УДК 519.6:[004.85 : 519.25]
DOI: https://doi.org/10.30837/IYF.CVSAMM.2024.287

ЗАСТОСУВАННЯ НЕЙРОННИХ МЕРЕЖ ДО РОЗВ'ЯЗАННЯ ЗАДАЧІ ТЕМАТИЧНОГО МОДЕЛЮВАННЯ

Стецун К.С.
Науковий керівник – канд. техн. наук, доц. Гибкіна Н.В.
Харківський національний університет радіоелектроніки, каф. ПМ,
м. Харків, Україна
e-mail: kateryna.stetsun@nure.ua

This work delves into the application of neural networks, particularly the BAT model, for addressing thematic modeling tasks. Thematic modeling revolves around constructing models that facilitate the organization, understanding, and interpretation of vast collections of textual data. The BAT model (Bidirectional Adversarial Topic model) has garnered attention due to its widespread use in neural networks for text partitioning. This work underscores the potential of neural networks, particularly the BAT model, in advancing thematic modeling techniques and their application across diverse domains.

Сучасні методи аналізу даних зіштовхуються із серйозними викликами у зв'язку із суттєвим зростанням обсягів інформації, яка генерується різними джерелами, збирається за допомогою різноманітних технологій і часто супроводжується обмеженнями у часі для аналізу. Ця проблема виникає у різних сферах діяльності. Так, у наукових дослідженнях наявність великої кількості публікацій, які виконують важливу роль для генерації нових знань, може ускладнити пошук необхідної для наукової роботи інформації. Однією з ключових задач в роботі зі збірками наукових робіт є їх систематизація, заснована на встановленні тематик та ключових слів, що дозволяє організувати та оптимізувати подальшу роботу з цими текстовими документами. Для вирішення цієї проблеми застосовуються методи тематичного моделювання.

Введемо позначення: D – колекція текстових документів, d – документи колекції, W – словник, тобто множина унікальних термів (слів) колекції D, w – окреме слово (терм) в словнику, t – теми документів з множини тем T колекції. Змінні w та d – спостережувані, змінна t – прихована. Кожна тема описується дискретним розподілом ймовірностей слів $p(w|t)$, кожен документ – дискретним розподілом ймовірностей тем $p(t|d)$.

Розподіл термів у документі $p(w|d)$ визначається розподілами термів за темами $p(w|t)$ та тем за документами $p(t|d)$:

$$p(w|d) = \sum_{t \in T} p(w|t,d)\, p(t|d) = \sum_{t \in T} p(w|t)\, p(t|d). \qquad (1)$$

Задача тематичного моделювання полягає в тому, щоб знайти такі тематики документів колекції у вигляді розподілу $p(t|d)$ та структуру кожної теми у вигляді розподілу слів $p(w|t)$, за яких тематична модель (1) як-

найкраще наближає частотні оцінки ймовірностей $\hat{p}(w|d)$, обчислені за заданою колекцією документів D. Результатом розв'язання задачі тематичного моделювання є поставлені у відповідність кожному документу колекції теми (можливо, не одна), яким він відповідає, та набір найхарактерніших слів для кожної теми.

Для застосування тематичного моделювання текстові дані повинні бути записані у цифровому вигляді, для чого використовується модель BagofWords («Мішок слів»). Ця модель дозволяє подати кожен документ з колекції D у вигляді вектора, у якому для кожного слова зі словника міститься інформація про частоту появи цього слова у даному документі, але не враховує порядок слів у документі.

Особливості, які притаманні сучасним текстовим даним, вимагають залучення потужних інформаційних технологій для їх обробки, зокрема, нейронних мереж. Різноманітність архітектур нейронних мереж, їх здатність до навчання на великих обсягах даних та автоматичного виявлення складних залежностей робить нейромережеві технології перспективним інструментом для виділення тематичних структур у тексті. Для розв'язання поставленої задачі досліджено застосування нейронної мережі BATM (Bidirectional Adversarial Topic model). Вона має таку структуру. Енкодер приймає на вхід V-вимірне представлення документа $\vec{d}_r$ з колекції документів та перетворює його у K-вимірний розподіл тем $\vec{\theta}_r$. Генератор приймає на вхід випадковий тематичний розподіл $\vec{\theta}_f$ з апріорним розподілом Діріхле і генерує V-вимірний розподіл fake-слів $\vec{d}_f$. Дискримінатор приймає дійсну пару розподілу $\vec{p}_r = [\vec{\theta}_r; \vec{d}_r]$ і fake-пару розподілу $\vec{p}_f = [\vec{\theta}_f; \vec{d}_f]$ на вхід та відокремлює справжні пари розподілу від fake-пар; вихідні сигнали дискримінатора використовуються як сигнали контролю під час змагального навчання [2].

Застосування BATM для розв'язання задачі тематичного моделювання досліджено на колекції текстів, що містить україномовні анотації до наукових статей (загалом 4608 документів за 7 темами: актуарна математика; кластерний аналіз, електродинаміка, інвестування, математична фізика, нейронні мережі, оптимальне керування. Перед навчанням нейронної мережі документи колекції були попередньо оброблені (стемінг, лематизація, видалення гіперпосилань, стоп-слів тощо) та токенізовані за допомогою токенізатора SpaCy.

Було проведено навчання нейронної мережі BATM за умови розділення документів колекції на 6, 7, 8 та 12 топіків. У кожному експерименті отримано набори ключових слів для кожного з топіків та ймовірнісний розподіл цих слів, а також ймовірнісний розподіл топіків для випадково обраних документів колекції.

Було проаналізовано значення похибки для кожного з модулів (Generator, Discriminator, Encoder) для розподілу датасету на 6, 7, 8 та 12

топіків. Аналіз значень похибок показав, що найкращі результати мережа видає при обранні кількості топіків, яка співпадає з реальною кількістю тем (7 топіків, 7 тем), при навчанні протягом 50 епох. При збільшенні кількості епох навчання результати дещо погіршуються, причиною чого може бути перенавчання мережі та особливості датасету.

Далі наведемо деякі результати роботи BATM та їх аналіз для випадку розбиття досліджуваної колекції на 7 топіків. Так, для топіку 7 першими п'ятнадцятьма найчастіше повторюваними (ключовими) словами мережа визначила такі (у дужках наведено відповідне значення ймовірності): «інвестиційної» (0,0170), «доходності» (0,0052), «страхової» («0,0052»), «інвесторів» (0,0051), «інвестиційних» (0,0039), «дохід» (0,0038), «фінансової» (0,0031), «бізнес» (0,0028), «інструмент» (0,0027), «фінансовий» (0,0027), «фінансових» (0,0027), «цінні» (0,0027), «ризиками» (0,0025), «інвестиції» (0,0025), «страхових» (0,0024). Аналіз цих ключових слів дозволяє зробити висновок про те, що виділений мережею топік 7 повністю відповідає одній з реальних тем колекції, а саме темі «інвестування».

Також було проаналізовано ймовірності належності деяких документів досліджуваної колекції до кожного з 7 топіків. Результати аналізу випадково обраних з датасету документів підтвердили високу якість обраного методу розподілу. Так, наприклад, для документа колекції за номером 2201 отримано такі ймовірності його віднесення до кожного з 7 топіків: (0,005; 0,005; 0,028; 0,009; 0,009; 0,012; 0,932), звідки видно, що найбільш ймовірним для нього є топік 7 (з ймовірністю 0,932). Аналіз змісту цього документа: «у статті розглядаються питання прибутковості цінних паперів, що враховують умови їх випуску та положення дивідендної політики емітентів, формування прибутковості диверсифікованого портфеля» підтверджує, що його реальною темою є тема «інвестування».

Зауважимо, що для навчання нейронної мережі були використані не самі наукові статті, а лише їх анотації, що могло знизити якість навчання через маленький розмір документів. Продовження досліджень передбачає використання повнотекстових документів, у тому числі іншомовних.

Список використаних джерел:

1. Knowledge discovery through directed probabilistic topic models: a survey / A. Daud, J. Li, L. Zhou, F. Muhammad. Front*iers of Computer Science in China*. 2010. V. 4, № 2. P. 280–301.

2. Wang R., Hu X., Zhou D., He Y., Xiong Y., Ye C., Xu H. Neural Topic Modeling with Bidirectional Adversarial Training. 2020.