DOI: 10.35681/1560-9189.2020.22.4.225914

УДК 004.912

О. О. Дмитренко
Інститут проблем реєстрації інформації НАН України
вул. М. Шпака, 2, 03113 Київ, Україна

Побудова направлених зважених мереж термінів із застосуванням Part-of-speech tagging

Розглянуто новий метод побудови термінологічних онтологій у вигляді мереж із ключових термінів (ключових слів і словосполучень) текстів, що змістовно пов'язані з певною предметною галуззю. Виокремлення ключових слів і словосполучень з тематичних текстових потоків і подальша побудова направленої зваженої мережі термінів здійснюється на основі застосування більш широкої обробки природної мови, що базується на розбитті на частини мови (Part-of-speech tagging). Комп'ютерну обробку текстових корпусів і побудову направлених зважених мереж термінів представлено у вигляді цілісної методики. Показано апробацію запропонованої методики на прикладі відомої народної європейської казки «Little Red Cap» і побудовано направлену зважену мережу зі слів і словосполучень, які відповідають окремим ключовим поняттям у досліджуваному творі.

Ключові слова: *текстовий корпус, обробка природної мови, Part-of-speech (PoS) tagging, термінологічна онтологія, мережа термінів.*

Вступ

З початком стрімкого розвитку інформаційно-комунікаційних технологій і глобалізацією інформаційного простору щодня на інформаційних ресурсах продукуються величезні масиви текстових даних. Звичайно, серед них зростає і частка неструктурованих даних, що ускладнює пошук необхідної і релевантної інформації. Тож величезні об'єми інформаційних потоків і динамічних текстових масивів, які пов'язані з певною проблемною предметною галуззю, обумовлюють актуальність процесу концептуалізації даних, що ними супроводжуються, та їхньої подальшої формалізації у вигляді певної онтологічної моделі. А отже, і актуальною є розробка нових та удосконалення існуючих методів, які застосовуються для вирішення цього завдання.

Враховуючи той факт, що багато задач, які виникають під час роботи з текстовими інформаційними потоками, лежать на перетині між математичними науками та лінгвістикою, то це відкриває широкі можливості для застосування потужно-

го математичного апарату та лінгвістичної теорії. Наприклад, застосування знань з області дискретної математики, зокрема теорії графів і складних мереж, дає можливість представити текстові дані у вигляді мережевої моделі, яка є зручною і ефективною у використанні. Якщо говорити в термінах теорії складних мереж, то тексти визначеної тематичної спрямованості можна представити у вигляді мережі зі слів і словосполучень, пов'язаних між собою формальним смисловим зв'язком. Одним із видів цієї мережевої моделі може бути мережа, що побудована з ключових термінів (далі — мережа термінів). У ній вузли відповідають окремим ключовим поняттям предметної галузі, а ребра — зв'язкам між ними. Аналіз таких мереж може бути основою для прийняття рішень в обраній проблемній предметній галузі, адже дозволяє робити конструктивні висновки щодо предметної галузі, з якою змістовно пов'язані тексти.

Та все ж під час побудови мереж термінів відкритою та до кінця не вирішеною проблемою є визначення та виокремлення базових об'єктів (ключових термінів — слів і словосполучень). Також у зв'язку зі складністю природної мови, не менш складною та відкритою проблемою концептуалізації є встановлення семантико-синтаксичних зв'язків між вузлами мережі, що відповідають термінам, визначення напрямків таких зв'язків і встановлення їхніх вагових значень. Не менш важливою є автоматизація вищезгаданих процесів і подальша візуалізація отриманих результатів.

Мета цієї роботи — запропонувати новий метод визначення напрямків зв'язків між вузлами ненаправленої мережі, побудованої зі слів і словосполучень тематичного текстового масиву, щоби будувати термінологічні онтології у вигляді направлених мереж термінів для того, щоб у подальшому робити конструктивні висновки щодо мережевої структури та її параметрів, і на основі цього приймати ефективні рішення в проблемних предметних галузях, з якими змістовно пов'язані тексти.

Основні прийоми обробки природної мови

Текстові дані є частиною природної мови, в процесі використання якої виникає ряд проблем, які пов'язані, в першу чергу, з її багатозначністю, некомпозиційністю та самозастосованістю. Адже природна мова містить різні форми слова (словоформи, що мають спільну основу), похідні від іншого слова, та мовні вирази, які використовуються для вираження різного змісту; тож їхнє значення в конкретній ситуації залежатиме від контексту [1]. Така мова ще називається переплетеною мовою (з англ. Inflected Language). Некомпозиційність викликана відсутністю в природній мові правил, які би дозволяли визначити точне значення складного висловлювання не знаючи контекст, хоч і знаючи значення всіх інших складових слів у висловлюванні, адже деякі фрази можуть тлумачитися двояко.

Під час побудови термінологічних онтологій предметних галузей на основі текстових документів визначеної тематики [2] важливо, щоб елементи цієї формальної схеми — терміни (слова та словосполучення), які використовуються як назви концептів, що супроводжують обрану предметну галузь, підпорядковувалися принципу однозначності: слово, що використовується як назва, має бути назвою тільки одного об'єкта, якщо це одинична назва; а якщо це загальна назва, то це словосполучення має бути загальною назвою для всіх об'єктів одного класу.

У даній роботі застосовуються деякі найбільш поширені прийоми попередньої обробки текстових даних, що включають токенізацію тексту та видалення стоп-слів.

Токенізація використовується для попереднього лексичного аналізу та сегментації вхідного тексту на елементарні одиниці (токени, лексеми). Під лексемою або, іншими словами, токеном прийнято розуміти певну форму слова (словоформу) як самостійну значеннєву одиницю, яку розглядають у сукупності всіх своїх можливих форм і значень. Токенізація є зазвичай початковим етапом обробки текстів, адже дає змогу працювати зі словом як з окремою сутністю, при цьому знаючи його контекст [3].

Видалення стоп-слів застосовується для того, щоб видалити з тексту зокрема всі прийменники (наприклад, «an», «the» тощо), які можна розглядати як джерело шуму в даних, а також інші слова, що не несуть додаткового інформативного навантаження. До загальних стоп-слів відносять прийменники, частки, вигуки, сполучники, прислівники, займенники, вступне слово, числа від 0 до 9 (однозначні), інші часто вживані службові, самостійні частини мови, символи, знаки пунктуації. Відносно недавно цей список поповнили такі часто використовувані в мережі Інтернет послідовності символів, як www, com, http та ін.

Усі вищезгадані методи попередньої обробки можна легко застосувати до різних типів текстів, використовуючи стандартні бібліотеки Python NLP (Natural Language Processing), зокрема NLTK [4].

Крім того, щоб екстраполювати синтаксис мови та структуру тексту, пропонується використовувати такий прийом як розмічування частин мови (англ. Part-of-speech tagging) або просто розмічування (англ. tagging) (рис. 1). Розмічування зазвичай є наступним кроком обробки природної мови, який застосовується після токенізації, і полягає у віднесенні слова в тексті (корпусі) до певної частини мови та заснований як на його визначенні, так і на його контексті — тобто, на його зв'язку із суміжними та спорідненими словами у фразі, реченні або абзаці. Також розмічування частин мови — одна із головних і базових складових практично будь-якого завдання NLP. Для цього завдання використовується колекція тегів, які ставляться у відповідність кожному слову в реченні. PoS tagging може бути використаний для індексації слів, пошуку інформації і має також багато інших застосувань. Особливо PoS tagging може бути дуже корисним, якщо є слова або токени, які можуть мати декілька тегів. І найголовніше, розмічування спрощує контекст, який відноситься до певної предметної галузі.

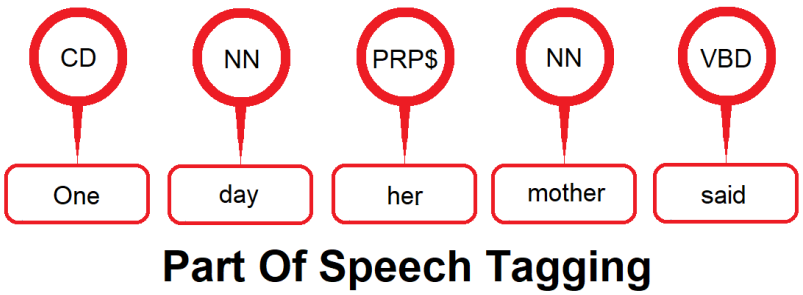

Рис. 1. Приклад розмічування частин мови [6]

Оскільки частини мови також відомі як класи слів або лексичні категорії (які базуються на синтаксичному контексті фрази), то використовуючи вищеназваний прийом класифікації слів за частинами мови, позначаємо кожне слово відповідно до його лексичної категорії.

Одним із перших і найбільш широко використовуваних англійських розбірників на частини мови є розбірник Е. Брілла [5], що використовує алгоритми на основі правил. Окрім групи алгоритмів на основі правил існують і стохастичні алгоритми.

Щоб виокремити ключові слова із тексту необхідно присвоїти їм певну числову оцінку — статистичний показник важливості. У роботі [7] показано, що під час роботи з текстовим корпусом ефективним є використання глобальної частоти терміну — GTF (Global Term Frequency). GTF визначається відношенням загальної кількості появи терміну у всіх документах корпусу до загальної кількості термінів у документах корпусу та показує, наскільки значимим є слово в глобальному контексті. Було показано, що на відміну від звичайного статистичного показника TF-IDF [8] запропонована оцінка важливості термінів дозволяє більш ефективно знаходити інформаційно-важливі елементи тексту під час роботи з текстовим корпусом заздалегідь визначеної теми, коли інформаційно важливий термін зустрічається майже в кожному документі корпусу.

Методика

Побудова направленої мережі термінів здійснюється в межах кожного окремого речення текстового корпусу.

У даній роботі для автоматичного розбиття на токени та розмічування тексту і присвоєння тегів кожному слову застосовуються відповідно окремі функції «word_tokenize» та «pos_tag» спеціалізованої надбудови — модуля NLTK (Natural Language Toolkit), що розроблений на мові програмування Python.

Також, окрім стандартних наборів стоп-слів, що доступні за посиланнями [9, 10], пропонується використовувати список стоп-слів, який сформований експертами в межах досліджуваної предметної галузі.

Запропонований у даній роботі метод визначення ключових слів і словосполучень, а також напрямків зв'язків базується на використанні результатів, отриманих за допомогою процесу класифікації слів за частинами мови та відповідним маркуванням — розмічуванням частин мови (Part-of-Speech tagging). Виходячи із практичних досліджень [4] можна помітити, що найбільш вживаними членами речення в англійській мові є артиклі (DT — determiner), іменники (NN — sing or mass noun, NNS — plural noun), займенники (PR — personal pronoun), дієслова (VB — verb base form), означення (JJ — adjectives) та прислівники (RB — adverb). Загалом ключовими словами являються окремі іменники, що зазвичай стосуються людей, місць, речей чи концептів, та іменники в парі з означеннями — словосполучення виду «JJ NN». Також у цій роботі вважається, що важливими можуть бути словосполучення виду «NN_1 NN_2», «JJ_1 JJ_2», «JJ_1 JJ_2 NN», «JJ_1 JJ_2 NN_1 NN_2». Хоча артиклі, прийменники (IN — preposition), сполучники (CC — conjunction, coordinating), окремі дієслова, прислівники та займенники являються стоп-словами, проте словосполучення виду «VV_1 to VV_2», «NN_1 IN/CC NN_2», «JJ_1 IN/CC JJ_2», «JJ NN_1 IN/CC NN_2», «JJ_1 IN/CC JJ_2 NN», «JJ_1 JJ_2 NN_1 IN/CC NN_2», «JJ_1 IN/CC

JJ_2 NN_1 IN/CC NN_2» можуть бути ключовими. Після формування вищеназваних термінів та упорядкування їх у певному порядку (формується послідовність, де словосполучення з більшою кількістю слів розташовуються перед словосполученнями та словами, які є їхньою частиною) здійснюється видалення одиничних стоп-слів (окремих артиклів, прийменників, сполучників, деяких дієслів, прислівників і займенників).

Далі за допомогою глобальної частоти терміну GTF, ідея якої описана вище, здійснюється статистичне зважування слів і словосполучень, що входять у сформовану на попередньому етапі послідовність.

Для кожного слова у порядку його зустрічання в тексті формується так званий кортеж. Кожен елемент кортежу складається з трьох значень: перше — термін (слово або словосполучення); наступне — тег, який присвоюється слову залежно від його приналежності до певної частини мови; останній елемент такого набору — числове значення GTF. Важливо зазначити, що GTF обчислюється з урахуванням двох попередніх значень — слова або словосполучення та частини мови, до якої воно належить. Кількість таких однакових кортежів у всьому тексті, що нормована на загальну кількість сформованих термінів, і визначає значення третього елемента.

На наступному кроці пропонується визначити ненаправлені зв'язки між термінами в тексті. Для досягнення цієї мети застосовується алгоритм графа горизонтальної видимості для часових рядів (Horizontal Visibility Graph algorithm — HVG) [11]. Часовим рядом у нашому випадку є послідовність числових значень GTF, що сформована на попередньому етапі. Ідея алгоритму полягає в тому, що два вузли t_i та t_j, які відповідають елементам часового ряду x_i та x_j, знаходяться в горизонтальній видимості тоді і тільки тоді, коли $x_k < \min(x_i; x_j)$ для всіх t_k таких, що $t_i < t_k < t_j$. У нашому випадку послідовність t_i, $i = 1,\ldots, n$ — це послідовність слів у межах речення (n — кількість слів, що залишились у реченні після вищеописаної попередньої обробки). HVG дозволяє будувати мережеві структури на основі текстів, у яких окремим словам або словосполученням деяким чином поставлені у відповідність числові вагові значення.

Якщо між вузлами t_i до t_j часового ряду існує ненаправлений зв'язок, встановлений за вищеописаним алгоритмом, то:

— напрямок зв'язку пропонується встановлювати від вузла t_i до t_j, якщо в реченні слово (не словосполучення), якому відповідає вузол t_i зустрічається раніше ніж термін (слово або словосполучення), якому відповідає вузол t_j;

— напрямок зв'язку пропонується встановлювати від вузла t_j до t_i, якщо в реченні словосполучення (не слово), якому відповідає вузол t_j зустрічається раніше ніж термін, якому відповідає вузол t_i (рис. 2).

Беручи до уваги принцип формування послідовності з термінів, що описаний вище, та запропоновані правила встановлення зв'язків, можна помітити, що слова та словосполучення будуть входити у відповідні словосполучення, що мають більшу кількість слів. Тобто значна частина словосполучень з більшою кількістю слів є розширенням відповідних їм словосполучень і слів. Подібний принцип побудови направлених мереж зі слів, побудова мереж природних ієрархій термінів, запропонований у роботі [12], де направлена мережа зі слів і словосполучень будується за принципом входження терміну у відповідне йому словосполучення.

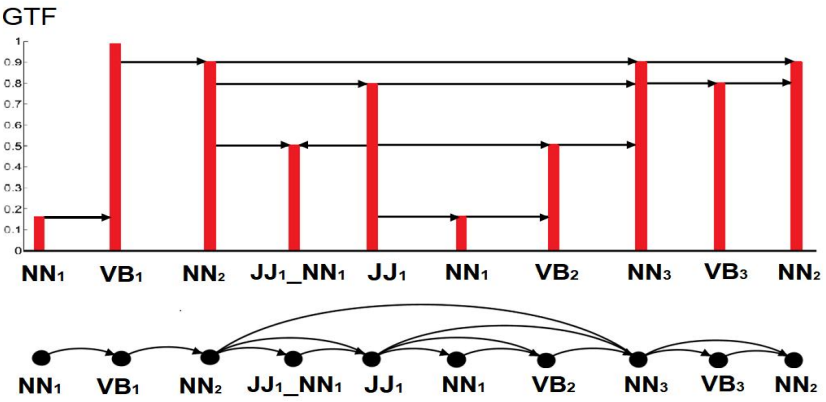

Рис. 2. Приклад побудови направленої мережі термінів

Вагові значення зв'язків між вузлами направленої мережі визначаються за запропонованим у роботі [2] принципом, який полягає у тому, що вузли, які відповідають однаковим термінам побудованої на попередньому етапі направленої мережі, об'єднуються («склеюються»), а кількість однаково направлених зв'язків між відповідними вузлами і визначає вагове значення зв'язку між цими вузлами.

Результати досліджень

Запропонована методика обробки текстових корпусів і побудови направлених зважених мереж термінів була апробована на прикладі відомої народної європейської казки «Червона шапочка» (англ. «Little Red Cap»), переказаної братами Грімм [14]. Відповідно до методики, що запропонована вище, було здійснено обробку обраного текстового документа та виокремлено ключові терміни (табл. 1).

Таблиця 1. Топ-19 ключових термінів для тексту 2Little Red Cap» та їхні числові значення GTF

№	Термін	Тег	Вагове значення
1	cap	NN	0,049
2	little	JJ	0,047
3	red	JJ	0,046
4	grandmother	NN	0,046
5	red_cap	JJ_NN	0,044
6	little_red	JJ_JJ	0,042
7	little_red_cap	JJ_JJ_NN	0,04
8	wolf	NN	0,023
9	woods	NNS	0,014
10	door	NN	0,012
11	bed	NN	0,011
12	time	NN	0,011
13	large	JJ	0,009
14	jumped	VB	0,007
15	trough	NN	0,007
16	house	NN	0,007
17	beautiful	JJ	0,007
18	big	JJ	0,007
19	flowers	NNS	0,007

Якщо упорядкувати всі ключові терміни за спаданням їхнього числового значення GTF, то на графіку (рис. 3) простежується закон Ципфа [13].

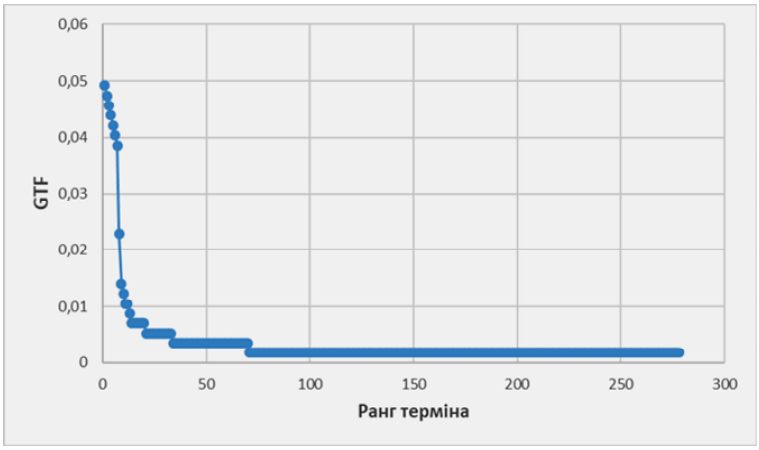

Рис. 3. Графічне зображення закону Ципфа для ключових термінів тексту «Little Red Cap»

Отримана направлена зважена мережа зі слів і словосполучень була візуалізована за допомогою засобів програмного забезпечення для моделювання та візуалізації графів — Gephi [15]. На рис. 4 представлено результати застосування запропонованої методики. Для побудованої мережі було видалено всі зв'язки, які мають вагове значення рівне 1 та вузли, вихідна та вхідна степінь яких дорівнює нулю.

Також за допомогою засобів програмного забезпечення Gephi було отримано такі параметри побудованої мережі: загальна кількість вузлів мережі — 79; зв'язків — 117; середній коефіцієнт кластеризації — 0,12; середня довжина шляху – 3,74; щільність мережі — 0.019; кількість зв'язаних компонент — 4; середня степінь — 1,48.

Список найбільш вагомих зв'язків між відповідними вузлами в мережі термінів представлений у табл. 2.

Таблиця 2. Топ-13 найбільш вагомих зв'язків для тексту «Little Red Cap»

№	Вихідний вузол	Цільовий вузол	Вагове значення
1	red	red_cap	25
2	red	cap	25
3	little	cap	25
4	little	little_red	24
5	little	red	24
6	little	red_cap	24
7	little_red	little_red_cap	23
8	cake	wine	6
9	leave	path	6
10	sick	sick_and_weak	4
11	beautiful	flowers	4
12	tasty	tasty_bite	4
13	tasty	bite	4
14	grandmother	flowers	4
15	cap	woods	3

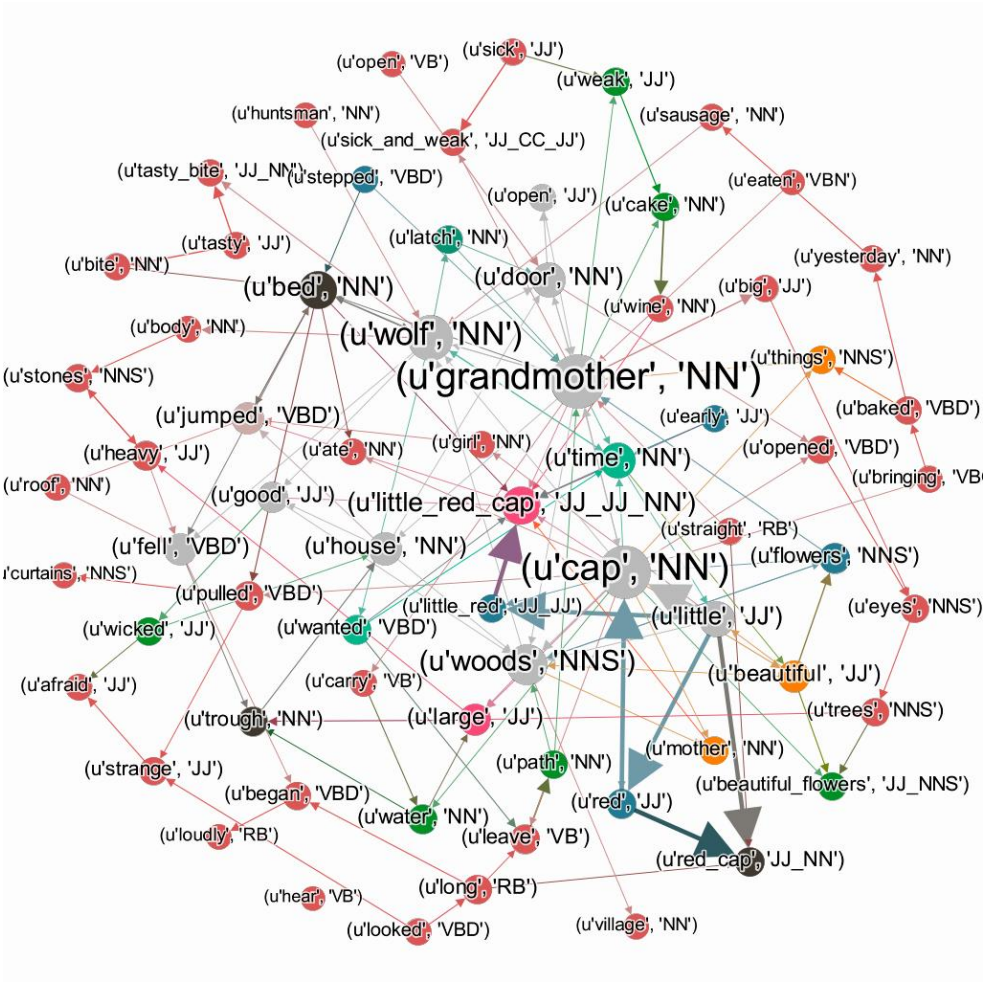

Рис. 2. Направлена зважена мережа термінів побудована для тексту «Little Red Cap» (мітки вузлів містять термін і відповідний йому тег)

Висновки

Запропоновано новий метод виокремлення ключових термінів і новий метод встановлення напрямків зв'язків із застосуванням більш широкої обробки природної мови, що базується на розбитті на частини мови (Part-of-speech tagging). Також представлено цілісну методику, що дозволяє будувати направлені зважені мережі з ключових слів і словосполучень текстового корпусу.

Апробацію запропонованої методики було проведено на прикладі відомої народної європейської казки «Червона шапочка» (англ. «Little Red Cap») братів Грімм. Проаналізувавши результати дослідження, було виявлено найбільш вагомі зв'язки між відповідними вузлами в мережі термінів, що відповідають окремим ключовим поняттям у досліджуваному творі. В межах запропонованої онтологічної моделі ключовими виявилися терміни «cap», «little» та «red», що відповідають назві твору, а найбільш вагомими, як і очікувалось, зв'язки між цими ж термінами «red → red_cap», «red → cap» та «little → cap».

У роботі для аналізу текстів застосовувалися виключно статистичні методи, проте надалі автором планується представити результати аналізу із застосуванням

більш широкої обробки природної мови, такої як виокремлення частин мови (Part-of-speech tagging), синтаксичний аналіз та інші типи лінгвістичного аналізу.

1. Никоненко А.О. Огляд комп'ютерно-лінгвістичних методів обробки природномовних текстів. *Штучний інтелект*. 2011. № 3. С. 174–181.

2. Lande D.V., Dmytrenko O.O. Creating the Directed Weighted Network of Terms Based on Analysis of Text Corpora. 2020 IEEE 2nd International Conference on System Analysis & Intelligent Computing (SAIC) (5–9 Oct. 2020, Kyiv). doi.org/10.1109/SAIC51296.2020.9239182.

3. Manning C.D., Raghavan P., & Schütze H. An Introduction to Information Retrieval. Cambridge University Press, 2009. P. 22–36.

4. Steven Bird, Ewan Klein, Edward Loper. Natural Language Processing with Python. O'Reilly Media, 2009. ISBN 0-596-51649-5.

5. Brill. E. A simple rule-based part of speech tagger. In Proceedings of the third conference on Applied natural language processing (ANLC '92). Association for Computational Linguistics, Stroudsburg. PA, USA, 1992. P. 152–155. doi:10.3115/974499.974526.

6. Extract Custom Keywords using NLTK POS tagger in python. URL: https://thinkinfi.com/extract-custom-keywords-using-nltk-pos-tagger-in-python/ (Last accessed 24.10.2020).

7. Lande D., Dmytrenko O., Radziievska O. Determining the Directions of Links in Undirected Networks of Terms. In: CEUR Workshop Proceedings (ceur-ws.org). Vol-2577 urn:nbn:de:0074-2318-4. Selected Papers of the XIX International Scientific and Practical Conference «Information Technologies and Security» (ITS 2019). 2019. Vol. 2577. P. 132–145. ISSN 1613-0073 [http://ceur-ws.org/Vol-2577/paper11.pdf].

8. Ramos J. Using tf-idf to determine word relevance in document queries. *In Proceedings of the first instructional conference on machine learning*. 2003. Vol. 242. P. 133–142.

9. Google Code Archive: Stop-words. URL: https://code.google.com/archive/p/stop-words/downloads (Last accessed 24.10.2020).

10. Text Fixer: Common English Words List. URL: http://www.textfixer.com/tutorials/commonenglishwords.php (Last accessed 24.10.2020).

11. Luque B., Lacasa L., Ballesteros F., & Luque J. Horizontal visibility graphs: Exact results for random time series. *Physical Review E*. 2009. **80**(4). doi.org/10.1103/PhysRevE.80.046103.

12. Lande D.V., Snarskii A.A., Yagunova E.V., & Pronoza E.V. The use of horizontal visibility graphs to identify the words that define the informational structure of a text. In: 2014 12th Mexican International Conference on Artificial Intelligence. 2014. P. 209–215. doi.org/10.1109/MICAI.2013.33.

13. Li Wentian. Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on information theory*. 1992. 38.6. P. 1842–1845.

14. Little Red Cap. URL: http://www.pitt.edu/~dash/type0333.html#grimm

15. Gephi. URL: https://gephi.org (Last accessed 02.12.2020).

Надійшла до редакції 04.12.2020