УДК 004.912 **DOI:** 10.31673/2412-9070.2025.027728
В. М. ДАНИЛЬЧЕНКО[1], PhD, доцент;
ORCID: 0009-0004-6839-2132
С. І. ОТРОХ[2], доктор техн. наук, професор;
ORCID: 0000-0001-9008-0902
М. О. ШАЛИГІН[2], студент;
ORCID: 0009-0009-6455-6888
А. Г. ДОНЕЦЬ[2], канд. техн. наук, доцент,
ORCID: 0000-0002-8122-051X
[1] Державний університет інформаційно-комунікаційних технологій, Київ
[2] Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

БАГАТОКРИТЕРІАЛЬНЕ РОЗПІЗНАВАННЯ ВІДПОВІДНОСТІ ТЕКСТІВ ТЕМІ НА ОСНОВІ АЛГОРИТМУ TF-IDF

Досліджено актуальність застосування статистичних методів для оцінки приналежності документів до тематики, обраної користувачем. Проаналізовано можливості та імплементовано використання модифікованого алгоритму на основі метрики TF-IDF для ефективного аналізу та класифікації текстових документів. Описано процес розроблення програми з використанням багатокритеріального підходу до розпізнавання відповідності тексту темі, а також розглянуто методи нормалізації та фільтрації текстових даних для вдосконалення точності класифікації. Запропоновано високоефективне вирішення для виявлення релевантних документів, яке може бути застосовано у різних сферах: від пошукових систем та інформаційних фільтрів до рекомендаційних платформ і аналітичних інструментів. Відображено важливість використання інноваційних методів для автоматизованої фільтрації даних.

Ключові слова: обробка; дані; метод TF-IDF; класифікація документів; багатокритеріальне розпізнавання; стоп-слова; нормалізація тексту; масив даних; глибоке навчання; технологія.

Вступ

Обробка надвеликих масивів даних є однією з ключових задач сучасних інформаційних технологій, яка знаходить застосування в найрізноманітніших галузях: від аналізу ринку до штучного інтелекту і роботи з великими текстовими корпусами. У рамках цієї дисципліни особливу увагу приділяють методам аналізу структурованих і неструктурованих даних, що дозволяють ефективно знаходити закономірності, класифікувати об'єкти та виконувати інші аналітичні завдання. Для обробки великих масивів текстових даних, які постійно генеруються в Інтернеті, соціальних мережах, електронних бібліотеках та інших джерелах, важливим є створення алгоритмів, що дозволяють не тільки зберігати і структурувати такі дані, але й швидко знаходити релевантну інформацію. Це є основою побудови інтелектуальних пошукових систем, рекомендаційних платформ і автоматизованих аналітичних інструментів.

Однією з ключових задач у цій сфері є оцінка приналежності документів до заданої тематики, що є критично важливим при розробці пошукових систем, автоматизованих фільтрів інформації та класифікаторів текстів. Розв'язання цієї проблеми ускладнюється через наявність так званих стоп-слів, омографів, а також можливу неоднозначність ключових термінів, яка виникає через багатозначність слів у різному контексті. Основним підходом до класифікації документів у сучасних системах є використання статистичних методів, зокрема метрики TF-IDF (Term Frequency-Inverse Document Frequency), яка дозволяє визначити значимість термів у контексті документа. Проте, стандартна реалізація цієї метрики не завжди забезпечує доста-

тню точність розпізнавання тематики документа, особливо в контексті багатокритеріального аналізу текстів.

У даній статті встановленно за мету розробити модифікований алгоритм виявлення документів за тематикою, описаною користувачем, що забезпечує багатокритеріальне розпізнавання відповідності тексту темі. Цей алгоритм є не просто імплементацією стандартної метрики TF-IDF, а її вдосконаленням, що дозволяє оцінювати релевантність документів за кількома параметрами одночасно: відповідність ключовим термам, загальна відповідність темі та врахування нормалізованих форм слів. Для досягнення цієї мети буде проведено аналіз проблематики класифікації текстів, розроблено програмний алгоритм на основі формальної оцінки концентрації певних термів та загального співпадіння за темою та його реалізування у вигляді програмного продукту. Очікується, що результати дослідження сприятимуть кращому розумінню ефективності підходів до класифікації текстів і дозволять створити більш досконалі рішення для роботи з великими текстовими даними.

Основна частина

Класифікація документів за допомогою статистичних методів стала важливим етапом у розвитку інформаційного пошуку та обробки текстових даних. Перші дослідження у цьому напрямку виникли у 1970-х роках, коли вчені почали розробляти методи автоматичного визначення релевантності документів запитам користувачів. Одним з ключових інструментів стала метрика TF-IDF (Term Frequency-Inverse Document Frequency), запропонована Карен Спарк Джонс як спосіб оцінки важливості слів у текстових документах. Згодом із розвитком технологій і збільшенням обчислювальних можливостей дослідники розширили цей підхід, додавши нормалізацію та модифікації, що дозволило більш точно визначати релевантність документів до заданих критеріїв пошуку.

Незважаючи на переваги статистичних методів, існують і альтернативні підходи до класифікації документів, такі як використання семантичного аналізу для визначення контекстуальних зв'язків між словами, застосування методів машинного навчання, зокрема, нейронних мереж і дерев рішень, а також імовірнісні моделі, такі як наївний байєсівський класифікатор. Однак метрика TF-IDF залишається одним з найбільш ефективних і широко застосовуваних підходів завдяки поєднанню простоти реалізації з високою точністю у багатьох задачах класифікації текстів.

Сучасні модифікації алгоритму TF-IDF часто включають багатокритеріальний підхід, який дозволяє одночасно враховувати кілька аспектів релевантності документів. Одним із сучасних прикладів систем класифікації документів, які активно використовуються в інформаційному пошуку, є модель Okapi BM25 (Best Matching) – удосконалена версія алгоритму TF-IDF, розроблена Стівеном Робертсоном та його колегами у 1990-х роках. Ця модель використовується для ранжування документів за релевантністю до запиту користувача та активно застосовується у сучасних пошукових системах, таких як Elasticsearch, Solr та Lucene. Алгоритм спеціалізується на великих вибірках текстових документів, які можуть містити сотні тисяч текстів різної тематики та структури. Особливістю BM25 є застосування насичення терм-частотності – коли частота слова перевищує певний поріг, його вага зростає нелінійно, що робить оцінку більш збалансованою.

Складова програми

Для реалізації цієї роботи було вибрано два основних інструменти: Python як базову мову програмування та NLTK (Natural Language Toolkit) для обробки природної мови та аналізу текстів.

Python є однією з найпоширеніших мов програмування для наукових досліджень та аналізу даних, особливо у сфері обробки текстової інформації. Вона забезпечує широкі можливості для роботи з різними типами даних, підтримує численні бібліотеки та має зрозумілий і лаконічний синтаксис. За допомогою цієї мови програмування реалізуються всі етапи роботи

алгоритму: від завантаження та попередньої обробки текстів до обчислення метрик TF-IDF та виведення результатів.

NLTK із свого боку є однією з найпопулярніших бібліотек для обробки природної мови. Вона надає широкий спектр інструментів для роботи з текстами, включаючи токенізацію, стемінг, видалення стоп-слів та інші операції. У нашій програмі NLTK використовується для роботи із стоп-словами, що є критично важливим для ефективного аналізу текстів та точного визначення важливих термів.

Взаємодія між цими двома інструментами реалізована у класі Document Analyzer, який об'єднує всі функціональні можливості програми. Python забезпечує базові операції для роботи з файлами та текстами, а NLTK доповнює їх спеціалізованими функціями для обробки природної мови. Програма також використовує вбудовані бібліотеки Python, такі як os для роботи з файловою системою, math для обчислення логарифмів при розрахунку IDF, re для пошуку слів за допомогою регулярних виразів та collections. Counter для ефективного підрахунку частоти слів. Така архітектура дозволяє створити потужний та ефективний інструмент для аналізу текстових документів та визначення їх релевантності заданій тематиці.

Алгоритм написання програми

Для написання програми було вибрано алгоритм на основі метрики TF-IDF - це статистичний метод, який спеціалізується на обробленні текстових даних та виявленні значущих слів у документах. Метод використовує частотний аналіз (який також називається статистичним) для оцінки важливості слів та визначення релевантності документів до заданої теми, що дає можливість ефективно фільтрувати текстові матеріали за різними критеріями пошуку.

Основні компоненти алгоритму класифікації документів охоплюють такі етапи:

- завантаження документів. Це перший крок у алгоритмі класифікації. Завантаження здійснюється шляхом зчитування текстових файлів із вказаної директорії, приведення тексту до нижнього регістру та збереження у внутрішніх структурах даних. Цей етап забезпечує підготовку текстів для подальшого аналізу;
- попередня обробка текстів. Після завантаження застосовується фільтрація стоп-слів та нормалізація тексту для очищення даних від малозначущих слів, які можуть спотворювати результати аналізу;
- аналіз частотності термів. Для кожного документа визначаються найчастіші слова та проводиться підрахунок входжень термів, які відшукуються, з урахуванням їх морфологічних варіацій. Це дозволяє визначити статистичні характеристики тексту;
- перевірка тематичної відповідності. Відбувається порівняння ключових слів теми з текстом документа, враховуючи поріг відповідності (понад 35% слів теми мають хоча б раз збігатися з текстом). Цей етап відсіює документи, що не відповідають заданій тематиці;
- обчислення метрик TF-IDF. На цьому етапі розраховуються показники TF-IDF та їх комбінація для кожного терму, що дозволяє оцінити його значущість у контексті документа та колекції.

Окрему увагу варто звернути на нормалізацію слів — це процес приведення різних форм слова до єдиної базової форми шляхом видалення суфіксів та аналізу коренів. У нашій програмі це реалізується через функцію get_word_root, яка обробляє суфікси 'ing', 'ed', 's', 'es', 'ies', дозволяючи розпізнавати різні форми одного й того ж слова.

У контексті аналізу текстів нормалізація часто використовується для покращення точності підрахунку частоти слів. Під час обробки програма перевіряє корені слів, щоб різні граматичні форми (наприклад, "dive", "dives", "diving") враховувались як одне й те ж слово. Це дає змогу підвищити точність оцінки важливості термів і забезпечити більш релевантні результати пошуку.

Основну формулу для обчислення TF можна подати у такому вигляді:

$$TF(t,d) = \frac{f(t,d)}{\max\{f(w,d): w \in} , \quad (1)$$

де TF(t,d) — частота терму t у документі d; f(t,d) — кількість входжень терму t у документ d; max{f(w,d): w ∈ d} — максимальна кількість входжень будь-якого слова у документ d.

Для обчислення IDF використовується формула:

$$IDF(t) = log_2(N/ı \ , \quad (2)$$

де IDF(t) — обернена частота документа для терму t; N — загальна кількість документів; n_t — кількість документів, що містять терм t.

Використовуючи ці показники, програма обчислює TF-IDF для кожного терму за формулою:

$$TF - IDF(t,d) = TF(t,d) \cdot IDF(ı \quad (3)$$

І нарешті середнє значення для всіх шуканих термів, що дозволяє ранжувати документи за релевантністю. Формула обчислення:

$$Avg\_TF - IDF(d) = \frac{\sum TF.IDF(t,}{|T|} \cdot \quad (4)$$

Отже, програма для класифікації документів працює у такий спосіб: коли користувач вказує директорію з документами, тему пошуку та ключові слова, алгоритм завантажує та аналізує всі тексти, вираховує їхню відповідність заданим критеріям та сортує за релевантністю. На екрані відображається результат аналізу, включно з найчастішими словами документів, середнім показником TF-IDF та детальною інформацією про кожен терм.

Розроблена програма класифікації документів забезпечує комплексний підхід до виявлення релевантних текстів через інтеграцію багатокритеріальної системи оцінювання, яка поєднує перевірку тематичної відповідності, нормалізацію слів, фільтрацію стоп-слів та гнучку модель розрахунку метрик TF-IDF, при цьому ключовим новаторським аспектом алгоритму є метод обчислення середнього значення TF-IDF для всіх відшукуваних термів як основного критерію порівняння документів, що, на відміну від традиційних підходів із використанням сумарного або максимального показника, забезпечує більш збалансовану оцінку релевантності, враховуючи важливість кожного терму рівномірно та запобігаючи домінуванню окремих високочастотних слів, що дозволяє точніше визначати тематичну приналежність документів та отримувати більш якісні результати пошуку навіть при наявності термів з різною значущістю.

У майбутньому, для подальшого вдосконалення цієї програми, можна розглянути такі напрямки оптимізації. Застосування більш ефективних алгоритмів нормалізації, зокрема використання лематизації замість стемінгу, може підвищити точність класифікації. Також можна впровадити врахування семантичного контексту слів через використання векторних представлень слів (word embeddings), що сприятиме покращенню здатності програми розпізнавати тематичну приналежність документів навіть за наявності синонімів та контекстуальних значень.

Висновки

У даній роботі було розроблено алгоритм класифікації документів за заданою користувачем тематикою з використанням метрики TF-IDF. Застосування Python та бібліотеки NLTK дало змогу створити потужний та ефективний інструмент, здатний точно визначати приналежність текстів до тематики з високою точністю та гнучкістю налаштувань.

Запропонований багатокритеріальний підхід до оцінки релевантності, що включає аналіз відповідності темі, врахування ключових термів та їхніх морфологічних варіацій, а також інноваційний метод обчислення середнього значення TF-IDF як основного критерію порівняння, демонструє значні переваги порівняно з традиційними методами класифікації текстів. Попередня обробка даних, включаючи фільтрацію стоп-слів та нормалізацію текстів, суттєво підвищує точність аналізу та забезпечує високу якість результатів.

Застосування цієї технології може мати значущий вплив на багато сфер, включно з інформаційним пошуком, автоматизованою фільтрацією даних, рекомендаційними системами, аналізом відгуків та моніторингом соціальних медіа. Ефективні методи оцінки

приналежності документів до тематики стають важливим інструментом для роботи з надвеликими масивами текстових даних, сприяючи оптимізації інформаційних потоків та підвищенню ефективності аналітичних систем у сучасному інформаційному просторі.

Список літератури

1. TF–IDF. [Електронний ресурс]. URL: https://en.wikipedia.org/

2. Introduction to Information Retrieval. [Електронний ресурс]. URL: https://nlp.stanford.edu/IR-book/information-retrieval-book.html

3. Understanding TF-IDF and How to Use it for Text Mining Using Python. [Електронний ресурс]. URL: https://towardsdatascience.com/understanding-tf-idf-and-how-to-use-it-for-text-mining-using-python-6517d6ac90bd

4. Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit. [Електронний ресурс]. URL: https://www.nltk.org/book/

5. Text Feature Extraction with Scikit-Learn. [Електронний ресурс]. URL: https://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction

6. Manning C. D., Raghavan P., Schütze H. Introduction to Information Retrieval [Електронний ресурс]. URL: https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf

7. Leskovec J., Rajaraman A., Ullman J. D. Mining of Massive Datasets [Електронний ресурс]. URL: http://www.mmds.org/

8. Jurafsky D., Martin J. H. Speech and Language Processing (3rd ed.) [Електронний ресурс]. URL: https://web.stanford.edu/~jurafsky/slp3/

9. Yang Y., Liu J. A New Term Weighting Scheme for Text Categorization [Електронний ресурс]. URL: https://arxiv.org/pdf/cs/0412098.pdf

10. Büttcher S., Clarke C. L. A., Cormack G. V. Information Retrieval: Implementing and Evaluating Search Engines [Електронний ресурс]. URL: http://ir.dcs.gla.ac.uk/resources/iet-book.html

V. Danylchenko, S. Otrokh, M. Shaligin, A. Donets

MULTI-CRITERIA RECOGNITION OF TEXT-TO-TOPIC CORRESPONDENCE BASED ON THE TF-IDF ALGORITHM

This research delves into the critical role of statistical methodologies in determining the thematic alignment of textual content with specific user interests. Recognizing the ever-increasing volume of digital information, the need for accurate and efficient document classification has become paramount across numerous applications. This study meticulously examines the potential and subsequently implements a refined algorithmic approach grounded in the Term Frequency-Inverse Document Frequency (TF-IDF) metric. The modifications introduced aim to optimize the algorithm's performance in analyzing and categorizing diverse sets of textual data.

A comprehensive account of the program development process is provided, emphasizing the integration of a multi-criteria decision-making framework to achieve a nuanced understanding of text-to-topic relevance. This multi-faceted approach considers various linguistic features and statistical indicators beyond the basic TF-IDF scores, thereby enabling a more robust and accurate classification outcome. Furthermore, the research addresses the crucial pre-processing steps involved in handling textual data. Detailed attention is given to the methodologies of text normalization, which includes techniques such as stemming, lemmatization, and case conversion, to reduce data dimensionality and improve the consistency of feature representation.

The study also rigorously explores the application of effective filtering techniques, particularly the identification and removal of stop words, which are common words that carry minimal semantic weight and can often introduce noise into the classification process. The strategic implementation of these normalization and filtering methods is shown to significantly contribute to the overall precision and recall of the proposed text classification system.

The outcome of this research is a highly effective and adaptable solution for the identifycation of documents that are genuinely relevant to a user's specified thematic focus. The potential applications of this solution span a wide spectrum of information management and retrieval systems. In the context of search engines, the enhanced classification capabilities can lead to more precise and contextually appropriate search results, improving user satisfaction and information discovery. Similarly, in information filtering systems, the proposed approach can facilitate the delivery of tailored content streams, reducing information overload and enhancing user engagement.

Ultimately, this research underscores the growing significance of employing innovative and statistically sound methods for the automated filtering and categorization of textual data in the age of big data. The proposed multi-criteria TF-IDF-based algorithm offers a valuable contribution to the field, providing a practical and efficient means of navigating and extracting relevant information from the vast digital landscape. The findings highlight the potential for significant advancements in various information-intensive applications through the intelligent automation of text analysis and classification processes.

Keywords: processing; data; TF-IDF method; document classification; multi-criteria recognition; stop words; text normalization; data set; deep learning; technology.