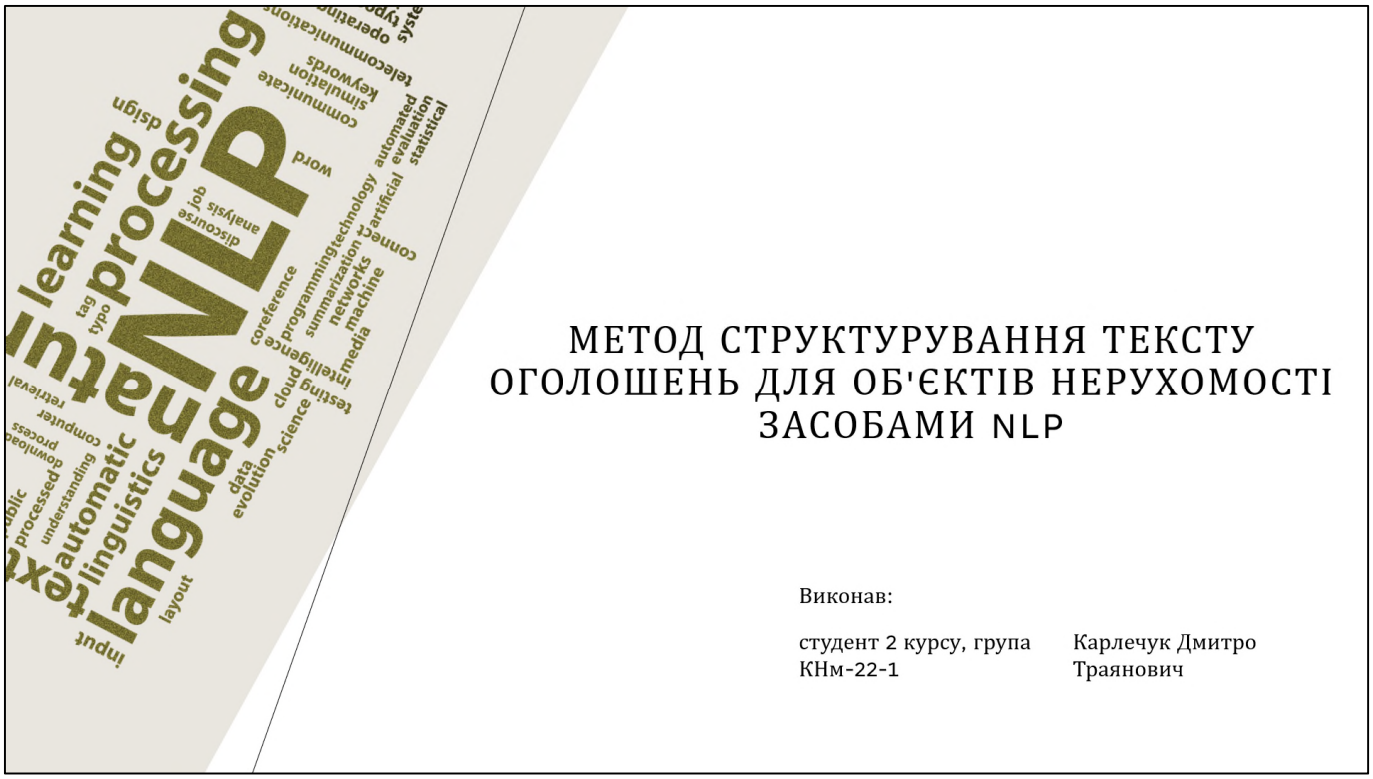
11. Killen C. P. The role of decision makers' use of visualizations in project portfolio decision making / C. P. Killen, J. Geraldi, A. Kock // International Journal of Project Management. – 2020. – № 38 (5). – P. 267–277. – DOI : https://doi.org/10.1016/j.ijproman.2020.04.002

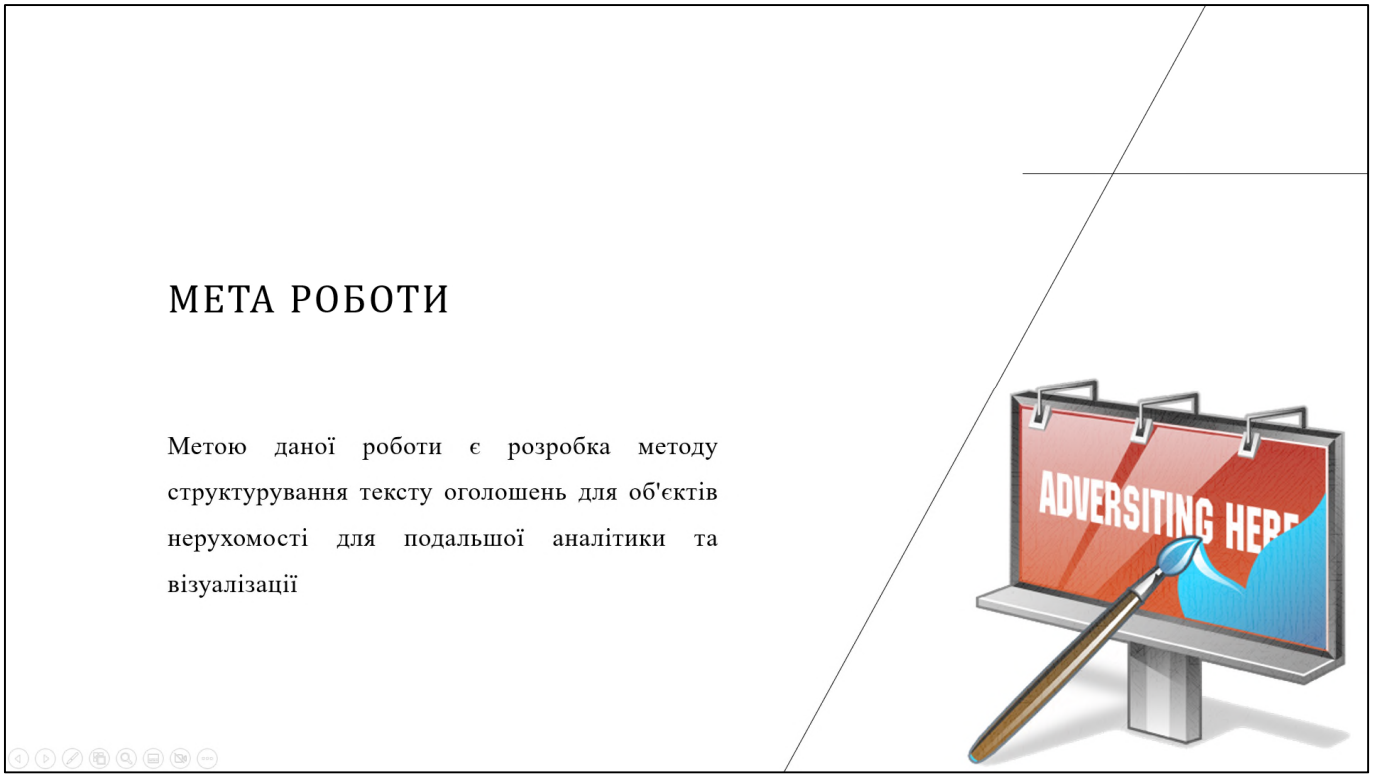
12. Radiuk P. M. A framework for exploring and modelling neural architecture search methods. / P. M. Radiuk, N. V. Hrypynska // The 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS-2020), Lviv, Ukraine, 23–24 April 2020. – CEUR-WS, 2020. – Vol. 2604. – P. 1060–1074.

Рецензія/Peer review : 02.09.2021 р. Надрукована/Printed :10.10.2021 р.

Додаток Д

Презентаційний матеріал





ЗАВДАННЯ

1 Провести аналіз методів обробки природної мови

2 Провести аналіз методу розпізнавання іменованих сутностей

3 Провести аналіз методологій та підходів з ідентифікації та класифікації об’єктів у тексті

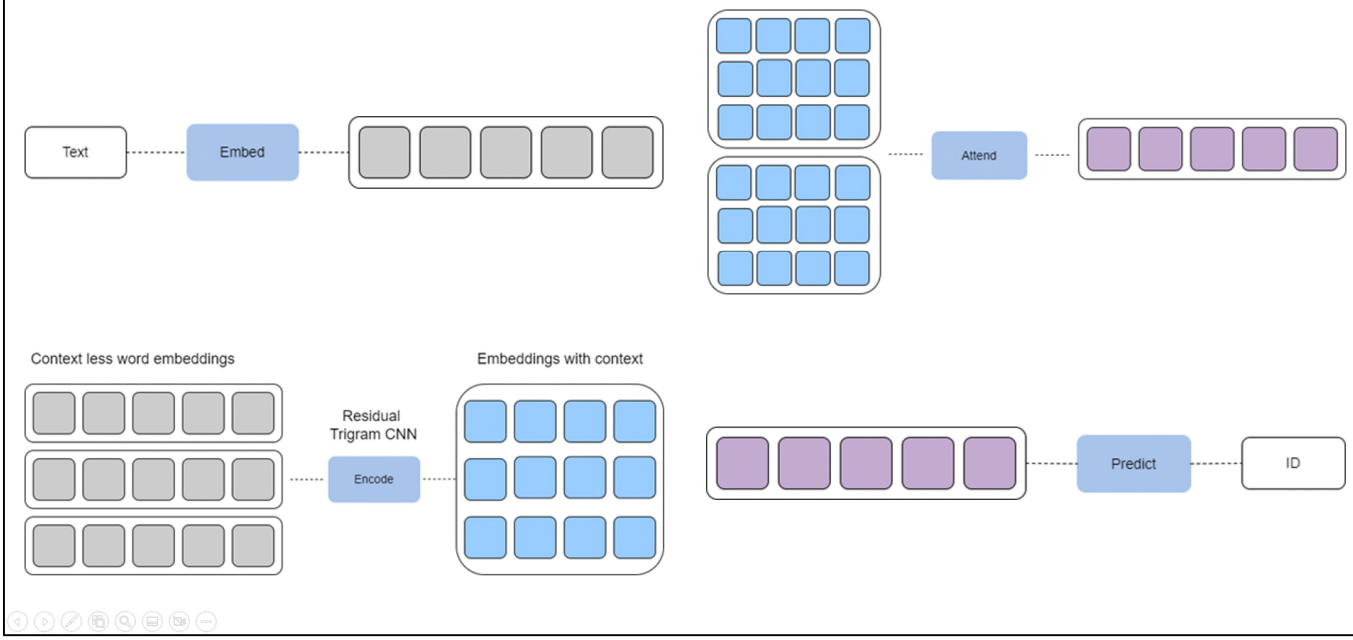
4 Розробити метод структурування тексту оголошень

Об’єкт та предмет дослідження

Об’єкт дослідження – процес визначення ключових слів для структурування текстової інформації

Предмет дослідження – моделі, методи та технології обробки текстових даних для структурування текстової інформації.

АЛГОРИТМ ІДЕНТИФІКАЦІЇ ТА КЛАСИФІКАЦІЇ СУТНОСТЕЙ



РЕЗУЛЬТАТ РОБОТИ МЕТОДУ З РОЗПІЗНАВАННЯ СУТНОСТЕЙ З ТЕКСТІВ ОГОЛОШЕНЬ ПО ОБ'ЄКТАХ НЕРУХОМОСТІ

```
text = """
Explore this stunning property in the heart of downtown, boasting a luxurious for sale property, a spacious and elegant mansion property for an attractiv
"""
doc = nlp(text)

spacy.displacy.render(doc, style="ent", jupyter=True)
```

Explore this stunning property in the heart of downtown, boasting a luxurious for sale PROPERTY STATUS property, a spacious and elegant mansion PROPERTY TYPE property for an attractive price of $1.5 million. PRICE The expansive house size of 5,000 square feet HOUSE SIZE accommodates numerous features and amenities, including a private pool FEATURES/AMENITIES , modern kitchen FEATURES/AMENITIES , and a cozy fireplace. FEATURES/AMENITIES Residents can enjoy the convenience of nearby landmarks like Central Park NEARBY LANDMARKS and breathtaking views of the city skyline. With its impeccable condition CONDITION and ample rooms, this 6-bedroom BED , 4-bath BATH gem on Main Street STREET , nestled in the vibrant city of New York CITY , is ready to be your new dream home. The property offers essential utilities such as central air conditioning UTILITIES and heating UTILITIES , ensuring year-round comfort. Located in the state of New York, this property combines luxury living with urban convenience. If you are interested please contact us at kidehen@optonline.net OWNER EMAIL or call us at (206) 342-8631 OWNER PHONE

ВИСНОВКИ

01	02	03	04	05
Проведено аналіз методів обробки природної мови в структуруванні текстів оголошень для об'єктів нерухомості засобами NLP	Розглянуто та основний підхід по ідентифікації та класифікації сутностей у неструткурованому тексті	Представлено стратегії та методології, які використовують для ідентифікації та категоризації об'єктів у текстах оголошень	На основі попердньої інформації розроблено та реалізовано метод структурування тексту оголошень	Розроблений метод дозволяє розпізнавати, та ідентифікувати, інформацію із звичайних текстів оголошень у сфері нерухомості.

ДЯКУЮ ЗА УВАГУ

Mon Dec 04 18:37:15 EET 2023, Петровський Сергій Степанович, Хмельницький національний університет, ХНУ

Anti-Plagiarism v-15.257

Максимальне співпадіння з одним документом 3.0%

Словники перевірки: en_US, ru_RU, ua_UA. **Помилок в документах: 10%**

	Документ		Сумарний збіг по Базі Даних	
ID: 121729 Назва: КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА на тему Метод структурування тексту оголошень для об'єктів нерухомості засобами NLP Додано в БД: 2023-12-04 Автора: Д.Т. Карлечук Керівники: Р.О. Багрій Консультанти: Опоненти:	Символи	Лексеми	Символи	Лексеми
	90195	1298	3784 (4%)	65 (5%)

Джерело плагіату

ID	Опис	Наявність плагіату в документі	
		Символи	Лексеми

UNICHECK
by Turnitin

Ім'я користувача:
Кафедра КН

ID перевірки:
1015968923

Дата перевірки:
04.12.2023 19:34:14 EET

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
04.12.2023 19:38:11 EET

ID користувача:
100005671

Назва документа: КНм-22-1 Карлечук

Кількість сторінок: 76 **Кількість слів:** 14138 **Кількість символів:** 108101 **Розмір файлу:** 2.74 MB **ID файлу:** 1015647786

Виявлено модифікації тексту (можуть впливати на відсоток схожості)

5.76%
Схожість

Найбільша схожість: 1.58% з джерелом з Бібліотеки (ID файлу: 1013021999)

5.43% Джерела з Інтернету	763	Сторінка 78
3.38% Джерела з Бібліотеки	164	Сторінка 83

0% Цитат

Вилучення цитат вимкнене

Вилучення списку бібліографічних посилань вимкнене

0%
Вилучень

Немає вилучених джерел

Модифікації

Виявлено модифікації тексту. Детальна інформація доступна в онлайн-звіті.

Підозріле форматування	18 сторінок

РІШЕННЯ ЕКСПЕРНОЇ КОМІСІЇ
КАФЕДРИ КОМП'ЮТЕРНИХ НАУК
ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ МАГІСТРА ДО ЗАХИСТУ ЗА РЕЗУЛЬТАТАМИ АНАЛІЗУ ЗВІТУ ПОДІБНОСТІ

Підтверджуємо ознайомлення з результатом звіту подібності щодо роботи, генерованого системою виявлення текстових збігів/ідентичності/схожості:

Назва: *Метод структурування тексту оголошень для об'єктів нерухомості засобами NLP*

Автор: *Карлечук Дмитро Траянович*

Спеціальність: 122 – Комп'ютерні науки

Освітня програма: освітньо-професійна

Науковий керівник: к.т.н., доц. Багрій Р.О.

Після аналізу звіту подібності зроблено такий висновок:

№	Висновок	Позначка про відповідність
1	Запозичення, виявлені в роботі, є законними і не є плагіатом. Робота приймається до захисту.	відповідає
2	Виявлені запозичення не є плагіатом, розміщені в розділах, які не описують безпосередньо авторське дослідження, але кількість цитат перевищує обсяг, виправданий поставленою метою роботи Робота приймається до захисту, але має бути відкоригована. Відкоригований варіант має бути поданий на кафедру за 2 дні до захисту, разом із заявою щодо самостійності виконання письмової роботи та ідентичності друкованої та електронної версії роботи	
3	Виявлені запозичення не є плагіатом, але частково розміщені в розділах, які описують безпосередньо авторське дослідження, а кількість цитат перевищує обсяг, виправданий поставленою метою роботи. В зв'язку з цим мета роботи та поставлені завдання не були досягнені. Робота може бути допущена до захисту (наступного року) після того як буде відкоригована та допрацьована і успішно пройде повторну перевірку на академічний плагіат.	
4	Робота містить навмисні текстові спотворення, передбачувані спроби укриття запозичень або інші прояви академічного плагіату. Робота містить фабрикацію або фальсифікацію даних. Робота не допускається до захисту.	

Підтвердження:

Запозичення, виявлені в роботі, є законними і не є плагіатом, оскільки:

1) за програмою Anti-Plagiarism виявлені 3% є фрагментарними – містять поширені конструкції, загальновідомі терміни, скорочення та визначення.
2) За програмою UNICHECK виявлені 5.76%, що є запозиченнями, які розміщені в розділах аналізу існуючих технологій та прототипів, які не описують безпосередньо авторське дослідження і не стосуються результатів роботи.

Сумарний обсяг всіх запозичень, визначений системою виявлення збігів/ ідентичності/схожості, складає 3% і 5.76% відповідно, що, з урахуванням наведених обґрунтувань, відповідає характеру наукового дослідження і свідчить на користь кваліфікаційної роботи.

Керівник роботи _____ Руслан Багрій

Гарант ОП _____ Руслан Багрій

Завідувач кафедри КН _____ Олександр Бармак

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ
кафедра комп'ютерних наук

ВІДГУК ОПОНЕНТА

на кваліфікаційну роботу магістра

гр. КНм-22-1 Карлечука Дмитра Траяновича за темою: *Метод структурування тексту оголошень для об'єктів нерухомості засобами NLP*

1. **Актуальність обраної теми**

Обрана тема кваліфікаційної роботи магістра є актуальною у сфері комп'ютерних наук, оскільки вона стосується потреби в ефективних та автоматизованих методах структурування текстової інформації щодо рекламних оголошень про нерухомість із використанням обробки природної мови. Актуальність теми обгрунтована проведеними дослідженнями в цьому напрямку та методами прийняття рішень згідно ідентифікації і класифікації об'єктів у текстовому наборі даних, що допомогло вирішити завдання з розпізнавання сутностей.

2. **Відповідність роботи предметній області спеціальності 122 Комп'ютерні науки та загальним вимогам до наукових робіт**

Тематика кваліфікаційної роботи відповідає вимогам предметної області спеціальності 122 Комп'ютерні науки, оскільки пов'язана з теоретичними та практичними задачами в галузі комп'ютерних наук та вимогам до наукових робіт освітньо-кваліфікаційного рівня магістра.

3. **Повнота розкриття мети та завдань дослідження**

Мета й завдання дослідження були докладно описані в магістерській роботі, забезпечивши чітке розуміння цілей та сформульованої проблеми дослідження в рамках обраної теми. У роботі детально описується основна мета й ретельно аналізується кожне завдання, що розкриває суть кваліфікаційної роботи в повному обсязі.

4. **Наявність наукової новизни**

Вдосконалено метод розпізнавання іменованих сутностей для задачі категоризації об'єктів у текстах оголошень. Це досягається за рахунок використання архітектурного підходу EEAP для ідентифікація та класифікації сутностей. Проведене дослідження було опубліковано на XV Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2023» 2023 р., м. Хмельницький, Україна.

5. **Зміст кожного розділу роботи**

Кваліфікаційна робота магістра складається із чотирьох розділів. У першому розділі, автором було проаналізовано предметну область структурування текстів оголошень засобами NLP, здійснено дослідження існуючих наукових робіт та публікацій близьких до тематики

даної роботи, та сформульовано постановку завдань для розробки методу структурування тексту оголошень для об'єктів нерухомості засобами NLP.

У другому розділі проведено детальний аналіз механізму та алгоритму роботи по розпізнаванню іменованих сутностей, де було розглянуто основні принципи керування та обробки текстової інформації необхідних при категоризації та ідентифікації даних, а також запропоновано використання архітектурного підходу EEAP для розпізнавання іменованих сутностей.

У третьому розділі розглянуто архітектуру з компонентами інформаційної моделі та визначено засоби розробки програмного застосунку з розпізнавання іменованих сутностей.

У четвертому розділі проведено дослідження ефективності роботи програмного застосунку із розпізнавання сутностей, представлено результати порівняння параметрів точності розпізнавання, та показано оцінку отриманих результатів.

6. Ступінь розкриття теми роботи

У кваліфікаційній роботі магістра демонструється повний рівень розкриття обраної теми дослідження. Структура роботи побудована логічно та послідовно, кожний розділ детально описано, а поставлені завдання успішно реалізовані, що підтверджується проведеним аналізом ефективності результатів.

7. Якість оформлення кваліфікаційної роботи

Оформлення кваліфікаційної роботи було здійснено відповідно до норм та вимог. Граматичних та мовних помилок не виявлено. Наявні незначні синтаксичні помилки.

8. Недоліки кваліфікаційної роботи

У роботі приділено недостатньо уваги процесу формування датасетів, які повинні включати різноманітні форми оголошень у сфері нерухомості. Це може суттєво впливати на точність та ефективність розпізнавання сутностей.

9. Загальний висновок (допускається чи не допускається до захисту), якої оцінки заслуговує кваліфікаційна робота.

Кваліфікаційна робота магістра виконана в повному обсязі, у якій продемонстровано успішне виконання всіх поставлених завдань – у результаті робота допускається до захисту. Незважаючи на присутність незначних помилок, кваліфікаційна робота заслуговує на оцінку «відмінно» з урахуванням загальної якості виконання та дотримання необхідних вимог.

Опонент _____ д. ф-м. н., професор Бедратюк Л.П.

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ
кафедра комп'ютерних наук

ВІДГУК НАУКОВОГО КЕРІВНИКА

на кваліфікаційну роботу магістра

гр. КНм-22-1 Карлечука Дмитра Траяновича за темою: Метод структурування тексту оголошень для об'єктів нерухомості засобами NLP

1. Актуальність теми

Актуальність теми достатньо обґрунтована, оскільки метод структурування тексту оголошень для об'єктів нерухомості дозволяє автоматично визначити важливі сутності в неструктурованому тексті, категоризувати їх та зберегти у структурованому форматі, що максимізує ефективність аналітичного процесу та поліпшує прийняття управлінських рішень.

2. Відповідність роботи предметній області спеціальності 122 Комп'ютерні науки та загальним вимогам до наукових робіт

Теми кваліфікаційної роботи "Метод структурування тексту оголошень для об'єктів нерухомості засобами NLP" відповідає предметній області спеціальності 122 Комп'ютерні науки та вимогам до кваліфікаційної роботи магістра, оскільки об'єктом дослідження є процес визначення ключових слів для структурування текстової інформації, предметом дослідження – моделі, методи та технології обробки текстових даних для структурування текстової інформації.

3. Професійні та особистісні якості магістранта

Карлечук Д. Т. під час роботи над кваліфікаційною роботою магістра продемонстрував високий рівень знань та умінь за спеціальністю "Комп'ютерні науки".

4. Ступінь самостійності під час виконання кваліфікаційної роботи

Робота виконана самостійно, академічного плагіату не виявлено, стосовно всіх запозичень наведено відповідні посилання на джерела.

5. Наукова новизна та оригінальність запропонованих підходів

Отримані такі результати: вдосконалено метод розпізнавання іменованих сутностей (NER) для задачі категоризації об'єктів у текстах оголошень. Це досягається за рахунок використання архітектурного підходу EEAP для ідентифікація та класифікація сутностей; розроблено метод структурування тексту оголошень для об'єктів нерухомості за допомогою методу розпізнавання іменованих сутностей. Це дозволило визначити послідовності слів, які представляють іменовані об'єкти в тексті оголошення.. Отримані результати оприлюднені на XIII всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2023», 17-17 листопада

результати оприлюднені на XIII всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2023», 17-17 листопада 2023 р., м. Хмельницький, Україна, доповідь на тему «Метод структурування тексту оголошень для об'єктів нерухомості засобами NLP».

6. Ступінь оволодіння методами дослідження

Студент Карлечук Д. Т. продемонстрував високий рівень оволодіння методами дослідження в ході виконання магістерської кваліфікаційної роботи. Його компетентність у використанні аналітичних і дослідницьких інструментів, також як і навички ефективного збору та обробки даних, свідчать про його високий рівень професійної підготовки та здатність до виконання наукових завдань у галузі "Комп'ютерні науки".

7. Повнота та якість розкриття теми роботи

Мета роботи повністю розкрита, отримані результати підтверджують достовірність наукових положень.

8. Логічність, послідовність, аргументованість, літературна грамотність викладу матеріалу

Викладення матеріалу логічне, послідовне та аргументоване. Мова і стиль викладення кваліфікаційної роботи магістра відповідають стандартам, що забезпечує доступність сприймання матеріалу і відповідає вимогам до сучасних наукових робіт.

9. Можливість практичного застосування кваліфікаційної роботи, окремих її частин

Може мати практичне значення для автоматичної ідентифікації та категоризації важливих даних про об'єкти нерухомості, сприяючи оптимізації маркетингу, покращенню взаємодії з користувачами та прийняттю важливих інвестиційних рішень на основі аналізу структурованих даних.

10. Висновок про можливість допуску кваліфікаційної роботи до захисту, на яку оцінку заслуговує робота

Вважаю, що кваліфікаційна робота студента Карлечука Дмитра Траяновича може бути рекомендована до захисту та заслуговує на оцінку "відмінно".

Науковий керівник _____ к.т.н., доц. Руслан Багрій