DOI 10.6084/m9.figshare.22886720
Challenges and Issues of Modern Science, Vol. 1 (2023)

Автоматична рубрикація текстів з використанням семантичних словників

Єгор Романович Ковилін
ORCID: https://orcid.org/0000-0002-2734-4095
Університет імені Альфреда Нобеля, Дніпро

Микита Олександрович Басистий
Університет імені Альфреда Нобеля, Дніпро

Вступ

У сучасному інформаційному суспільстві величезна кількість текстового контенту стає доступною користувачеві щодня. У таких умовах актуальність та ефективність подання інформації стають критично важливими. Люди витрачають все більше часу на пошук та фільтрацію інформації, щоб знайти те, що їм справді потрібно. Вирішенням цієї проблеми може стати рубрикація тексту, яка дозволяє організувати інформацію у структуровану форму, що значно полегшує її пошук та засвоєння. Читач може швидко переглянути рубрики і вибрати тему, що цікавить, що економить час і покращує його сприйняття інформації. У деяких областях, таких як новини, наука чи технології, де актуальність інформації є вкрай важливою, рубрикація дозволяє швидко знайти потрібну інформацію та бути в курсі останніх подій та розробок у конкретній галузі. Це особливо корисно для професіоналів, дослідників та активних користувачів, яким важливо бути в курсі подій. Врешті, рубрикація тексту сприяє поліпшенню користувальницького досвіду та задоволенню потреб читачів. Вона допомагає скоротити час, що витрачається на пошук та читання тексту, а також підвищує рівень розуміння та засвоєння інформації. Це може призвести до збільшення задоволеності користувачів та їхнього повернення до джерела інформації в майбутньому. Виходячи з цього, актуальність теми автоматичної рубрикації текстів є досить високою.

Проведене у роботі дослідження існуючих розробок показало, що існуючі механізми автоматичної рубрикації текстів, які гуртуються на методах штучного інтелекту, хоча і дозволяють вирішити порушену у роботі проблему, зазвичай вимагають складних налаштувань і глибоких предметних знань у користувачів таких доробок. Для відносно невеликих інформаційних множин із сталим або слабо змінним переліком рубрик, якими наприклад можуть бути бібліотеки або довідкові каталоги, більш привабливим виглядає перспектива використання рубрикації на основі семантичних словників, де контроль за якістю

рубрикації можливо покращувати самому користувачу, шляхом наповнення та правки словників за його безпосередньою потребою.

Саме тому, метою даної роботи є розробка алгоритму автоматичної рубрикації тексту на основі семантичних словників, та побудова інформаційної системи, яка реалізує створений алгоритм.

Результати

Результатом проведеного дослідження стала розробка алгоритму, схема якого зображена на рис. 1. Розглянемо його більш детально.

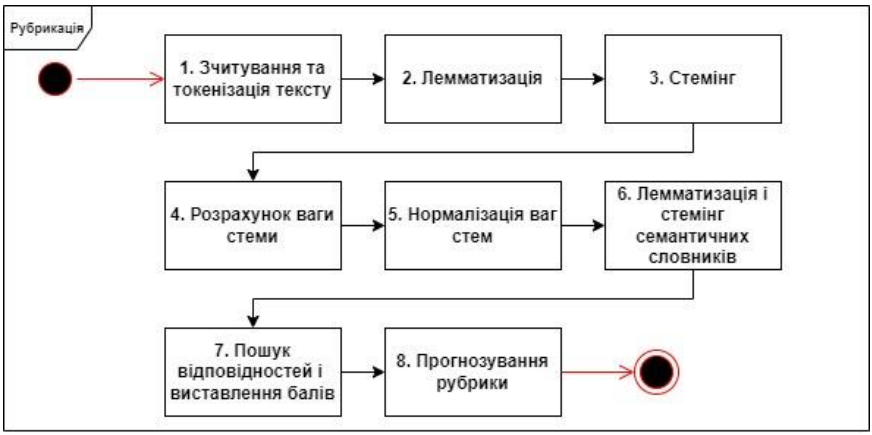

Рис. 1. Алгоритм автоматичної рубрикації тексту на основі семантичного словника

Найпершим кроком розробленого алгоритму є отримання вхідного тексту у форматі «plain text», без попередньої обробки і семантичної розмітки. Над отриманим таким чином текстом проводиться операція токенізації, яка спрямована на парсинг слів із тексту у форматі токенів, доступних до подальшої програмної обробки. Окрім того, на цьому етапі відбувається фільтрація тексту від стоп-слів, таких як союзи, займенники та інші, які не мають значення для подальшого статистичного аналізу. Фільтрація відбувається за допомогою словника стоп-слів, який було складено заздалегідь для англійської мови та є сталим і незмінним для кожної мови.

Для кожного токена-слова, отриманого на попередньому етапі, необхідно виконати приведення слів до єдиної форми – стеми, для подальшого розрахунку їх статистичної ваги. Однак, існуючі алгоритми стемінгу не дозволяють достатньо якісно привести слова до єдиної

форми, особливо у випадках словозміни через використання суфіксу або закінчення [1]. Тому, було прийнято рішення про проведення попереднього процесу лематизації для кожного слова-токена, який полягає у приведення слів до їх лем - основних форм або словникових форм. Лематизація [2] допомагає уніфікувати різні словоформи, наводячи їх до єдиної базової форми. Це дозволяє скоротити різноманітність словоформ та обробляти текст більш ефективно. Наприклад, слова "біжить" і "біжать" можуть бути лематизовані до загальної форми "бігти".

Над отриманими лемами відбувається процес стемінгу, який полягає у обрізанні словоформи до її основи (стеми) шляхом видалення суфіксів та закінчень. У якості алгоритму стемінгу був обраний стемінг Портера [3] - один із найпоширеніших методів стемінгу, який є універсальним і широко застосовуваним методом стеммінгу для багатьох мов, враховує різні типи словозміни, такі як суфікси, закінчення та приставки, та застосовує відповідні правила для їх видалення та приведення слова до єдиної основи, і врешті, відрізняється відносною простотою реалізації та високою швидкістю роботи, через що він може бути легко впроваджений у системи обробки тексту та використаний для обробки великих обсягів даних.

Над отриманим масивом стем слів, відбувається процес розрахунку статистичної ваги стем, яка представляє собою кількість повторень стеми у тексті. Отриману таким чином вагу стеми необхідно зробити незалежною від загальної кількості слів у тексті, для чого була використана формула 1:

$$\text{НВС} = \frac{\text{ВС}}{\text{ЗКС}} \tag{1}$$

де *ВС* - кількість повторень слова (вага слова), *НВС* - нормована вага слова, *ЗКС* - загальна кількість слів у тексті.

Заповнені заздалегідь семантичні словники рубрик, які містять перелік слів-маркерів, що найкраще, з точки зору користувача описують рубрику тексту, так само проходять через процеси лематизації та стемінгу. Для кожної стеми із тексту, що аналізується, відбувається її зіставлення із стемами словника. У разі співпадіння стем, нормована вага стеми додається до ваги рубрики, з якою відбувається порівняння стем з тексту. Таким чином, ми отримаємо чіткі цифри ваги кожної існуючої у системі рубрики для конкретного тексту. Рубрика із максимальною сумарною нормалізованою вагою стає результатом рубрикації нашого тексту.

Для перевірки адекватності розробленого алгоритму було проведено тестування на складеному корпусу із 70 текстів і словників на три семантично незалежних теми: кулінарія, кіно та відеоігри. Мануальне тестування отриманих результатів рубрикації, яке полягало у ручному зіставленню отриманої рубрики із очікуваною, показало точність рубрикації у 92%.

Висновки

У роботі розроблено підхід до рубрикації текстів, заснований на аналізі статистичних характеристик тексту і використання семантичних словників. На основі побудованого алгоритму, був розроблений додаток автоматизованого рубрикації тексту на англійській мові. Серед методів обробки природньої мови використані токенізація, стемінг і лематизація. Формалізована задача статистичної класифікації неструктурованих текстових даних, створено алгоритм рубрикації, що включає позначення тематичного класу тексту на основі складання семантичних словників з такими даними як назва теми та слова що їй належать.

Проведене мануальне тестування розробленого додатку показало точність визначення рубрики у 92%, на 70 текстах різної тематики, що є цілком задовільним для першочергової задачі роботи.

Посилання

1. Jasmeet Singh and Vishal Gupta. Text Stemming: Approaches, Applications, and Challenges. ACM Comput. Surv., 2016 - 1-46 pp. https://doi.org/10.1145/2975608.

2. Siddhartha B S. An Interpretation of Lemmatization and Stemming in Natural Language Processing - Shanghai Ligong Daxue Xuebao/Journal of University of Shanghai for Science and Technology 22(10), 2021 - 350-357 pp.

3. Willett, P. (2006) The Porter stemming algorithm: then and now. Program: Electronic Library and Information Systems, 40 (3). pp. 219-223. ISSN 0033-0337 https://doi.org/10.1108/00330330610681295.