УДК 004.8+004.93

В.Б. Мокін

М.А. Гораш

Є.М. Крижановський

ІНФОРМАЦІЙНА ІНТЕЛЕКТУАЛЬНА ТЕХНОЛОГІЯ АВТОМАТИЗОВАНОЇ ОБРОБКИ ТЕКСТОВОЇ ПРИРОДНО-МОВНОЇ ІНФОРМАЦІЇ

Вінницький національний технічний університет;

Анотація

Запропоновано інформаційну інтелектуальну технологію автоматизованої обробки текстової природно-мовної інформації для розпізнавання природномовної інформації за допомогою технології розпізнавання іменованих сутностей NER та технологій опрацювання природної мови NLP та з використанням геоінформаційних технологій для формування датасету.

Ключові слова: технологія розпізнавання іменованих сутностей, NER, технологія опрацювання природної мови, NLP, машинне навчання, штучний інтелект, ГІС.

Abstract

The information intelligent technology of automated processing of textual natural language information for recognition of natural language information by means of technology of recognition of the named essences of NER and technologies of processing of natural language NLP and with use of geoinformation technologies for formation of a dataset is offered.

Keywords: named entity recognition technology, NER, natural language processing technology, NLP, machine learning, artificial intelligence, GIS.

Вступ

З кожним днем все більше і більше формується інформації про світ, яку потрібно опрацювати, формалізувати та структурувати для пошуку у ній. Зі збільшенням її кількості все більшою стає проблема з пошуком методів і технологій, які дозволять її швидко і якісно класифікувати в певні категорії за відповідними ознаками. На даний момент вже є досить багато інформаційних технологій, якій дозволяють розпізнавати природномовну інформацію [1], але є проблеми з формуванням датасету та налаштуванням параметрів моделей.

Результати дослідження

Побудова інформаційної технології буде проходити в такі етапи [2]:

1. Першим етапом з використанням геоінформаційних технологій, різних кадастрів, баз даних водокористувачів та ін. формується вибірка слів, які характеризують кожен масив вод (усі географічні назви, назви підприємств-водокористувачів, назви річок та ін.)

2.Другим етапом є формування навчальної і тестової вибірок.

3. Розв’язуємо задачу розпізнавання природної мови. Для реалізації поставленої задачі скористаємося бібліотекою spaCy [3].

4. Проводимо навчання моделі за допомогою spaCy на навчальній вибірці. Для даного етапу потрібно буде конвертувати навчальну вибірку в формат json. Це конвертування є необхідним на вимогу технології spaCy.

5. Тестуємо розроблену технологія на тестовій вибірці, яку, як і навчальну, теж було конвертовано в json.

Було реалізовано пілотну версію цієї технології. Робота велась з англомовними назвами (name entity).. Вдалось розпізнати текст і конкретно кожного слова як сутності(рис. 1).

```
[('There', 'PROPN','O'),
('are', 'VERB','O'),
('6594', 'NUM', 'O'),
('rivers', 'PROPN', 'GEO'),
('flowing', 'VERB', 'O'),
('in', 'PART', 'O'),
('Southern', 'PROPN', 'GEO'),
('Bug', 'PROPN','GEO'),
('basin', 'PROPN','O')]
```

Рисунок 1 - Результат обробки інформації.

Висновки

Розроблено інформаційну інтелектуальну технологію автоматизованої обробки текстової природно-мовної інформації за допомогою технологій розпізнавання природньої мови NLP та технології NER, яка розпізнає іменовані сутності та дозволяє пришвидшувати пошук даних в масивах даних, а також географічно прив'язувати дані до відповідних об'єктів на карті.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Interoperable cross-domain semantic and geospatial framework for automatic change detection / Chiao-Ling Kuo, Jung-Hong Hong // Journal Computers & Geosciences. — Volume 86 Issue C, January 2016. – Pages 109-119 , DOI 10.1016/j.cageo.2015.10.011.

2. В.Б. Мокін, М.А. Гораш, Д. Пасічнюк, О. Радецький. Концепція інтелектуальної NLP технології для геоприв'язки та класифікації відкритої текстової інформації про масиви вод // Матеріали XV міжнародної конференції" Контроль і управління в складних системах (КУСС-2020)", м. Вінниця, 8-10 жовтня 2020 р. - Вінниця, 2020. [Електронний ресурс]. Режим доступу: http://ir.lib.vntu.edu.ua/bitstream/handle/123456789/30607/KUSS%202020%20MHPR%20-%20NLP.pdf?sequence=1

3. Industrial-Strength Natural Language Processing [Електронний ресурс] – Режим доступу до ресурсу: https://spacy.io/.

Мокін Віталій Борисович — доктор технічних наук, професор кафедри системного аналізу та інформаційних технологій, комп'ютерного моніторингу та інженерної графіки, Вінницький національний технічний університет, м. Вінниця, vbmokin@gmail.com.

Гораш Микола Анатолійович — аспірант кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, м. Вінниця, kolia28011994@gmail.com.

Крижановський Євгеній Миколайович — кандидат технічних наук, доцент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, м. Вінниця, kruzhan@gmail.com.

Horash M. A. - postgraduate student of the Department of Systems Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia.

Kryzhanovsky Y.M. - Candidate of Technical Sciences, Associate Professor of the Department of Systems Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia.

Mokin V.B. - Doctor of Technical Sciences, Professor, Department of Systems Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia.