DOI 10.31891/2307-5732-2020-283-2-112-116
УДК 004.91

О.В. МАЗУРЕЦЬ, Т.К. СКРИПНИК, В.А. ЖИТНЯКІВСЬКИЙ
Хмельницький національний університет

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АВТОМАТИЗОВАНОГО ВИЗНАЧЕННЯ КЛЮЧОВИХ СЛІВ У ПОВІДОМЛЕННЯХ ДЛЯ СОЦІАЛЬНИХ МЕРЕЖ

У статті розглянуто інформаційну технологію автоматизованого визначення ключових слів у текстових повідомленнях для соціальних мереж, яка проводить аналіз текстового повідомлення із використанням методів оцінки TFIDF, дисперсійної оцінки та оцінки TFIDF з використанням NLP. Розроблена інформаційна технологія автоматизованого визначення ключових слів була реалізована в тестовому програмному продукті, який відтворює роботу соціальної мережі. Вхідними даними для системи є текстове повідомлення із цифровим текстом, а вихідними даними є текстове повідомлення з множиною ключових термінів. Під час розробки соціально орієнтованого сервісу для спілкування за інтересами на платформі IOS, відповідно до визначених функцій, виділено наступні групи користувачів: зареєстрований користувач, адміністратор соціальної мережі, незареєстрований користувач. За допомогою розробленого тестового програмного забезпечення були проведені дослідження, що підтвердили можливість ефективно автоматизовано визначення множини ключових слів у текстових повідомленнях з показниками точності для методу оцінки TFIDF – 27,1% та методу дисперсної оцінки – 45,5%, методу оцінки TFIDF з NLP 88,3%. Перевагами розробленої інформаційної технології автоматизованого визначення ключових слів у текстових повідомленнях для соціальних мереж, яка проводить аналіз текстового повідомлення із використанням методів оцінки TFIDF, дисперсійної оцінки та оцінки TFIDF з використанням NLP, є відсутність необхідності використання лексичних баз даних корпусів слів, суттєве прискорення швидкодії, можливість використання для текстів на різних мовах, можливість використання для текстів з кількома мовами. Дана інформаційна технологія може бути ефективно використана для аналізу текстових повідомлень із невідомими властивостями тематики та мови.

Ключові слова: *текстові повідомлення, оцінка TFIDF з використанням NLP, ключові слова.*

O. MAZURETS, T. SKRYPNYK, V. ZHYTNIAKIVSKYI
Khmelnytskyi National University

INFORMATION TECHNOLOGY FOR AUTOMATED DETERMINATION OF KEYWORDS IN MESSAGES FOR SOCIAL NETWORKS

The article discusses automated keyword definition in text messaging for social networks that analyzes text message using the methods of TFIDF estimation, variance estimation, and TFIDF estimation using NLP. The developed automated keyword definition information technology was implemented in the test software. The input data of the system is a text message with digital text, and the output data is a text message with a set of key terms. In developing a socially-oriented interest-based communication service on the iOS platform, according to the defined functions, the following groups of users are identified: registered user, social network administrator, unregistered user. A registered user works with the system via the IOS Mobile App - this group includes users who are logged in and have access to photo and video messaging, location sharing, search for other users, viewing other users' news feeds, commenting on news from others users, create their own news feed, view their own news feed, and track other users' news. The social network administrator works with the system through a browser interface on an arbitrary platform - this group includes users whose function is to backup the database, view the news feed of any user, edit all spreadsheets, exchange messages, lock the user and password reset. The unregistered user only has the option to register. With the help of the developed test software, studies were carried out, which confirmed the possibility of effectively automated determination of a set of keywords in text messages with accuracy indicators for the TFIDF estimation method – 27,1% and the dispersive estimation method – 45,5%, the TFIDF estimation method with NLP 88,3%. Advantages of the developed information technology of automated definition of keywords in text messages for social networks, which conducts the analysis of text message using the methods: TFIDF estimation, variance estimation and TFIDF estimation using NLP, there is no need to use lexical databases of corpora of words, significant acceleration possibility to use for texts in different languages, possibility to use for texts with several languages. This information technology can be effectively used to analyze text messages with unknown subject and language properties.

Keywords: *text messages, TFIDF estimates using NLP, keywords.*

Постановка проблеми в загальному вигляді

За останні роки використання мережі Інтернет значно зросло, збільшивши кількість постійних користувачів. Можливості всесвітньої мережі широко використовуються у різних сферах діяльності людини, а особливої популярності за останні роки здобули соціальні мережі [1]. Сьогодні соціальні мережі широко використовуються для особистого спілкування, ведення блогів, реклами та навіть ведення бізнесу. Майже кожна компанія, від маленьких стартап-проектів і невеликих крафтових виробництв до величезних корпорацій та лідерів індустрії, має свою сторінку у соціальній мережі. Соціальні мережі стали місцем спілкування та об’єднання людей за інтересами. Там, де є велике скупчення людей, є і великі обсяги неструктурованих даних.

Тому розробка спеціалізованих соціально орієнтованих сервісів, що можуть бути майданчиком для спілкування й взаємодії окремих груп людей та пошуку інформації за сферами їх інтересів та діяльності, є актуальною на сучасному етапі. В ході розробки такої системи варто орієнтуватися на мобільні платформи, оскільки вони набирають все більшої популярності, вже ставши невід’ємною частиною повсякденного життя багатьох людей. В соціальних сервісах для спілкування відправлення та обробка повідомлень є одними із найбільш важливих функцій, тому пошук ключових слів в текстових повідомленнях та новинах соціальної мережі є актуальним.

Аналіз останніх досліджень

На даний час актуальним залишається питання визначення важливих структурних елементів тексту, що виявляються інформаційно-значущими, визначають інформаційну структуру. Використання таких елементів дозволяє формувати тезауруси, пошукові образи документів, онтології. Ключові слова – розряд високочастотної автосемантичної лексики тексту, яка складає його семантичне ядро на лексичному рівні і виступає як вектор інтерпретації тексту. Ключові слова для пошуку в тексті, опорні слова для автоматичного екстрагування значущих фрагментів текстів чи формування автоматичних рефератів, обираються з урахуванням такої властивості слів, як «дискримінантна сила». Однак, незважаючи на досить велику кількість досліджень, щоб автоматично завантажувати ключових слів є проблемою, яка донині остаточно не вирішена.

Для автоматизації пошуку ключових слів використовуються різноманітні методи аналізу текстів, таких як частотна оцінка TF, оцінка TFIDF й дисперсійна оцінка. Ці методи дозволяють співставити окремим словам чи словосполученням тексту деякі певним чином поставлені в відповідність числові вагові значення, що вказують на міру їх важливості у досліджуваному тексті. Статистичні методи пошуку ключових слів ґрунтуються на численних даних про частоту зустрічі слова в тексті. У літературі відзначається, що перевагами статистичних методів є відсутність необхідності у трудомістких процедурах побудови лінгвістичних баз знань, простота реалізації, універсальність алгоритмів вилучення ключових слів [2]. Але статистичні методи часто не забезпечують достатньої якості результатів. У ряді робіт наводяться результати досліджень, згідно яким метод дисперсійного оцінювання дозволяє одержувати найбільш релевантну множину ключових термінів у цифрових текстах [3, 4]. Є й інші статистичні підходи для виділення термінологічних сполучень. Наприклад, один з варіантів полягає в знаходження *n*-слівних поєднань за заданими частотними характеристиками. Це можуть бути значення абсолютних або відносних частот для даних словосполучень чи значення деякої статистичної міри, згідно із якою дана конструкція була знайдена і видана серед результатів.

Постановка задачі

Мета роботи полягає в розробці інформаційної технології семантичного аналізу текстових повідомлень, яка поєднує результати пошуку ключових термінів ефективними методами (TF-IDF, дисперсійного оцінювання та TFIDF з NLP), і програмній реалізації відповідної тестової інформаційної системи для дослідження практичної ефективності розробленої інформаційної технології.

Викладення основних матеріалів дослідження

Клас обробки природних мов NSLinguisticTagger (NLP) [5] у IOS SDK є інструментом штучного інтелекту та обчислювальної лінгвістики, що стосується взаємодії між комп'ютерами та людськими природними мовами. NPL пов'язана з областю взаємодії людини з комп'ютером і здатністю комп'ютерної програми розуміти людську мову. Клас NSL доступний як в Swift, так і в Objective-C, використовується для аналізу тексту природною мовою для позначення частин мови і лексичного класу, визначення імен, виконання лематизації і визначення мови й сценарію.

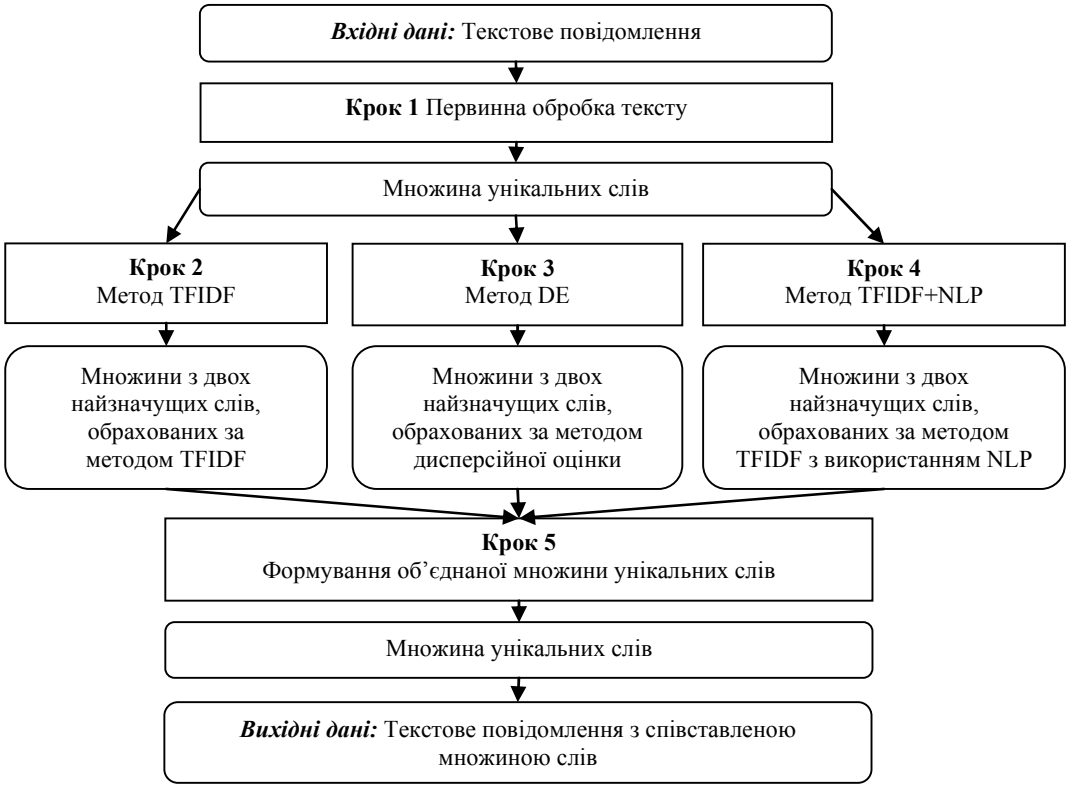

Рис. 1. Загальна схема інформаційної технології

Обробки природних мов з використанням NLP розпочинається з концепції «токенізації», тобто сегментування тексту в певну одиницю, яка може бути абзацом, реченням або словом. «Токенізація» дозволяє виконувати інші завдання, зокрема:

- розпізнавання домінуючої мови;
- частинну ідентифікацію мови – визначення того, чи може конкретне слово бути іменником або дієсловом тощо;
- лематизацію – визначення початкової кореневої форми слова;
- розпізнавання сутності назв – визначення того, чи відповідає слово чи набір слів людині, організації чи компанії або, можливо, місцезнаходженню.

Схему інформаційної технології автоматизованого пошуку ключових слів у текстових повідомленнях для соціальних мереж зображено на рис. 1. На початку виконання інформаційної технології вхідні дані отримують у вигляді текстового повідомлення, після чого виконується первинна обробка тексту (Крок 1). На даному етапі текст позбавляється розділових знаків та інших символів, формується загальна множина слів, після чого формується множина унікальних слів, обраховується кількість унікальних слів та кількість появ кожного слова в текстовому документі, також обраховується загальна кількість текстових документів, в яких дане слово зустрічається. На наступному етапі обраховується оцінка методом TFIDF (Крок 2). За допомогою сортування отримують два слова з найбільшою оцінкою.

На наступному етапі оцінки важливості слова застосовується метод TFIDF з використанням NLP (Крок 4). Даний метод відрізняється від класичного TFIDF тим, що спершу за допомогою фреймворка Apple NLP видаляються стоп-слова.

Схема методу TDIDF з використанням NLP, який в інформаційній технології викликає найбільшу цікавість [6] – рис. 2. На початковому етапі виконується частинна ідентифікація мови, відбувається визначення належності кожного слова певній частині мови (Крок 1). Також відбувається видалення тих частин мови, які не несуть важливого значення, після чого отримують множину унікальних слів.

Під час виконання наступного кроку відбувається приведення кожного слова до початкової кореневої форми, за допомогою чого вдається позбутись повторювання слів (Крок 2). Також на даному етапі відбувається видалення слів, які повторюються. На наступному етапі відбувається видалення загальновідомих слів: відомих персон, назв об’єктів тощо (Крок 3).

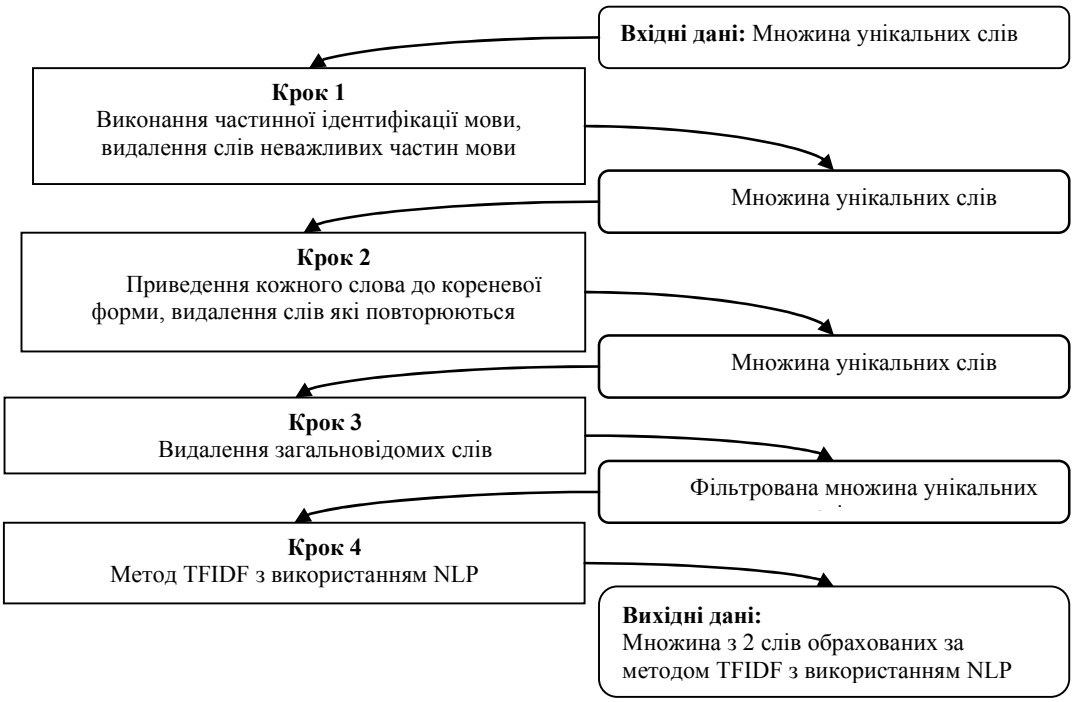

Рис. 2. Схема методу TDIDF+NLP

Наступним етапом є обрахування важливості термінів методом TDIDF. За допомогою фреймворка NLP відкидаються другорядні частини мови, всі слова приводяться до називного відмінку, а також відкидаються відомі назви, імена та об’єкти. Наступні дії відбуваються по аналогії з Кроком 2. Вихідними даними методу є множина з двох слів, обрахованих за методом TFIDF+NLP.

Дослідження ефективності інформаційної технології

Розроблена інформаційна технологія автоматизованого визначення ключових слів у текстових повідомленнях для соціальних мереж була реалізована в тестовому програмному продукті, архітектурним шаблоном якого є MVP (Model View Presenter) [7], й який відтворює роботу соціальної мережі (рис. 3). Для дослідження, вхідними даними для системи є текстове повідомлення із цифровим текстом (рис. 4), а вихідними даними є текстове повідомлення з множиною ключових термінів, відповідних досліджуваному

текстову повідомленню (рис. 5). Для написання програмного продукту на платформі IOS було використано мову програмування Swift та використано розширення NSLinguisticTagger для видалення незначущих слів. При розробці соціально орієнтованого сервісу для спілкування по інтересах на платформі IOS, відповідно до визначених функцій, виділено наступні групи користувачів: зареєстрований користувач, адміністратор соціальної мережі, незареєстрований користувач.

За результатами застосування розробленого тестового програмного продукту, що виконаний на засадах розробленої інформаційної технології було виконано дослідження ефективності інформаційної технології автоматизованого визначення ключових слів у текстових повідомленнях для соціальних мереж.

В процесі обробки контенту три переліки ключових слів, отримані за відповідними методами (TDIDF, дисперсійна оцінка, TFIDF+NLP), обмежуються за кількісним порогом й формують множини B_1, B_2, B_3. В подальшому ці множини порівнюються із множиною B_A, утвореною переліком ключових термінів, який сформовано автором. Перетин цих множин $B_k \cap B_A$ визначає ефективність відповідного методу k.

Максимальна область перетину авторського переліку зі сформованими автоматично переліками $B_k \cap B_A \rightarrow \max$ визначає найбільш ефективний метод автоматизації пошуку ключових семантичних термінів у текстових повідомленнях соціальних мереж.

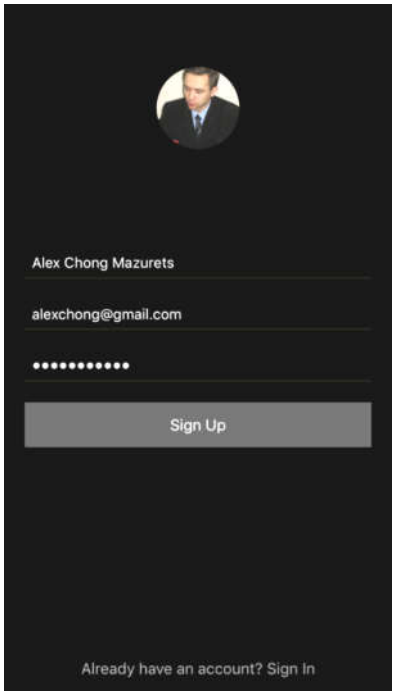

Рис. 3. Авторизація у соціальній мережі

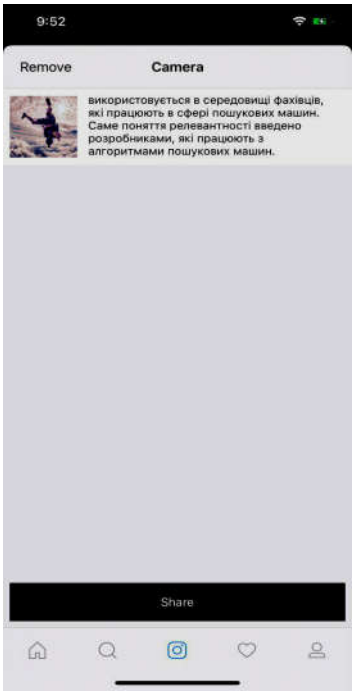

Рис. 4. Сторінка створення нового повідомлення

Рис. 5. Сторінка зі знайденими ключовими термінами

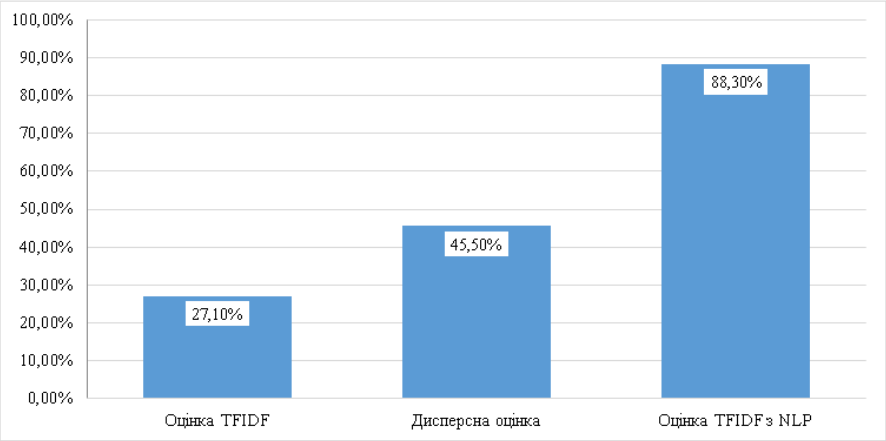

Рис. 6. Діаграма середньої ефективності методів пошуку ключових семантичних термінів у текстових повідомленнях соціальних мереж

Ефективність наведених методів пропонується визначати за наступною формулою:

$$E_k = \frac{N_{Ak}}{N_A} \cdot 100\% ,$$

де N_{Ak} – кількість термінів у авторському (B_A) та сформованому за k-м методом (B_k) переліками

термінів, що співпали ($B_k \cap B_A$); N_A – кількість термінів у переліку термінів B_k, сформованому експертом (автором).

В результаті тестування розробленим програмним забезпеченням отримують три переліки ключових термінів за відповідними методами аналізу та проводиться їх порівняння у сукупності з авторським переліком.

Загалом було досліджено 214 випадків й обраховано середню ефективність кожного із методів. Середня ефективність методу частотної оцінки склала 27,1%, методу дисперсної оцінки – 45,5% та методу TFIDF+NLP – 88,3% (рис. 6).

Отже, метод оцінки TFIDF з використанням NLP продемонстрував найвищу ефективність серед досліджуваних методів, показавши при цьому мінімальну ефективність 67,7%, максимальну – 100%.

Висновки

У статті розглянуто інформаційну технологію автоматизованого визначення ключових слів у текстових повідомленнях для соціальних мереж. При цьому використано розроблений метод TFIDF з NLP, що з використанням розширення NSLinguisticTagger дозволяє автоматизовано видаляти допоміжні та семантично незначущі слова, а також враховувати різні форми слів в одному документі.

Одержані результати дослідження ефективності інформаційної технології показали, що в переважній більшості випадків програмна система, виконана відповідно до запропонованої інформаційної технології автоматизованого визначення ключових слів у текстових повідомленнях, коректно виконала пошук ключових слів із середнім показником точності пошуку 88,3%

Література

1. Internet World Stats [Електронний ресурс]. – Режим доступу : https://www.internetworldstats.com/stats.htm.

2. Мазурець О. В. Інформаційна технологія автоматизованого визначення семантичних термінів в елементах навчальних матеріалів / О. В. Мазурець // Вісник Хмельницького національного університету. Серія: Технічні науки. – 2018. – № 3. – С. 223–230.

3. Ландэ Д. В. Компактифицированный горизонтальный граф видимости для сети слов / Д. В. Ландэ, А. А. Снарский // Труды Международной научной конференции «Интеллектуальный анализ информации ИАИ-2013. Знания и рассуждения». – Киев : КПИ, 2013. – С. 158–164.

4. Бармак О. В. Методи автоматизації визначення семантичних термінів у навчальних матеріалах / О. В. Бармак, О. В. Мазурець // Вісник Хмельницького національного університету. Сер.: Технічні науки. – 2015. – № 2(223). – С. 209–213.

5. Stanford NLP [Електронний ресурс]. – Режим доступу : https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html.

6. Житняківський В. А. Інформаційна технологія автоматизованого визначення ключових слів у текстових повідомленнях для соціальних мереж / В. А. Житняківський, О. В. Мазурець // Збірник наукових праць за матеріалами XI всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2019». – Хмельницький, 2019. – Т. 1. – С. 89–93.

7. JetBrains AppCode [Електронний ресурс]. – Режим доступу : https://www.jetbrains.com/objc/?fromMenu.

References

1. Internet World Stats [Elektronnyi resurs]. – Rezhym dostupu : https://www.internetworldstats.com/stats.htm.

2. Mazurets O. V. Informatsiina tekhnolohiia avtomatyzovanoho vyznachennia semantychnykh terminiv v elementakh navchalnykh materialiv / O. V. Mazurets // Herald of Khmelnytskyi National University. – 2018. – № 3. – S. 223–230.

3. Lande D. V. Kompaktificirovannyj gorizontalnyj graf vidimosti dlya seti slov / D. V. Lande, A. A. Snarskij // Trudy Mezhdunarodnoj nauchnoj konferencii «Intellektualnyj analiz informacii IAI-2013. Znaniya i rassuzhdeniya». – Kiev : KPI, 2013. – C. 158–164.

4. Barmak O. V. Metody avtomatyzatsii vyznachennia semantychnykh terminiv u navchalnykh materialakh / O. V. Barmak, O. V. Mazurets // Herald of Khmelnytskyi National University. – 2015. – № 2(223). – S. 209–213.

5. Stanford NLP [Elektronnyi resurs]. – Rezhym dostupu : https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html.

6. Zhytniakivskyi V. A. Informatsiina tekhnolohiia avtomatyzovanoho vyznachennia kliuchovykh sliv u tekstovykh povidomlenniakh dlia sotsialnykh merezh / V. A. Zhytniakivskyi, O. V. Mazurets // Zbirnyk naukovykh prats za materialamy XI vseukrainskoi naukovo-praktychnoi konferentsii «Aktualni problemy kompiuternykh nauk APKN-2019». – Khmelnytskyi, 2019. – T. 1. – S. 89–93.

7. JetBrains AppCode [Elektronnyi resurs]. – Rezhym dostupu : https://www.jetbrains.com/objc/?fromMenu.

Рецензія/Peer review : 29.11.2019 р. Надрукована/Printed : 16.6.2020 р.
Рецензент: д.т.н., проф. Сорокатий Р. В.