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 знахолиться на Трипільській*» розбивається на список слів <w_n>, де w – означає слово з речення, а n – порядковий номер слова.

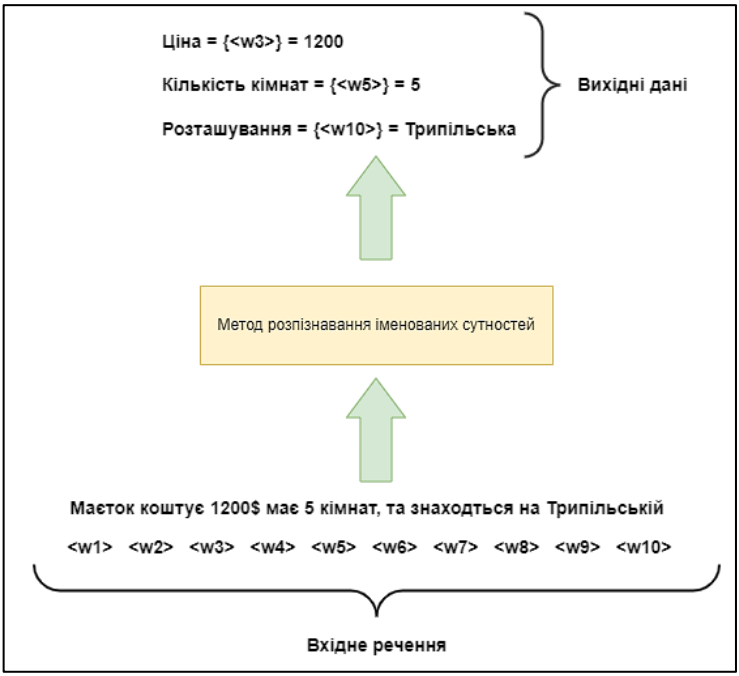

Рисунок 2 – Схема роботи методу розпізнавання сутностей в оголошенні

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 знахолиться 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.

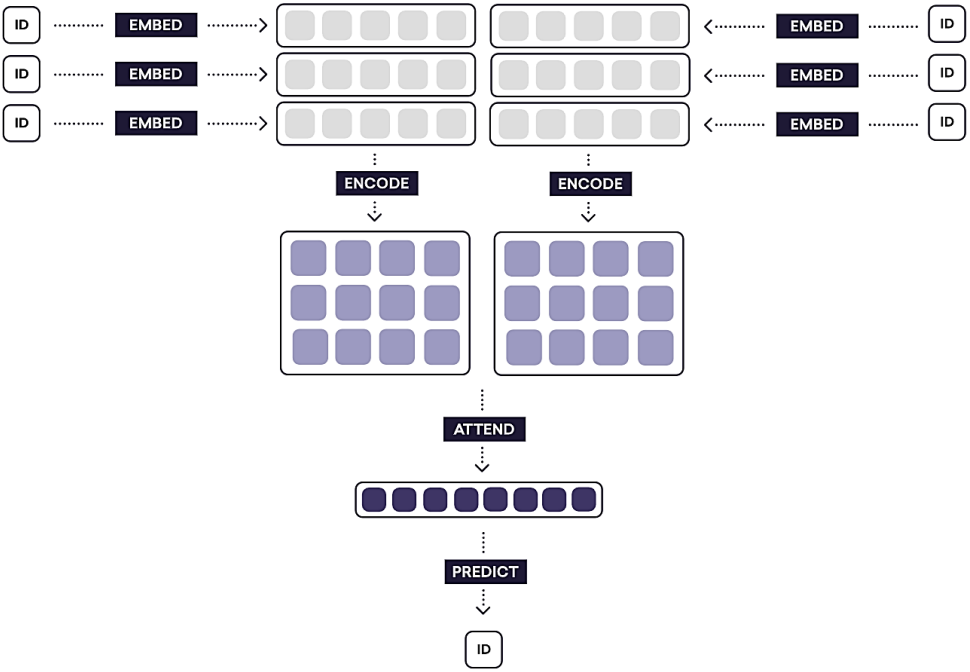

Рисунок 3 – Фреймворк «Embed, Encode, Attend, Predict»

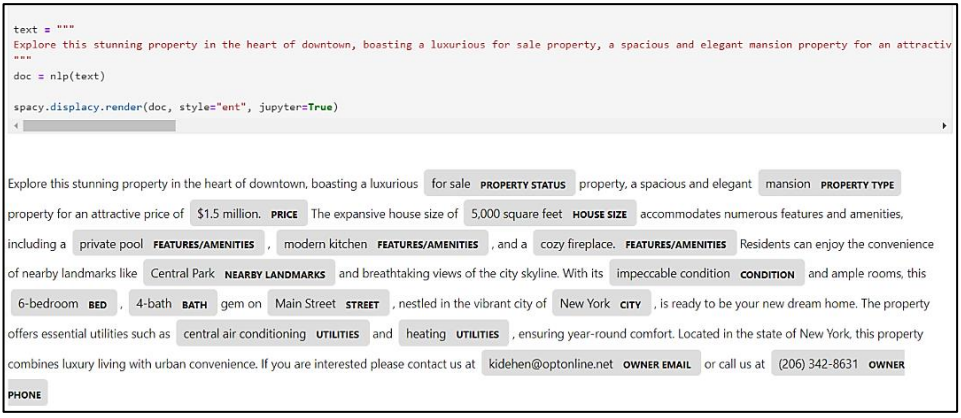

Рисунок 3 – Результат роботи методу з розпізнавання сутностей з текстів оголошень по об'єктах нерухомості

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 оголошенн. [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