УДК 004.912

РОЗВ'ЯЗАННЯ ЗАДАЧІ КЛАСТЕРИЗАЦІЇ НАУКОВИХ ТЕКСТІВ

Стецун К.С.
Науковий керівник – канд. техн. наук, доц. Гибкіна Н.В.
Харківський національний університет радіоелектроніки, каф. ПМ,
м. Харків, Україна
тел. +38(067) 731-49-09, email: kateryna.stetsun@nure.ua

This study delves into the formal and substantive aspects of the problem of thematic modeling of scientific texts through the division of probability distribution mixtures. The key objective is to identify the subject of each document from a collection of scientific documents. The collection is comprised of textual scientific documents sourced from a variety of places. The simulation will yield a set of documents, with each marked by multiple topics from a general list. These topics will each be associated with a set of keywords that best represent them.

Враховуючи постійне збільшення потоків даних у мережі Інтернет, зокрема у вигляді текстової інформації різноманітної направленості (як то тематичні статті на веб-ресурсах, пости та відгуки у соціальних мережах, художні твори за жанрами, наукові роботи у різних галузях знань тощо), можна стверджувати, що задача пошуку, аналізу та систематизації інформації за заданою тематикою ставатиме все актуальнішою, а її реалізація буде пов'язана з все більшими труднощами через збільшення обсягів даних, які потрібно для цього обробляти. Перспективними для розв'язання подібних задач можна вважати математичні методи, зокрема, методи теорії ймовірностей, математичної статистики та статистичного аналізу, оскільки вони, з одного боку, забезпечують наукову обґрунтованість результатів, а, з іншого, програмні реалізації таких методів дозволяють зручно обробляти великі набори даних.

Однією з найважливіших задач у цьому напрямку є задача кластеризації текстових документів, а саме, визначення тематичної спрямованості окремих документів колекції. Зазначимо, що на відміну від задачі класифікації, кількість тематичних напрямків та їх назви є невідомими, отже, використання методів кластеризації може бути корисним, наприклад, під час рубрикації текстів з цифрового сховища інформації тощо.

Одним з підходів до розв'язання задачі кластеризації документів може стати метод тематичного моделювання [1]. Перевагою цього методу є те, що тематичне моделювання дозволяє не лише групувати документи колекції за тематиками, а також визначає направленість кожної з виділених тем у вигляді ключових слів.

Отже, розглянемо ймовірнісну тематичну модель, де окремий документ колекції x подається вектором Bag of Words: $x \in \mathbb{N}^V$, що має поліноміальний розподіл ймовірностей з певними параметрами, де V – розмір від-

повідного розглядуваній колекції документів словника. Виходячи з припущення, що документи колекції належать до K різних тем і кожній з тем відповідає свій поліноміальний розподіл $p_k(x)$, $k = 1,...,N$, можна записати загальний розподіл документів у колекції як суміш K поліноміальних розподілів окремих тем у вигляді $p(x) = \sum_{k=1}^{K} \pi_k p_k(x)$, де π_k – вага k-го розподілу, $\pi_k \geq 0$ для всіх k та $\sum_{k=1}^{K} \pi_k = 1$.

За таких умов документ x_i, $i = 1,...,M$ (де M – обсяг колекції) може бути поданий як вектор у кодуванні Bag of Words з параметрами $n_i \in \mathbb{N}$ (кількість слів) та $\beta_k \in [0,1]^V$ (вектор ймовірностей k-го поліноміального розподілу). Належність документа до тієї чи іншої тематики визначається прихованою змінною t, що у даній задачі має сенс індикатора компонент суміші $t \in \{0,1\}^K \sim \text{Polinomial}(1,\pi)$.

Задача кластеризації документів за темами полягає у знаходженні за даною колекцією документів векторів β_k, відповідних окремим розподілам суміші, $k = \overline{1,K}$, та вектора π, який визначає розподіл компонентів у суміші. Оптимальні значення цих параметрів, які найточніше описують досліджувану колекцію, можуть бути визначені як розв'язок задачі максимізації логарифмічної функції правдоподібності:

$$L(\pi, B) = \sum_{i=1}^{M} \log p(x_i; \pi, B) \to \max_{\pi \in \Theta_1, B \in \Theta_2}, \quad (1)$$

де $\Theta_1 = \left\{ \pi \in \mathbb{R}_+^K : \sum_{i=1}^{K} \pi_i = 1 \right\}$, $\Theta_2 = \left\{ B \in \mathbb{R}_+^{V \times K} : \sum_{i=1}^{M} \beta_{ik} = 1, k \in \{1,...,K\} \right\}$. (2)

Задача кластеризації розв'язана на текстах наукової спрямованості. Підготовчий етап полягає у передобробці текстів, зокрема, видаленні гіперпосилань, дат, лематизації тощо. На основі оброблених текстів складаються їх моделі Bag of Words, які є вхідною інформацією у задачі (1), (2). Розв'язання задачі (1), (2) через її складність реалізується за допомогою ЕМалгоритму [2]. Під час розв'язання поставленої задачі використані бібліотеки Python, такі як NumPy, Pandas, CountVectorizer та Matplotlib/wordcloud.

Список використаних джерел:

1. Vorontsov, K.V., & Potapenko A.A. (2014). Additive regularization of topic models. *Machine Learning, Speial Issue on Data Analysis and Intelligent Optimization*, 303–323.

2. Стецун К., Гибкіна Н., & Шпакович М. (2022). Розв'язування задачі тематичного моделювання наукових текстів шляхом розділення сумішей ймовірнісних розподілів. *Матеріали статей Міжнар. наук.-практ. конф. «Інформаційні технології та комп'ютерне моделювання»*, 74-76.