УДК 004.021

**О.В. Бісікало**
**І.В. Богач**
**П.О. Петрук**

# Побудова термінологічного словника української мови

Вінницький національний технічний університет

***Анотація***

*У даній науково-дослідницькій роботі досліджено методи відокремлення термінологічних конструкцій з метою побудови термінологічного словника української мови певної галузі. У роботі проведено аналіз предметної області дослідження, розглянуто відомі методи та засоби виокремлення термінів, наведено алгоритм одного з існуючих методів та розроблено власний алгоритм для досягання поставленої мети. На його основі було розроблено програмне забезпечення та отримано результати для предметної області машинне навчання*.

**Ключові слова:** NLP, термін, словник.

***Abstract***

*In this research work the methods of separation of terminological constructions for the purpose of construction of the terminological dictionary of the Ukrainian language of a certain branch are investigated. The analysis of the subject area of research is carried out in the work, the known methods and means of separation of terms are considered, the algorithm of one of the existing methods is resulted and the own algorithm for achievement of the set purpose is developed. Based on it, software was developed and the results for the subject area of machine learning were obtained..*

**Keywords**: NLP, term, dictionary.

## Вступ

У сучасний час стрімкого розвитку інформаційних технологій обсяг корисної інформації у світі щосекунди зростає. Будуються нові дата-центри для зберігання великої кількості даних у хмарних сховищах. Але, на жаль, людський мозок має лімітований об'єм пам'яті і його не можна масштабувати як хмарні сервери. Тому кількість інформації, яку ми отримаємо з навколишнього світу, потрібно ретельно фільтрувати для виявлення найважливішої.

Безперечно, найкращим джерелом пізнання світу людство вважає книги і споконвіків передає найважливіші знання саме через них. У наші дні роль паперової книги стала не настільки важливою [1]. Але, попри це, в освітньому процесі тексти залишаються фундаментальним джерелом знань, незалежно від форми подання та предметної області навчання.

Відомо, що досить багато статей та відео на ту чи іншу тематику будуються на репрезентації знань, отриманих з книг, та, в більшій мірі, пропускають деталі, які описані в книгах. У кожній книзі (підручнику, монографії) по професійній тематиці знайдеться значна кількість термінів, які не завжди вдається правильно інтуїтивно інтерпретувати. А шукати їх кожного разу при зустрічі в тексті не досить зручно, адже шукаючи один термін можна рекурсивно переключитися на інший, який лежить в основі попереднього, зрештою на це буде згаяно значна кількість часу. Існує ряд програм, які дозволяють виділяти ключові терміни, однак вони не є досконалими та універсальними.

Постановка задачі дослідження полягає в створені алгоритмічного та програмного забезпечення для побудови термінологічного словника української мови з метою виявлення ключових термінів у будь-якому тексті та інтерпретації їх належності до певної предметної галузі.

## Результати дослідження

Для реалізації поставленої задачі пропонується використовувати мова програмування Python. З цією метою спочатку було розроблено клас, у якому реалізований парсер, що вилучає усю текстову інформацію з файла формату html. Наступним етапом є фільтрація текстів. Для цього було розроблено відповідний клас, який видаляє слова, що не належать до української мови, за допомогою регулярних виразів. Також здійснюється видалення стоп-слів, які не несуть вагомої змістовної інформації, а тільки засмічують текст з огляду на пошук термінів. Після цього отримані тексти передаються в модель PYATE, у якій реалізовано клас, що надає змогу вилучати терміни [2]. Слід зауважити, що на

цьому етапі виникли певні трудності, адже бібліотека, яка мала обробляти ці тексти, за умовчанням не підтримувала української мови [3]. Але, з іншого боку, існує змога додавати модель будь якої мови, попередньо перевизначивши відповідний клас, що і було зроблено.

Після проходження всіх текстів через дану модель було зібрано разом всі терміни для загальної колекції текстів. Для виявлення більш реалістичних було видалено однакові терміни та ті, які мають спільні слова з іншими. При виконанні такої умови видаляються ті терміни, у яких вага менша. Таким чином було відфільтровано значну частину термінологічних кандидатів. Слід зазначити, що даний алгоритм оперує тільки біграмами. При використані його для n-грам вищого порядку даний алгоритм працює некоректно, що є обмеженням запропонованого підходу. Всі терміни, які залишилися, зберігаються у файл формату csv.

Дане програмне забезпечення було апробовано на предметній області «машинне навчання». На основі порівнянні отриманих ваг даних термінологічних кандидатів з відповідними експертними оцінками виявлено, що чисельний показник адекватності запропонованого алгоритму сягає 75.163%, що є достатнім для більшої кількості задач.

## Висновки

У результаті проведеного дослідження було проаналізовано існуючі методи та інструментальні засоби побудови термінологічних словників, визначено основні поняття, які визначають дану предметну область. На основі аналізу відомих підходів було розроблено алгоритм для виокремлення термінологічних конструкцій з заданих однотематичних текстів. Було проаналізовано спосіб очистки текстів з метою вилучення невалідної та нерепрезентативної інформації. Особливістю алгоритму є отримання вихідної інформації для побудови термінологічного словника шляхом парсингу пов'язаних україномовних сторінок Вікіпедії. Недоліком даного алгоритму є те, що термінологічні кандидати можуть бути лише біграмами. На основі запропонованого алгоритму було розроблено та апробовано на предметній області «машинне навчання» відповідне програмне забезпечення на мові програмування Python.

## СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ


1. Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit / Steven Bird, Ewan Klein, and Edward Loper. – Режим доступу: https://www.nltk.org/book.
2. GitHub : веб-сайт. URL: https://github.com/kevinlu1248/pyate.
3. LINDAT CLARIAH-CZ : веб-сайт. URL: https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3131.



***Петрук Петро Олександрович*** — студент групи 1ІСТ-18б, факультет комп'ютерних систем та автоматики, Вінницький національний технічний університет, Вінниця, e-mail: petr.square@gmail.com

***Богач Ілона Віталіївна*** — канд. техн. наук, доцент кафедри автоматики та інтелектуально-інформаційних технологій, Вінницький національний технічний університет, Вінниця, e-mail: ilona.bogach@gmail.com

***Бісікало Олег Володимирович*** — д-р техн. наук, декан факультету КСА, Вінницький національний технічний університет, м. Вінниця e-mail: obisikalo@vntu.edu.ua

Науковий керівник: ***Бісікало Олег Володимирович*** — д-р техн. наук, декан факультету КСА, Вінницький національний технічний університет

***Petruk Petro O.*** — Department of Computer System and Automatics, Vinnytsia National Technical University, Vinnytsia, email : petr.square@gmail.com

***Bogach Ilona V.*** — Cand. Sc. (Eng), Assistant Professor of Computer System and Automatics, Vinnytsia National Technical University, Vinnytsia

***Bisikalo Oleg V.*** — Dr.Sc. (Eng.), Professor, Dean of the Faculty for Computer Systems and Automatic, Vinnytsia National Technical University, Vinnytsia

Supervisor: ***Bisikalo Oleg V.*** — Professor, Dean of the Faculty for Computer Systems and Automatic, Vinnytsia National Technical University, Vinnytsia.