УДК 004.5 + 004.89

М.П. Урлапова

Д.П.Урлапова

О.В. Бісікало

АНАЛІЗ МОВНИХ ОБРАЗІВ ТЕКСТУ БІБЛІЇ ЗАСОБАМИ СЕМАНТИЧНОГО ТА ГРАФОВОГО АНАЛІЗУ

Вінницький національний технічний університет

Анотація

У роботі розглянуто підхід до аналізу мовних образів тексту Біблії засобами семантичного аналізу, text mining та графового моделювання. Актуальність дослідження зумовлена потребою формалізації смислових зв'язків у великих природномовних текстах та можливістю застосування методів обробки тексту і візуалізації для виявлення ключових сутностей та асоціативних відношень між ними. У межах роботи реалізовано програмний засіб мовою Python, який виконує завантаження біблійного тексту з файлу формату DOCX, токенізацію, концептуальну класифікацію мовних образів, побудову асоціативної мережі та формування графічних результатів аналізу. Отримані результати демонструють можливість виокремлення ключових образів, оцінювання їх смислової ваги та візуального подання зв'язків між ними.

Ключові слова: *мовні образи; текст Біблії; семантичний аналіз; text mining; асоціативна мережа; обробка природної мови; граф знань.*

Abstract

The paper considers an approach to the analysis of linguistic images in the text of the Bible using semantic analysis, text mining and graph modeling. The relevance of the study is determined by the need to formalize semantic relations in large natural language texts and by the possibility of applying text processing and visualization methods to identify key entities and associative relations between them. A Python-based software tool was developed to load biblical text from a DOCX file, tokenize it, classify linguistic images by concepts, build an associative network and generate graphical analysis results. The obtained results demonstrate the possibility of identifying key images, estimating their semantic weight and visually representing the relations between them.

Keywords: *linguistic images; Bible text; semantic analysis; text mining; associative network; natural language processing; knowledge graph.*

Вступ

Сучасні методи обробки природної мови дають змогу аналізувати не лише частотні характеристики тексту, а й виявляти смислові зв'язки між поняттями, образами та контекстами. Особливий інтерес становлять великі структуровані тексти, для яких важливо не тільки визначити ключові слова, а й побудувати модель смислових відношень між ними. Одним із таких текстів є Біблія, яка водночас є значним культурним, релігійним і мовним корпусом, придатним для комп'ютерного аналізу [1]. У computational linguistics Біблія вже розглядається як цінний корпус для мовних і семантичних досліджень, зокрема для побудови багатомовних текстових ресурсів та аналізу структури тексту [2].

Обробка текстів із урахуванням семантики залишається актуальною науковою задачею, оскільки text mining повинен враховувати не лише частотні ознаки, а й змістові зв'язки між мовними одиницями [3, 5]. Окремим напрямом є graph-based NLP, у межах якого текст подається у вигляді мережі сутностей і відношень, що дозволяє застосовувати графові алгоритми для виявлення центральних понять та побудови знань [6]. Такі підходи близькі до knowledge graph, які широко використовуються для представлення семантичних відношень між сутностями тексту [7]. Для формалізації категорій мовних образів у роботі також використано підхід до формальних методів образного аналізу та синтезу природно-мовних конструкцій [4].

Актуальність дослідження полягає в тому, що формалізація мовних образів у біблійному тексті дозволяє поєднати класичний текстовий аналіз із графовим представленням знань. Це дає змогу перейти від простого перегляду тексту до побудови асоціативної моделі, у якій можна виявляти центральні образи, відношення між ними та їх відносну смислову вагу.

Метою роботи є розроблення програмного засобу для аналізу мовних образів тексту Біблії засобами семантичного та графового аналізу, а також візуалізація отриманих зв'язків у вигляді графа образів.

Результати дослідження

У ході дослідження було розроблено програмний засіб мовою Python для аналізу мовних образів біблійного тексту. Вхідним джерелом даних обрано файл nevidomyy-avtor-bibliia10806.docx, сформований на основі електронного тексту Біблії українською мовою, завантаженого з ресурсу UkrLib [1]. Програмна реалізація забезпечує повний цикл обробки тексту та побудови результатів аналізу.

1. Реалізовано завантаження тексту Біблії з файлу формату DOCX та автоматичне вилучення текстового вмісту документа.
2. Виконано токенізацію тексту та нормалізацію словоформ для подальшого аналізу.
3. Сформовано словники концептуальних категорій мовних образів, зокрема для об'єктів, понять, дій, якостей, місця та часу, з урахуванням підходу, описаного у [4].
4. Реалізовано класифікацію токенів за цими категоріями та обчислення частоти входження мовних образів у корпусі.
5. Побудовано асоціативну мережу образів у вигляді графа, де вузли відповідають ключовим образам, а ребра відображають синтагматичні, синонімічні, антонімічні та відношення типу частина-ціле.
6. Виконано ранжування образів за показником смислової ваги та визначено центральні вузли графа на основі алгоритму PageRank.
7. Сформовано графічні результати аналізу, які подано на рис. 1 та рис. 2.

У результаті виконання програми було отримано кількісні характеристики корпусу та виявлено домінантні мовні образи, серед яких найбільш вагомими виявилися образи Господь, Бог, Слово, Син, Ісус, а також предикативні одиниці Сказав і Прийшов. На рис. 1 подано статистичні результати аналізу мовних образів тексту Біблії, зокрема піраміду сенсу та розподіл концептуальних категорій. На рис. 2 наведено граф образів Біблії, який відображає центральні смислові зв'язки між ключовими релігійними, антропологічними та предикативними одиницями тексту. Отримані результати підтверджують, що поєднання методів text mining, формального образного аналізу [4] і graph-based NLP є придатним для формального подання образної структури тексту та її візуального аналізу.

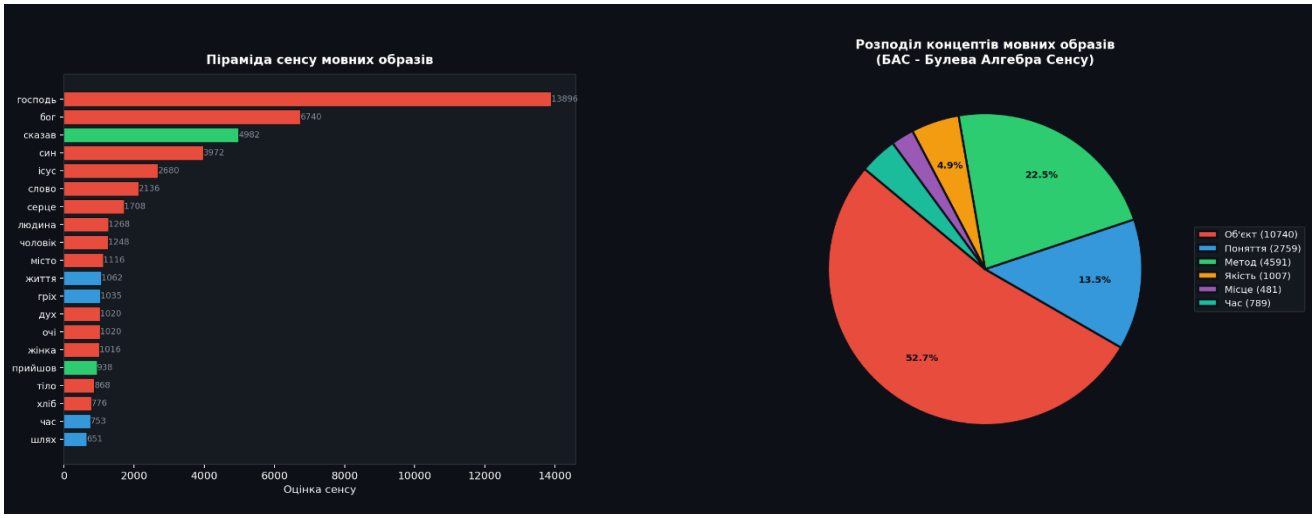

Рис. 1 - Статистичні результати аналізу мовних образів тексту Біблії: піраміда сенсу та розподіл концептуальних категорій.

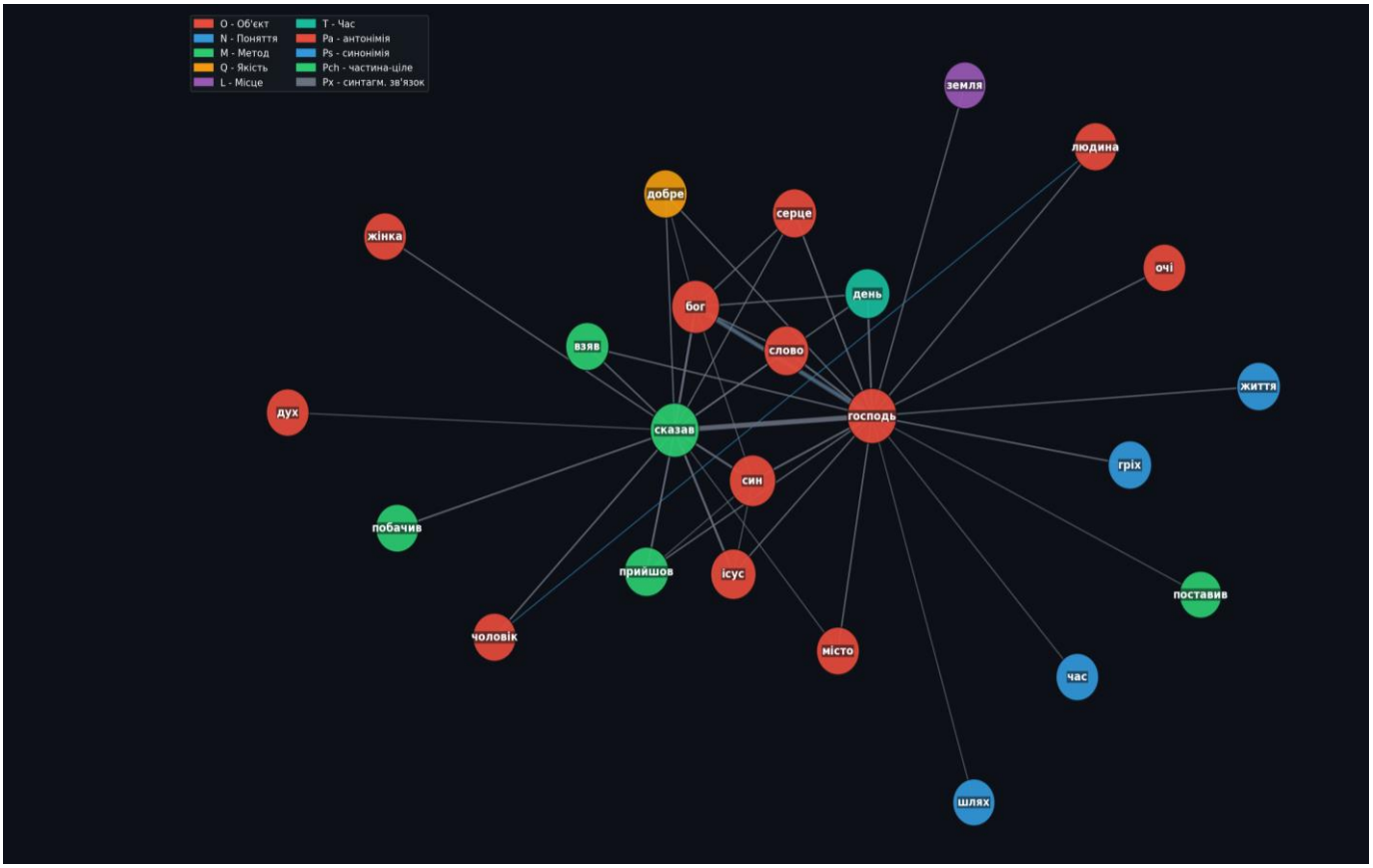

Рис. 2 - Граф образів Біблії, побудований за результатами семантичного та графового аналізу.

Висновки

У роботі реалізовано програмний засіб для аналізу мовних образів тексту Біблії засобами семантичного та графового аналізу. Запропонований підхід поєднує токенізацію, концептуальну класифікацію образів, частотний аналіз та побудову асоціативного графа, що дозволяє формалізувати смислові зв'язки в природномовному тексті.

Практичним результатом є програмна система, яка працює з текстом Біблії у форматі DOCX, автоматично формує статистичні та графові характеристики і створює наочні візуалізації. Отримані результати підтверджують доцільність використання методів text mining, semantic analysis та graph-based NLP для дослідження складних текстових корпусів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Біблія. Переклад Івана Огієнка в редагуванні Івана Модлінського. URL: https://www.ukrlib.com.ua/world/printitzip.php?tid=10806

2. Christodouloupoulos C., Steedman M. A massively parallel corpus: the Bible in 100 languages. Language Resources and Evaluation. 2015. Vol. 49. P. 375-395. URL: https://link.springer.com/article/10.1007/s10579-014-9287-y

3. Cohen K. B., Hunter L. Getting Started in Text Mining. PLoS Computational Biology. 2008. Vol. 4(1). e20. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0040020

4. Бісікало О. В. Формальні методи образного аналізу та синтезу природно-мовних конструкцій : монографія / О. В. Бісікало. - Вінниця : ВНТУ, 2013. - 316 с. - ISBN 978-966-641-528-1.

5. «Sinoara R. A., Antunes J., Rezende S. O. Text mining and semantics: a systematic mapping study. Journal of the Brazilian Computer Society. 2017. Vol. 23, Article 9. URL: https://link.springer.com/article/10.1186/s13173-017-0058-7

6. Mihalcea R., Radev D. Graph-based Natural Language Processing and Information Retrieval. Cambridge University Press, 2011. URL: https://www.cambridge.org/core/books/graph-based-natural-language-processing-and-information-retrieval/216B4D2C3F82BF04C0CC383CD3760C19
7. Schneider P., Schopf T., Vladika J., Galkin M., Simperl E., Matthes F. A Decade of Knowledge Graphs in Natural Language Processing: A Survey. AACL-IJCNLP 2022. P. 601-614. URL: https://aclanthology.org/2022.aacl-main.46/

Бісікало Олег Володимирович – д-р техн. наук, професор, завідувач кафедри АІІТ, Вінницький національний технічний університет, м. Вінниця, e-mail: obisikalo@vntu.edu.ua

Урлапова Марія Павлівна – студентка групи 1іст-23б, факультет автоматизації та інтелектуальних інформаційних технологій, Вінницький національний технічний університет, Вінниця, e-mail: mashaurlapova@gmail.com

Урлапова Дар'я Павлівна – студентка групи 1іст-23б, факультет автоматизації та інтелектуальних інформаційних технологій, Вінницький національний технічний університет, Вінниця, e-mail: dashaurlapova@gmail.com

Bisikalo Oleg V. – Dr.Sc. (Eng.), Professor of the Faculty for Computer Systems and Automatic, Vinnytsia National Technical University, Vinnytsia, e-mail: obisikalo@vntu.edu.ua

Urlapova Maria P. – Department of Automation and intelligent information technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: mashaurlapova@gmail.com

Urlapova Daria P. – Department of Automation and intelligent information technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: dashaurlapova@gmail.com