УДК 004.4 http://doi.org/10.15407/pp2024.01.12

Д.О. Бухаленков, Т.М. Заболотня

МОДИФІКОВАНИЙ МЕТОД ПОШУКУ КЛЮЧОВИХ СЛІВ ТА ТЕРМІНІВ У ТЕКСТОВИХ ДАНИХ

У даній статті розглядається питання автоматизованого пошуку ключових слів та термінів у текстових даних. Для підвищення ефективності засобів автоматизованого пошуку ключових слів у тексті за критеріями абсолютної точності та повноти за Жаккаром розроблено модифікацію одного з найсучасніших методів для пошуку ключових слів. Запропоновано модифікацію існуючого гібридного методу пошуку ключових слів, що враховує складні залежності між парами слів у тексті для визначення багатослівних виразів, що, на відміну від оригінального методу, дозволяє знаходити ключові терміни, які складаються з кількох слів. Здійснені випробування створеної модифікації гібридного методу пошуку ключових термінів показали ефективність її використання для пошуку ключових термінів у текстах у порівнянні з існуючими аналогами.

Ключові слова: ключові слова, ключові терміни, оброблення текстових даних, Python, стенфордська класифікація.

Вступ

Основний зміст будь-якого тексту можна описати одним або кількома словами з цього тексту, що відображають його суть. Такі слова називають ключовими словами. В більшості випадків кількість таких слів становить близько десяти [1]. Іноді ключовими можуть вважати не тільки слова, а й цілі словосполучення і речення.

Незважаючи на просте визначення поняття ключового слова, процес пошуку ключових слів є складним аналітичним завданням. Не існує ідеального способу визначення переліку ключових термінів для довільного тексту будь-якої тематики. Кожен текст має свою структуру, стиль викладення, стилістичні особливості написання [2].

Слід зазначити, що задача пошуку ключових слів виникає у багатьох сферах, пов'язаних з обробленням текстових даних. Так інформація про ключові слова використовується у інформаційному текстовому пошуку, класифікації, кластеризації даних. Водночас важливою вимогою до методів визначення ключових слів є можливість їх автоматизації, адже обсяги даних, які проходять через сучасні електронні пристрої та системи, неможливо ефективно обробляти вручну.

Упродовж багатьох років досліджень спеціалістами було запропоновано методи, різні за ефективністю та умовами застосування. Одні методи – добре налаштовані на оброблення текстів вузької тематики, зокрема, технічної літератури, інші – можуть бути методами ширшого застосування. Однак досі триває пошук шляхів покращення точності методів для пошуку ключових слів та підвищення ступеня їхньої універсальності щодо застосування до різних типів текстів.

Метою даної статті є підвищення ефективності засобів для автоматизованого пошуку ключових слів у тексті за критеріями абсолютної точності та повноти за Жаккаром шляхом модифікації одного з сучасних методів пошуку ключових слів та за допомогою використання сучасних лінгвістичних програмних пакетів.

Існуючі методи

Не існує ідеального "золотого правила", за яким можна було б визначити набір ключових слів для будь-якого тексту. Відомі методи можна умовно поділити на кілька основних груп:

– Статистичні методи – найстарішими можна назвати методи, що ґрунтуються на використанні статистичних даних, отриманих під час аналізу тексту [3]. Одними із перших статистичних закономірностей, виявлених для природномовних текстових даних, можна вважати здобутки

американського лінгвіста та економіста Джорджа Ципфа. Виведені ним закони про розподіл слів у текстах є основами багатьох відомих статистичних методів.

Статистичні методи в основному аналізують частоту входжень слів у тексті, їхню довжину, відстань у тексті між ними. Приклади відомих статистичних метрик та методів – метрика **TF-IDF**, методи виявлення **семантичного поля**, метод **системного зважування слів за частотою та довжиною**, метод **k-factor**, метод **C-value**, **TERMC** та інші.

Основними **перевагами** статистичних методів можна вважати їхню **швидкість** та **мовонезалежність**, адже статистичні закономірності майже однаково прослідковуються у більшості природних мов.

Характерні **недоліки** таких методів полягають у великому обсязі вербального шуму в отриманих результатах та недостатній точності у застосуванні на текстах невеликих розмірів.

– Словникові методи – методи, що використовують заздалегідь зібрані словникові дані, або тезауруси з деяких тематик [4]. На відміну від статистичних, такі методи здатні надавати **більш точні** результати з меншою кількістю вербального шуму. Однак для отримання точних результатів треба мати дуже докладні тезауруси з відповідної до тексту тематики. З цього випливає, що методи, побудовані на основі використання словників, складно застосовувати до текстів **вузької або нової тематики**.

– Гібридні методи – сучасні розробки використовують поєднання особливостей статистичних та словникових методів для найбільш ефективного пошуку ключових слів [5]. **Статистичні закономірності** допомагають швидко знайти основний масив потенційних ключових слів, а моделі машинного навчання, натреновані на словникових даних, **збільшують точність і відсіюють вербальний шум**, в результаті чого на виході отримується набір ключових слів. Серед популярних сучасних програмних інструментів для оброблення природномовних текстових даних: Python NLTK, Stanford NLP, Keras, spaCy тощо.

Обраний метод для модифікації

Гібридний метод пошуку ключових слів в англомовних текстах, що був запропонований українським фахівцем О.В. Яхимовичем [6], належить до останньої з трьох вищезазначених груп методів. Метод використовує інструменти сучасних програмних синтаксичних аналізаторів для оброблення текстів і отримання необхідних даних для подальшого зважування слів-кандидатів у ключові слова. Наведемо основні етапи методу:

1. Синтаксичний аналіз тексту і отримання даних про зв’язки між парами слів і частини мови, до яких належать слова тексту.
2. Фільтрування пар слів, зв’язки між якими належать до переліку неінформативних.
3. Заміна займенників у парах слів відповідними іменниками.
4. Відсіювання слів, які при синтаксичному аналізі було зараховано до неінформативних частин мови.
5. Фільтрування стоп-слів.
6. Визначення кількості зв’язків для кожного слова з пари.
7. Прийняття перших **n** слів з найбільшою кількістю зв’язків як ключових (де n - бажана кількість шуканих ключових слів).

Для отримання пар слів використовується **стенфордська класифікація** [7] зв'язків між лексичними одиницями речень тексту. Розробниками методу шляхом численних випробувань було визначено 7 типів зв’язків, що не несуть суттєвого змістовного навантаження і не відіграють важливої ролі в контексті пошуку ключових слів. Це зв’язки: CC, DET, EXPL, FIXED, PUNCT, REF, ROOT.

Для фільтрації слів, що належать до неінформативних частин мови, автори гібридного методу використовують **класифікацію Пенна** [8] і виділяють 21 тег з цієї класифікації: CC, CD, DT, EX, IN, LS, MD, PDT, POS, PRP, PRP$, RP, SYM, TO, UH, WDT, WP, WP$, WRB, -LRB-, -RRB-.

Заміна займенників на відповідні іменники відбувається за допомогою ана-

лізу кореференційних зв'язків між словами в тексті.

Розробники методу запевняють, що запропонований гібридний метод має приріст повноти у межах від 8,1% до 12,7% **за метрикою Жаккара**, та від 9,1% до 14,3% **абсолютної точності** пошуку ключових слів у порівнянні з існуючими аналогами. Отже, метод можна вважати одним із найбільш точних серед сучасних розробок, і його модифікація для отримання ще більш точних результатів вбачається перспективною.

Гіпотеза №1 про підвищення точності "Гібридного методу"

Головною особливістю оригінального методу можна вважати визначення кількості синтаксичних зв'язків між словами і відсіювання пар слів із неінформативними типами зв'язків та слів, що належать до неінформативних частин мови. Таким чином, на якість результатів пошуку ключових слів насамперед впливає вміст списків неінформативних типів зв'язків та частин мови, що були визначені авторами заздалегідь. Отже, **модифікація цих списків** потенційно може покращити кількісні характеристики якості отримуваних результатів.

Авторами даної статті було вирішено зробити наступні модифікації списків:

– виключити зі списку неінформативних частин мови тег **CD**, або **cardinal number** (кардинальне число). Було висунуте припущення, що важливі для змісту ключові слова або фрази можуть містити конкретні числа, як-от у виразі "Order **767**". Якщо є якийсь загальновідомий історичний наказ під номером 767, внесення такого числа до списку ключових слів покращить результати пошуку і ймовірність знаходження даного ресурсу, де описано цей наказ. За оригінальним методом числівник 767 було б повністю вилучено зі списку потенційних ключових слів, що, можливо б, погіршило якість інформаційного пошуку;

– вилучити зі списку неінформативних частин мови тег **RP**, або **particle** (частка). В деяких специфічних поняттях чи термінах може міститися слово-частка. Наприклад, у реченні "Located right on the airfield, guests can watch other planes take **off** and land." "off" є частиною "take off", тобто злітати. Якщо відфільтрувати цю частку зі списку ключових слів, то пошуковий запит, що містить "take off", не буде чітко відповідати набору ключових слів;

– включити до списку неінформативних типів зв'язків тип **ccomp**, або **clausal complement** (комплементна клаузальна конструкція). Найчастіше такий зв'язок трапляється між дієсловом або прикметником і додатком. Зв'язок є досить специфічним і часто не несе достатнього інформаційного змісту, щоб додавати ваги словам-кандидатам у ключові слова.

Були проведені випробування із запропонованими модифікаціями списків. Результати тестування на чисельних текстах тез статей з наукових журналів показали, що списки неінформативних частин мов та типів зв'язків, виділені авторами оригінального методу, є ефективними. Майже в усіх випадках результати не змінювалися, спостерігалися лише невеликі відхилення в той чи інший бік на одне слово. Середні значення абсолютної точності та повноти за Жаккаром виявилися майже однаковими (рис.1).

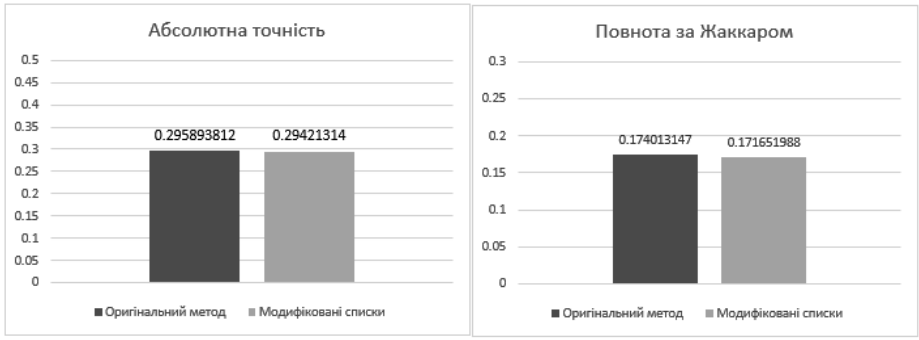

Рис.1. Результати тестування методу з модифікованими "неінформативними" списками

Отож, кількісні показники якості результатів роботи методу зі зміненими списками за абсолютною точністю знайдених ключових слів та повнотою Жаккара були практично однакові з результатами оригінального методу. Отже, перша гіпотеза була спростована.

Гіпотеза №2 про використання інформації про багатослівні вирази для пошуку ключових термінів

У оригінального методу є суттєвий недолік – він дозволяє шукати **лише однослівні ключові терміни**. Під час виконання алгоритму, що реалізує метод, пари слів роз'єднуються в окремі слова, після чого для кожного слова окремо визначається кількість зв'язків, отже, на виході можна отримати лише окремі ключові слова. Використання однослівних ключових термінів сприяє більш загальному пошуку, але недостатньо добре покриває специфічні і конкретні запити.

Авторами цієї статті було вирішено модифікувати метод таким чином, щоб додати можливість пошуку ключових термінів, що складаються з кількох слів.

Проаналізуємо, які типи зв'язків зустрічаються між словами багатослівних ключових термінів. Для отримання "еталонного" переліку ключових термінів у рамках даного дослідження було вирішено використовувати статті наукових журналів, де до кожної з них є наданий авторами набір ключових слів та виразів, який можна вважати довідковим для проведення порівнянь [9].

Для тексту анотації до статті [10] дослідимо зв'язки між словами ключового терміна "ammonium perchlorate" (рис.2).

compound(perchlorate-6, ammonium-5)
compound(perchlorate-25, ammonium-24)
compound(perchlorate-32, ammonium-31)

Рис.2. Зв'язки між словами ключового терміна "ammonium perchlorate" для тексту анотації до статті [10]

Аналізатор визначив, що такий ключовий термін можна відшукати в тексті за зв'язками типу compound. Згідно таблиці типів синтаксичних зв'язків **Universal Dependencies** (стенфордська класифікація) [11] тип зв'язку compound належить до категорії **MWE (Multiword Expressions)**, тобто багатослівних виразів. Розглянемо, які типи синтаксичних зв'язків містить у собі категорія MWE.

Fixed. Використовується для позначення спеціальних службових конструкцій, фіксованих виразів, зворотів. Слова, що мають зв'язок fixed, не мають зв'язків інших типів з іншими словами. Приклади таких конструкцій в англійській мові: *as well as*, *because of*, *rather than*. На рис.3 наведено приклади речень, де визначено зв'язки типу fixed [12].

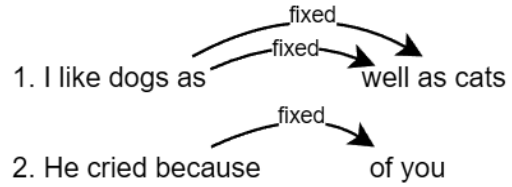

Рис. 3. Приклади речень зі зв'язком типу fixed

Flat. Цей тип зв'язку використовується для екзоцентричних виразів, тобто таких, де немає головного слова. До таких належать імена і дати. На рис.4 наведено приклади визначення зв'язків для імен.

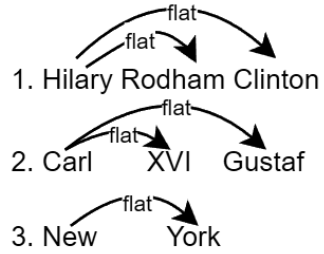

Рис. 4. Приклади визначення зв'язку flat для імен

Зв'язок flat також застосовується до виразів, де згадується титул або звання персони разом з ім'ям (рис.5).

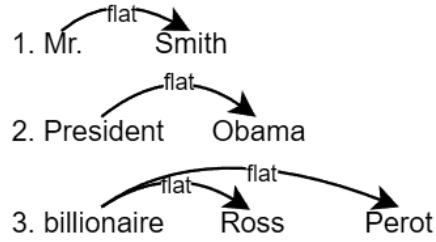

Рис. 5. Зв'язки flat для виразів з титулами або званнями

Для складених числових виразів також застосовується зв'язок flat [13]. На рис.6 наведено приклад виразів з датами або складними числівниками, де визначено цей тип зв'язку.

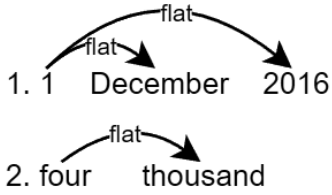

Рис. 6. Вирази з датами та складними числівниками, де визначено тип зв'язку flat

Compound. Цей тип зв'язку визначено у виразах ендоцентричного типу, де, на відміну від екзоцентричних, є головне слово. Такі вирази є сполученнями з кількох частин мови: іменникові сполучення, дієслівні, прикметникові, їхні комбінації та іноді серійні дієслівні конструкції. Більшою мірою виражені [14]:

– складними іменниками. Це можуть бути **Іменник** + **Іменник** (bus stop, fire-flies, football), **Прикментик** + **Іменник** (full moon, blackboard, software), **Дієслово** + **Іменник** (breakfast, washing machine, swimming pool), **Іменник** + **Дієслово** (sunrise, haircut) та інші сполучення (USB cell phone chargers);

– серійними дієслівними конструкціями. Синтаксична конструкція, де представлена послідовність двох або більше дієслів, які функціонують як один предикат та описують одну подію. В сучасній англійській мові майже не зустрічаються, але збереглися деякі вирази: let's go eat, come live with me.

– фразовими дієсловами. Комбінація дієслова і прийменника, або дієслова і прислівника, або одночасно дієслова і прийменника з прислівником, яка є окремим членом речення і утворює окрему семантичну одиницю. Найчастіше складаються із власне смислового дієслова та одного або декількох прийменників (рідше прислівників): keep on, pass out, look up, give up, put off, come across.

Декілька прикладів, наведених в довіднику Universal Dependencies, зображено на рис.7.

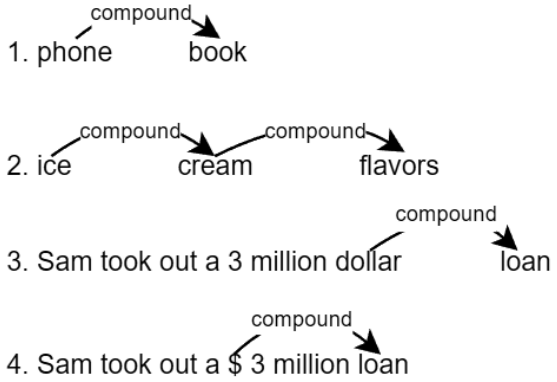

Рис. 7. Приклади виразів, де присутній зв'язок типу compound

Із перелічених трьох типів зв'язків fixed становить найменший інтерес з точки зору вирішення задачі пошуку ключових термінів. Fixed позначає спеціальні звороти і усталені сполучення слів, що утворилися історично або іншим шляхом, такі як instead of, rather than та інші. Вони слугують радше пов'язуючими функціональними частинами речень, але смислового навантаження самі собою не несуть.

Таким чином, з огляду на вищезазначене, для подальшого дослідження було обрано можливості використання знайдених зв'язків типу **flat** та **compound** для отримання списку ключових словосполучень із текстів.

Було знайдено наступні вирази: ['Honolulu', 'Hawaii'], ['Harvard', 'Law', 'Review'], ['Columbia', 'University'], ['Harvard', 'Law', 'School'], ['community', 'organizer'], ['law', 'degree']. Знайдені словосполучення дійсно певною мірою відображають зміст тексту, і теоретично підходять як ключові сліва.

Очевидно, що вибір з тексту всіх виразів з типами зв'язків flat та compound без якої-небудь фільтрації не є повноцінним рішенням задачі пошуку ключових виразів.

Як запевняють автори оригінального методу, його використання дозволяє досить точно знаходити окремі ключові слова [6]. Звідси можна використати визначений список окремих ключових слів для фільтрування отриманих багатослівних виразів. Адже логічно припустити, що якщо текст містить деякі ключові словосполучення або терміни з кількох слів,

окремі слова з них ймовірно будуть знайдені алгоритмом оригінального методу.

Таким чином, можна "відфільтрувати" всі знайдені багатослівні вирази, отримані з пошуку в тексті зв'язків flat та compound, використовуючи отримані раніше поодинокі ключові слова. Тобто, якщо в деякому знайденому з тексту виразі, що мав між словами зв'язки flat чи compound, міститься хоча б одне слово з набору ключових слів, такий вираз з великою ймовірністю буде ключовим для даного тексту. Проведені випробування на текстах тез статей наукових журналів підтверджують ефективність даного припущення за критеріями абсолютної точності та повноти пошуку ключових слів, про що йтиметься далі в цій статті.

Модифікований метод пошуку ключових слів та виразів

Із урахуванням вищенаведеного, пропонується модифікація оригінального методу з додатковими кроками для отримання списку ключових виразів та термінів до тексту. А саме:

– на етапі номер 1) оригінального методу збирається інформація про всі зв'язки типу flat і compound, з чого отримується набір усіх багатослівних виразів з такими зв'язками в тексті;

– після етапу номер 7) оригінального методу, враховуючи отримані результати, відбувається фільтрація отриманих на етапі 1) багатослівних виразів наступним чином: якщо ключове слово міститься у виразі, він потрапляє до списку ключових, інакше – відсіюється;

– в результаті отримується список ключових слів і список ключових виразів із двох або більше слів.

Список ключових виразів можна буде використовувати у процесі індексації ресурсів у пошукових системах, надаючи їм більший пріоритет збігу із пошуковим запитом користувача, адже ключові терміни точніше відображають зміст тексту, аніж поодинокі ключові слова.

Отже, 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 у парах слів відповідними іменниками.
5. Відсіювання слів, які під час синтаксичного аналізу 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.**

На рис.8 наведено загальну схему запропонованого модифікованого методу.

Вибір метрик для визначення кількісних характеристик ефективності модифікованого методу

Враховуючи постановку задачі, а саме модифікацію методу з метою уможливлення отримання в результаті аналізу не тільки ключових слів, а й багатослівних ключових термінів та виразів, що може давати точніші за існуючі аналоги результати, ніж, час виконання нового модифікованого методу має не таке вирішальне значення, як точність і якість отриманих результатів. Отже, потрібно обрати метрики, що покажуть переваги використання модифікованого методу замість існуючих аналогів саме за знайденими ключовими словами та виразами.

Складно правильно оцінити точність знайдених ключових слів чи виразів для певного довільно взятого тексту, адже

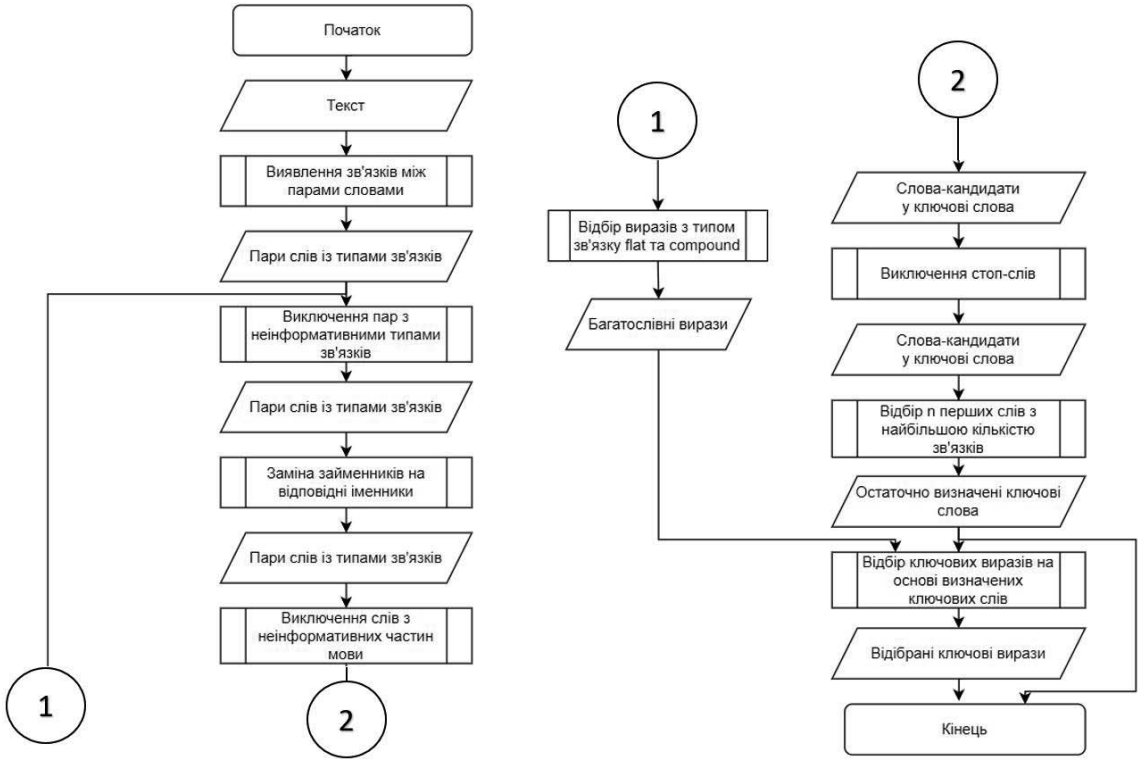

Рис.8. Схема етапів модифікованого методу для пошуку ключових виразів та термінів

хоча у ключових слів є деякі визначені характеристики, це доволі суб'єктивна оцінка. Тож різні спеціалісти можуть сперечатися щодо правильності визначення того чи іншого слова як ключового. Тому для досягнення максимально можливої об'єктивності було вирішено використовувати для випробування тексти із наперед визначеним набором ключових слів. Серед таких текстів є наукові статті, до яких при публікації в журналі автор сам підбирає набір ключових слів чи понять.

Для обґрунтування доцільності використання гібридного методу для пошуку ключових слів автори обрали дві метрики: **абсолютну точність** та **повноту за Жаккаром**. Коротко опишемо суть цих метрик.

Абсолютна точність визначається як відношення кількості правильно знайдених ключових слів за допомогою використання програмної реалізації методу до кількості ключових слів, визначених автором тексту.

Наприклад, якщо взяти множину еталонних ключових слів до тексту як A, а множину ключових слів, що було знайдено програмою як B, тоді абсолютну точність a пошуку ключових слів можна обчислити за формулою:

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

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

У свою чергу, **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:

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

де $n(B)$ – кількість програмно знайдених ключових слів; $n(A \cup B)$ – кількість елементів об'єднання обох множин [16].

Ці дві метрики достатньо легко використати для порівняння двох множин окремих слів, адже тоді можна точно визначити, чи входить слово до переліку визначених автором, чи ні. Але з виразами і

ключовими термінами з кількох слів це складніше застосувати. Буде не зовсім справедливо порівнювати вирази один з одним, адже так втратиться велика кількість правильно знайдених ключових виразів, які лише трохи відрізняються. Наприклад, авторами статті [17] було визначено ключовий термін “reynolds number test”, а автоматизований метод знайшов вираз “reynolds number”. Порівнюючи вирази за логікою “один до одного” результат буде визначено помилковим, незважаючи на збіг 2 з 3 слів, і однаковий сенс двох виразів. Звідси постає питання, як порівняти два вирази із деяким порогом допустимої різниці складу виразів.

Є сфери, де науковці стикаються з подібною задачею оцінки точності збігу словесних виразів. До них можна віднести розроблення і тестування систем автоматичного розпізнавання мовлення або систем машинного перекладу. В обох випадках в результаті роботи систем на виході отримується текст або ж набір словесних виразів чи фраз, які треба порівняти з “еталонним” набором для оцінки якості роботи системи чи методу.

Однією із найвідоміших метрик для оцінювання роботи систем розпізнавання мовлення та машинного перекладу є **Word Error Rate (WER)**, або ж Частота Помилкових Слів [18]. Метрика базується на понятті відстані Левенштейна, але працює на рівні слів, а не символів. Визначається як відношення кількості замін слів, видалення слів, або додавання слів до довідкового варіанту для приведення його до вигляду отриманого автоматичною системою, до кількості слів у довідковому значенні.

Значення метрики **WER** може бути обчислене за наступною формулою:

$$WER = \frac{S + D + I}{N} \qquad (3)$$

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

Існує також і обернена до WER метрика – **WAcc**, або ж **Word Accuracy**, яка обчислюється як різниця одиниці і значення WER.

$$WAcc = 1 - WER \qquad (4)$$

Фактично це та сама метрика, що і WER, але кількісно відображається не помилка, а точність.

Експериментальним шляхом було вирішено встановити поріг значення WAcc, за якого вирази будуть вважатися рівними, у 66,66%. Із зниженням цього порогу спостерігалося збільшення вербального шуму в результатах пошуку, а із збільшенням – втрачалося більше ключових виразів, визначених авторами. Це означає, що за умови існування еталонного ключового виразу *reynolds number test*, знаходження алгоритмом виразу *reynolds number* буде вважатися успішним. Таким чином, ми можемо використовувати модифіковані метрики **абсолютної точності** та **повноти за Жаккаром** для оцінювання результатів роботи модифікованого методу.

Особливості програмної реалізації методу

Запропонований модифікований метод було реалізовано у вигляді консольного додатку на Python з можливістю взаємодії за CLI. Для виконання основних операцій оброблення природномовних текстів було використано платформу Python NLTK, а також допоміжний лінгвістичний пакет AllenNLP [19]. Для обчислення метрики WER використано пакет JiWER [20].

Використовуючи Python NLTK для пошуку зв’язків між парами лем в тексті рекомендується розглядати кожне речення окремо. Для цього необхідно розбити текст на речення, для чого використовується функція *sent_tokenize* з модуля *nltk.tokenize*.

Після того, як текст було розбито на речення, необхідно проаналізувати кожне і отримати список усіх зв’язків, або залежностей (dependency) між парами слів. Для цього застосовується модуль *StanfordDependencyParser*, а саме його метод *raw_parse*, який приймає на вхід речення в строковому вигляді, і на вихід віддає складну структуру-дерево з усіма зв’язками між парами слів.

Для отримання більш точних результатів за алгоритмом необхідно приводити слова до основної форми перед визначенням кількості зв'язків. Це прибере похибку результатів. У випадку, якщо одна й та сама сутність або поняття малися на увазі в різних частинах тексту, і вживалися в різних формах, вони будуть розцінені як різні слова при визначенні кількості зв'язків і зважуванні. Це може значно знизити вагу поняття в кінцевому випадку. Для приведення слів до основної форми використовується модуль *WordNetLemmatizer*, що імпортується з *nltk.stem*, а саме його метод *lemmatize*, який, однак, потребує дані про частину мови. Для цього використовується метод *pos_tag*, що дозволяє при надаванні тексту отримати для кожного слова інформацію, який тег частини мови має кожне слово з цього тексту.

Для фільтрації стоп-слів використовується словник стоп-слів, що надається модулем *stopwords* із набору *nltk.corpus*.

Для запуску додатку необхідно задати такі параметри за допомогою рядка CLI:

1. Шлях до файлу з вхідним текстом.
2. Шлях до файлу з результатами, що буде створено.
3. Кількість бажаних окремих ключових слів в результаті n. *Необов'язковий параметр*. У разі відсутності параметру за ключові будуть узяті слова-кандидати, що лишилися після кроків фільтрування.
4. Файл із переліком еталонних ключових слів, якщо такі є. Необов'язковий параметр. Використовується для перевірки точності роботи методу для тексту з наперед визначеними ключовими словами.

Випробування розробленого модифікованого методу

Програмну реалізацію розробленого модифікованого методу було протестовано на 50 довільних текстах тез до статей з наукових журналів [9]. Для порівняння було обрано існуючий сервіс для пошуку ключових слів **MonkeyLearn** [21], що є одним із найбільш популярних і ефективних. Розробники сервісу не розкривають, який саме метод пошуку ключових слів використовують, адже сервіс має багато платних функцій для аналізу контенту. Але згідно з інформацією про сервіс [22], використовується гібридний метод, що поєднує статистичні підходи та можливості машинного навчання.

Результати випробування у вигляді діаграм порівнянь середніх значень для метрик абсолютної точності пошуку ключових слів та повноти за Жаккаром зображено на рис.9.

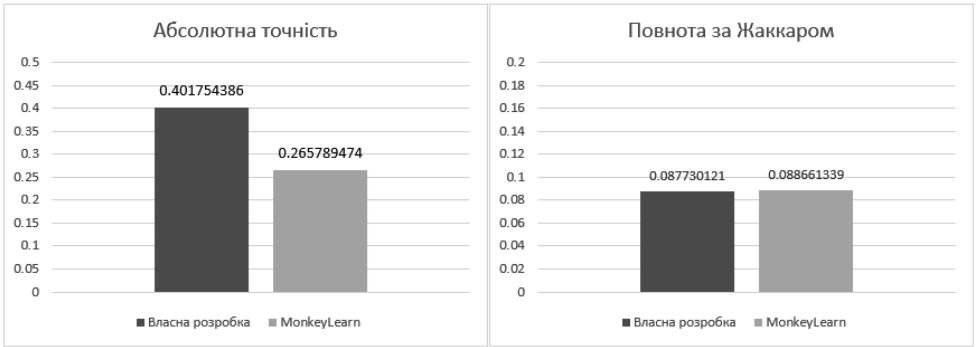

Рис.9. Результати випробування власної модифікації та сервіса MonkeyLearn

Середнє значення абсолютної точності для власної розробки – **0,402**, для сервісу MonkeyLearn – **0,266**, отже, власна розробка збільшує абсолютну точність пошуку ключових слів на **13,6%**. Середнє значення повноти за Жаккаром для власної розробки – **0,088**, для сервісу MonkeyLearn – **0,089**, отже маємо зменшення повноти пошуку ключових слів за Жаккаром на **0,1%**. Аналізуючи результати, можна стверджувати, що за невеликого зменшення повноти пошуку, модифікований метод має суттєве підвищення абсолютної точності порівняно з аналогом.

Висновки

У даній статті обґрунтовано актуальність проблеми пошуку ключових слів в тексті. Коротко описано та проаналізовано існуючі типи методів пошуку ключових слів, їхні переваги та недоліки.

Проаналізовано та обґрунтовано вибір для подальшого дослідження гібридного методу пошуку ключових слів за авторством О.В. Яхимовича. Висвітлено недоліки цього методу, та важливість їх уникнення. Висунуто гіпотези щодо підвищення ефективності методу та усунення недоліків.

На основі гіпотези про використання даних щодо багатослівних виразів у тексті для пошуку ключових термінів із кількох слів побудовано модифікацію оригінального методу. Це дозволяє шукати не лише окремі ключові слова, а й ключові терміни, що складаються з кількох слів.

Для випробування розробленого методу реалізовано програмне забезпечення у вигляді додатку мовою Python із використанням сучасних лінгвістичних програминих пакетів. Протестовано програмну реалізацію модифікованого гібридного методу на текстах тез статей із наукових журналів, отримані результати порівняно з результатами існуючого популярного сервісу MonkeyLearn.

Запропонована модифікація методу пошуку ключових слів збільшує абсолютну точність пошуку ключових слів у англомовних текстах із невеликим зменшенням повноти за Жаккаром.

Автори статті вважають за доцільне проведення подальших досліджень за такими напрямками:

–збільшення кількості випробувань на текстах різних розмірів та тематик;

– зменшення вербального шуму серед багатослівних ключових термінів, отриманих у результаті роботи методу;

– оформлення розробленого програмного забезпечення у вигляді простої для використання бібліотеки.

Література

1. Shibamouli Lahiri, Sagnik Ray Choudhury, Cornelia Caragea. Keyword and Keyphrase Extraction Using Centrality Measures on Collocation Networks, 2014.
2. H. M. Mahedi Hasan, Falguni Sanyal, Dipankar Chaki, Md. Haider Ali. An empirical study of important keyword extraction techniques from documents. 2017. In Proceedings of the 2017 1st International Conference on Intelligent Systems and Information Management, 91–94.
3. 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 Inteligˆencia Artificial e Computacional (ENIAC), Fortaleza, Brazil, 2013, 1–12.
4. Takashi Yamauchi, Dongshik Kang, Hayao Miyagi. The Keyword Search Using Thesaurus Concept, 2002 [Online] – Available from: https://koreascience.kr/article/CFKO200211921321260.pdf, last accessed 2024/01/08.
5. K. S. Sampada, N Kavya. Machine Learning Methods for Keyword extraction and Indexing, 2019.
6. Яхимович О.В., "Інформаційна технологія пошуку ключових слів на основі парсингу англомовних текстів", Вінниця, 2021.
7. Marie-Catherine de Marneffe, Christopher D. Manning (2008). Stanford typed dependencies manual [Online] – Available from: https://downloads.cs.stanford.edu/nlp/software/dependencies_manual.pdf, last accessed 2024/01/08.
8. Beatrice Santorini (1990). Part-of-Speech Tagging Guidelines for the Penn Treebank Project [Online] – Available from: https://www.cis.upenn.edu/~bies/manuals/tagguide.pdf, last accessed 2024/01/08.
9. Journal of Aerospace Technology and Management [Online] – Available from: https://jatm.com.br/jatm/issue/archive, last accessed 2024/01/08.
10. Rene Gonçalves, Koshun Iha, Francisco Machado, José Rocco. (2012). Ammonium Perchlorate and Ammonium Perchlorate-Hydroxyl Terminated Polybutadiene Simulated Combustion. Journal of Aerospace Technology and Management. 4.

11. Universal Dependency Relations [Online] – Available from: https://universaldependencies.org/u/dep/, last accessed 2024/01/08.
12. Fixed dependency [Online] – Available from: https://universaldependencies.org/u/dep/fixed.html, last accessed 2024/01/08.
13. Flat dependency [Online] – Available from: https://universaldependencies.org/u/dep/flat.html, last accessed 2024/01/08.
14. Compound dependency [Online] – Available from: https://universaldependencies.org/u/dep/compound, last accessed 2024/01/08.
15. Steven Bird, Ewan Klein, Edward Loper. (2009). Natural Language Processing with Python.
16. NC Chung, B. Miasojedow, M. Startek, A. Gambin (2019). "Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data". BMC Bioinformatics.
17. Maurício Silva, Victor Gamarra, Koldaev Vitor. (2009). Control of Reynolds number in a high speed wind tunnel. Journal of Aerospace Technology and Management. 1.
18. Dietrich Klakow, Peters Jochen (2002). "Testing the correlation of word error rate and perplexity". Speech Communication. 38 (1–2), 19–28.
19. AllenNLP Library [Online] – Available from: https://allenai.org/allennlp/software/allennlp-library, last accessed 2024/01/08.
20. JiWER [Online] – Available from: https://jitsi.github.io/jiwer/, last accessed 2024/01/08.
21. Keyword Extractor – MonkeyLearn [Online] – Available from: https://monkeylearn.com/keyword-extractor-online/, last accessed 2024/01/08.
22. Keyword Extraction: A Guide to Finding Keywords in Text – MonkeyLearn [Online] – Available from: https://monkeylearn.com/keyword-extraction/, last accessed 2024/01/08.

Одержано: 23.02.2024

Про авторів:

Бухаленков Дмитро Олександрович,
магістрант НТУУ "КПІ імені
Ігоря Сікорського",
https://orcid.org/0009-0001-0224-8873
E-mail: 3a43mka@gmail.com

Заболотня Тетяна Миколаївна
Кандидат технічних наук
Кафедра програмного забезпечення
комп'ютерних систем
Національний технічний університет
України «Київський політехнічний
інститут імені Ігоря Сікорського»
Кількість статей в загальнодержавних
базах даних: 27
Кількість статей в міжнародних
базах даних: 2
H-index за Scopus: 2
ResearchGate: - ID Scopus: 6507406568
ResearcherID: J-2245-2017

Місце роботи авторів:

Національний технічний університет
України «Київський політехнічний
інститут імені Ігоря Сікорського»,
Берестейський проспект, 37, м. Київ,
Україна, індекс 03056
E-mail: zabolotnia@pzks.fpm.kpi.ua
ORCID: 0000-0001-8570-7571
Контактний тел.: +38-066-369-93-63