УДК 519-7.51
DOI: 10.20535/SRIT.2308-8893.2020.4.02

РОЗРОБЛЕННЯ ІНСТРУМЕНТАРІЮ ДЛЯ АНАЛІЗУ ТЕКСТІВ ПУБЛІЧНИХ ТА СПЕЦІАЛІЗОВАНИХ ДЖЕРЕЛ У ЗАВДАННЯХ ПЕРЕДБАЧЕННЯ ТА СИСТЕМНОГО АНАЛІЗУ

В.В. САВАСТЬЯНОВ

Анотація. Розроблено комбінований підхід до вилучення понять і побудови класифікаторів та онтологій за допомогою відкритих і пропрієтарних пакетів програмного забезпечення. Досліджено сучасні підходи, методи та моделі збереження великих обсягів слабко структурованої інформації з наборів програмного забезпечення OpenSource. Побудовано онтологію, у листях якої реалізовано класифікатор на булевих правилах із застосуванням програмного забезпечення SAS(R) Content Categorization Software. Для побудови онтології використано підхід до побудови векторів близьких понять за допомогою бібліотеки OpenSource програмного забезпечення Gensim — модель Word2Vec. Розроблено типовий алгоритм щодо побудови класифікуючої онтології. Результати дослідження можуть бути використані для побудови онтології предметних галузей, створення класифікуючих онтологій та розмічення корпусів текстів.

Ключові слова: системний аналіз, передбачення, textmining, NLP, класифікатори, онтології, OpenSource, Python, Gensim.

ВСТУП

Сьогодні кожен з нас стикається із впливом різноманітних каналів надходження інформації різної природи: голосової, візуальної, текстової. Джерела інформації не завжди є надійними або такими, що не намагаються нав'язати чиюсь суб'єктивну або комусь вигідну точку зору на ситуації або факти. Інформація сучасних інформаційних джерел викривлена, містить домисли, сплутує інформацію з гео-, спатіал- та причинно-наслідковими фактами.

На жаль, така інформація суттєво впливає не тільки на звичайних членів суспільства, а й на осіб, що приймають рішення. Особливо критичним такий вплив відчувається на території України, де є наслідки гібридних та інформаційних війн [1], тобто послідовного керованого інформаційного впливу. Іншим прикладом критичного впливу є ситуації глобальних катастроф або епідемій, коли за швидкий проміжок часу виникає великий масив протилежних думок, знань, фактів, домислів, які поширюються медіаджерелами, особливо не офіційними, проте «соціально заразними», оскільки сформовані так, що викликають емоційні почуття, а тому швидко поширюються.

Однією з причин такого інформаційного впливу є наявність у джерелах у своєму масиві аналогу «методу брейнштормy», коли у медіа конструюються будь-які «дивовижні» комбінації факторів, ситуацій, причин та наслідків. Така інформація, особливо текстова, може бути зібрана й оброблена сучасними машинними засобами оброблення мови для ідентифікації вияв-

лених факторів різної природи та використання цих знань у завданнях передбачення та системного аналізу. Пришвидшення вивчення предметної галузі, особливо у кризовій ситуації, дає змогу швидко побудувати платформу прийняття рішень, наприклад, на базі методології передбачення, для якісного супроводження проти кризових рішень та дій [2].

ЗМІСТОВНА ПОСТАНОВКА ЗАВДАННЯ

Завдання вилучення та ідентифікації факторів різної природи починається з лавиноподібного надходження інформації про ще не виявлену/формульовану предметну галузь з її взаємозв'язками, асоціативними поняттями та ситуаціями.

Інакше кажучи, потрібно виявити типові поняття, асоціації та взаємозв'язки стосовно масштабного явища чи проблеми з текстового масиву, що їх згадують або на них посилаються та обсяг яких швидко зростає.При цьому бракує експертів, або експертизи у досліджуваному явищі. Прикладом такого явища може слугувати ситуація глобальної епідемії COVID-19.

Іншим варіантом постановки завдання є типове завдання дослідницького характеру щодо змін у функціонуванні або структурі складного об'єкта/системи, як то, наприклад, мегаполіс. Вирішення питання проведення структурних/функціональних змін за наявності багатьох джерел централізованих, децентралізованих та самокерування/регулювання породжує багато явищ, сподівань, думок і спекуляцій, а також об'єктивних знань. Уся ця інформація висвітлюється у різноманітних джерелах і потребує швидкого вивчення з метою вилучення та ідентифікації факторів різної природи для побудови стратегії раціональних змін у функціонуванні або структурі складного об'єкта/системи.

Для аналізу текстів публічних та спеціалізованих джерел з метою ідентифікації факторів різної природи у завданнях передбачення та системного аналізу пропонується застосувати такі прийоми.

Етап 1. Класифікація та збирання джерел.

1. Збирання джерел (посилань на джерела).
2. Класифікація джерел.
3. Збирання примірників інформації з кожного джерела.
4. Визначення наборів метаданих, які можна вилучити з джерела.
5. Неперервне збирання інформації з джерел, вилучення змісту та метаданих, збереження.

Етап 2. Очищення та оброблення інформації з джерел.

1. Очищення текстів джерел.
2. Попереднє оброблення текстів.
3. Вилучення та вивчення концептів та понять.

Етап 3. Вилучення знань з текстів.

1. Побудова класифікуючої онтології.
2. Створення правил класифікації текстів на базі вилучених концептів та понять.
3. Статистичне оброблення результатів з метою подальшого вивчення впливу/згадування вилучених факторів на ситуації, що їх описано у текстах.

4. Передавання статистичної інформації та переліків факторів у методи якісного аналізу (наприклад, у методи якісного аналізу у складі методології передбачення).

КЛАСИФІКАЦІЯ ТА ЗБИРАННЯ ДЖЕРЕЛ: ПОТЕНЦІЙНІ ДЖЕРЕЛА СЛАБКО СТРУКТУРОВАНОЇ ІНФОРМАЦІЇ ТА ТИПИ ДОКУМЕНТІВ, ЩО МОЖУТЬ НАДХОДИТИ ЦИМИ ДЖЕРЕЛАМИ

Потенційними джерелами слабко структурованої інформації можуть бути:

– документи про аналіз стану досліджуваної системи;

– плани розвитку досліджуваної системи (у вигляді таблиці: проблема, захід, результат, бюджет);

– стенографовані аудіозаписи круглих столів та конференцій з питань розвитку;

– перелік інвестиційних проєктів для створення та/або розвитку;

– профільні публікації та огляди стану системи або схожих систем;

– витяги із засідань рад щодо системи або адміністративного чи законодавчого поля, що стосується об’єктів досліджуваної системи;

– плани розвитку об’єктів та підсистем схожих систем;

– план розвитку галузі у межах розглядуваної країни/регіону;

– паспорти передових інноваційних технологій, що застосовані або потенційно можуть бути застосовані у межах об’єктів та підсистем розглядуваної системи;

– переліки класифікаторів, статистичні таблиці характеристик та показників об’єктів та підсистем розглядуваної системи;

– новини з медіаресурсів;

– блоги компаній-розробників у досліджуваній галузі;

– патенти;

– сторінки соціальних мереж;

– публікації твітеру;

– стенограми відеоматеріалів;

– інші джерела.

Указаний набір матеріалів та звітів складається із чотирьох типів джерел:

A. Документи, які ідентифікують тип даних і переліки об’єктів, суб’єктів та систем (предметний домен).

B. Документи та звіти, що визначають, яким чином дані можуть зберігатися за допомогою стандартизованих метаданих, відомих баз даних, таксономій.

C. Документи та звіти, які визначають переліки можливих факторів, наслідків і наперед визначають тенденції у проблемних галузях.

D. Інші документи, що мають спекулятивний характер, як то блоги, твіти, новини, інтерв’ю.

Документи типу A надають набір визначень для предметної галузі: типової термінології, типових ключових слів і описують:

- концепції, стратегії і фактори, важливі для опису ситуацій у досліджуваній галузі;
- показники основних характеристик досліджуваних систем;
- фактори ризику;
- стратегії та алгоритми дій;
- визначення та пояснення, надані організаціями, урядовими установами та науковцями.

Документи типу B надають набір класифікаторів та стандартів для метаданих під час дослідження проблемних галузей, серед яких:

- стандарти баз даних предметної галузі;
- коди об'єктів;
- класифікатор IPTC;
- класифікатор КВЕД;
- класифікатори доменних галузей.

Документи типу C надають набір моделей, списків, факторів, наслідків і наперед визначають тенденції у проблемних галузях:

- моделі соціальної поведінки;
- економічні тренди;
- концепції про технологічні, соціальні та економічні уклади;
- інформаційні явища;
- негативні тренди та явища.

Документи типу D надають набір спекуляцій, що викликають явища інформаційного шуму або визначають передові ідеї у проблемних галузях:

- новини;
- спекуляції;
- інтерв'ю;
- твіти;
- негативні тренди та явища.

Документи типу D є важливими, саме вони є джерелом інформації як найбільш численне у глобальному інформаційному просторі.

АНАЛІЗ ЛЕГАЛЬНОСТІ ЩОДО ЗЧИТУВАННЯ ЗМІСТУ ДОКУМЕНТІВ ІЗ ДЖЕРЕЛ СЛАБКО СТРУКТУРОВАНОЇ ІНФОРМАЦІЇ

Можливість вилучення знань з публічних джерел слабко структурованої інформації було завжди конфліктним питанням: чи легально вилучати, копіювати, оброблювати та/або зберігати інформацію, що захищена копірайтом, чи можна взагалі переміщувати контент не у браузері і читати, а у скрипті з метою оброблення даних.

Кримінальний та адміністративний кодекси встановлюють відповідальність за порушення авторських і суміжних прав [3]. Проте 9 вересня 2019 р. у Сан-Франциско суд прийняв важливе рішення, що скрапінг публічних

сайтів не суперечить закону CFAA (Computer Fraudand Abuse Act). Інцидент було створено у ході конфлікту компаній LinkedIn та hiQ, що скрапила дані профілей LinkedIn для аналізу та консалтингу HR-агентств [4].

Постанова суду є такою: не можна перешкоджати збиранню інформації. Це накладається і на новини, оскільки ця інформація має фактуальний, а не творчий характер. Крім того, інформація не перепубліковується, а виконується аналіз та відбувається збагачення зібраної інформації.

ІНФОРМАЦІЙНА МОДЕЛЬ І КОНЦЕПТУАЛЬНА АРХІТЕКТУРА ПЛАТФОРМИ ЗБИРАННЯ ТА ЗБЕРІГАННЯ ВЕЛИКИХ ОБСЯГІВ СЛАБКО СТРУКТУРОВАНОЇ ІНФОРМАЦІЇ

Створено інформаційну модель і концептуальну архітектуру платформи збирання та зберігання великих обсягів слабко структурованої інформації (рис. 1).

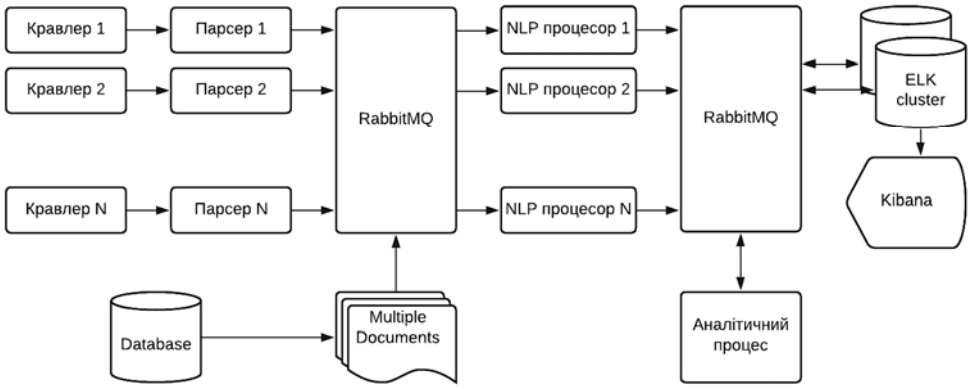

Рис. 1. Узагальнена архітектура платформи збирання та зберігання великих обсягів слабко структурованої інформації

Склад платформи:

1. Набір кравлерів — скрипти, що містять правила сканування сайтів, включаючи правила перебирання посилань.
2. Набір парсерів — скрипти будь-якою мовою програмування, що містять правила розбору текстів сайтів з метою перетворення тексту в набір первісних артефактів-метаданих: тему, текст, дату, реферат і т.ін.
3. Платформа розподілення інформаційних потоків (RabbitMQ) [5].
4. Платформа оброблення/сховище даних (Elasticsearch) [6].
5. Підсистема візуалізації даних (Elasticsearch Kibana).
6. Аналітичний процес та NLP-процесор: містить реалізацію етапів 2 і 3.

Інформація з джерел надходить:

а) через кравлери, потім парсери;

б) з бази даних.

Далі інформація надходить до черг платформи розподілення інформаційних потоків і розподілено обробляється. Потім вона знову надходить до черг і зберігається у сховищі. За необхідності візуалізація інформації відбувається через підсистему візуалізації даних (Elasticsearch Kibana).

Платформу RabbitMQ пропонується використовувати як універсальний механізм обміну та розподілення даних між модулями платформи. Вона забезпечує безпечний гнучкий механізм взаємодії між модулями, що добре масштабується за рахунок універсального механізму обміну на базі черг та джерел.

Платформа Elasticsearch — це розподілений механізм пошуку та аналітики, доступної через RESTful API. Програмне забезпечення у складі Elastic Stack централізовано зберігає дані для швидкого пошуку з широкими надбудовами релевантності, надає можливості для аналітики через додаток Kibana. Платформа Elastic Stack легко масштабується штатними вбудованими засобами і має вбудовані засоби контролю доступу.

Використовуються також засоби NLP та аналітичні процеси, що побудовані на базі бібліотек мови Python. Для аналізу використовуються пакети NLTK, Pandas, NumPy та інші, такі як gensim.

Для первісного анотування елементів слабко структурованої інформації застосовуються елементи метаданих та класи онтологій. Схема метаданих Dublin Core є найбільш вживаною і відомою [7]. Серед інших схем семантичного анотування елементів неструктурованих даних можна виокремити SIOC (Semantically-Interlinked Online Communities) [8] та SKOS (Simple Knowledge Organization System) [9]. Тому пропонується використовувати саме ці елементи метаданих та класи онтологій для первісного анотування елементів слабко структурованої інформації.

ОЧИЩЕННЯ ТА ОБРОБЛЕННЯ ІНФОРМАЦІЇ, ЩО НАДХОДИТЬ ІЗ ДЖЕРЕЛ

Очищення корпусу (скрипт мовою Python). Очищення корпусу є дуже важливим кроком. Експериментально опрацьовано три етапи, що надали найбільший результат за найменших затрат процесорного часу для різних корпусів. Ця процедура складається з таких етапів:

1. Розбиття за роздільниками — поділення тексту на фрагменти.
2. Очищення по довжині.
3. Вилучення зайвих букв.

Лематизація текстів корпусу (pymorphy2) з очищенням. У результаті процедури очищення видаляються обрані частини мови (за допомогою бібліотеки pymorphy2) [10], а саме: PREP, CONJ, PRCL, IINTJ. Приклад фрагмента обробленого тексту (у вигляді списків слів у реченнях):

['сталии', 'розвиток', 'київ', 'визначаймося', 'збалансовані', 'функціонування', 'забезпечення', 'економічні', 'зростання', 'потреба', 'населення', 'одночасні', 'поліпшення', 'екологічні', 'стан', 'міські', 'середовище', 'ціле', 'раціональні', 'використання', 'ресурс', 'число', 'природні', 'технологічні'], ….. , ['розвиток', 'економічні', 'комплекс'], ['розвиток', 'ринкова', 'інфраструктура', 'необхідна', 'формування', 'забезпечення', 'ефективні', 'функціонування', 'ринкова', 'економіка']

На жаль, відкриті засоби оброблення текстів, навіть із застосуваннями власного розробленого набору словників української мови, не такі доскона-

лі, тому серед речень трапляються артефакти. Утім вдосконалення словників не є метою цієї роботи.

Побудова моделі Word2Vec (libgensim). За допомогою libgensim [11] побудовано та порівняно 6 моделей з різними параметрами фільтрації слів за частотою, довжиною контексту/вікна пошуку, кількістю термінів, кількістю ітерацій пошуку:

1) model1 = Word2Vec(txts, min_count=1);

2) model2 = Word2Vec(txts, min_count=3);

3) model3 = Word2Vec(txts, min_count=10, size=300, iter=50, window=12);

4) model4 = Word2Vec(bitxts, min_count=10) (на основі біграм);

5) model5 = Word2Vec(txt, min_count=30, size=300, iter=50, window=22) (з вилученням додаткових грамем — ['INTJ', 'PRCL', 'CONJ', 'PREP', 'PRED', 'NPRO', None (не визначено)]);

6) model6 = Word2Vec(bitxts, min_count=3, size=300, iter=50, window=2) (біграми з вилученням додаткових грамем — ['INTJ', 'PRCL', 'CONJ', 'PREP', 'PRED', 'NPRO', None (не визначено)]).

Порівняння моделей виконується експертним методом виділення асоціацій щодо понять обраного домену (коронавірус). Експерти назвали слово і перевірялись слова-асоціації, що їх згенерувала модель. Найбільш вдалою (корисною) щодо нагадування зв'язаних слів у формуванні таксономій/онтологій виявилась шоста модель model6.

ВИЛУЧЕННЯ ЗНАНЬ З ТЕКСТІВ

У ході дослідницьких робіт проаналізовано знання щодо двох предметних доменів:

1) підземна та наземна інфраструктури мегаполіса;

2) COVID-19.

Вилучення концептуальних понять домену «Підземна та наземна інфраструктури мегаполіса». Експериментально виявлено, що найбільш зрозумілі вихідні словосполучення припадають саме на біграми як концептуальні поняття:

bigram2 = phrases.Phrases(txt5, min_count=3, threshold=10).

Результат: ['автомобіль_стоянка', 'авторськии_колектив', 'автостоянка_гараж', 'адміністративні_раион', 'аналіз_дані', 'база_геодані', 'база_дані', 'баль_фактор', 'благоустріи_облаштування', 'бортницькоа_станції', 'ботанічнии_сад', 'брикет_неутилізовані', 'будівельні_комплекс', 'будівельні_матеріал', 'будівництво_архітектура', 'будівництво_експлуатація',

...,

'якість_вода', 'якість_життя', 'імовірнісні_метод', 'імітаційні_моделювання', 'індустріальне_домобудування', 'інженерне_захист', 'інженерне_обладнання', 'інженерноа_підготовка', 'інженерноа_інфраструктура', 'інженерні_підготовка', 'інститут_київгенплан', 'інтенсивність_використання', 'ініціація_зсувні', 'існуючоа_забудова', 'існуючі_межа', 'існуючіи_забудова', 'історії_культура', 'ґрунтові_масив']

Результати виведення моделі model6 для вивчення асоціацій наведено у табл. 1.

Таблиця 1. Деякі результати виведення моделі для вивчення асоціацій і концептів

ПОНЯТТЯ	АСОЦІАЦІЇ (ВАГИ)
ТРАНСПОРТНІ_ЗАСІБ	[('резервування', 0.7325636744499207), ('рівня_автомобілізації', 0.7049006223678589), ('паркування', 0.6886059641838074), ('менше', 0.6751060485839844), ('зберігання', 0.6734187602996826), ('постіинога', 0.6668195724487305), ('тимчасові_зберігання', 0.629106879234314), ('розрахунок', 0.6270085573196411), ('легкові_автомобіль', 0.605026125907898), ('вимога', 0.6034138202667236)]
ТУНЕЛЬ	[('вплив_побудова', 0.9952846169471741), ('діоксид_азот', 0.770656943321228), ('економічність', 0.735840916633606), ('повітря', 0.7345056533813477), ('стабілізація', 0.7330552339553833), ('лення', 0.7189959287643433), ('деформація', 0.7173842191696167), ('відповідаймо_вимога', 0.7098426818847656), ('відсутні', 0.7091654539108276), ('існуючии', 0.7089160680770874)]
ПІДЗЕМНІ	[('гаражістоянка', 0.8044840097427368), ('напівпідземні', 0.7562092542648315), ('гараж', 0.7020224332809448), ('наземні', 0.6915863752365112), ('паркінг', 0.6592525243759155), ('тощо', 0.6541709899902344), ('правові', 0.6475299596786499), ('відношення', 0.643531084060669), ('пішохідні', 0.6306630373001099), ('ґрунтові', 0.6259087324142456)]

Вилучення концептуальних понять домену «COVID-19». Для корпусу текстів домену «COVID-19» найбільш якісними після перегляду виявились моделі аналізу біграм з тими ж параметрами, що і для корпусу домену «Підземна та наземна інфраструктури мегаполіса»:

bigram2 = phrases.Phrases(txt5, min_count=3, threshold=10).

Результат: ['боротися_пандемія', 'борімося_життя', 'випадок_захворювання', 'володимир_ватрас', 'встановлено_перш', 'віко_група', 'віко_рік', 'голови_київські', 'госпіталізація_хворий', 'допомога_хворий', 'екс-заступник_голови', 'ексзаступник_київські', 'закарпаття_тест', 'закупімо_тест', 'київські_ода', 'клінічні_сортування', 'коронавірусні_хвороба', 'країна_носімо', 'кількість_підтверджені', 'кількість_хворий', 'кіровоградські_область', 'лабораторно_підтверджені', 'легкі_форма', ……, 'людина_контактуймо', 'міські_клінічні', 'надання_медичні', 'нардеп_скороход', 'нардеп_слуга', 'офіційні_дан', 'перевищмо_тисяча', 'повернути_квиток', 'позитивні_результат', 'помрімо_ексзаступник', 'приватні_клініка', 'підтверджені_випадок']

Результати виведення моделі model6 для вивчення асоціацій наведено у табл. 2.

Таблиця 2. Результати виведення моделі model6 для вивчення асоціацій і концептів

ПОНЯТТЯ	АСОЦІАЦІЇ (ВЕСА)
КОРОНАВІРУС	[('година', 0.9996013641357422), ('пишімо', 0.9995890855789185), ('мена', 0.9995628595352173), ('повернімо', 0.9995517134666443), ('заходи', 0.9995511174201965), ('медицина', 0.9995511174201965), ('зайві', 0.9995459914207458), ('кашель', 0.9995441436767578), ('коронавірусні_хвороба', 0.999543309211731), ('апарат', 0.999542236328125)]
ЛІКУВАТИСЯ	[('сидімо', 0.9997162818908691), ('легкі_форма', 0.9996713399887085), ('друз', 0.9996439218521118), ('лікуймося', 0.9996432662010193), ('нея', 0.9996223449707031), ('просто', 0.9996215105056763), ('смертність', 0.9996176362037659), ('тварина', 0.9996176362037659), ('пандемія', 0.9996066093444824), ('тип', 0.9996057152748108)]
МАСКА	[('захист', 0.9989341497421265), ('пройдімо', 0.9989300966262817), ('смертність', 0.9989254474639893), ('новий', 0.9989224672317505), ('медицина', 0.9989136457443237), ('фейсбук', 0.9988991618156433), ('кордон', 0.9988906383514404), ('епідемія', 0.9988903403282166), ('легкі_форма', 0.9988902807235718), ('імунітет', 0.9988836050033569)]
КІЛЬКІСТЬ_ ПІДТВЕРДЖЕНІ	[('сягнімо_тисяча', 0.999620795249939), ('сша', 0.9995124936103821), ('київщина', 0.9994584321975708), ('катастрофа', 0.999439001083374), ('відомо', 0.9994375705718994), ('троя', 0.9994269609451294), ('країна', 0.9994232654571533), ('українське', 0.9994181394577026), ('відома', 0.9994150400161743), ('апарат', 0.9994133710861206)]
ВЛАДА	[('держслужбовець', 0.9997319579124451), ('київські', 0.9997277855873108), ('дружина', 0.9997268319129944), ('вирішімо', 0.9997234344482422), ('уряд', 0.9997172355651855), ('заразімося', 0.9997122287750244), ('більш', 0.9997105002403259), ('дорога', 0.9997104406356812), ('грош', 0.9997095465660095), ('коронавірусні_хвороба', 0.9997024536132812)]

Побудова класифікуючої онтології. Під час побудови класифікуючої онтології вирішується проблема формування структури домену нової проблеми зі словами-синонімами та асоціаціями для розміщення їх у правилах класифікатора.

Приклад виведення асоціацій і концептів, що зв'язані зі словом «захворювання» у проблемному домені «коронавірус», ілюструє рис. 2.

```
In [336]: 1 model6.wv.most_similar('захворювання')

Out[336]: [('інфекція', 0.7220960855484009),
           ('смерть', 0.7193725109100342),
           ('зараження', 0.7098100185394287),
           ('інфіковані', 0.6770030856132507),
           ('коронавірус', 0.6726216077804565),
           ('стан_березень', 0.6338338851928711),
           ('тестування', 0.6336660385131836),
           ('вірус', 0.6335515975952148),
           ('випадок', 0.6324660181999207),
           ('хвороба', 0.6273006200790405)]
```

Рис. 2. Асоціації зі словом «захворювання» у проблемному домені «коронавірус».

На основі запитів до моделі можна досить швидко побудувати класифікуючу онтологію (табл. 3) у проблемному домені «коронавірус» для подальшої генерації правил класифікації.

Таблиця 3. Приклад аналізу для побудови класифікуючої онтології

Клас	Слова/поняття/концепти
Захворювання	інфекція, зараження, вірус, хвороба, інфіковані, пандемія, поширення, спалах
Смерть	смерть, померти, вмирати
Паніка	захворіймо, помремо, черга, закупай
Обмеження	карантин, транспорт, режим, закон, заборона, поліція, надзвичайні_стан, міжнародні_перевезення

Імплементація правил у SAS® Content Categorization Studio. Приклад формування правил для класифікації у SAS® Content Categorization Studio [12] наведено на рис. 3, 4. За допомогою моделі model6 виявлено найближчі асоціації та синоніми щодо понять із предметної галузі «Епідемія коронавірусу».

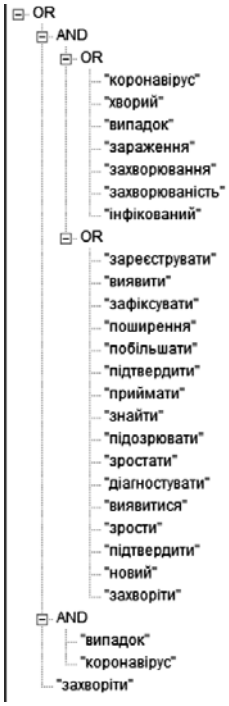

Рис. 3. Правило класифікації нових випадків захворювання

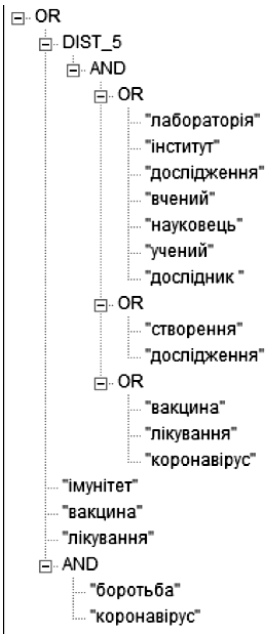

Рис. 4. Правило класифікації ситуацій щодо розроблення вакцини

Завантаження моделі до SAS® Content Categorization Server. Після формування моделі її можна скомпілювати у бінарний вигляд та завантажити до SAS® Content Categorization Server.

Це робиться через меню у SAS® Content Categorization Studio.

Маркування текстів. Кожний текст з корпусу передається до SAS® Content Categorization Server, який згідно з моделлю категоризує тексти.

Кожне правило моделі зіставляється із завантаженим текстом відповідно до моделі вилучення фактів, а потім скорингується. На виході отримуємо тексти, марковані метаданими.

Приклад роботи категоризатора на базі згенерованих правил ілюструє рис. 5.

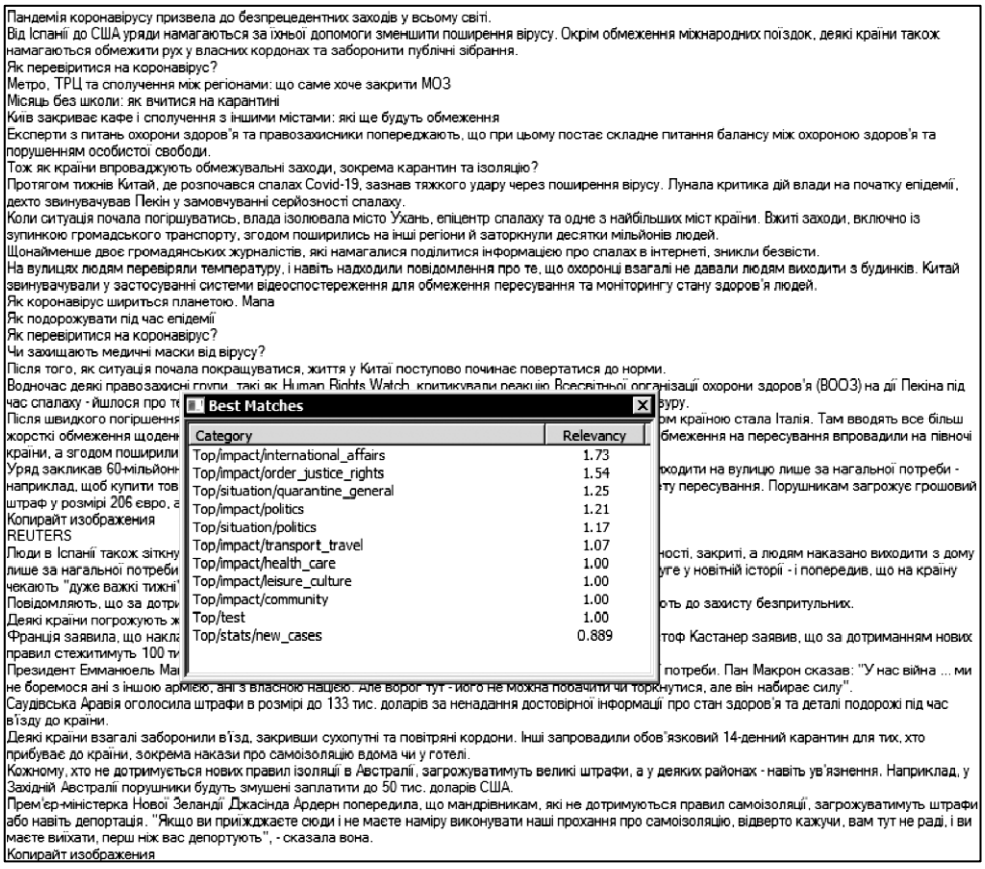

Рис. 5. Приклад роботи категоризатора: маркування тексту категоріями предметного домену «коронавірус»

АВТОМАТИЗАЦІЯ ПІДХОДУ НА ВЕЛИКИХ ОБСЯГАХ ДАНИХ (НА ПРИКЛАДІ ПРЕДМЕТНОГО ДОМЕНУ COVID)

Використання мови Python, у тому числі як клієнтської частини API до SAS® Content Categorization Server, дозволяє побудувати будь-яку архітектуру для оброблення вхідних текстів. Вихідний формат після оброблення категоризатором на базі синтезованої моделі має такий формат:

Кількість категорій, що ввійшла у документ:

Numberofcategories = 7.

Виявлена релевантна категорія із скорингом:

Category: Нові зараження (new_cases) (relevance = 11.0). Ключові слова, що було виявлено із їхніми позиціями у документі: Match (1429-1438): "вірус", Match (1429-1447): "короновірусна хвороба", Match (1622-1632): "заразився", Match (2383-2393): "випадок", Match (2509-2519): "інфіковано".

Category: Одужали (recovered) (relevance = 9.0): Match (203-215): "одужало", Match (679-696): "побороти", Match (806-822): "побороти", Match (1131-1143): "перехворіти", Match (1317-1329): "Перехворіти ", Match (3304-3325): " одужали", Match (3378-3392): "виліковано", Match (3503-3514): "Побороти", Match (4521-4533): "виписані".

Зручний вихідний формат дає змогу не тільки класифікувати фрагменти, а ще й локалізувати номери символів у тексті.

У підсумку всі дані (текст, клас, скоринговий бал) заносяться до таблиці у базу даних.

ВИСНОВКИ

Розроблено методики та прийоми щодо аналізу текстів публічних та спеціалізованих джерел з метою ідентифікації факторів різної природи у завданнях передбачення та системного аналізу. Розроблено комбінований підхід до вилучення понять і побудови класифікаторів та онтологій за допомогою відкритих і пропрієтарних пакетів програмного забезпечення.

Досліджено сучасні підходи, методи та моделі збереження великих обсягів слабко структурованої інформації з наборів програмного забезпечення OpenSource. Вивчено останні публікації за три роки за тематикою підходів до структуризації джерел слабко структурованої інформації.

Вивчено та структуровано потенційні джерела слабко структурованої інформації та типи документів, що можуть надходити цими джерелами. Вивчено можливості щодо зчитування змісту документів з джерел слабко структурованої інформації та сформовано вимоги й обмеження відповідно до форматів, способів доступу та авторських прав/ліцензій щодо легальності зчитування та копіювання інформації.

Як засоби побудови онтологій використовується ієрархічний класифікатор із набору SAS® Textual Analytics, а саме SAS® Content Categorization Studio. У складі пакета SAS® Textual Analytics є також пакет SAS® Ontology Management Studio, проте він дозволяє побудувати тільки загальну онтологію, але не має можливості застосовувати її як класифікуючу систему.

У відкритому програмному забезпеченні для класифікації, що має у складі підходи ML, є проблеми наявності початкового розміченого корпусу для тренування класифікатора й обмеженість потужності понять та концептів у корпусі типу «що маємо, те й на виході».

Досліджений комбінований підхід має переваги використання навіть малого корпусу (спеціально для дослідження на прикладі COVID-19 взято всього 300 текстів новин). Побудовано онтологію, у листя якої імплементовано класифікатор на булевих правилах. Для побудови онтології використано підхід побудови векторів близьких понять за допомогою бібліотеки OpenSource програмного забезпечення Gensim — модель Word2Vec. Протестовано декілька предметних галузей, зокрема тематику COVID-19 і «Підземну та наземну інфраструктури мегаполіса».

Розроблено типовий алгоритм побудови класифікуючої онтології.

Результати дослідження можуть використовуватись для побудови онтології предметних галузей, створення класифікуючих онтологій та марку-

вання корпусів текстів для подальшого використання для моделювання та дослідження предметних галузей, зокрема для завдань планування та передбачення в різних предметних галузях в умовах надходження великих обсягів слабко структурованої інформації.

ЛІТЕРАТУРА

1. I.V. Feskov, "NU "OUA" Basic methods of hybrid warfare in the modern information society", *Current policy issues*, vol. 58, pp. 66–76, 2016.
2. Mikhail Z. Zgurovsky and N.D. Pankratova, *System Analysis: Theory and Applications*. Springer, 2007.
3. *Articles 164-9, 164-13 Code of Ukraine on Administrative Offenses.*
4. Judge Berzon, "hiQ Labs, Inc. vs. LinkedIn Corporation Opinion", *United States Court of Appeals for the Ninth Circuit*, September 9, 2019. Available: http://cdn.ca9.uscourts.gov/datastore/ opinions/2019/09/09/17-16783.pdf.
5. *RabbitMQ*. Available: https://www.rabbitmq.com
6. *Elasticsearch*. Available: https://www.elastic.co
7. *DCMI Dublin Core Metadata Initiative.* Available: http://dublincore.org.
8. *SIOC Project.* Available: http://sioc-project.org.
9. *SKOS Simple Knowledge Organization System.* Available: http://www.w3.org/ 2004/02/skos/.
10. M. Korobov, "Morphological Analyzer and Generator for Russian and Ukrainian Languages", *Analysis of Images, Social Networks and Texts*, pp. 320–332, 2015.
11. R. Rehu˚rek and P. Sojka, "Software framework for topic modelling with large corpora", *LREC*, 2010.
12. G. Chakraborty, M. Pagolu, and S. Garla, *Text Mining and Analysis. Practical Methods, Examples, and Case Studies Using SAS®.* SAS Institute Inc., 2013.

Надійшла 16.11.2020

INFORMATION ON THE ARTICLE

Volodymyr V. Savastiyanov, ORCID: 0000-0002-2052-0420, Educational and Scientific Complex "Institute for Applied System Analysis" of the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine, e-mail: vvs.in.ua@gmail.com

DEVELOPMENT OF TEXTUAL ANALYTICS TOOLS FOR ANALYSIS OF PUBLIC AND SPECIALIZED SOURCES IN THE TASKS OF FORESIGHT AND SYSTEM ANALYSIS / V.V. Savastiyanov

Abstract. A combined approach to extracting concepts and constructing classifiers and ontologies using open and proprietary software packages has been developed. Modern approaches, methods and models of storing large amounts of poorly structured information from Open Source software sets are studied. An ontology was built, in the leaves of which a classifier based on Boolean rules was implemented using SAS(R) Content Categorization Software. To build the ontology, the approach of constructing vectors of related concepts is employed using the Open Source library of Gensim software, namely the Word2Vec model. A typical algorithm for constructing a classifying ontology has been developed. The results of the research can be used to build an ontology of subject areas, create classification ontologies and mark corpora of texts.

Keywords: systems analysis, foresight, text mining, NLP, classifiers, ontologies, Open Source, Python, Gensim.

РАЗРАБОТКА ИНСТРУМЕНТАРИЯ ДЛЯ АНАЛИЗА ТЕКСТОВ ПУБЛИЧНЫХ И СПЕЦИАЛИЗИРОВАННЫХ ИСТОЧНИКОВ В ЗАДАЧАХ ПРЕДВИДЕНИЯ И СИСТЕМНОГО АНАЛИЗА / В.В. Савастьянов

Аннотация. Разработан комбинированный подход по извлечению понятий и построения классификаторов и онтологий с помощью открытых и проприетарных пакетов программного обеспечения. Исследованы современные подходы, методы и модели хранения больших объемов слабоструктурированной информации из наборов программного обеспечения OpenSource. Построена онтология, в листьях которой реализован классификатор на булевых правилах с применением программного обеспечения SAS (R) Content Categorization Software. Для построения онтологии используется подход построения векторов близких понятий с помощью библиотеки Open Source Gensim, а именно модель Word2Vec. Разработан типовой алгоритм построения классифицирующей онтологии. Результаты исследования могут быть использованы для построения онтологий предметных областей, создания классифицирующих онтологий и разметки корпусов текстов.

Ключевые слова: системный анализ, предвидение, text mining, NLP, классификаторы, онтологии, OpenSource, Python, Gensim.