УДК 004.04 DOI: 10.31673/2412-4338.2020.011673

Голуб Т.В., Зеленьова І.Я., Грушко С.С., Луценко Н.В. *Національний університет «Запорізька політехніка», Запоріжжя*

ПРОГРАМНА РЕАЛІЗАЦІЯ АВТОМАТИЧНОГО КЛАСИФІКАТОРА ТЕКСТІВ НА ОСНОВІ УТОЧНЕНОГО МЕТОДУ ФОРМУВАННЯ ПРОСТОРУ ОЗНАК КАТЕГОРІЙ

У статті запропоновано рішення однієї із задач комп'ютерної лінгвістики, описана теоретична розробка і програмна реалізація уточненого методу формування простору ознак категорій при класифікації текстів за тематикою. Виконано дослідження ефективності запропонованого методу при використанні в процесі класифікації текстових документів.

В рамках однієї тематики спостерігається використання однакової термінології в декількох категоріях, що підвищує складність процесу класифікації. Особливість уточненого методу полягає в тому, що він дозволяє виконати класифікацію документів за категоріями загальної тематики і таким чином отримати більш точний результат.

Уточнений метод формування простору ознак категорій включає в себе етапи попередньої обробки тексту і формування простору ознак. Етап попередньої обробки тексту характеризується залежністю даного процесу від мови тексту, що зумовлює використання алгоритмів, спеціалізованих для окремих мов. В даному дослідженні розглядаються тексти українською мовою. Стемінг, як один з кроків попередньої обробки тексту, побудований на основі адаптованого методу для текстів українською мовою. Він враховує особливості синтаксису і словотворення в даній мові. Формування простору ознак категорій виконується на основі методу TF-SLF, який враховує входження слів в кожну категорію, а також подальшої фільтрації отриманого простору на основі порогового значення, що відображує важливість кожного слова для певної категорії.

В результаті послідовного виконання всіх етапів уточненого методу формується простір ознак окремих категорій, з яких виключаються малоінформативни терми. Це призводить до зменшення кількості ітерацій і розрахунків при подальшій класифікації, що в свою чергу веде до загального скорочення часових витрат на рішення задачі.

На основі запропонованого авторами уточненого методу формування простору ознак розроблено програмний комплекс, за допомогою якого підтверджено доцільність практичного застосування методу.

Ключові слова: *класифікація тексту, попередня обробка тексту, стемінг, фільтрація, простір ознак категорій.*

Golub T.V.,. Zeleneva I.Ya, Hrushko S.S., Lutsenko N.V. *National University "Zaporizhzhya Polytechnic", Zaporizhzhia*

SOFTWARE IMPLEMENTATION OF THE AUTOMATIC TEXT CLASSIFIER BASED ON THE SPECIFIED METHOD OF FORMING A FEATURES SPACE FOR CATEGORIES

The article proposes a solution to one of the tasks of computer linguistics such as a text classification. Theoretical development and software implementation of the specified method for forming the space of category attributes is considered in this article. A study of the effectiveness of this method when used in the classification of text documents is carried out.

The peculiarity of this method is that it allows the classification of documents into categories of general subjects and thus clarifies the result. Within the framework of one subject, the use of the same terminology in several categories is observed. This increases the complexity of the classification process.

The specified method for creating the category attribute space includes two stages: preliminary processing of the text; creation of the attribute space. The stage of preliminary text processing is characterized by the dependence on the language of initial text, which determines the use of algorithms

specialized for individual languages. This study examines texts in Ukrainian. Stemming, as one of the steps of text preprocessing, is built on the basis of an adapted method for texts in Ukrainian. It takes into account the peculiarities of syntax and word formation (morphology) in a given language. The creation of the category attribute space is performed on the basis of the TF-SLF method, which takes into account the occurrence of words in each category, and further filtering the resulting space based on a threshold value that reflects the importance of the word for the category.

As a result of consistent implementation of the specified method stages, a space of features for different categories is reduced, and the little-informative terms are excluded from this space. This allows decreasing in the number of iterations and calculations upon further classification, which leads to a reduction in the total time spent on solving the problem.

Based on the specified method proposed by the authors, a software package was developed, which is then used to confirm the effectiveness of this method.

Keywords: *text classification, word preprocessing, stemming, filtering, category attribute space.*

Голуб Т.В., Зеленева И.Я., Грушко С.С., Луценко Н.В. *Национальный университет «Запорожская политехника», Запорожье*

ПРОГРАММНАЯ РЕАЛИЗАЦИЯ АВТОМАТИЧЕСКОГО КЛАССИФИКАТОРА ТЕКСТОВ НА ОСНОВЕ УТОЧНЕННОГО МЕТОДА ФОРМИРОВАНИЯ ПРОСТРАНСТВА ПРИЗНАКОВ КАТЕГОРИЙ

В статье предложено решение одной из задач компьютерной лингвистики, описана теоретическая разработка и программная реализация уточненного метода формирования пространства признаков категорий при классификации текстов по тематике, выполнено исследование его эффективности относительно процесса классификации текстовых документов.

Особенностью данного метода является его направленность на классификацию документов по категориям общей тематики. В рамках одной тематики наблюдается использование одинаковой терминологии в нескольких категориях. Это обуславливает сложность процесса классификации.

Уточненный метод формирования пространства признаков категорий включает в себя этапы предварительной обработки текста и непосредственное создание пространства признаков. Этап предварительной обработки текста характеризуется зависимостью данного процесс от языка текста, что обуславливает использование алгоритмов, специализированных для отдельных языков. В данном исследовании рассматриваются тексты на украинском языке. Стемминг, как один из шагов предварительной обработки текста, построен на основе адаптированного метода для текстов на украинском языке. Он учитывает особенности синтаксиса и словообразования в данном языке. Создание пространства признаков категорий выполняется на основе метода TF-SLF, который учитывает вхождение слов в каждую категорию, а также дальнейшей фильтрации полученного пространства на основе порогового значения, учитывающего важность слова для категории.

На основе предложенного авторами метода формирования пространства признаков разработано программный комплекс, тестирование которого подтвердило целесообразность практического использования метода.

Ключевые слова: *классификация текста, предварительная обработка текста, стемминг, фильтрация, пространство признаков категорий*

Вступ

Час інформаційних технологій характеризується подальшим зростанням обсягів інформації, в тому числі в текстовій формі. Для аналізу такого різноманіття існує потреба в систематизації цих даних. Одним із способів такої систематизації є класифікація.

Класифікація текстів за змістом є однією із ключових задач комп'ютерної лінгвістики. Вона залишається актуальною і надалі, незважаючи на значний внесок науковців світу у її вирішення. Це пов'язане зі складністю процесів попередньої підготовки та безпосереднього процесу класифікації текстів.

Окремим питанням стоїть класифікація текстів в одну з категорій спільної тематики. В

рамках однієї тематики спостерігається використання однакової термінології в декількох категоріях, що зумовлює ускладнення процесу визначення приналежності тексту до певної категорії.

В результаті, для досягнення необхідної точності результатів роботи класифікатора, використовуються складніші правила прийняття рішень. Така складність правил висуває високі вимоги до апаратних ресурсів і потребує значних часових витрат на виконання задачі. Тому для забезпечення роботи класифікаторів в реальному масштабі часу потрібно вживати певних заходів щодо прискорення комп'ютерної обробки текстової інформації.

1. Аналіз літератури

На даний момент основними напрямками по прискоренню роботи класифікаторів текстів є вдосконалення процесів попередньої обробки вхідного тексту [1], виділення основних ознак для прийняття рішення [2] та саме класифікації [3]. Кожен з напрямків характеризується своїми задачами і вимагає окремих підходів для їх вирішення.

Попередня обробка тексту складається з етапів токенізації, видалення малоінформативних слів, стемінгу.

Токенізація [4] виконує розподіл тексту на структурні елементи. Для текстових документів такими елементами можуть бути абзаци, речення, словосполучення, слова. В даному випадку виділяються слова.

Видалення малоінформативних слів [5, 4] полягає в «очищенні» тексту від сполучників, прийменників, займенників та інші. В результаті розмір тексту скорочується, що дозволяє підвищити швидкість його обробки.

Під стемінгом [5] розуміється перетворення слова шляхом відсікання закінчень та суфіксів у відповідності до граматичних особливостей мови написання похідного тексту. В результаті виконується виділення основи слова (терм), яка буде незмінною для ряду форм незалежно від відмінку та наявних суфіксів. При цьому отримана форма слова не обов'язково відповідатиме його кореню. Для реалізації даного методу необхідно описати правила словотворення для певної мови, тексти якої аналізуються. У зв'язку з цим виникає необхідність в розробці алгоритму стемінгу для кожної мови окремо, в тому числі для української мови [6].

Отримані в результаті попередньої обробки тексти використовуються для формування простору ознак категорій. На основі зазначеного простору ознак виконується подальша класифікація текстів.

Для формування простору ознак кожної категорії використовується опорна вибірка [5, 7, 8], яка представляє собою сукупність текстів, категорії яких заздалегідь визначені експертами.

На основі опорної вибірки спочатку визначається ступінь інформативності кожного терма для окремої категорії. Рішення про ступінь інформативності терма приймається на основі величини його вагового коефіцієнта [8, 9].

Серед найбільш розповсюджених методів визначення вагових коефіцієнтів термів [2, 7, 10] для розробки методу програмного прискорення класифікації текстів в даній роботі обрано метод TF-SLF (формула 1) [7, 11, 10], який враховує важливість кожного окремого терма для певної категорії.

$$F_{TSL}^{t} = TF_{tc} \cdot SLF_{t}, \tag{1}$$

де TF_{tc} — частота терма в рамках категорії (формула 2), визначається як відношення числа входжень терма t_i в категорію с (n_i) до загальної кількості слів в категорії ($\sum_k n_k$) [10]:

$$TF_{tc} = \frac{n_i}{\sum_k n_k}, \tag{2}$$

SLF_t — логарифмована сума частот терма t (формула 3). Даний параметр показує усереднене значення SLF для кожного терма в рамках всієї опорної вибірки, а використання

логарифмованого значення дозволяє компенсувати дисбаланс між категоріями з малою кількістю документів та категоріями з великою кількістю документів.

$$SLF_t = \log \frac{|C|}{\sum NDF_{tc}}, \quad (3)$$

де NDF_{tc} — нормалізована частота зустрічаємості терма t в окремій категорії (формула 4). Дана оцінка є локальною для категорії.

$$NDF_{tc} = \frac{DF_{tc}}{d_c}, \quad (4)$$

де DF_{tc} — число документів категорії, в яких зустрічається хоча б один раз терм t; d_c – кількість документів в категорії.

В результаті підрахунку параметра TF-SLF (F_{TSL}) визначаються вагові коефіцієнти термів з урахуванням їх входження в категорії [11, 10]. Таким чином, формується перелік термів з їхніми ваговими коефіцієнтами для кожної категорії окремо.

На основі вагових значень термів кожної категорії виконується класифікація текстів. Існує багато методів побудови класифікаторів. В загальному плані їх можна поділити на дві групи: основані на штучному інтелекті та аналітичні [12].

Методи класифікації, основані на штучному інтелекті, в основному реалізовані з використанням штучних нейронних мережах (ШНМ, artificial neural networks, ANN) [14, 15]. Застосування ШНМ характеризується складністю визначення вагових значень зв'язків між нейронами, великою кількістю нейронів і, відповідно, великою кількістю розрахунків. [16]

До найпоширеніших аналітичних методів, які використовуються для класифікації текстів і показують хороші результати [17], згідно з описами в літературних джерелах, належать наївний метод Баєса (Naive Bayes, NB) [3, 13] і метод опорних векторів (Support Vector Machine, SVM) [18].

При цьому, метод Баєса в порівнянні з методом опорних векторів показав кращі результати при роботі з невеликим обсягом документів [19], що часто спостерігається при необхідності побудови класифікатора по близьким категоріям спільної тематики.

2. Постановка задачі

Метою даного дослідження є скорочення часових витрат при класифікації текстових документів по категоріям спільної тематики.

Задача дослідження полягає в розробці та програмній реалізації уточненого методу формування простору ознак категорій, а також в дослідженні його ефективності стосовно прискорення процесу класифікації текстових документів.

3. Розробка уточненого методу формування простору ознак категорій

Особливістю даного методу є те, що класифікація виконується «прецезійно», тобто уточнено по декількох категоріях в межах однієї спільної тематики.

Уточнений метод формування простору ознак категорій має комплексну структуру та включає в себе елементи, розроблені та опубліковані авторами в [23, 20, 21, 22]. До них відносяться адаптований алгоритм стемінгу україномовних текстів та спосіб уточнення порогового значення при фільтрації простору ознак категорій.

Надамо коротку характеристику основної сутності зазначених методів.

3.1 Адаптований алгоритм стемінгу україномовних текстів

Для формування простору ознак категорій та подальшої безпосередньої класифікації необхідно провести попередню обробку тексту.

При цьому одним з етапів є стемінг, особливістю якого є залежність від мови тексту. В даній роботі в якості опорної мови було обрано українську. Зміна направленості стемінгу на іншу мову супроводжується лише відповідною зміною алгоритму стемінгу.

Адаптований алгоритм стемінгу україномовних текстів побудований на основі аналізу описів, наведених в роботах [5, 23] та враховує особливості словотворення в українській мові. Суттєві особливості даного алгоритму полягають в наступному:

– реалізовано вилучення апострофа з метою зменшення кількості символів в словах для подальшої обробки;

– змінено послідовність етапів відносно відомого алгоритму, а саме спочатку виконується етап перевірки на наявність постфіксу зворотної форми, потім перевіряється наявність закінчень дієприслівника, що дозволяє оптимізувати процес аналізу слів у відповідності з особливостями словотворення в українській мові та зменшити перелік отриманих форм слів (термів);

– вилучено етап видалення подвоєних приголосних, що зумовлено наявністю подвоєнь приголосних лише при співпадінні літер закінчення кореня і початку суфіксу і дозволяє скоротити час на виконання стемінгу.

В результаті використання даної модифікації скорочуються часові витрати на етапі обробки слів шляхом скорочення кроків при виконанні стемінгу.

3.2 Спосіб уточнення порогового значення при фільтрації простору ознак категорій

Для скорочення часових витрат на подальшу класифікацію текстів доцільно також зменшити простір ознак кожної категорії. Для цього пропонується виконання додаткової фільтрації на основі розробленого авторами уточненого порогового значення, описаного в [21, 22]. Метою такої фільтрації є видалення з простору ознак термів, які не дозволяють відокремити специфіку даної категорії серед інших, близьких за тематикою, тобто низько інформативних термів.

В якості порогового значення використано емпіричну величину, зворотну до кількості документів опорної вибірки для категорії, яка аналізується (формула 5).

$$k_j = \frac{1}{d_j}, \tag{5}$$

де d_j – кількість документів опорної вибірки для категорії j.

Після цього значення вагових коефіцієнтів термів для окремої категорії порівнюються з пороговим значенням для тієї ж категорії (формула 6).

$$\psi(e_i, c_j) = \begin{cases} 0, & якщо \quad \psi(e_i, c_j) < k_j \\ \psi(e_i, c_j), & якщо \quad \psi(e_i, c_j) \geq k_j \end{cases}, \tag{6}$$

де $\psi(e_i, c_j)$ – вагове значення *i*-го терма із загального словника *e* в рамках *j*-ї категорії множини всіх категорій *c*.

Якщо ваговий коефіцієнт певного терма більше порогового значення, він залишається без змін в рамках простору ознак даної категорії. Якщо цей параметр менше порогового значення, то він для даного терма в межах даної категорії замінюється на нульове значення, тобто, таким чином, виключається із простору ознак категорії.

Таке зменшення простору ознак кожної категорії опосередковано призводить до скорочення часових витрат на процес класифікації текстових документів в цілому.

3.3 Уточнений метод формування простору ознак категорій

Враховуючи описані вище елементи, які є важливими складовими, уточнений метод формування простору ознак категорій включає наступні етапи:

1. Виконання токенізації і видалення малоінформативних слів.
2. Стемінг україномовного тексту за адаптованим алгоритмом.
3. Формування простору ознак:

3.1. Визначення вагового коефіцієнту кожного терма простору ознак за методом TF-SLF для кожної категорії окремо (формула 1).

3.2. Фільтрація простору ознак для категорій на основі уточненого порогового значення ступеню інформативності термів для кожної категорії.

Таким чином, виконання зазначених етапів формує простір ознак окремих категорій, з

яких виключаються малоінформативні терми. Це призводить до зменшення кількості ітерацій та розрахунків при подальшій класифікації, що опосередковано скорочує зальні часові витрати на вирішення задачі.

З метою дослідження запропонованого методу виконано програмну реалізацію автоматичного класифікатора текстів, побудовану на його основі.

4. Програмна реалізація автоматичного класифікатора текстів на основі уточненого методу формування простору ознак категорій

Програмна реалізація класифікатора текстових документів, виконана авторами, структурно та функціонально відповідає трьом основним, доволі складним етапам: попередня обробка тексту, формування простору ознак категорій та класифікація (рис.1).

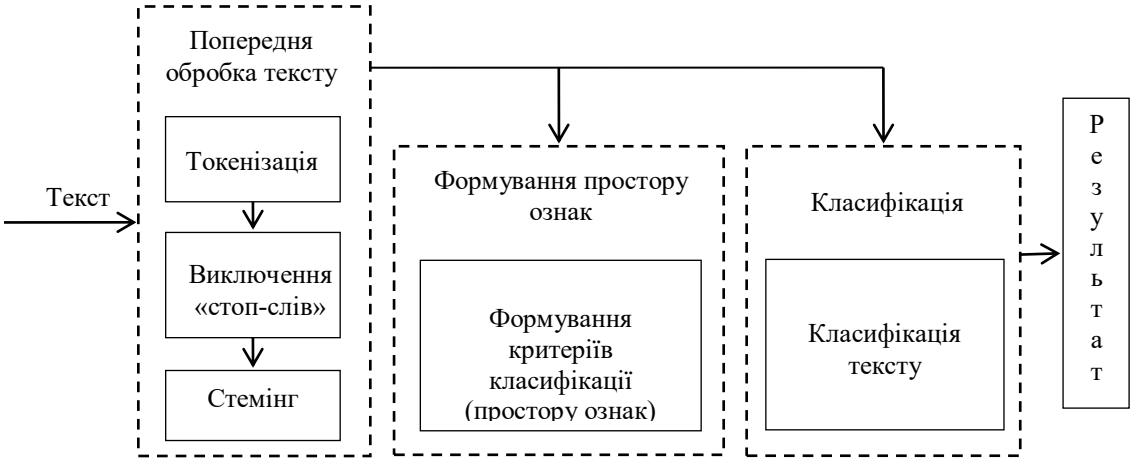

Рис. 1. Структура програмного комплексу реалізації класифікатору

На етапі формування простору ознак категорій в якості вхідних даних виступають текстові документи, категорії яких визначені заздалегідь експертами. При цьому над кожним документом виконується попередня обробка тексту і після зчитування всіх документів на основі отриманих формується простір ознак категорій.

Етап класифікації виконується для кожного документу окремо, шляхом визначення його приналежності певній категорії. Такий документ також підлягає попередній обробці тексту, після чого визначається категорія.

Таким чином, в обох випадках перший крок виконується однаково, що є умовою подальшої коректної класифікації. В даній роботі одним із етапів попередньої обробки тексту є адаптований алгоритм стемінга для україномовних текстів [20], коротко охарактеризований вище.

Програмна реалізація процесу формування простору ознак категорій побудована на основі розробленого уточненого методу. Вона складається з наступних блоків: блок переліку документів; блок попередньої підготовки тексту; блок формування і фільтрації простору ознак для кожної категорії. Алгоритм процесу наведено на рисунку 2.

На вхід програми, в блок переліку документів, подається перелік категорій спільної тематики (процес 1) і документів, які до них належать (цикл 3-6).

Блок попередньої підготовки тексту (цикл 8-18) виконує попередню обробку тексту для кожного документу даної категорії (цикл 10-15) і формування переліку термів цієї категорії. Попередня обробка текстів полягає в зчитуванні кожного тексту в рамках категорій (блок 11), його обробки (блок 12) та формування переліку термів категорій (блок 13). Обробка тексту включає розділення вхідного тексту на частини (токенізацію) з паралельним видаленням не літерних символів. Токенізація виконується за словами. Наступним етапом є стемінг, який застосовується для кожного слова окремо. Стемінг виконується у відповідності із запропонованим алгоритмом. По завершенню цього етапу текст представляє собою перелік термів, які складаються з літерних символів. На основі результатів етапу стемінга

формується перелік термів документу (блок 13). Даний перелік включає всі варіації термів та скільки разів вони зустрічаються в тексті. Після завершення обробки поточного тексту отриманий перелік термів додається до переліку термів відповідної категорії із збереженням інформації про зустрічаємість термів в документі (блок 16). В результаті формується перелік термів категорій з частотою їх зустрічаємості в кожному документі категорії.

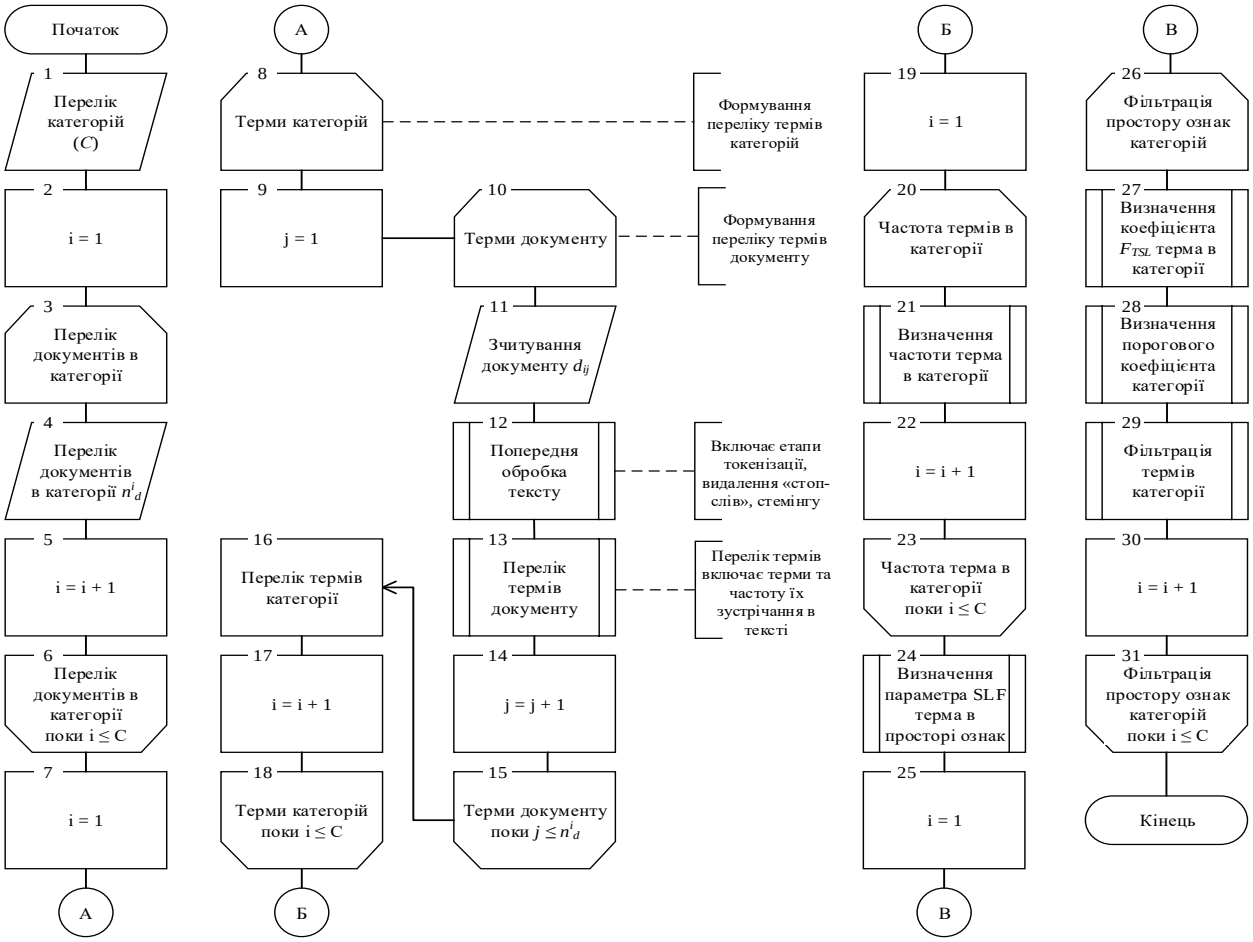

Рис. 2. Алгоритм процесу формування простору ознак

Наступний блок формування і фільтрації простору ознак складається з двох етапів: отримання параметра SLF для кожного терма простору ознак всіх категорій (блоки 20-24); формування і фільтрація простору ознак кожної категорії окремо (цикл 26-31). На першому етапі на основі отриманого переліку термів кожної категорії (цикл 20-23) виконується розрахунок параметра SLF (блок 24), який використовується для врахування важливості окремих термів для категорій. Другий етап забезпечує безпосереднє формування простору ознак окремих категорій (цикл 26-31) з виконанням фільтрації цього простору у відповідності із запропонованим уточненим методом (блок 28-29).

Процес класифікації текстових документів реалізовано на основі принципів наївного методу Баєса [3], які полягають в тому, що кожний терм зустрічається в категорії незалежно від інших термів, а також важливість термів для категорії характеризується їх ваговими коефіцієнтами. Тоді класифікація тексту в одну з описаних категорій виконується на основі обчислювання суми добутків вагових значень термів, які співпали в документі та в опорній вибірці категорії. Алгоритм даної частини програми наведено на рисунку 3.

Перед початком процесу класифікації в програму завантажується перелік категорій та опорні вибірки кожної категорії (блоки 1, 2). Далі задається перелік документів для класифікації (блок 3). Після цього кожний документ аналізується окремо для визначення його категорії (цикл 5-17). Для певного документу виконуються зчитування, попередня

обробка і формування переліку термів документу (блоки 6-8). Отриманий перелік термів використовується при класифікації (цикл 10-13). Для цього кожний терм документу порівнюється з переліком термів простору ознак певної категорії. Якщо терм в просторі ознак присутній, вагове значення терма в документі помножується на вагове значення терма в категорії. Результати добутку термів, які співпали, додаються. Тоді текст документу належить до тієї категорії, отримана сума добутків якої буде максимальною (блок 14).

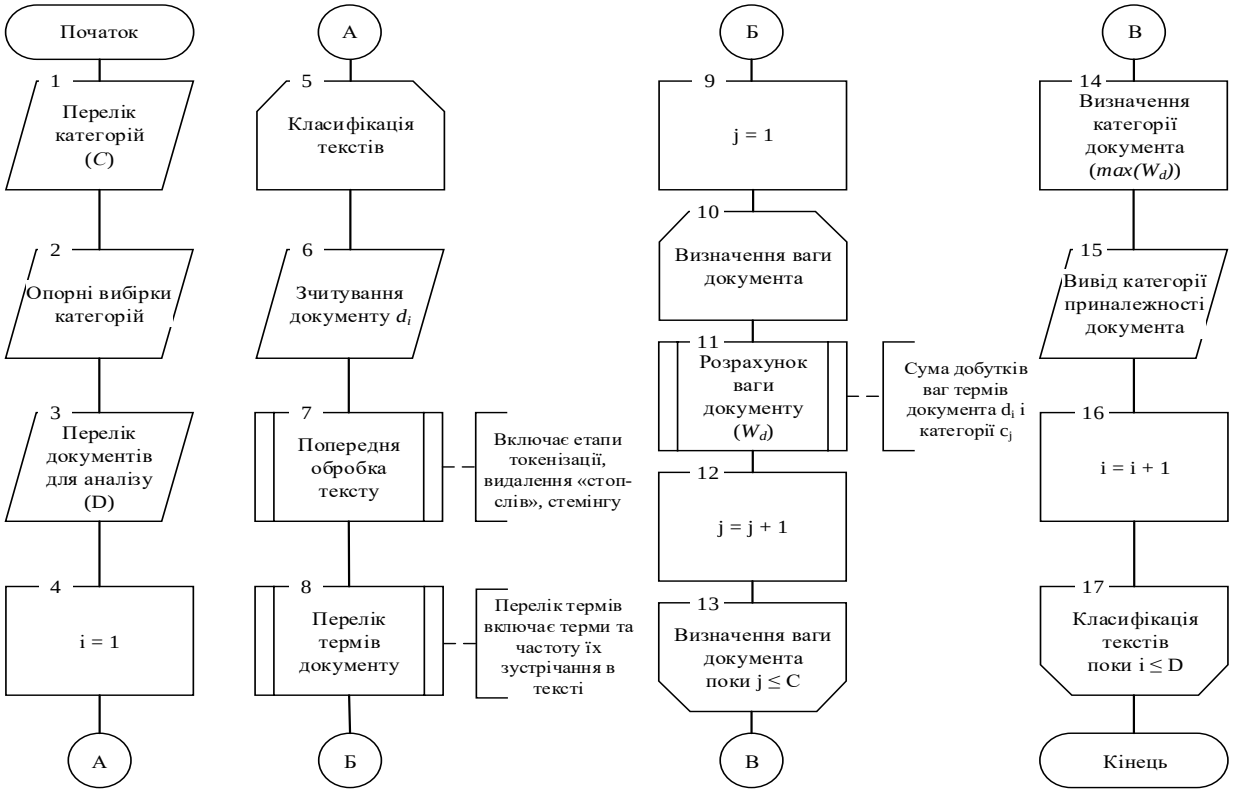

Рис. 3. Алгоритм процесу класифікації текстів

Дана програма була реалізована на мові програмування Python. Зазначена мова програмування характеризується [24, 25] наступними властивостями, зручними для даної задачі: мова є кросплатформовою, що дозволяє запускати програмні рішення на більшості сучасних операційних системах; доступна вбудована бібліотека для виконання основних функцій обробки текстових документів, адаптована до більшості розповсюджених мов.

5. Основні результати досліджень

В якості вхідних даних для тестування програми було використано документи однієї тематики УДК [26] 004 «Комп'ютерна наука та технологія. Застосування комп'ютера», а саме: 004.0 «Спеціальні визначники для позначення застосування комп'ютера», 004.7 «Комп'ютерний зв’язок. Комп’ютерні мережі», 004.8 «Штучний інтелект», 004.9 «Прикладні інформаційні (комп’ютерні) системи». Формування простору ознак виконувалося на основі україномовних текстових документів із загальною кількістю слів в середньому близько 20000 для кожної категорії. Кожен документ представляє собою статтю, категорія якої була визначена її авторами на момент публікації. Тестування проводилося на основі текстових документів різних категорій приналежності із зазначеного переліку.

В таблиці 1 наведено отримані результати класифікації десяти текстових документів різного розміру і категорії та відповідні часові витрати на даний процес.

Середній час на виконання класифікації відповідних текстових документів при використанні уточненого методу визначення простору ознак скоротився на 20%. Далі наведено графічне представлення порівняння часових витрат на класифікацію окремо кожного документу (рис. 4) та середнє значення часових витрат (рис. 5).

Таблиця 1

Результати порівняння часових витрат на класифікацію

Категорія тестового документу	Кільк. слів в докум., шт.	Відомий метод		Уточнений метод		Відсоток зменшення часових витрат
		Визначена категорія	Витрачений час, с	Визначена категорія	Витрачений час, с	
1	2	3	4	5	6	7
004.032.26	891	004.0	0,004002	004.0	0,003602	-11%
004.93	1362	004.9	0,003803	004.9	0,003402	-12%
004.94	1645	004.9	0,005604	004.9	0,005203	-8%
004.056.53	2189	004.0	0,009606	004.0	0,007405	-30%
004.032.26	2193	004.0	0,011006	004.0	0,007404	-49%
004.056.5	2478	004.0	0,010007	004.0	0,008005	-25%
004.075	2503	004.0	0,009607	004.0	0,008205	-17%
004.93	2712	004.9	0,010006	004.9	0,009006	-11%
004.93.1	2835	004.9	0,011208	004.9	0,010006	-12%
004.94	3261	004.9	0,011007	004.9	0,009806	-12%
Середнє значення			0,009756		0,00813	-20%

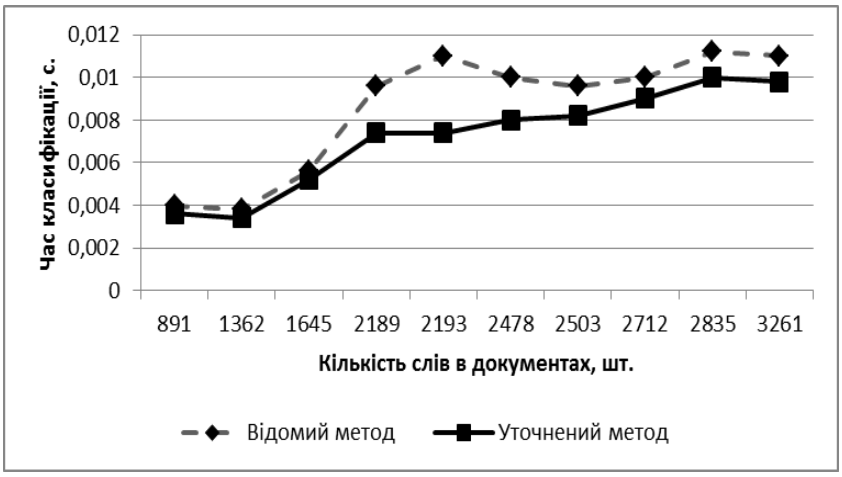

Рис. 4. Порівняння часових витрат на класифікацію окремих документів

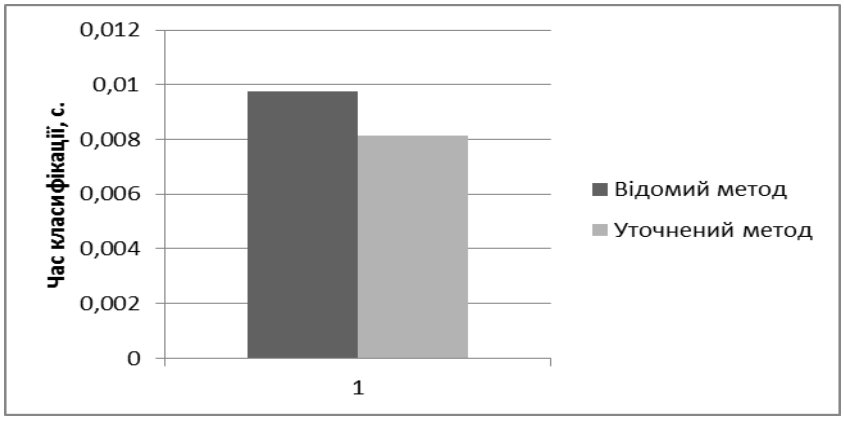

Рис. 5. Порівняння середніх значень часових витрат на класифікацію документів

В таблиці 2 та на рисунку 6 наведено порівняння часових витрат на процеси формування простору ознак за категоріями та класифікацію 10 текстових документів двома методами.

Як видно із таблиці 2 та рисунку 6, часові витрати на підготовчий етап збільшилися на 17%. При цьому витрати часу на класифікацію 10 текстових документів зменшилися на 30%. Так, як підготовчий етап виконується один раз, то збільшення часових витрат на його

реалізацію нівелюється багаторазовим використанням отриманих результатів на етапі безпосередньої класифікації.

Таблиця 2

Порівняння часових витрат на створення простору ознак і класифікацію 10 текстів

Простір ознак категорій	Час на створення простору ознак, с.	Час на класифікацію, с.
За відомим методом	0,16792	0,12508
За уточненим методом	0,20193	0,09646
Відсоток відхилення	+17%	-30%

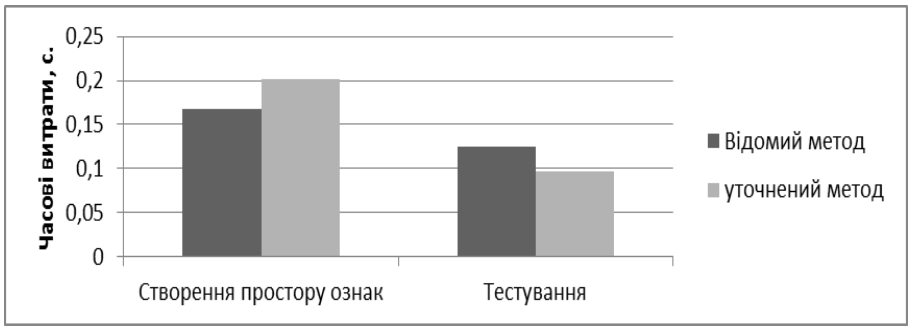

Рис. 6. Порівняння часових витрат на різні процеси при використанні обох методів

Таким чином, за результатами тестування програмного комплексу, використання запропонованих методів призвело до суттєвого зменшення часу, який витрачається на класифікацію текстових документів.

Висновки

В даній статті розроблено уточнений метод формування простору ознак категорій спільної тематики з метою прискорення класифікації текстових документів, виконано його програмну реалізацію і досліджено її ефективність. Складність уточненої класифікації полягає в тому, що береться до уваги факт використання однакових термів для ідентифікації різних категорій. При цьому такі терми є практично несуттєвими для визначення категорій, тому їхнє вилучення не впливатиме на якість класифікації. В програмній реалізації, яка побудована на основі даного методу, це враховано у вигляді фільтрації простору ознак категорій.

Мета даного дослідження була досягнена завдяки саме фільтрації простору ознак категорій, що дозволило зменшити обсяг масиву термів категорій, який підлягає обробці. Це, в свою чергу, призвело до скорочення часових витрат на класифікацію текстів.

Тестування програми на основі описаного методу показало, що час на формування простору ознак категорій збільшився на 17%. Це пояснюється більшою деталізацією процеса визначення порогового значення, але ж цей процес виконується тільки один раз – для декількох категорій та будь-якої кількості текстів, що підлягають подальшому аналізу. Саме цьому, незважаючи на таке збільшення зазначеного часу на підготовчому етапі, результати тестування показали, що часові витрати на класифікацію окремого тексту скоротилися в середньому на 20%, а сумарний час класифікації 10 документів – на 30%.

Наукова новизна розробленого уточненого методу полягає в урахуванні морфологічних особливостей української мови на етапі попередньої обробки тексту, а також у додаткової фільтрації простору ознак категорій. Ці два фактори дозволяють, по-перше, суттєво зменшити обсяг масиву термів категорій, і по-друге – отримати уточнену класифікацію текстів по різним категоріям в межах спільної тематики.

Практична значимість метода полягає в скороченні комп’ютерних часових витрат на класифікацію текстових документів по категоріям спільної тематики, що було підтверджено

дослідженнями програмної реалізації даного методу

Перспективний спосіб подальшого прискорення процесу автоматичної класифікації текстів полягає в розробці програмно-апаратного комплексу, в якому передбачається перенесення етапів стемінгу і класифікації на апаратну базу, з можливістю подальшої оптимізації даного процесу за критерієм часу.

Список використаної літератури

1. Безверхий О. А., Самохвалова С. Г. Кластеризация большого объема текстовых поисковых запросов. Ученые заметки ТОГУ. 2016. Том 7, № 3. С. 104 – 110.

2. Labani, M., Moradi, P., Jalili, M. A multi-objective genetic algorithm for text feature selection using the relative discriminative criterion. Expert Systems with Applications. 2020. Vol. 149. Access mode: https://doi.org/10.1016/j.eswa.2020.113276

3. Karpovich, S.N., Smirnov, A.V., Teslya, N.N. Classification of Text Documents Based on a Probabilistic Topic Model. Scientific and Technical Information Processing. 2019. Vol. 46, Issue 5. P. 314-320

4. Глибовець А. М., Точицький В. В. Алгоритм токенізації та стемінгу для текстів українською мовою. Наукові записки НаУКМА. Комп'ютерні науки. 2017. Т. 198. С. 4-8.

5. Бісікало О. В. Висоцька В. А. Виявлення ключових слів на основі методу контент-моніторингу україномовних текстів. Радіоелектроніка, інформатика, управління. 2016. № 1. С. 74-83.

6. Moral Cristian, Angélica de Antonio, Imbert Ricardo, Ramírez Jaime. A survey of stemming algorithms in information retrieval. Information research. 2014. Vol. 19, no. 1. P. 605-625.

7. Hassanein A.M.D.E. Nour, M A Proposed model of selecting features for classifying Arabic text. Jordanian Journal of Computers and Information Technology. 2019. Vol. 5, issue 3. P. 275-290

8. Alper Kursat Uysal. An improved global feature selection scheme for text classification. Expert Systems with Applications. 2016. Vol. 43. P. 82-92

9. Pouramini Jafar, Behrouze Minaei-Bidgoli Dr., Mahdi Esmaeili Dr. A Novel One Sided Feature Selection Method for Imbalanced Text Classification. JSDP. 2019. Vol. 16, Issue 1 (5). P. 21-40.

10. Ferreira Charles Henrique Porto, Debora Maria Rossi de Medeiros, Fabricio Olivetti de Franc DCDistance: A Supervised Text Document Feature extraction based on class labels. Computer Science. 2018. Vol.2. P.23-31.

11. Doan Son, Horiguchi Susumu. Dynamic Feature Selection in Text Classification. Part of book Intelligent Control and Automation, Lecture Notes in Control and Information Sciences. 2006. P. 664-675

12. Котельников Е.В. Методология интеллектуального анализа мнений при обработке текстовой информации на основе правдоподобного вывода : автореф. дис. … канд. техн. наук : 05.13.17. Нижний Новгород, Россия. 2019. 39 с.

13. Chen, J., Dai, Z., Duan, J., Matzinger, H., Popescu, I. Naive bayes with correlation factor for text classification problem. 18th IEEE International Conference on Machine Learning and Applications, ICMLA, Boca Raton, United States. 16 - 19 December 2019. Boca Raton, United States. P. 1051-1056

14. Yampolsky L.S. Analytical approach to the choice of neural network topologies to solve the applied problems. Adaptive systems of automatic control. 2012. Vol. 20. P. 159-179

15. А.Ю.Кононюк. Нейронні мережі і генетичні алгоритми. К.:«Корнійчук», 2008. 446 с.

16. Краснянский М. Н., Обухов А. Д., Соломатина Е. М., Воякина А. А. Сравнительный анализ методов машинного обучения для решения задачи классификации документов научно-образовательного учреждения. Вестник ВГУ, Серия: Системный анализ и информационные технологи. 2018. № 3. С. 173-182.

17. Акбархужаев С. А., Абдурахманова Н. Н. Сравнительный анализ методов Наивного

Байеса и SVM алгоритмов при классификации текстовых документов. Молодой ученый. 2019. №29. С. 8-10.

18. Mbaikodzi E., Dral' A. A., Sochenko I. V. The method of automatic classification of short text messages. Information technologies and computer systems. 2012. Vol. 3. P. 93-102

19. Tehseen Zia, Muhammad Pervez Akhter Qaiser Abbas. Comparative Study of Feature Selection Approaches for Urdu Text Categorization. Malaysian Journal of Computer Science. 2015. Vol. 28(2). P. 93-109

20. Голуб Т.В., Тягунова М.Ю. Метод стемінгу україномовних текстів для класифікації документів на базі алгоритму Портера. Наукові праці Донецького національного технічного університету. Серія : Інформатика, кібернетика та обчислювальна техніка. 2017. №1. С.59-63.

21. Golub T. Modernized Mathematical Model of Text Document Classification. The Second International Workshop on Computer Modeling and Intelligent Systems (CMIS-2019), Zaporizhzhia, Ukraine, April 15-19, 2019. Zaporizhzhia, Ukraine. P. 607-617. Access mode: http://ceur-ws.org/Vol-2353/paper48.pdf

22. Голуб Т.В., Тягунова М.Ю. Метод уменьшения размера вектора термов для классификации текстовых документов по категориям. Проблемы региональной энергетики. 2019. № 1–2(41). С. 84–94. DOI: 10.5281/zenodo.3240216

23. Глибовець А. М., Точицький В. В. Алгоритм токенізації та стемінгу для текстів українською мовою. Наукові записки НаУКМА. Комп'ютерні науки. 2017. Т. 198. С. 4-8.

24. Bird S., Klein E., Loper E. Natural Language Processing with Python. Sebastopol (USA): O'Reilly Media. 2009. 504p.

25. Perkins J. Python 3 Text Processing with NLTK 3 Cookbook. Birmingham (UK): Packt Publishing Ltd. 2014. 304 p.

26. Універсальний десятковий класифікатор. Режим доступу: http://www.udcsummary.info/php/index.php?id=13358&lang=uk

References

1. Bezverkhiy O. A. and Samokhvalova S. G. (2016), "Clustering of a large volume of text search queries". *Scientific notes Togu*, 7(3). P. 104 - 110.

2. Labani, M., Moradi, P. and Jalili, M. (2020) "A multi-objective genetic algorithm for text feature selection using the relative discriminative criterion". *Expert Systems with Applications*. 149. Access mode: https://doi.org/10.1016/j.eswa.2020.113276

3. Karpovich, S.N., Smirnov, A.V. and Teslya, N.N. (2019) "Classification of Text Documents Based on a Probabilistic Topic Model". *Scientific and Technical Information Processing*. 46(5). P. 314-320

4. Glibovets A. M. and Tochitsky V. V. (2017) "Algorithm of tokenization and stemming for texts in Ukrainian". *Science notes of NaUKMA. Computer science*. 198. P. 4-8.

5. Bisikalo O. V. And Visotska V. A. (2016) "Viyavlennya key words based on the method of content monitoring of Ukrainian new texts". *Radio electronics, informatics, control*. 1. P. 74-83.

6. Moral Cristian, Angélica de Antonio, Imbert Ricardo and Ramírez Jaime (2014) "A survey of stemming algorithms in information retrieval". *Information research*. 19(1). P. 605-625.

7. Hassanein A.M.D.E. and Nour M.A (2019) "Proposed model of selecting features for classifying Arabic text". *Jordanian journal of computers and information technology*.5(3).P.275-290.

8. Alper Kursat Uysal (2016) "An improved global feature selection scheme for text classification". *Expert Systems with Applications*. 43. P. 82-92

9. Pouramini Jafar, Behrouze Minaei-Bidgoli Dr. and Mahdi Esmaeili Dr. (2019) "A Novel One Sided Feature Selection Method for Imbalanced Text Classification". *JSDP*. 16(1). P. 21-40.

10. Ferreira Charles Henrique Porto, Debora Maria Rossi de Medeiros and Fabricio Olivetti de Franc (2018) "DCDistance: A Supervised Text Document Feature extraction based on class labels". *Computer Science*. 2. P.23-31.

11. Doan Son and Horiguchi Susumu. (2006) "Dynamic Feature Selection in Text Classification". Part of book Intelligent Control and Automation, Lecture Notes in Control and

Information Sciences. P. 664-675.

12. Kotelnikov E.V. (2019) “The methodology of the intellectual analysis of opinions in the processing of textual information based on a plausible vivod”: author. dis. ... cand. tech. Sciences: 05.13.17. Nizhny Novgorod, Russia. 39 s.

13. Chen, J., Dai, Z., Duan, J., Matzinger, H. and Popescu, I. (2019) “Naive bayes with correlation factor for text classification problem”. *18th IEEE International Conference on Machine Learning and Applications, ICMLA*, Boca Raton, United States. 16 - 19 December 2019. Boca Raton, United States. P. 1051-1056

14. Yampolsky L.S. (2012) “Analytical approach to the choice of neural network topologies to solve the applied problems”. *Adaptive systems of automatic control*. 20. P. 159-179

15. A.Yu. Kononyuk(2008) “Neural measures and genetic algorithms”. K.: Korniychuk, 446p.

16. Krasnyansky M. N., Obukhov A. D., Solomatina E. M. and Voyakina A. A. (2018) “Comparative analysis of machine learning methods to solve the problem of classifying documents of a scientific and educational institution”. *Vestnik VGU, Series: System analysis and information technology*. 3. P. 173-182.

17. Akbarhuzhayev S. A. and Abdurakhmanova N. N. (2019) “Comparative analysis of the methods of Naive Bayes and SVM algorithms for the classification of text documents”. *Young scientist.* 29. P. 8-10.

18. Mbaikodzi E., Dral’ A. A. and Sochenko I. V. (2012) “The method of automatic classification of short text messages”. *Information technologies and computer systems*. 3. P. 93-102

19. Tehseen Zia and Muhammad Pervez Akhter Qaiser Abbas (2015) “Comparative Study of Feature Selection Approaches for Urdu Text Categorization”. *Malaysian Journal of Computer Science.* 28(2). P. 93-109

20. Golub T.V. and Tyagunova M.Yu. (2017) “A method of stemming Ukrainian-language texts for classifying documents based on Porter's algorithm”. *Scientific works of Donetsk National Technical University. Series: Computer Science, Cybernetics and Computer Engineering*.1.P.59-63.

21. Golub T. (2019) “Modernized Mathematical Model of Text Document Classification”. *The Second International Workshop on Computer Modeling and Intelligent Systems (CMIS-2019),* Zaporizhzhia, Ukraine, 15-19 April 2019. Zaporizhzhia, Ukraine. P. 607-617. Access mode: http://ceur-ws.org/Vol-2353/paper48.pdf

22. Golub T.V. and Tyagunova M.Yu. (2019) “Method for reducing the size of the term vector for classifying text documents into categories”. *Problems of regional energy*. 1–2 (41). P. 84–94. DOI: 10.5281 / zenodo.3240216

23. Glibovets A. M. and Tochitsky V. V. (2017) “Algorithm of tokenization and stemming for texts in Ukrainian”. *Science notes of NaUKMA. Computer science*. 198. P. 4-8.

24. Bird S., Klein E. and Loper E. (2009) “Natural Language Processing with Python”. Sebastopol (USA): O’Reilly Media. 504p.

25. Perkins J. (2014) “Python 3 Text Processing with NLTK 3 Cookbook”. Birmingham (UK): Packt Publishing Ltd. 304 p.

26. The universal ten-year classifier. Access mode: http://www.udcsummary.info/php/index.php?id=13358&lang=uk