УДК 004.4
DOI https://doi.org/10.32689/maup.it.2024.1.4

Дмитро БУХАЛЕНКОВ
магістрант, НТУУ «КПІ імені Ігоря Сікорського», 3a43mka@gmail.com
ORCID: *0009-0001-0224-8873*

Тетяна ЗАБОЛОТНЯ
кандидат технічних наук, доцент,
доцент кафедри програмного забезпечення комп'ютерних систем,
НТУУ «КПІ імені Ігоря Сікорського», zabolotnia@pzks.fpm.kpi.ua
ORCID: *0000-0001-8570-7571*

ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ МОДИФІКОВАНОГО МЕТОДУ АВТОМАТИЗОВАНОГО ПОШУКУ КЛЮЧОВИХ СЛІВ У ТЕКСТІ

Анотація. *В умовах невпинного зростання обсягу текстових даних, які доводиться обробляти людині майже в усіх сферах її діяльності, непересічної важливості набуває задача забезпечення швидкого доступу до необхідної інформації. Для вирішення цієї задачі наявні пошукові системи, як правило, проводять індексацію даних: спеціальні боти сканують ресурси і намагаються відшукати пов'язані з ними ключові слова. Від коректності знайдених ключових слів напряму залежить релевантність результатів пошуку, що будуть видані користувачу пошукової системи.*

В даній статті розглянуто модифікований метод автоматизованого пошуку ключових слів у природномовних текстових даних. Він ґрунтується на аналізі складних синтаксичних зв'язків між словами в реченнях тексту та здатний шукати ключові терміни, що складаються з кількох слів.

Метою дослідження *є програмна реалізація та експериментальне дослідження ефективності модифікованого методу автоматизованого пошуку ключових слів у тексті.*

Методика реалізації*. Для випробувань модифікований метод було реалізовано на платформі Python NLTK. У якості тестового масиву даних було обрано два набори текстів: тексти невеликого обсягу (до 400 слів) та тексти більшого обсягу (до 2500 слів). Порівняння проводилися з трьома популярними аналогами, кожен з яких реалізовано на основі різних підходів (машинне навчання, аналіз N-грам, статистичний аналіз). Для кількісного вимірювання ефективності та порівняння з існуючими аналогами запропоновано використовувати метрики абсолютної точності та повноти за Жаккаром.*

Висновки*. Результати випробувань продемонстрували перевагу запропонованого методу над аналогами в точності пошуку ключових слів. Відмічено, що зі збільшенням обсягу текстів абсолютна точність зростає майже в усіх випадках, втім повнота за Жаккаром зменшується. На основі результатів випробувань сформульовано подальші напрямки роботи над покращенням запропонованого методу.*

Ключові слова: *ключові слова, аналіз ефективності, оброблення текстових даних, Python NLTK, стенфордська класифікація.*

Dmytro BUKHALENKOV, Tetiana ZABOLOTNIA. STUDY OF THE EFFECTIVENESS OF THE MODIFIED METHOD OF AUTOMATED SEARCH FOR KEYWORDS IN TEXT

Abstract*. In the conditions of constant growth of the volume of text data, which a person has to process in almost all spheres of his activity, the task of ensuring quick access to the necessary information becomes extremely important. To solve this problem, existing search engines, as a rule, perform data indexing: special bots scan resources and try to find keywords related to them. The relevance of the search results that will be issued to the user of the search engine directly depends on the correctness of the keywords found.*

This article discusses a modified method of automated search for keywords in natural language text data. It is based on the analysis of complex syntactic relationships between words in the sentences of the text and is able to search for key terms consisting of several words.

The research objective *is the programmatic implementation and experimental study of the effectiveness of the modified method of automated search for keywords in text data.*

Methodology of implementation*. For testing, the modified method was implemented on the Python NLTK platform. Two sets of texts were chosen as a test dataset: texts of a small volume (up to 400 words) and texts of a larger volume (up to 2500 words). Comparisons were made with three popular analogues, each of which is implemented on the basis of different approaches (machine learning, N-gram analysis, statistical analysis). For quantitative measurement of efficiency and comparison with existing analogues, it is proposed to use absolute accuracy and completeness metrics according to Jaccard.*

Conclusions*. The results of the tests demonstrated the superiority of the proposed method over analogues in the accuracy of searching for keywords. It was noted that with an increase in the volume of texts, the absolute accuracy increases in almost all cases, but the completeness according to Jaccard decreases. Based on the test results, further directions of work on improving the proposed method are formulated.*

Key words: *keywords, performance analysis, text data processing, Python NLTK, Stanford classification.*

Вступ. Задача пошуку ключових слів виникає у багатьох сферах роботи з текстовими даними. Інформація про ключові слова використовується при інформаційному текстовому пошуку, класифікації,

кластеризації даних тощо. За багато років досліджень спеціалістами було запропоновано методи, різні за точністю, ефективністю та можливістю застосування. Але на сьогоднішній день досі не існує універсального способу визначити перелік ключових термінів для довільного тексту будь-якої тематики. Кожен текст має свою структуру, стиль викладення, стилістичні особливості написання. Тож тривають пошуки нових шляхів вирішення задачі автоматизованого визначення ключових слів в текстових даних, а також спроби підвищити ефективність уже існуючих методів.

Аналіз останніх досліджень і публікацій. Дана стаття присвячена дослідженню ефективності модифікованого методу автоматизованого пошуку ключових слів у природномовних текстових даних [1]. В основу даного методу покладено сучасний гібридний метод пошуку ключових слів в англомовних текстах, що був запропонований українським фахівцем О.В. Яхимовичем [2] в 2021 році. Він застосовує можливості сучасних програмних лінгвістичних пакетів для побудови розмітки тексту і аналізу слів. Головною з особливостей цього методу, окрім фільтрації вербального шуму, можна назвати використання даних залежностей між парами слів та даних про частини мови, що отримуються за допомогою синтаксичного аналізатора. Але цей гібридний метод має суттєвий недолік: він здатен шукати лише поодинокі ключові слова, що погіршить точність видачі результатів пошуку. Запропонована модифікація усуває даний недолік і дозволяє знаходити ключові терміни, що складаються з кількох слів.

Постановка завдання. Метою даної роботи є експериментальне дослідження ефективності модифікованого методу автоматизованого пошуку ключових слів у текстових даних, сформульованого та теоретично обґрунтованого у [1].

У відповідності до поставленої мети задачами дослідження є:

– програмна реалізація модифікованого методу автоматизованого пошуку ключових слів;

– дослідження ефективності роботи розробленої програмної реалізації запропонованого методу автоматизованого пошуку ключових слів за критеріями абсолютної точності та повноти пошуку ключових слів за Жаккаром;

– визначення подальших кроків щодо підвищення ефективності модифікованого методу автоматизованого пошуку ключових слів.

Виклад основного матеріалу дослідження. Ключовими словами називають такі слова або вирази, якими можна описати основний зміст деякого тексту. Іноді ключовими словами, що відображають суть тексту, називають цілі словосполучення. В більшості випадків для одного тексту наводять близько десяти ключових слів [3].

Задача визначення ключових слів в тексті є складною і нетривіальною задачею, адже знайдені ключові слова повинні найбільш точно передавати тематику тексту. Таким чином, бажано не обирати в якості ключових загальновживані слова, або такі, що не несуть змістового навантаження. Точно визначеного алгоритму пошуку ключових слів людиною, нажаль, не існує, що робить цей процес складним для автоматизації [4].

Методи пошуку ключових слів

Більшість відомих методів автоматизованого пошуку ключових слів поділяється на кілька типів:

1. Статистичні методи – такі, що ґрунтуються на законах статистики [5].

2. Словникові методи – використовують наперед зібрані словникові дані, або тезауруси з деяких тематик [6].

3. Гібридні методи – поєднання особливостей статистичних та словникових методів для найбільш ефективного пошуку ключових слів [7].

Статистичні та словникові методи мають свої переваги і недоліки, тож сучасні дослідження проводяться в напрямках розроблення і покращення гібридних методів.

Модифікований метод автоматизованого пошуку ключових слів у тексті

Запропонований у [1] метод створений на базі гібридного методу, сформульованого у [2] і використовує інструменти сучасних програмних синтаксичних аналізаторів для оброблення текстів і отримання необхідних даних для подальшого зважування слів-кандидатів у ключові слова.

В загальному вигляді запропонований авторами модифікований метод є таким:

1. Синтаксичний аналіз тексту і отримання даних про зв’язки між парами слів і частини мови, до яких належать слова тексту.

2. Отримання з тексту набору всіх виразів з типами зв’язків flat та compound.

3. Фільтрування пар слів, зв’язки між якими належать до переліку неінформативних.

4. Заміна займенників в парах слів відповідними іменниками.

5. Відсіювання слів, які при синтаксичному аналізі було віднесено до неінформативних частин мови.

6. Фільтрування стоп-слів.

7. Визначення кількості зв’язків для кожного слова з пари.

8. Прийняття перших n слів з найбільшою кількістю зв’язків як ключові (де n – бажана кількість шуканих ключових слів).

9. Фільтрація отриманих багатослівних виразів за допомогою попередньо отриманих ключових слів.

Загальну схему запропонованого модифікованого методу наведено на рис. 1.

Для отримання пар слів використовується стенфордська класифікація [8] зв'язків між лексичними одиницями речень тексту. Для фільтрації слів, що відносяться до неінформативних частин мови, автори гібридного методу використовують класифікацію Пенна [9].

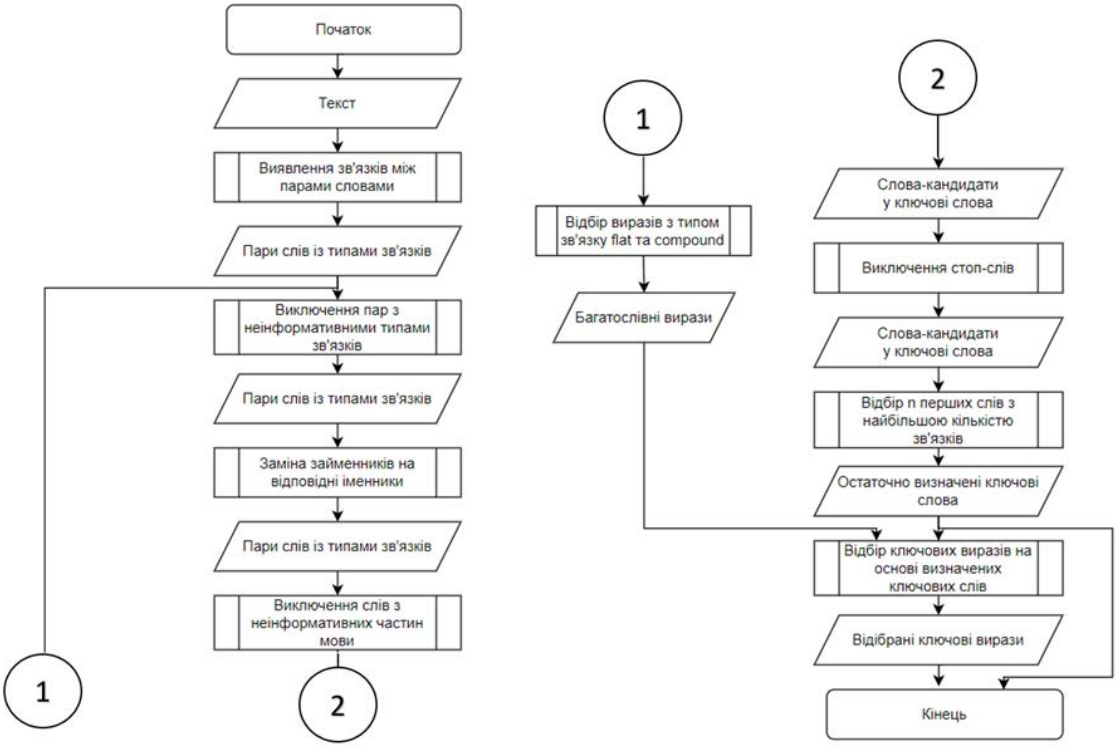

Рис. 1. Загальна схема запропонованого модифікованого методу [1]

Зазначимо, що, на відміну від гібридного методу [2], запропонована модифікація здатна знаходити ключові терміни з кількох слів. З точки зору використання таких методів в якості складових комплексної пошукової системи, однослівні ключові терміни сприяють більш загальному пошуку, але недостатньо добре покривають специфічні і конкретні запити. Таким чином, модифікований метод може покращити релевантність результатів пошуку, виданих пошуковою системою.

Кількісне оцінювання ефективності модифікованого методу автоматизованого пошуку ключових слів

Перш, ніж проводити практичне дослідження ефективності модифікованого методу автоматизованого пошуку ключових слів в тексті, необхідним є обрання способу кількісного оцінювання якості результатів роботи методу. У даній статті дослідження ефективності модифікованого методу проводиться із застосуванням двох метрик: абсолютної точності та повноти пошуку за Жаккаром. З огляду на постановку задачі, час знаходження ключових слів для вибраного тексту не має такого вирішального значення, як якість знайдених слів. Окрім того, більшість існуючих інструментів для пошуку ключових слів є онлайн сервісами, обчислювальні потужності яких можуть складатися із мережі серверів. В свою чергу, розроблене програмне забезпечення з реалізації запропонованого методу можливо протестувати лише на одному комп’ютері, тож порівняння часу виконання з хмарними серверами не є доречним.

Абсолютна точність визначається як відношення кількості правильно знайдених ключових слів за допомогою використання програмної реалізації методу до кількості ключових слів, визначених автором тексту. Якщо взяти множину еталонних ключових слів до деякого тексту як A, а множину ключових слів, що було знайдено програмою як B, тоді абсолютну точність а пошуку ключових слів можна обчислити за формулою:

$$a = \frac{n(A \cap B)}{n(A)}\# \tag{1}$$

де $n(A \cap B)$ – кількість правильно знайдених ключових слів; $n(A)$ – кількість еталонних ключових слів.

Повнота за Жаккаром визначається як відношення кількості правильно знайдених ключових слів до загальної кількості еталонних ключових слів і знайдених ключових слів мінус кількість правильно знайдених ключових слів. Повнота за Жаккаром J обчислюється за формулою:

$$J = \frac{n(A \cap B)}{n(A) + n(B) - n(A \cap B)} = \frac{n(A \cap B)}{n(A \cup B)}\# \tag{2}$$

де $n(B)$ – кількість програмно знайдених ключових слів; $n(A \cup B)$ – кількість елементів об'єднання обох множин [10].

Для застосування вищенаведених метрик до термінів, що складаються з кількох слів, у [1] пропонується використовувати метрику Word Accuracy (WAcc) з пороговим значенням 66,66%, що є оберненою до метрики Word Error Rate (WER) [11]. Значення метрики WER може бути обчислене за наступною формулою:

$$WER = \frac{S + D + I}{N}\# \tag{3}$$

де S – кількість замін; D – кількість видалень; I – кількість вставлень; N – кількість слів в "еталонному", або довідковому варіанті.

Значення метрики WAcc є оберненим до WER і обчислюється за формулою:

$$WAcc = 1 - WER\# \tag{4}$$

Для перевірки ефективності модифікованого методу було розроблено програмну реалізацію у середовищі Python за допомогою платформи Python NLTK та допоміжних пакетів AllenNLP та JiWER, особливості якої наведено в [1].

Аналіз результатів експериментальних досліджень ефективності модифікованого методу автоматизованого пошуку ключових слів в текстових даних

Порівняння ефективності роботи розробленого програмного забезпечення проводилися з наступними існуючими сервісами, що надають подібні можливості з пошуку ключових слів у тексті:

1. MonkeyLearn на основі машинного навчання [12].
2. WordCount, що використовує аналіз N-грам [13].
3. Komprehend, що побудований на статистичному аналізі [14].

Для експериментальної апробації ефективності модифікованого методу автоматизованого пошуку ключових слів в тексті було взято 200 довільних текстів тез до статей з наукового технічного журналу [15]. Це невеликі тексти обсягом 150-400 слів, із наперед зазначеним переліком ключових слів, тож їх зручно використовувати для швидкого тестування. Спочатку наведемо порівняння значень абсолютної точності для власної розробки та аналогів.

Рис. 2. Гістограма порівняння значень абсолютної точності пошуку ключових слів для текстів обсягом 150-400 слів

Результати випробувань (рис. 2) демонструють, що середнє значення абсолютної точності пошуку ключових слів для програмної реалізації модифікованого методу становить 0,402, для сервісу MonkeyLearn – 0,266, WordCount – 0.34, Komprehend – 0.294. Отже, власна розробка збільшує абсолютну точність пошуку ключових слів у межах від 6.2% до 13,6% у порівнянні з аналогами в текстах обсягом 150-400 слів.

Додатково були проведені випробування на текстах більших розмірів, близько 1500-2500 слів. Для цього було взято 100 довільних текстів статей з того ж наукового журналу.

Результати випробувань (рис. 3) демонструють, що середнє значення абсолютної точності пошуку ключових слів для власної розробки становить 0,592, для сервісу MonkeyLearn – 0,296, WordCount – 0.251, Komprehend – 0.575. Отже, власна розробка збільшує абсолютну точність пошуку ключових слів у межах від 1.7% до 34,1% у порівнянні з аналогами в текстах обсягом 1500-2500 слів.

Рис. 3. Гістограма порівняння значень абсолютної точності пошуку ключових слів для текстів обсягом 1500-2500 слів

За результатами отриманих значень абсолютної точності пошуку ключових слів можна відмітити наступне:

1. На текстах невеликого обсягу (150-400 слів) запропонований модифікований метод має явну перевагу над іншими аналогами.

2. На текстах більшого обсягу (1500-2500 слів) запропонований метод все ще має найкращі результати, однак сервіс Komprehend, що базується на методах статистики, майже його наздоганяє.

3. Збільшення середньої точності майже всіх 4 інструментів зі збільшенням обсягу тексту можна пояснити більш явним проявом статистичних закономірностей. Таким чином, за достатньо великого обсягу тексту, ключові вирази будуть неодноразово повторюватися, що збільшить їх вагу при зважуванні кандидатів у ключові терміни.

На основі отриманих результатів порівнянь можна стверджувати, що запропонований модифікований метод є кращим за критерієм абсолютної точності. Гіпотеза про використання інформації, отриманої з синтаксичного аналізатора, для пошуку багатослівних виразів в тексті надає модифікованому методу можливість шукати ключові терміни, які складаються з кількох слів, що і було очікуваним. Причому зі збільшенням обсягу тексту збільшується і кількість правильно знайдених ключових термінів, що можна пояснити більш частим входженням ключового терміну в текст. Виявлення ключових термінів з кількох слів було неможливим для методу, взятого за основу модифікації.

Для порівняння значень повноти пошуку ключових слів за Жаккаром випробування були проведені на тих самих 200 довільно обраних текстах тез статей з наукового технічного журналу [15].

Результати отриманих значень повноти пошуку ключових слів за Жаккаром для запропонованої модифікації та існуючих аналогів (рис. 4) демонструють, що середнє значення для власної розробки становить 0,088, для сервісу MonkeyLearn – 0,089, WordCount – 0.087, Komprehend – 0.048. Отже маємо зменшення повноти пошуку ключових слів за Жаккаром на 0,1% у порівнянні з найкращим результатом іншого метода при застосуванні на текстах обсягом 150-400 слів.

Аналогічно були проведені випробування на текстах обсягом 1500-2500 слів.

Рис. 4. Гістограма порівняння значень повноти пошуку ключових слів за Жаккаром для текстів обсягом 150-400 слів

Рис. 5. Гістограма порівняння значень повноти пошуку ключових слів за Жаккаром для текстів обсягом 1500-2500 слів

Результати отриманих значень повноти пошуку ключових слів за Жаккаром (рис. 5) демонструють, що середнє значення для власної розробки становить 0,058, для сервісу MonkeyLearn – 0,108, WordCount – 0.065, Komprehend – 0.012. Отже маємо зменшення повноти пошуку ключових слів за Жаккаром на 5% у порівнянні з найкращим результатом іншого метода при застосуванні на текстах обсягом 1500-2500 слів.

За результатами отриманих значень повноти пошуку ключових слів за Жаккаром можна відмітити наступне:

1. На текстах невеликого обсягу (150-400 слів) значення повноти майже однакові, окрім значень сервісу Komprehend, що базується на використанні методів статистики. Це можна пояснити тим, що даний сервіс видавав помітно більше ключових слів у результатах, чим значно понижував повноту через збільшення вербального шуму.

2. На текстах більшого обсягу (1500-2500 слів) запропонована модифікація має третій результат, що можна пояснити більшою кількістю ключових слів на виході, у порівнянні з такими аналогами як MonkeyLearn та WordCount. MonkeyLearn має найкращий результат, адже сервіс завжди обмежує кількість ключових слів до 10. Найгірший результат – у сервіса Komprehend, який отримує в результаті занадто багато вербального шуму і має повноту пошуку близько 1%.

На основі отриманих результатів порівнянь можна стверджувати, що запропонований модифікований метод має меншу повноту пошуку ключових слів за Жаккаром у порівнянні з аналогами, однак не найменшу. На етапі пошуку ключових термінів, що складаються з кількох слів, дещо збільшується вербальний шум, що зменшує значення повноти пошуку, причому кількість вербального шуму збільшується зі збільшенням обсягу тексту.

Виходячи з вищенаведених результатів дослідження ефективності модифікованого методу автоматизованого пошуку ключових слів у тексті, можна визначити такі подальші напрямки роботи над підвищенням ефективності роботи запропонованого методу та розширенням його можливостей: проведення додаткового тренування моделей для синтаксичних парсерів; використання синонімічних та тематичних словників для пошуку ключових термінів, які не зустрічаються в тексті; зменшення кількості вербального шуму в результатах пошуку; проведення більшої кількості досліджень на текстах різних тематик та стилю; додавання підтримки пошуку ключових слів в текстах інших природних мов.

Висновки. В рамках даного дослідження виконано програмну реалізацію модифікованого методу автоматизованого пошуку ключових слів в тексті. Проведені порівняння ефективності розробленої програми та існуючих сервісів MonkeyLearn (машинне навчання і статистичний аналіз), WordCount (аналіз N-грам) та Komprehend (статистичний аналіз) за критеріями абсолютної точності та повноти пошуку за Жаккаром показали, що запропонований модифікований метод має кращі показники абсолютної точності, але повнота може бути нижчою, ніж в інших аналогів. На основі отриманих результатів дослідження ефективності модифікованого методу для автоматизованого пошуку ключових слів запропоновано подальші кроки щодо підвищення його ефективності та розширення можливостей його застосування.

Список використаних джерел:

1. Бухаленков Д.О., Заболотня Т.М. Модифікований метод пошуку ключових слів та термінів у текстових даних. *Проблеми програмування* № 1 (2024). С. 12–22. Київ, 2024.

2. Яхимович О.В. Інформаційна технологія пошуку ключових слів на основі парсингу англомовних текстів. Вінниця, 2021.

3. Shibamouli Lahiri, Sagnik Ray Choudhury, Cornelia Caragea. Keyword and Keyphrase Extraction Using Centrality Measures on Collocation Networks, 2014.

4. C. Zhang, H. Wang, Y. Liu, D. Wu, Y. Liao, and B. Wang, «Automatic keyword extraction from documents using conditional random fields», *Journal of Computational Information Systems* №4, pp. 1169–1180, 2008.

5. Rafael Geraldeli Rossi, Ricardo Marcondes Marcacini, Solange Oliveira Rezende. Analysis of Statistical Key-word Extraction Methods for Incremental Clustering. Proceedings of the 10th of the Encontro Nacional de Intelig^encia Artificial e Computacional (ENIAC), Fortaleza, Brazil, 2013, 1–12.

6. Takashi Yamauchi, Dongshik Kang, Hayao Miyagi. The Keyword Search Using Thesaurus Concept, 2002 [Електронний ресурс] URL: https://koreascience.kr/article/CFKO200211921321260.pdf (дата звернення 27.03.2024).

7. K. S. Sampada, N Kavya. Machine Learning Methods for Keyword extraction and Indexing, 2019.

8. Marie-Catherine de Marneffe, Christopher D. Manning (2008). Stanford typed dependencies manual [Електронний ресурс] URL: https://downloads.cs.stanford.edu/nlp/software/dependencies_manual.pdf (дата звернення 27.03.2024).

9. Beatrice Santorini (1990). Part-of-Speech Tagging Guidelines for the Penn Treebank Project [Електронний ресурс] URL: https://www.cis.upenn.edu/~bies/manuals/tagguide.pdf (дата звернення 27.03.2024).

10. NC Chung, B. Miasojedow, M. Startek, A. Gambin (2019). «Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data». BMC Bioinformatics.

11. Klakow, Dietrich; Jochen Peters (September 2002). «Testing the correlation of word error rate and perplexity». Speech Communication. 38 (1–2): 19–28. doi:10.1016/S0167-6393(01)00041-3. ISSN 0167-6393

12. Keyword Extractor – MonkeyLearn [Електронний ресурс] URL: https://monkeylearn.com/keyword-extractor-online/ (дата звернення 27.03.2024).

13. Keyword Extractor – WordCount [Електронний ресурс] URL: https://wordcount.com/keyword-extractor (дата звернення 27.03.2024).

14. Keyword Extractor – Komprehend [Електронний ресурс] URL: https://komprehend.io/keyword-extractor (дата звернення 27.03.2024).

15. Journal of Aerospace Technology and Management [Електронний ресурс] URL: https://jatm.com.br/jatm/issue/archive (дата звернення 27.03.2024).