Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет інформаційно-аналітичних технологій та менеджменту
(повна назва)

Кафедра прикладної математики
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Застосування автоматичного варіаційного виводу
для розв'язання задачі тематичного моделювання
(тема)

Виконав:
студент _2_ курсу, групи _ПМм-20-1
Деркач О.С.
(прізвище, ініціали)
Спеціальність
113 Прикладна математика
(код і повна назва спеціальності)
Тип програми _освітньо-професійна
(освітньо-професійна або освітньо-наукова)
Освітня програма
Прикладна математика
(повна назва освітньої програми)
Керівник ____доц. Гибкіна Н.В.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри ПМ _____ Тевяшев А.Д.
(підпис) (прізвище, ініціали)

2021 р.

Харківський національний університет радіоелектроніки

Факультет _інформаційно-аналітичних технологій та менеджменту
Кафедра _прикладної математики
Рівень вищої освіти _другий (магістерський)
Спеціальність _113 Прикладна математика
(код і повна назва)
Тип програми _освітньо-професійна
(освітньо-професійна або освітньо-наукова)
Освітня програма _Прикладна математика
(повна назва)

ЗАТВЕРДЖУЮ:

Зав. кафедри ПМ _____
(підпис)
"____" _____ 2021 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові _Деркачу Олексію Сергійовичу
(прізвище, ім'я, по батькові)

1. Тема роботи _Застосування автоматичного варіаційного виводу для розв'язання задачі тематичного моделювання

затверджена наказом по університету від 05 листопада 2021 р. № 1641 Ст

2. Термін подання студентом роботи до екзаменаційної комісії 10 грудня 2021 р.

3. Вихідні дані до роботи _тематична модель, тренувальний набір даних (колекція документів)

4. Перелік питань, що потрібно опрацювати в роботі
1. Аналіз предметної області
2. Вибір і обґрунтування методу розв'язання
3. Програмна реалізація
4. Результати обчислювального експерименту

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій

1. Актуальність теми роботи
2. Постановка задачі
3. Аналіз предметної області
4. Метод чисельного аналізу
5. Результати обчислювального експерименту

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Підбір та вивчення технічної літератури за темою роботи	8 – 14 листопада 2021 р.	виконано
2	Вибір та обґрунтування методу	15 – 21 листопада 2021 р.	виконано
3	Розробка алгоритму і програми	22 – 28 листопада 2021 р.	виконано
4	Проведення аналітичних досліджень та розрахунків	29 листопада – 5 грудня 2021 р.	виконано
5	Робота над текстом пояснювальної записки	6 – 9 грудня 2021 р.	виконано
6	Представлення роботи на рецензію в ЕК	10 грудня 2021 р.	виконано

Дата видачі завдання _8_ листопада 2021 р.

Студент _____
(підпис)

Керівник роботи _____ доц. Гибкіна Н.В.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка: 57 с., 2 табл., 7 рис., 1 дод., 11 джерел.

ТЕМАТИЧНА МОДЕЛЬ, АВТОЕНКОДЕР, ВАРІАЦІЙНИЙ ВИВІД, ДИВЕРГЕНЦІЯ, СПРЯЖЕНІ РОЗПОДІЛИ, БАЄСОВИЙ ВИВІД, ЯКІСТЬ ТЕМАТИЧНОЇ МОДЕЛІ.

Об’єкт дослідження – текстові документи з визначеними темами.

Мета роботи – дослідження можливості застосування автоматичного варіаційного виводу для визначення тем текстів.

Методи дослідження – варіаційний автоенкодер, методи попередньої обробки документів, оцінювання якості тематичної моделі.

У роботі проведено аналіз проблеми і методів тематичного моделювання. Було розглянуто декілька тематичних моделей, а також альтернативні методи розв’язання задачі тематичного моделювання.

Розв’язано задачу тематичного моделювання за допомогою варіаційного автоенкодера. Розроблено програмний продукт, який дозволяє визначати теми вхідного набору документів. За допомогою програмного продукту проведений обчислювальний експеримент, були виявлені теми, поставлено у відповідність кожному документу набір тем, які найгарніше описують тематичне направлення документа.

ABSTRACT

Introductory note: 57 pages, 2 tables, 7 figures, 1 appendixes, 11 sources.

TOPIC MODEL, AUTOENCODER, VARIATIONAL INFERENCE, DIVERGENCE, CONJUGATE DISTRIBUTIONS, BAYESIAN INFERENCE, QUALITY OF TOPIC MODEL.

Object of research – text documents with specific topics.

Purpose of work – the research of the automated variational inference application for the documents topics determining.

Methods of research – the variational autoencoder, text processing, quality assessment of the topic model.

The analysis of problems and methods regarding topic modeling is performed in this work. Several topic models were reviewed, as well as alternative methods for solving the topic modeling problem.

The probabilistic modeling problem was solved, using variational autoencoder. A software product, which allows us to determine topics from the source set of documents has been developed. A computational experiment was performed using the software product, topics were defined, sets of topics that describe the topic direction of the document the best way were placed in accordance with each document.

ЗМІСТ

ВСТУП

Актуальність теми. У зв'язку з розвитком інтернету в наш час з'являється все більше джерел інформації, тому стають дуже популярними алгоритми вилучення та обробки інформації. Алгоритми обробки текстових документів є одними з найпопулярніших і є досить поширеними. Одним із чудових прикладів обробки документів є визначення тематики. Подібні задачі називаються тематичним моделюванням і є надзвичайно корисними, бо завдяки ним знаходити потрібну інформацію стає набагато легше. Наприклад, заздалегідь визначивши тематики документів, можна подавати інформацію користувачам за запитами, які відносяться до конкретної або одразу до декількох тем. Актуальність цієї галузі тільки зростає, з'являються нові типи задач, нові методи обробки текстової інформації.

Мета і завдання кваліфікаційної роботи. Метою кваліфікаційної роботи є розв'язання задачі тематичного моделювання та оцінювання якості побудованої моделі. Для досягнення поставленої мети необхідно виконати наступні завдання:

– провести огляд і аналіз сучасного стану задачі «тематичного моделювання»;

– розв'язати задачу обраним методом;

– оцінити якість побудованої математичної моделі обраним критерієм оцінювання.

Об'єктом дослідження є текстові документи з визначеними темами.

Предметом дослідження є методи визначення тематики документів.

Методи дослідження. У кваліфікаційній роботі використовуються модель варіаційного автоенкодера для розв'язання задачі тематичного моделювання та перплексія як метод оцінювання якості моделі.

Публікації. Результати, отримані у кваліфікаційній роботі, було представлено на 25-му Міжнародному молодіжному форумі «Радіоелектроніка та молодь у XXI столітті» (м. Харків, 20-22 квітня 2021 р.) [11].

1 АНАЛІЗ ПРОБЛЕМИ І МЕТОДІВ ТЕМАТИЧНОГО МОДЕЛЮВАННЯ

1.1 Математичні моделі тематичного оцінювання

1.1.1 Попередня обробка документів

Перед побудовою тематичних моделей текст піддається серії перетворень. Ці перетворення усувають надмірність даних, що покращує роботу алгоритмів, допомагають будувати кращу модель, робити її якомога більш інформативною, а також дозволяють економити комп’ютерну пам’ять.

Лематизація – це приведення кожного слова в документі до нормальної форми. Наприклад, в українській мові нормальними формами вважаються: для іменників – називний відмінок, однина. Для прикметників – називний відмінок, однина, чоловічий відмінок. Для дієслів, дієприкметників та дієприслівників – дієслово в інфінітиві.

Стемінг – це відкидання закінчень та інших змінних частин слів.

Видалення стоп-слів. До них відносяться слова, які зустрічаються дуже часто у всіх документах будь-якої тематики. Вони не є інформативними для тематичного моделювання і їх можна відкинути. Прикладом таких слів можуть бути прийменники, числівники, займенники, деякі дієслова, прикметники та прислівники. Їх відкидання майже ніяк не впливає на обсяг словника, але може сприяти скороченню обсягу деяких документів.

Видалення рідкісних слів. Маються на увазі слова, які не є словами природньої мови (наприклад, які містять цифри або спецсимволи). Такі рідкісні слова зазвичай не впливають на тематику колекції, та їх видалення допомагає в багато разів зменшувати розмір словника, знижуючи витрати часу та пам’яті на побудову моделей.

Існує ще багато методів попередньої обробки текстів, такі як виділення ключових фраз, розпізнавання іменованих сутностей, але вони є не такими важливими, тому в роботі ми будемо користуватися методами, які перелічені вище.

1.1.2 Модель «мішку слів»

Введемо означення. Терм – це слово, нормальна форма слова, словосполучення, або терміни, в залежності від того, які види попередньої обробки текстів були виконані.

Порядок термів в документах не важливий для виявлення тематики документів, тобто тематику колекції можна розпізнати навіть після повної довільної перестановки термів, хоча для людини такий текст втратить суть. Це припущення називають гіпотезою «мішку слів». Порядок документів також не має значення – це припущення називають гіпотезою «мішку документів». Гіпотеза «мішку слів» дозволяє нам перейти до компактного представлення документа як мультимножини – підмножини термів, в якій кожний терм може включатися декілька разів.

1.1.3 Модель документів

Позначимо D – множина (колекція) текстових документів, W – множина (словник) всіх термів, які вживаються в документах. Кожний документ $d \in D$ являє собою послідовність слів $w_1,...,w_{n_d} \in W$, де n_d – довжина документа d в термах. Кожне входження терма w в документ d пов'язано з деякою темою t зі скінченної множини T. Наразі будемо вважати, що кожен документ d представляється вектором «мішку слів», розподіленим за поліноміальним законом розподілу $d \sim Multinomial(n, p)$ з параметрами $n \in \mathbb{N}$, який вказує, яка кількість слів необхідна для генерації документа, та $p \in [0,1]^{|W|}$ – вектор ймовірностей для кожного слова в словнику.

Отже, якщо припустимо, що $D = \{d_i\}_{i=1}^{|D|}$ – множина реалізацій випадкової величини $d \sim Multinomial(n, p)$ з умовною щільністю ймовірності $p(d \mid t; \theta)$, де t – прихована змінна з відомою апріорною щільністю розподілу $p(t)$, тоді бу-

демо вважати, що $T = \{t_i\}_{i=1}^{|T|}$ – множина прихованих змінних, яка використовується для генерації множини документів D.

1.1.4 Дивергенція Кульбака-Лейбнера

Використовуючи формулу Баєса та представлену тематичну модель можна отримати апостеріорну щільність розподілу тем по документах $p(t \mid d;\theta)$:

$$p(t \mid d;\theta) = \frac{p(d \mid t;\theta)\, p(t)}{\int p(d \mid t;\theta)\, p(t)\, dt}. \tag{1.1}$$

Зазвичай обчислення знаменнику (1.1) є складною задачею, яку іноді неможливо розв'язати, тому альтернативним підходом до обчислення апостеріорної щільності розподілу (1.1) є апроксимація його іншою щільністю розподілу $q(t \mid d;\phi) \approx p(t \mid d;\theta)$ [1]. Для вимірювання схожості двох функцій можна використати дивергенцію Кульбака-Лейбнера.

Нехай $q(x)$ та $p(x)$ – деякі функції щільності ймовірності. Тоді дивергенція Кульбака-Лейбнера матиме вигляд:

$$D_{KL}\left[p(x) \| q(x)\right] = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)}\, dx. \tag{1.2}$$

Якщо $q(x)$ та $p(x)$ – дискретні функції щільності розподілу, то

$$D_{KL}\left[p(x) \| q(x)\right] = \sum_{x} p(x) \ln \frac{p(x)}{q(x)}.$$

Для знаходження $q(t|d;\phi)$ треба мінімізувати обернену дивергенцію $D_{KL}\left[q(t|d;\phi)\|\,p(t|d;\theta)\right]$, але через те, що $p(t|d;\theta)$ невідоме, ми не можемо зробити це безпосередньо. Переформуємо (1.2) в термінах $q(t|d;\phi)$ та $p(t|d;\theta)$ для всіх документів D та прихованих змінних T, позначивши $q(T|D;\phi)=\prod_{i=1}^{N} q(t_i|d_i;\phi)$, $p(T|D;\theta)=\prod_{i=1}^{N} p(t_i|d_i;\theta)$:

$$D_{KL}\left[q(T|D;\phi)\|\,p(T|D;\theta)\right]=\int_{-\infty}^{\infty} q(T|D;\phi)\ln\frac{q(T|D;\phi)}{p(T|D;\theta)}dT=$$

$$=\int_{-\infty}^{\infty} q(T|D;\phi)\ln\frac{q(T|D;\phi)}{p(T,D;\theta)}dT+\int_{-\infty}^{\infty} q(T|D;\phi)\ln p(D;\theta)dT=$$

$$=\int_{-\infty}^{\infty} q(T|D;\phi)\ln\frac{q(T|D;\phi)}{p(T,D;\theta)}dT+\ln p(D;\theta).$$

Виразимо з цього рівняння $\ln p(D;\theta)$ [1]:

$$\ln p(D;\theta)=D_{KL}\left[q(T|D;\phi)\|\,p(T|D;\theta)\right]-\int_{-\infty}^{\infty} q(T|D;\phi)\ln\frac{q(T|D;\phi)}{p(T,D;\theta)}dT=$$

$$=D_{KL}\left[q(T|D;\phi)\|\,p(T|D;\theta)\right]+ELBO\left(q(T|D;\phi)\right),\qquad(1.3)$$

де $ELBO\left(q(T|D;\phi)\right)=-\int_{-\infty}^{\infty} q(T|D;\phi)\ln\frac{q(T|D;\phi)}{p(T,D;\theta)}dT$ – нижня межа правдоподібності (evidence lower bound) [1]. Зауважимо, що $\ln p(D;\theta)$ – константа, це означає, що мінімізуючи $D_{KL}(q\|\,p)$, ми максимізуємо $ELBO(q)$, і навпаки.

Отже, розв’язуючи задачу

$$q^{*}\in\arg\max_{q\in Q} ELBO(q),$$

де Q – деяка сім'я функцій щільності ймовірностей, ми знаходимо $q^*(t|d;\phi)$, яка апроксимує апостеріорну щільність $p(t|d;\theta)$.

1.2 Огляд методів розв'язання задачі тематичного оцінювання

1.2.1 Варіаційний вивід

Запропонуємо наступну апроксимацію: $q(T;\phi) \approx p(T|D;\theta)$, а також припустимо, що розподіл $q(T;\phi)$ можна записати наступним чином:

$$q(T;\phi) = q\left(t_1,...,t_{|T|};\phi\right) = \prod_{i=1}^{|T|} q_i(t_i;\phi). \tag{1.4}$$

Використовуючи (1.4) спробуємо побудувати $ELBO(q)$ та використаємо варіаційне числення для знаходження оптимальної функції. Перепишемо $ELBO$, додавши залежність від апроксимуючих щільностей $q_1,...,q_{|T|}$:

$$ELBO\left(q_1,...,q_{|T|}\right) = -\int_{t_1,...,t_{|T|}} \left[\prod_{i=1}^{|T|} q_i(t_i;\phi)\right] \ln \frac{\left[\prod_{k=1}^{|T|} q_k(t_k;\phi)\right]}{p(T,D;\theta)} dt_1,...,t_{|T|} =$$

$$= \int_{t_1,...,t_{|T|}} \left[\prod_{i=1}^{|T|} q_i(t_i;\phi)\right]\left[\ln p(T,D;\theta) - \sum_{k=1}^{|T|} q_k(t_k;\phi)\right] dt_1,...,t_{|T|} =$$

$$= \int_{t_j} q_j(t_j;\phi) \int_{t_{m\neq j}} \left[\prod_{i\neq j} q_i(t_i;\phi)\right]\left[\ln p(T,D;\theta) - \sum_{k=1}^{|T|} q_k(t_k;\phi)\right] dt_1,...,t_{|T|} =$$

$$= \int_{t_j} q_j(t_j;\phi) \int_{t_{m\neq j}} \left[\prod_{i\neq j} q_i(t_i;\phi)\right] \ln p(T,D;\theta)\, dt_1,...,t_{|T|} -$$

$$-\int_{t_j} q_j\left(t_j;\phi\right)\int_{t_{m\neq j}}\left[\prod_{i\neq j} q_i\left(t_i;\phi\right)\right]\sum_{k=1}^{|T|} q_k\left(t_k;\phi\right)dt_1,...,t_{|T|}. \quad (1.5)$$

Позначимо математичне сподівання для всіх змінних, окрім j-ої змінної:

$$E_{m\neq j}\left[\ln p\left(T,D;\theta\right)\right]=\int_{t_{m\neq j}}\left[\prod_{i\neq j} q_i\left(t_i;\phi\right)\right]\ln p\left(T,D;\theta\right)dt_1,...,dt_{j-1},dt_{j+1},...,d_{|T|}. \quad (1.6)$$

Використовуючи (1.6) запишемо (1.5) в наступному вигляді:

$$ELBO\left(q_1,...,q_{|T|}\right)=\int_{t_j} q_j\left(t_j;\phi\right)E_{m\neq j}\left[\ln p\left(T,D;\theta\right)\right]dt_j -$$

$$-\int_{t_j} q_j\left(t_j;\phi\right)\ln q_j\left(t_j;\phi\right)\int_{t_{m\neq j}}\left[\prod_{i\neq j} q_i\left(t_t;\phi\right)\right]dt_1,...,dt_{|T|} -$$

$$-\int_{t_j} q_j\left(t_j;\phi\right)dt_j\int_{t_{m\neq j}}\left[\prod_{i\neq j} q_i\left(t_t;\phi\right)\right]\sum_{k\neq j}\ln q_k\left(t_k;\phi\right)dt_1,...,dt_{j-1},dt_{j+1},...,dt_{|T|}=$$

$$=\int_{t_j} q_j\left(t_j;\phi\right)E_{m\neq j}\left[\ln p\left(T,D;\theta\right)\right]dt_j-\int_{t_j} q_j\left(t_j;\phi\right)\ln q_j\left(t_j;\phi\right)dt_j -$$

$$-\int_{t_{m\neq j}}\left[\prod_{i\neq j} q_i\left(t_t;\phi\right)\right]\sum_{k\neq j}\ln q_k\left(t_k;\phi\right)dt_1,...,dt_{j-1},dt_{j+1},...,dt_{|T|}=$$

$$=\int_{t_j} q_j\left(t_j;\phi\right)\left[E_{m\neq j}\left[\ln p\left(T,D;\theta\right)\right]-\ln q_j\left(t_j;\phi\right)\right]dt_j-G\left(q_1,...,q_{j-1},q_{j+1},...,q_{|T|}\right), \quad (1.7)$$

де $G\left(q_1,...,q_{j-1},q_{j+1},...,q_{|T|}\right)=\int_{t_{m\neq j}}\left[\prod_{i\neq j} q_i\left(t_t;\phi\right)\right]\sum_{k\neq j}\ln q_k\left(t_k;\phi\right)dt_1,...,dt_{j-1},dt_{j+1},...,dt_{|T|}$.

Як можна побачити, ми виділили функціонал, в якому розділили залежність на залежність одного доданка від $q_j\left(t_j;\phi\right)$ та другого доданка від інших $q_{i\neq j}\left(t_{i\neq j};\phi\right)$. Запишемо лагранжіан, використовуючи (1.7):

$$L\left(q_1,...,q_{|T|},\lambda_1,...,\lambda_{|T|}\right)=ELBO\left(q_1,...,q_{|T|}\right)-\sum_{i=1}^{|T|}\lambda_i\left(\int_{t_i}q_i\left(t_i;\phi\right)dt_i-1\right),\qquad(1.8)$$

де $\int_{t_i}q_i\left(t_i;\phi\right)dt_i$ – ліві частини рівняння умови нормування щільності розподілу.

Обчисливши похідну функціоналу (1.8) відносно $q_j\left(t_j;\phi\right)$, використовуючи рівняння Ейлера-Лагранжа, отримаємо:

$$\frac{\partial L\left(q_1,...,q_{|T|},\lambda_1,...,\lambda_{|T|}\right)}{\partial q_j\left(t_j;\phi\right)}=$$

$$=\frac{\partial}{\partial q_j}\left[q_j\left(t_j;\phi\right)\left[E_{m\neq j}\left[\ln p\left(T,D;\theta\right)\right]-\ln q_j\left(t_j;\phi\right)\right]-\lambda_j q_j\left(t_j;\phi\right)\right]=$$

$$=E_{m\neq j}\left[\ln p\left(T,D;\theta\right)\right]-\ln q_j\left(t_j;\phi\right)-1-\lambda_j.$$

Прирівнявши останнє співвідношення до нуля, отримаємо:

$$\ln q_j\left(t_j;\phi\right)=E_{m\neq j}\left[\ln p\left(T,D;\theta\right)\right]-1-\lambda_j=E_{m\neq j}\left[\ln p\left(T,D;\theta\right)\right]+const\,.$$

Розв’язавши це рівняння відносно $q_j\left(t_j;\phi\right)$, остаточно отримаємо:

$$q_j\left(t_j;\phi\right)=\frac{e^{E_{m\neq j}\left[\ln p\left(T,D;\theta\right)\right]}}{Z_j},\qquad(1.9)$$

де Z_j – деяка нормалізуюча константа.

Для обчислення оптимальної $q_j\left(t_j;\phi\right)$ нам також потрібно знати всі значення інших функцій $q_i\left(t_i;\phi\right)$ для обчислення $e^{E_{m\neq j}\left[\ln p\left(T,D;\theta\right)\right]}$. Тому пропонується використовувати наступний ітеративний алгоритм [2]:

1) розпочати з деякими випадковими значеннями параметрів для $q_j\left(t_j;\phi\right)$;

2) обчислити кожну $q_j\left(t_j;\phi\right)$ для мінімізації дивергенції за допомогою (1.9).

Повторювати алгоритм треба до збіжності дивергенції.

1.2.2 Варіаційний автоенкодер

Альтернативною можливістю знаходження апроксимуючої щільності $q\left(t\,|\,d;\phi\right)\approx p\left(t\,|\,d;\theta\right)$ є використання варіаційного автоенкодера.

Нехай апроксимуюча функція щільності має наступний вигляд:

$$q\left(t\,|\,d;\phi\right)=N\left(g_1(d;\varphi_1),g_2\left(d;\varphi_2\right)I\right),$$

тобто функція є щільністю нормального розподілу з параметрами $g_1(d;\varphi_1),g_2\left(d;\varphi_2\right)I$, де I – одинична матриця. Позначимо $g\left(d;\varphi\right)=\left(g_1(d;\varphi_1),g_2\left(d;\varphi_2\right)\right)$ – деяка нейронна мережа, де $\varphi=\left(\varphi_1,\varphi_2\right)$ – параметри навчання мережі. Також позначимо

$$p\left(d\,|\,t;\theta\right)=p\left(t;f\left(t;\gamma\right)\right),$$

де $f\left(t;\gamma\right)$ – нейронна мережа, з параметром навчання γ, яка обчислює вектор параметрів θ для щільності розподілу $p\left(d\,|\,t;\theta\right)$. Функції $g\left(d;\varphi\right)$ і $f\left(t;\gamma\right)$ будемо називати енкодером та декодером відповідно. Передбачається, що апріорна щільність розподілу t – щільність стандартного нормального розподілу:

$$p\left(t\right)=N\left(0,I\right).$$

Оптимальні значення $\left(\varphi^*, \gamma^*\right)$ можуть бути отримані максимізацією $ELBO$. Після отримання оптимальних значень параметрів можна показати процес генерації документів:

1) реалізувати випадкову величину t з розподілу $p(t)$;

2) реалізувати випадкову величину d з розподілу $p\left(d \mid t; \theta^*\right)$, де $\theta^* = f\left(t; \gamma\right)$.

Для обчислення $ELBO$ нам також знадобиться наступна реалізація випадкової величини з розподілу $q(t \mid d; \phi)$, яка називається репараметризацією:

$$t = \mu + \sigma \odot \varepsilon, \tag{1.10}$$

де $\mu = g_1(d)$, $\sigma = g_2(d)$ і $\varepsilon \sim N(0, I)$, $\odot$ – поелементний добуток.

Тож спочатку ми реалізовуємо випадкову величину ε, а потім випадкову величину t, використовуючи значення виходу енкодера (μ, σ).

Для даного методу використовується алгоритм:

1) розпочати ітеративно проходити по множині документів D;

2) на кожному кроці подавати документ d на вхід енкодера для отримання множини параметрів μ та σ апроксимуючої апостеріорної щільності $q(t \mid d; \phi) = N(\mu, \sigma I)$;

3) виконати репараметризацію для реалізації випадкової величини з розподілу $q(t \mid d; \phi)$;

4) подати репараметризовані параметри t до декодера $f\left(t; \gamma\right)$ для отримання параметрів $\theta = f\left(t; \gamma\right)$ генеративної щільності розподілу $p\left(d \mid t; \theta\right)$;

5) обчислити градієнт $ELBO$ відносно φ та γ і виконати крок оптимізації.

1.2.3 ЕМ-алгоритм для ймовірнісної тематичної моделі

Ймовірнісна тематична модель свідчить, що поява термів в документі залежить тільки від теми, залежність від документа виключається. Це припущення можна описати так:

$$p(w|d,t) = p(w|t).$$

Також ця модель передбачає, що ймовірність появи документа d, який пов'язаний з термом w і який відноситься до теми t, залежить тільки від самої теми, тобто

$$p(d|w,t) = p(d|t).$$

З цих викладок можемо отримати, що $p(d,w|t) = p(d|t)\,p(w|t)$.

Позначимо розподіл термів в документі $p(w|d)$ за допомогою формули повної ймовірності:

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

Використовуючи гіпотезу умовної незалежності в результаті отримаємо [3]:

$$p(w|d) = \sum_{t\in T} p(w|t)\,p(t|d) = \sum_{t\in T} \varphi_{wt}\theta_{td}\,.$$

Цей розподіл термів в документі описується ймовірнісною сумішшю розподілів термів в темах $\varphi_{wt} = p(w|t)$ з вагами $\theta_{td} = p(t|d)$. В даному випадку задача тематичного моделювання є оберненою задачею, тобто метою є знаходження φ_{wt} та θ_{td}, і вона зводиться до задачі знаходження наближеного матри-

чного розкладання $F \approx \Phi\Theta$.

Виразимо ймовірності через d, w. Ці ймовірності можна пов'язати з частотами відповідних подій:

$$p(d,w)=\frac{n_{dw}}{n},\ p(d)=\frac{n_d}{n},\ p(w)=\frac{n_w}{n},\ p(w|d)=\frac{n_{dw}}{n_d}, \qquad (1.11)$$

де n_{dw} – число входжень терма w в документ d;

n_d – довжина документа в термах;

n_w – число входжень терма у всіх документах;

n – довжина колекції документів в термах.

Визначимо також ймовірності, пов'язані з прихованою змінною t, через частоти:

$$p(t)=\frac{n_t}{n},\ p(w|t)=\frac{n_{wt}}{n_t},\ p(t|d)=\frac{n_{td}}{n_d},\ p(t|d,w)=\frac{n_{tdw}}{n_{dw}}, \qquad (1.12)$$

де n_{tdw} – число трійок, в яких терм w документа d пов'язаний з темою t,

n_{td} – число трійок, в яких терм документа d пов'язаний з темою t,

n_{wt} – число трійок, в яких терм w пов'язаний з темою t,

n_t – число трійок, пов'язаних з темою t.

Згідно з законом великих чисел при $n \to \infty$ частотні оцінки (1.11), (1.12) наближуються до відповідних ймовірностей у просторі Ω.

Використовуючи основний принцип ЕМ-алгоритму, можна оцінити шукані параметри тематичної моделі φ_{wt} та θ_{td}, використовуючи умовний розподіл $p(t|d,w)$, та навпаки, знаючи φ_{wt} та θ_{td}, можна обчислити $p(t|d,w)$, користуючись формулою Баєса. Е-крок алгоритму для тематичної моделі матиме вигляд:

$$p(t|d,w)=\frac{p(t,w|d)}{p(w|d)}=\frac{p(w|t)p(t|d)}{p(w|d)}=\frac{\varphi_{wt}\theta_{td}}{\sum_s \varphi_{ws}\theta_{sd}}.$$

Щодо М-кроку для тематичної моделі, то можна помітити, що параметри моделі можна знайти, розв'язуючи систему нелінійних рівнянь відносно самих параметрів та допоміжних змінних [3]:

$$p_{tdw}=\frac{\varphi_{wt}\theta_{td}}{\sum_s \varphi_{ws}\theta_{sd}};$$

$$\varphi_{wt}=\frac{n_{wt}}{\sum_{w'} n_{w't}},$$

$$n_{wt}=\sum_{d\in D} n_{dw}p_{tdw};$$

$$\theta_{td}=\frac{n_{td}}{\sum_{t'} n_{t'd}},$$

$$n_{td}=\sum_{w\in W} n_{dw}p_{tdw}.$$

Обчислення виконуються до збіжності Φ та Θ.

1.3 Змістовна та формальна постановка задачі

1.3.1 Змістовна постановка задачі

Розв'язується задача визначення тем для кожного документа з набору документів. Розглядається набір документів з різних джерел. Теми документів заздалегідь відомі, що дозволить дати оцінку якості моделі.

У роботі пропонується використовувати модель, описану в пункті 1.1.3. Кожна тема в представленій тематичній моделі описується дискретним розпо-

ділом ймовірностей слів. Аналогічно кожний документ описується дискретним розподілом ймовірностей тем, тому можна дізнатись тематичну направленість кожного документа.

Задача полягає в знаходженні розподілу термів по темам та тем по документах при відомому розподілі документів по темам та відомому розподілі тем, кількість тем заздалегідь відома.

1.3.2 Формальна постановка задачі

Позначимо D – множина (колекція) текстових документів, W – множина (словник) всіх термів, які вживаються. Термами можуть бути нормальні форми слів, словосполучення або терміни. Кожний документ $d \in D$ є послідовністю слів $w_1,...,w_{n_d} \in W$, де n_d – довжина документа d в термах. Кожне входження терма w в документ d пов'язано з деякою темою t зі скінченної множини T. Для нашої задачі вважатимемо, що документи породжені щільністю розподілу $p(d \mid t;\theta)$, та вважатимемо, що t – приховані випадкові величини з відомою функцією щільністю розподілу. Нехай $T = \{t_i\}_{i=1}^{|T|}$ – множина прихованих змінних, яка використовується для генерації множини документів D.

Задача полягає в знаходженні розподілу $p(d \mid t;\theta)$, який теоретично можна отримати за допомогою формули Баєса (1.1), але на практиці будемо використовувати дивергенцію Кульбака-Лейбнера для знаходження апроксимуючої щільності розподілу $q(t \mid d;\phi) \approx p(t \mid d;\theta)$, тобто задача зводиться до пошуку функції $q(t \mid d;\phi)$, яка мінімізує функціонал $D_{KL}(q \| p)$:

$$q^* \in \arg\min_{q \in Q} D_{KL}(q \| p),$$

де Q – деяка сім'я функцій щільності ймовірностей.

1.4 Постановка задач дослідження

Отже, виходячи з розглянутих моделей, методів та сформульованої постановки задачі, можемо визначити перелік задач дослідження:

– сформулювати задачу тематичного моделювання для визначення тем документів;

– розв’язати задачу тематичного моделювання, використовуючи варіаційний автоенкодер;

– провести обчислювальні експерименти з різними вхідними даними, обраними з різних текстових джерел, використовуючи методи попередньої обробки тексту;

– проаналізувати результати обчислювальних експериментів, оцінити якість моделі автоенкодера.

2 ВИБІР ТА ОБҐРУНТУВАННЯ МЕТОДУ РОЗВ'ЯЗАННЯ

2.1 Метод головних компонент

Нехай $V = \mathbb{R}^n$ – векторний простір, $X = \{x_1, ..., x_N\} \in \mathbb{R}^{n \times N}$ – множина об'єктів з цього простору.

Метод головних компонент – це процес знаходження ортогонального базису для деякого підпростору $U \subset V$ з розмірністю $m < n$ такого, що дисперсія ортогональної проекції множини X на підпростір U – максимальна. З іншого боку, це процес знаходження підпростору U такого, що похибка між ортогональною проекцією множини X на підпростір U та X – мінімальна. Головними компонентами називаються вектори базису підпростору U.

Теорема 2.1 [1]. Нехай $1 \le m < n$, тоді головними компонентами будуть перші m власних векторів коваріаційної матриці, які відповідають m найбільшим власним значенням.

Позначимо $U = \{u_1, ..., u_m\} \subset \mathbb{R}^{n \times m}$ – матриця, стовпці якої – перші m власних векторів коваріаційної матриці S. Представлення множини X з простору V в прихований підпростір U з максимальною дисперсією задається відображенням $e : \mathbb{R}^n \to \mathbb{R}^m$ наступним чином:

$$e(x) = U^T x.$$

Ортогональна проекція множини X з простору V на простір U $p : \mathbb{R}^n \to \mathbb{R}^m$ матиме вигляд:

$$p(x) = UU^T x.$$

Матриця U – розв'язок задачі

$$\min_{U\in\mathbb{R}^{n\times m}} \left\| X_0 - UU^T X_0 \right\|_F^2, \qquad (2.1)$$

$$UU^T = I,$$

де $\| \cdot \|_F$ – матрична норма;

$X_0 = \left\{ \left(x_i - \bar{X} \right) \right\}_{i=1}^{N}$.

Відображення $e(x)$ будемо називати відображенням енкодера, а відображення декодера – це відображення $d : \mathbb{R}^m \to \mathbb{R}^n$:

$$d(y) = Uy.$$

Оператор проекції $p(x) = UU^T x$ будемо називати автоенкодером, де параметри U знаходяться з задачі (2.1).

2.2 Модель автоенкодера

Розглянемо поняття енкодера та декодера. Кодуванням будемо називати процес породження нових ознак шляхом перетворення вихідних ознак. Декодуванням будемо називати зворотний процес. Зменшення розмірності енкодером можна інтерпретувати як стискання даних з вихідного простору в закодований простір, який будемо називати прихованим. Процес декодування можна інтерпретувати як процес розпаковування ознак в вихідний простір. Звичайно, залежно від початкового розподілу даних, розмірності прихованого простору та енкодера частина інформації може бути втрачена під час кодування, і не може бути відновлена під час декодування.

Головна ідея автоенкодера – отримання на виході відгуку, найбільш схожого на вхід з прихованим простором меншої розмірності, ніж вхідний та вихідний простори.

Нехай $enc: \mathbb{R}^n \times W_e \to \mathbb{R}^m$ та $dec: \mathbb{R}^m \times W_d \to \mathbb{R}^n$ – відображення енкодера та декодера відповідно, де $m < n$, W_e, W_d – простори параметрів енкодера та декодера відповідно. Позначимо $X = \{x_1, ..., x_N\} \in \mathbb{R}^{n \times N}$ – множина об'єктів. Параметри моделі автоенкодера можуть бути знайдені з наступної задачі [1]:

$$\min_{(w_e, w_d) \in W_e \times W_d} \left\| X - dec\left(enc\left(X, w_e\right), w_d\right)\right\|_F^2. \tag{2.2}$$

Зауважимо, що на відміну від методу головних компонент, в якому відбувається лінійне відображення, автоенкодер може бути нелінійним за рахунок нелінійної архітектури нейронної мережі, що може покращити результати роботи автоенкодера.

Найпростіша модель автоенкодера – модель, яка складається з одношарових моделей енкодера та декодера:

$$enc\left(x, w_e, b_e\right) = w_e x + b_e,$$
$$dec\left(x, w_d, b_d\right) = w_d x + b_d,$$

де $b_e \in \mathbb{R}^m$, $b_d \in \mathbb{R}^n$ – зміщення мереж енкодера та декодера відповідно;

$w_e \in \mathbb{R}^{m \times n}$ – параметри енкодера;

$w_d \in \mathbb{R}^{n \times m}$ – параметри декодера.

Виходячи з цих моделей, перепишемо задачу (2.2) наступним чином:

$$\min_{(w_e, b_e, w_d, b_d) \in W_e \times W_d} \left\| X - w_d\left(w_e X + b_e \nu_N^T\right) + b_d \nu_N^T \right\|_F^2. \tag{2.3}$$

де ν_N – вектор одиниць розмірності N.

2.3 Зв’язок між методом головних компонент та автоендкодером

Оптимальний розв’язок (2.3) має вигляд:

$$b_d = \frac{1}{N}\left(X - w_d \left(w_e X + b_e v_N^T \right) v_N \right), \tag{2.4}$$

$$w_d = \arg \min_{w \in \mathbb{R}^{n \times m}} \left\| X_0 - w w^+ X_0 \right\|_F^2, \tag{2.5}$$

де w^+ – псевдообернена матриця;

$X_0 = X - \bar{x} v_N^T$;

$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$.

Зважаючи на те, що $\|A\|_F^2 = tr\left(AA^T \right)$, можна перевірити, що мінімізація (2.3) дозволяє нам позначити b_d як (2.4). Після підстановки (2.4) в (2.3) задача може бути сформульована наступним чином:

$$\min_{(w_e, w_d) \in \mathbb{R}^{m \times n} \times \mathbb{R}^{n \times m}} \left\| X_0 - w_d w_e X_0 \right\|_F^2. \tag{2.6}$$

Звідси, для будь-якого b_e оптимальне значення b_d не залежить від b_e. Отже, можемо сфокусуватися тільки на знаходженні ваг w_d, w_e. Продиференціювавши (2.6) відносно w_d, w_e і прирівнявши до нуля, отримаємо:

$$w_d = \left(\left(w_e w_e^T \right)^{-1} w_e \right)^T,$$

$$w_e = \left(w_d^T w_d \right)^{-1} w_d^T.$$

Обравши останній варіант, отримаємо (2.5).

Лема 2.1 (QR-факторізація) [2]. Позначимо матрицю $A \in \mathbb{R}^{n \times m}$, $n \geq m$. Існують така матриця $Q \in \mathbb{R}^{n \times m}$ з ортонормальними стовпцями та верхня трикутна матриця $R \in \mathbb{R}^{n \times m}$ з невід'ємною діагоналлю, такі, що $A = QR$.

Використовуючи лему 2.2 можна показати, що псевдообернена матриця A^+ може буде обчислена за допомогою QR-факторизації наступним чином:

$$A^+ = \left(A^T A\right)^{-1} A^T = \left(R^T Q^T QR\right)^{-1} R^T Q^T = R^{-1} Q^T .$$

Теорема 2.2 [2]. Простір, натягнутий на стовпці матриці w_d, співпадає з простором, натягнутим на перші m головних компонент, обчислених відносно множини X.

Якщо переписати w в рівнянні (2.5) через QR-розкладання та додати обмеження $Q^T Q = I$, тоді отримаємо задачу (2.1), яка відноситься до методу головних компонент.

2.4 Спряжені розподіли

Розглянемо задачу знаходження умовного розподілу випадкової величини x за наявності спостережуваних даних θ:

$$p(x \mid \theta) = \frac{p(\theta \mid x)\, p(x)}{\int\limits_{\theta} p(\theta \mid x)\, p(x)\, dx} .$$

Якщо апостеріорний розподіл $p(x \mid \theta)$ належить тій же сім'ї ймовірнісних розподілів, що і апріорний розподіл $p(x)$, тобто має той же вид розподілу, але з іншими параметрами, то ця сім'я розподілів називається спряженою сім'ї фун-

кцій правдоподібності $p(\theta|x)$. При цьому розподіл $p(x)$ називається спряженим апріорним розподілом до сім'ї функцій правдоподібності $p(\theta|x)$.

Знання спряжених сімей розподілів спрощує обчислення апостеріорних ймовірностей, так як дозволяє замінити обчислення знаменника в формулі Баєса простими алгебраїчними маніпуляціями над параметрами розподілів [5]. Далі ми будемо використовувати цей факт для спрощення побудови варіаційного автоенкодера для тематичної моделі.

2.5 Пряма та обернена дивергенція Кульбака-Лейбнера

Дивергенція Кульбака-Лейбнера не є симетричним функціоналом, тобто $D_{KL}\left[p(x)\|q(x)\right] \neq D_{KL}\left[q(x)\|p(x)\right]$, тому перед нами стоїть додаткова задача вибору однієї з цих дивергенцій. Дивергенцію $D_{KL}\left[p(x)\|q(x)\right]$ будемо називати прямою, а $D_{KL}\left[p(x)\|q(x)\right]$ – оберненою дивергенцією [5].

Розглянемо пряму дивергенцію, яка має вигляд:

$$D_{KL}\left[p(x)\|q(x)\right] = \sum_x p(x)\ln\frac{p(x)}{q(x)}. \tag{2.7}$$

При значеннях $q(x)\to 0$ та великих значеннях $p(x)$ логарифм в рівнянні приймає великі значення. Це означає, що при виборі нашого апроксимуючого розподілу $q(x)$, щоб мінімізувати дивергенцію, нам потрібно покрити всі ненульові значення $p(x)$. На рисунку 2.1 представлені приклади дивергенції функцій.

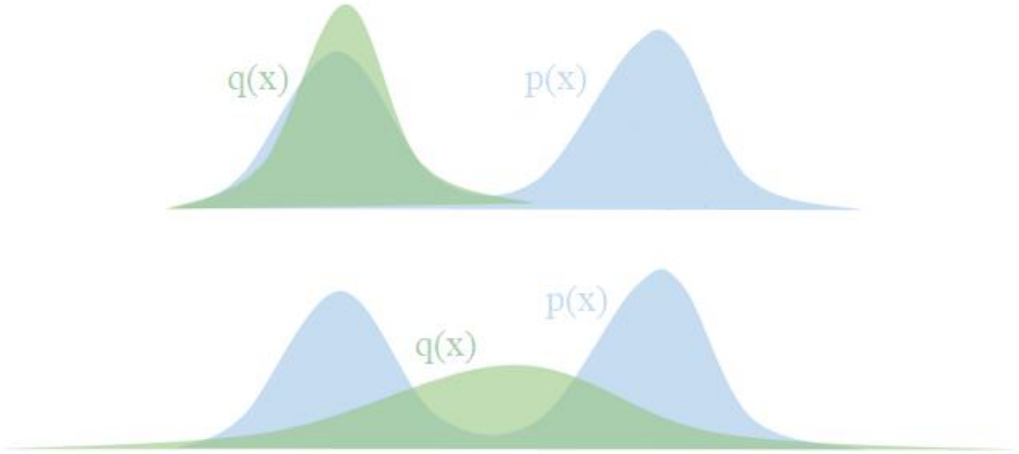

Рисунок 2.1 – Приклади дивергенції Кульбака-Лейбнера

На цьому рисунку можемо побачити приклад, коли $p(x)$ – мультимодальний розподіл, а $q(x)$ – унімодальний. З першого графіку видно, що якщо ми намагатимемось покрити один з «горбів» розподілу $p(x)$ розподілом $q(x)$ – матимемо багато майже нульових значень $q(x)$ та багато великих значень $p(x)$ на іншому «горбі», що призведе до великих значень дивергенції.

На другому графіку можемо побачити, що якщо намагатимемось покрити всі «горби» мультимодального розподілу $p(x)$, розташувавши $q(x)$ всередині, отримаємо менші значення прямої дивергенції відносно першого випадку. Звичайно, це рішення має інші проблеми, наприклад, максимальна щільність $q(x)$ знаходиться в точках, в яких щільність вихідного розподілу $p(x)$ – найменша.

Тепер розглянемо обернену дивергенцію Кульбака-Лейбнера, вона має вигляд:

$$D_{KL}\left[q(x)\| p(x)\right] = \sum_{x} q(x) \ln \frac{q(x)}{p(x)}. \tag{2.8}$$

З рівняння (2.8) видно, що ми отримали зворотну ситуацію: якщо $p(x)$ має малі значення, ми хочемо, щоб функція $q(x)$ мала також відносно малі зна-

чення, інакше дивергенція матиме великі значення. Додатково, коли $p(x)$ має великі значення, логарифм матиме малі значення, що не додає ніяких проблем.

Розглядаючи рисунок 2.1 для оберненої дивергенції, можемо побачити, що помістивши розподіл $q(x)$ всередині між модами $p(x)$, як показано на другому графіку, можемо зіштовхнутися з проблемою, бо на хвостах розподілу $q(x)$ значення $p(x)$ можуть бути малими, що призведе до великих значень дивергенції. Якщо ж розглянемо перший графік для оберненої дивергенції, можемо помітити, що $q(x)$ гарно покриває один з «горбів» $p(x)$, для цих значень логарифм матиме значення, близькі до нуля, а хвости розподілів мають майже однакові значення, на відміну від попередньої ситуації.

Отже, можемо зробити висновок, що для нас найкращим варіантом буде обернена дивергенція, ми будемо її використовувати для знаходження апроксимуючого розподілу.

2.6 Варіаційний автоенкодер для тематичного моделювання

Нагадаємо, ми припускаємо, що кожен документ $d \in D$ представляється моделлю «мішку слів», розподіленою за поліноміальним розподілом:

$$d \sim Multinomial(n, p),$$

де $n \in \mathbb{N}$ – кількість слів, необхідна для генерації документа;

$p \in [0,1]^{|W|}$ – вектор ймовірностей для кожного слова в словнику.

Додатково припустимо, що приховані параметри t також розподілені за поліноміальним розподілом:

$$t \sim Multinomial(1, \pi),$$

де $\pi \in [0,1]^{|T|}$.

В пункті 1.2.2 був наведений алгоритм знаходження параметрів розподілів апроксимуючої апостеріорної щільності $q(t \mid d;\phi) = N(\mu, \sigma I)$ та параметрів $\theta = f(t;\gamma)$ генеративної щільності розподілу $p(d \mid t;\theta)$, в якому одним із пунктів була вказана репараметризація t. Але, використовуючи розподіл $t \sim Multinomial(1,\pi)$, ми не маємо змоги виконати репараметризацію.

По-перше, будемо використовувати розподіл Діріхле замість поліноміального розподілу прихованих параметрів, виходячи з того, що розподіл Діріхле є спряженим до поліноміального розподілу, тобто матимемо [5]:

$$t \sim Dir(\alpha),$$

де α – вектор параметрів розподілу Діріхле.

По-друге, ми можемо апроксимувати розподіл Діріхле логнормальним розподілом, тобто матимемо [5]:

$$t \sim LN(\mu, \Sigma),$$

де $LN(\mu, \Sigma)$ – логнормальний розподіл з параметрами μ та Σ.

З рівняння (1.3) можемо записати $ELBO(q)$ наступним чином [9]:

$$ELBO(q) = -\int_{-\infty}^{\infty} q(T \mid D;\phi) \ln \frac{q(T \mid D;\phi)}{p(T, X;\theta)} dT =$$

$$= \int_{-\infty}^{\infty} \prod_{i=1}^{|D|} q(t_i \mid d_i;\phi) \ln \frac{\prod_{i=1}^{|D|} p(t_i, d_i;\theta)}{\prod_{i=1}^{|D|} q(t_i \mid d_i;\phi)} dt_1, ..., dt_{|D|} =$$

$$= \int_{-\infty}^{\infty} \prod_{i=1}^{|D|} q(t_i \mid d_i;\phi) \sum_{i=1}^{|D|} \ln \frac{p(t_i, d_i;\theta)}{q(t_i \mid d_i;\phi)} dt_1, ..., dt_{|D|} =$$

$$= \sum_{i=1}^{|D|} \int_{-\infty}^{\infty} \prod_{i=1}^{|D|} q\left(t_i \mid d_i;\phi\right) \ln \frac{p\left(t_i, d_i;\theta\right)}{q\left(t_i \mid d_i;\phi\right)} dt_1,...,dt_{|D|} =$$

$$= \sum_{i=1}^{|D|} \int_{-\infty}^{\infty} q\left(t_i \mid d_i;\phi\right) \ln \frac{p\left(t_i, d_i;\theta\right)}{q\left(t_i \mid d_i;\phi\right)} dt_i \cdot \int_{-\infty}^{\infty} \prod_{k \neq i} q\left(t_k \mid d_k;\phi\right) dt_1,...,dt_{k-1},dt_{k+1},...,dt_{|D|} =$$

$$= \sum_{i=1}^{|D|} E_{z_i \sim q(t|d_i;\phi)} \left[\ln p\left(t_i, d_i;\theta\right) - \ln q\left(t_i \mid d_i;\phi\right) \right] =$$

$$= E_{\left\{z_i \sim q(t|d_i;\phi)\right\}_{i=1}^{|D|}} \left[\sum_{i=1}^{|D|} \ln p\left(d_i \mid t_i;\theta\right) + \sum_{i=1}^{|D|} \ln p\left(t_i\right) - \sum_{i=1}^{|D|} \ln q\left(d_i \mid t_i;\phi\right) \right] =$$

$$= E_{\left\{z_i \sim q(t|d_i;\phi)\right\}_{i=1}^{|D|}} \left[\ln \prod_{i=1}^{|D|} p\left(d_i \mid t_i;\theta\right) + \ln \prod_{i=1}^{|D|} p\left(t_i\right) - \ln \prod_{i=1}^{|D|} q\left(t_i \mid d_i;\phi\right) \right] =$$

$$= E_{\left\{z_i \sim q(t|d_i;\phi)\right\}_{i=1}^{|D|}} \left[\ln p\left(D \mid T;\theta\right) + \ln p\left(T\right) - \ln q\left(T \mid D;\phi\right) \right] =$$

$$= E_{\left\{z_i \sim q(t|d_i;\phi)\right\}_{i=1}^{|D|}} \left[\ln p\left(D \mid T;\theta\right) \right] - D_{KL} \left[\ln q\left(T \mid D;\phi\right) \| \, p\left(T\right) \right], \qquad (2.9)$$

де ϕ – параметри, які знаходяться енкодером $\phi = dec\left(d\right)$;

θ – параметри, які знаходяться декодером $\theta = dec\left(t\right)$.

Саме через таке завдання $ELBO\left(q\right)$ виникала необхідність репараметризації (1.10).

Функція розподілу прихованих параметрів має також логнормальний розподіл $p\left(T\right) = LN(\mu, \Sigma)$, параметри розподілу задаються наступним чином [1]:

$$\mu_k = \ln \alpha_k - \frac{1}{|T|} \sum_{i=1}^{|T|} \ln \alpha_i ,$$

$$\Sigma_{kk} = \frac{1}{\alpha_k} \left(1 - \frac{2}{|T|} \right) + \frac{1}{|T|^2} \sum_{i=1}^{|T|} \frac{1}{\alpha_i},$$

де α – вектор параметрів дійсного розподілу тем. Наразі ми припускаємо, що

кожна тема має однакову ймовірність появи в документі, тобто $\alpha_k = \frac{1}{|T|}$.

Апроксимуючий апостеріорний розподіл $q(T \mid D;\phi)$ є логнормальним розподілом, параметри якого знаходяться за допомогою енкодера. Розподіл документів по темам $p(D \mid T;\theta)$ є поліноміальним:

$$p(d \mid t;\theta) = \frac{\Gamma\left(\sum_{i=1}^{|W|} d_i + 1\right)}{\prod_{i=1}^{|W|} \Gamma(d_i + 1)} \prod_{j=1}^{|W|} \theta_j^{d_j}, \tag{2.10}$$

$$p(D \mid T;\theta) = \prod_{i=1}^{|D|} p(d_i \mid t_i;\theta_i).$$

Можна показати, що дивергенція Кульбака-Лейбнера для двох логнормальних розподілів дорівнює дивергенції двох нормальних розподілів. Нехай $p_1(x) = LN(\mu_1, \Sigma_1)$ та $p_2(x) = LN(\mu_2, \Sigma_2)$ – щільності нормальних розподілів, для $x \in \mathbb{R}^n$, тоді дивергенція цих функцій матиме вигляд:

$$D_{KL}[p_1(x) \| p_2(x)] = \frac{1}{2}\left[\ln\frac{|\Sigma_1|}{|\Sigma_2|} - n + tr\left(\Sigma_2^{-1}\Sigma_1\right) + (\mu_2 - \mu_1)^T \Sigma_2^{-1}(\mu_2 - \mu_1)\right]. \tag{2.11}$$

2.7 Критерій якості тематичних моделей

Щоб зрозуміти, наскільки адекватною є представлена модель та чи можна її використовувати в подальшому на інших документах, використовуються методи оцінки якості таких тематичних моделей. Існує декілька типів оцінювання тематичних моделей: внутрішні та зовнішні. Внутрішні методи оцінювання якості оцінюють моделі на початковій колекції документів. Зовнішні методи

оцінюють релевантність моделі з точки зору застосунків, в яких використовуються моделі, та з точки зору користувачів. Для зовнішнього оцінювання зазвичай необхідно збирати додаткові дані.

2.7.1 Інтерпретованість тем

Розрідженість моделі вимірюється відношенням нульових елементів у шуканих матрицях розподілу тем по документах та слів по темам. Припускається, що інтерпретована тема повинна мати лексичне ядро – таку множину слів, які з великою ймовірністю використовуються в даній темі та рідко використовуються в інших темах.

Введемо ядро $W_t = \{w \in W \mid p(t \mid w) > z\}$, де $z \in \mathbb{R}$ – деякий гіперпараметр. Таке ядро теми t визначається як множина термів з великою умовною ймовірністю $p(t \mid w)$ для даної теми. Використовуючи це ядро можемо обчислити показник інтерпретованості теми t [3]:

$$\ker_t = |W_t|,$$

де $|W_t|$ – розмір ядра; його оптимальним значенням є $\frac{|W|}{|T|}$.

Саме розмір ядра може використовуватися для контролю адекватності моделі.

2.7.2 Перплексія

Найбільш розповсюдженим методом оцінювання є перплексія [3]. Перплексія – це міра невідповідності моделі $p(w \mid d)$ термам w, спостережуваним в

документах d колекції D, яка визначається за формулою:

$$P(D;p)=e^{\left(-\frac{1}{n}\sum_{d\in D}\sum_{w\in d}n_{dw}\ln p(w|d)\right)}, \qquad (2.12)$$

де p – модель $p(w|d)$;

n – довжина колекції документів в термах;

n_{dw} – число входжень терма w в документ d.

Якщо терми w породжуються з рівномірного розподілу $p(w|d)=\frac{1}{V}$ на словнику довжини V, тоді перплексія моделі p збігається до V зі збільшенням його довжини. Чим більше розподіл p відрізняється від рівномірного розподілу, тим менше значення перплексії, і навпаки.

3 ПРОГРАМНА РЕАЛІЗАЦІЯ

3.1 Мова Python 3

Мова програмування Python 3 зараз є однією з найчастіше використовуваних мов програмування. Згідно з відомими рейтингами Python 3 входить в трійку найпопулярніших мов і стрімко прагне зайняти перше місце. Ця мова програмування є однією з найпростіших для засвоєння, але також є однією з мов, яку використовують для створення різних додатків.

Велика популярність Python зумовлена своєю універсальністю, завдяки великим можливостям стандартної бібліотеки. Її використовують в найрізноманітніших областях, таких як веб-розробка, машинне навчання, розробка ігор, комп'ютерна безпека, та в наукових дослідженнях. Існує також багато додаткових бібліотек, встановлення яких не викликає труднощів, а документації багатьох бібліотек зрозумілі не тільки для професіоналів, а також і для новачків.

Також однією з переваг мови Python є швидкість розробки. За рахунок динамічної типізації, а також стандартних, широко поширених конструкцій стає легше створювати додатки, а також підтримувати їх. Завдяки віртуальному оточенню можливо запускати додаток на будь-якій ЕОМ без додаткової компіляції програми, що буває дуже зручно, особливо коли швидкість написання програмного продукту є з одним з основних критеріїв вибору мови програмування.

Тема роботи відноситься до задач обробки природньої мови, які є задачами машинного навчання. Як було описано вище, Python є однією з найчастіше використовуваних мов програмування для розв'язання подібних задач, тому в подальшому пропонується використовувати саме Python. Для розв'язання задачі тематичного оцінювання є декілька бібліотек з дуже зручним і зрозумілим інтерфейсом, що також впливає на швидкість розробки та розуміння роботи програмного продукту. Попередня обробка документів є дуже важливою для визначення тематики документу, на щастя, Python дозволяє нам з легкістю використовувати інструменти для цього.

3.2 Алгоритм розв’язання задачі тематичного моделювання

Для розв’язання задачі тематичного моделювання маємо модифікований алгоритм автоенкодера.

Крок 1. Розпочати ітеративно проходити по множині документів D.

Крок 2. На кожному кроці подавати документ d на вхід енкодера для отримання множини параметрів μ та Σ апроксимуючої апостеріорної щільності $q(t \mid d;\phi) = LN(\mu,\Sigma)$.

Крок 3. Виконати репараметризацію за формулою (1.10) для реалізації випадкової величини з розподілу $q(t \mid d;\phi)$, використовуючи параметри, отримані на другому кроці.

Крок 4. Подати репараметризовані параметри t до декодера для отримання параметрів $\theta = dec(t)$ генеративної щільності розподілу $p(d \mid t;\theta)$.

Крок 5. Обчислити $E_{\{z_i \sim q(t|d_i;\phi)\}_{i=1}^{|D|}}\left[\ln p(D \mid T;\theta)\right]$ за формулою (2.10), а також $D_{KL}\left[\ln q(T \mid D;\phi) \| p(T)\right]$, використовуючи формулу (2.11).

Крок 6. Обчислити градієнт $ELBO$ за формулою (2.9) відносно параметрів енкодера φ та параметрів декодера γ і виконати крок оптимізації.

3.3 Опис програми

Програма визначення тематики документів реалізована мовою програмування Python 3 в хмарному середовищі розробки Google Colab. При розробці програмного продукту була використана ключова бібліотека для нейронних мереж tensorflow, бібліотеки для обробки документів: tqdm, CountVectorizer, sklearn.feature_extraction.text, spacy.lang, а також бібліотеки загального призначення для машинного навчання: numpy, sklearn, skelearn.datasets.

Першим кроком мусимо завантажити документи, використовуючи

sklearn.datasets. На вході матимемо 7931 документів, кількість тем – 10. Далі, для застосування алгоритмів необхідно виконати попередню обробку даних. Після попередньої обробки необхідно представити документи в числовому вигляді, щоб мати можливість виконувати математичні операції з документами та термами, для цього використаємо для всіх документів модель «мішку слів». Додатково для видаляємо документи, кількість термів в яких менше десяти.

В роботі описана модель варіаційного автоенкодера, тому на даному етапі побудована нейронна мережа автоенкодера за допомогою tensorflow. Автоенкодер містить дві нейронні мережи: енкодер та декодер. Енкодер містить 4 шари:

– вхідний шар, на нього подаються документи;

– проміжний шар з функцією активації Softplus;

– проміжний шар з функцією активації Softplus;

– вихідний шар з двома виходами: μ та Σ апроксимуючої апостеріорної щільності $q(t|d;\phi)$.

Декодер містить також 4 шари:

– вхідний шар, на нього подаються приховані змінні, які попередньо отримуються з репараметризації з параметрами μ та Σ з енкодера;

– проміжний шар в комбінації з вхідним шаром є матрицею розподілу слів по темам;

– проміжний шар з пакетною нормалізацією, цей шар також додає декореляцію між темами;

– вихідний шар з функцією активації Softmax, цей шар містить параметри θ щільності розподілу $p(d|t;\theta)$.

Для навчання нейронної мережі необхідно розділити дані на тренувальні та тестові. Для цього використано train_test_split з модуля sklearn.model_selection. Корисним методом для запобігання перенавантаження пам’яті, а також збільшення швидкості навчання є розбиття тренувальних даних на пакети.

Для знаходження параметрів нейронних мереж необхідно мінімізувати

ELBO, тому, побудувавши ELBO за запропонованими формулами в роботі, ітеративно мінімізовано ELBO, заздалегідь була обрана кількість епох, яка дорівнює 2000. Для мінімізації використаний алгоритм Adam з модуля tf.keras.optimizers.

В результаті отримуємо матрицю розподілу слів по темам, для кожної теми виводиться 10 слів з найбільшими ймовірностями – саме ці слова найкраще описують відповідну тему. Використовуючи натренований енкодер, отримуємо вектори параметрів μ для кожного документа, за допомогою цих параметрів та функції softmax з бібліотеки numpy обчислюємо розподіл тем по документам. Для опису документів використовуємо 10 тем з найбільшими ймовірностями – саме вони будуть вказувати тематичний напрям документа.

На завершення за допомогою перплексії обчислюється показник якості моделі, чим менше значення перплексії, тим кращою вважається модель.

4 РЕЗУЛЬТАТИ ОБЧИСЛЮВАЛЬНОГО ЕКСПЕРИМЕНТУ ТА ЇХ АНАЛІЗ

Розглянемо результати застосування автоматичного варіаційного виводу для визначення тематики документів з досліджуваного набору даних. За допомогою Python була реалізована програма для визначення ймовірнісного розподілу тем по документах та розподілу слів по темам. Використовуючи бібліотеку sklearn, було завантажено датасет [10] з 7931 англомовних документів, розподілених на 10 тем. Список представлених тем:

– baseball;

– hockey;

– autos;

– motorcycles;

– crypt;

– electronics;

– space;

– med.

Перед застосуванням алгоритму застосуємо методи первинної обробки тексту: лематизацію, стемінг, видалення стоп-слів, видалення рідкісних слів. Приклад вхідного документа, а також результат первинної обробки представлені на рисунках 4.1 та 4.2 відповідно.

```
This is also commonly seen in new teachers.  The first few years, they're
sick a lot, but gradually seem to build up immunities to almost everything
common.  Come to think of it, I was about my healthiest when I was
working in a pathogens lab, exposed to who-knows-what all the time.  Pre-OSHA,
of course.
```

Рисунок 4.1 – Приклад документа

```
commonly seen new teachers years sick lot gradually
build immunities common come think healthiest working
pathogens lab exposed knows time pre osha course
```

Рисунок 4.2 – Результат первинної обробки

З набору документів були виділені унікальні слова, кількість унікальних слів дорівнює 944. Після проведення попередньої обробки були видалені деякі документи, кількість слів в яких менше десяти. Такі документи не дуже інформативні, вони не підходять для навчання моделі, і, навіть людина не завжди зможе визначити тематичне направлення таких документів. В результаті кількість документів для навчання моделі зменшилась до 5158.

Перед навчанням моделі потрібно також задати кількість прихованих тем. В нашому випадку кількість тем буде дорівнювати кількості тем датасету, тобто десяти. Тематичне направлення документа описується розподілом тем, тому продемонструємо 10 документів з розподілами тем. Кожна тема описується розподілом слів, тому продемонструємо отримані теми за допомогою слів з найбільшими ймовірностями. Отримані за допомогою автоенкодера теми представлені в таблиці 4.1. Рисунки 4.1 – 4.3 ілюструють ймовірнісні розподіли слів по темам, зокрема, за даними таблиці 4.1.

На рисунку 4.3 представлені графіки розподілів ймовірностей слів по темам, на рисунку 4.4 – ті ж ймовірності слів впорядковані за спаданням, на рисунку 4.5 наведені накопичені ймовірності слів за кожною темою.

Деякі приховані теми дають чітке розуміння про сфери, про які йде мова. В таких темах найпоказовіші слова мають більшу ймовірність у порівнянні з темами, в яких мова йде про різні сфери, і ймовірності слів, які до них відносяться, є більш розрідженими. Можна також поставити у відповідність отриманим темам вхідні теми з датасету. В нашому випадку для багатьох тем доволі легко можна поставити у відповідність вхідні теми:

– темі № 10 відповідає тема motorcycles;

– темі № 9 відповідає тема crypt;

– темі № 8 відповідає тема hockey;

– темі № 7 відповідає тема space;

– темі № 6 відповідає тема baseball;

– темі № 5 відповідає тема autos;

– темі № 4 відповідає тема electronics;

– темі № 2 відповідає тема med.

Таблиця 4.1 – Розподіл слів по темам

Тема	Слово 1	Слово 2	Слово 3	Слово 4	Слово 5	Слово 6	Слово 7	Слово 8	Слово 9	Слово 10
1	like $2,74 \cdot 10^{-3}$	people $2,59 \cdot 10^{-3}$	know $2,53 \cdot 10^{-3}$	think $2,44 \cdot 10^{-3}$	time $2,35 \cdot 10^{-3}$	good $2,34 \cdot 10^{-3}$	space $2,28 \cdot 10^{-3}$	years $2,21 \cdot 10^{-3}$	use $2,21 \cdot 10^{-3}$	things $2,20 \cdot 10^{-3}$
2	doctor $3,11 \cdot 10^{-3}$	disease $3,10 \cdot 10^{-3}$	patients $3,10 \cdot 10^{-3}$	people $2,78 \cdot 10^{-3}$	medical $2,72 \cdot 10^{-3}$	food $2,70 \cdot 10^{-3}$	medicine $2,69 \cdot 10^{-3}$	patient $2,65 \cdot 10^{-3}$	treatment $2,62 \cdot 10^{-3}$	know $2,59 \cdot 10^{-3}$
3	people $2,91 \cdot 10^{-3}$	mail $2,74 \cdot 10^{-3}$	list $2,55 \cdot 10^{-3}$	group $2,53 \cdot 10^{-3}$	post $2,52 \cdot 10^{-3}$	know $2,52 \cdot 10^{-3}$	like $2,47 \cdot 10^{-3}$	read $2,35 \cdot 10^{-3}$	space $2,33 \cdot 10^{-3}$	think $2,28 \cdot 10^{-3}$
4	use $2,61 \cdot 10^{-3}$	like $2,56 \cdot 10^{-3}$	know $2,52 \cdot 10^{-3}$	radio $2,47 \cdot 10^{-3}$	mail $2,37 \cdot 10^{-3}$	etc $2,35 \cdot 10^{-3}$	want $2,23 \cdot 10^{-3}$	electricity $2,22 \cdot 10^{-3}$	thanks $2,21 \cdot 10^{-3}$	need $2,21 \cdot 10^{-3}$
5	like $2,75 \cdot 10^{-3}$	car $2,51 \cdot 10^{-3}$	good $2,44 \cdot 10^{-3}$	think $2,42 \cdot 10^{-3}$	know $2,38 \cdot 10^{-3}$	time $2,36 \cdot 10^{-3}$	use $2,26 \cdot 10^{-3}$	problem $2,15 \cdot 10^{-3}$	power $2,07 \cdot 10^{-3}$	new $2,05 \cdot 10^{-3}$
6	game $3,73 \cdot 10^{-3}$	hit $3,73 \cdot 10^{-3}$	players $3,45 \cdot 10^{-3}$	year $3,44 \cdot 10^{-3}$	ball $3,19 \cdot 10^{-3}$	league $3,10 \cdot 10^{-3}$	good $3,07 \cdot 10^{-3}$	time $3,05 \cdot 10^{-3}$	baseball $3,04 \cdot 10^{-3}$	runs $3,03 \cdot 10^{-3}$
7	space $3,65 \cdot 10^{-3}$	data $3,22 \cdot 10^{-3}$	use $2,81 \cdot 10^{-3}$	current $2,78 \cdot 10^{-3}$	output $2,76 \cdot 10^{-3}$	input $2,74 \cdot 10^{-3}$	power $2,63 \cdot 10^{-3}$	available $2,60 \cdot 10^{-3}$	shuttle $2,59 \cdot 10^{-3}$	low $2,59 \cdot 10^{-3}$
8	game $5,29 \cdot 10^{-3}$	team $5,17 \cdot 10^{-3}$	hockey $4,37 \cdot 10^{-3}$	games $4,22 \cdot 10^{-3}$	win $3,99 \cdot 10^{-3}$	year $3,91 \cdot 10^{-3}$	period $3,72 \cdot 10^{-3}$	nhl $3,66 \cdot 10^{-3}$	toronto $3,62 \cdot 10^{-3}$	series $3,56 \cdot 10^{-3}$
9	key $5,00 \cdot 10^{-3}$	chip $4,06 \cdot 10^{-3}$	encryption $3,91 \cdot 10^{-3}$	clipper $3,90 \cdot 10^{-3}$	goverment $3,90 \cdot 10^{-3}$	keys $3,47 \cdot 10^{-3}$	nsa $3,39 \cdot 10^{-3}$	system $3,13 \cdot 10^{-3}$	use $3,10 \cdot 10^{-3}$	escrow $3,08 \cdot 10^{-3}$
10	car $4,92 \cdot 10^{-3}$	bike $3,88 \cdot 10^{-3}$	engine $3,41 \cdot 10^{-3}$	cars $2,91 \cdot 10^{-3}$	new $2,87 \cdot 10^{-3}$	miles $2,70 \cdot 10^{-3}$	dealer $2,69 \cdot 10^{-3}$	like $2,69 \cdot 10^{-3}$	riding $2,66 \cdot 10^{-3}$	oil $2,65 \cdot 10^{-3}$

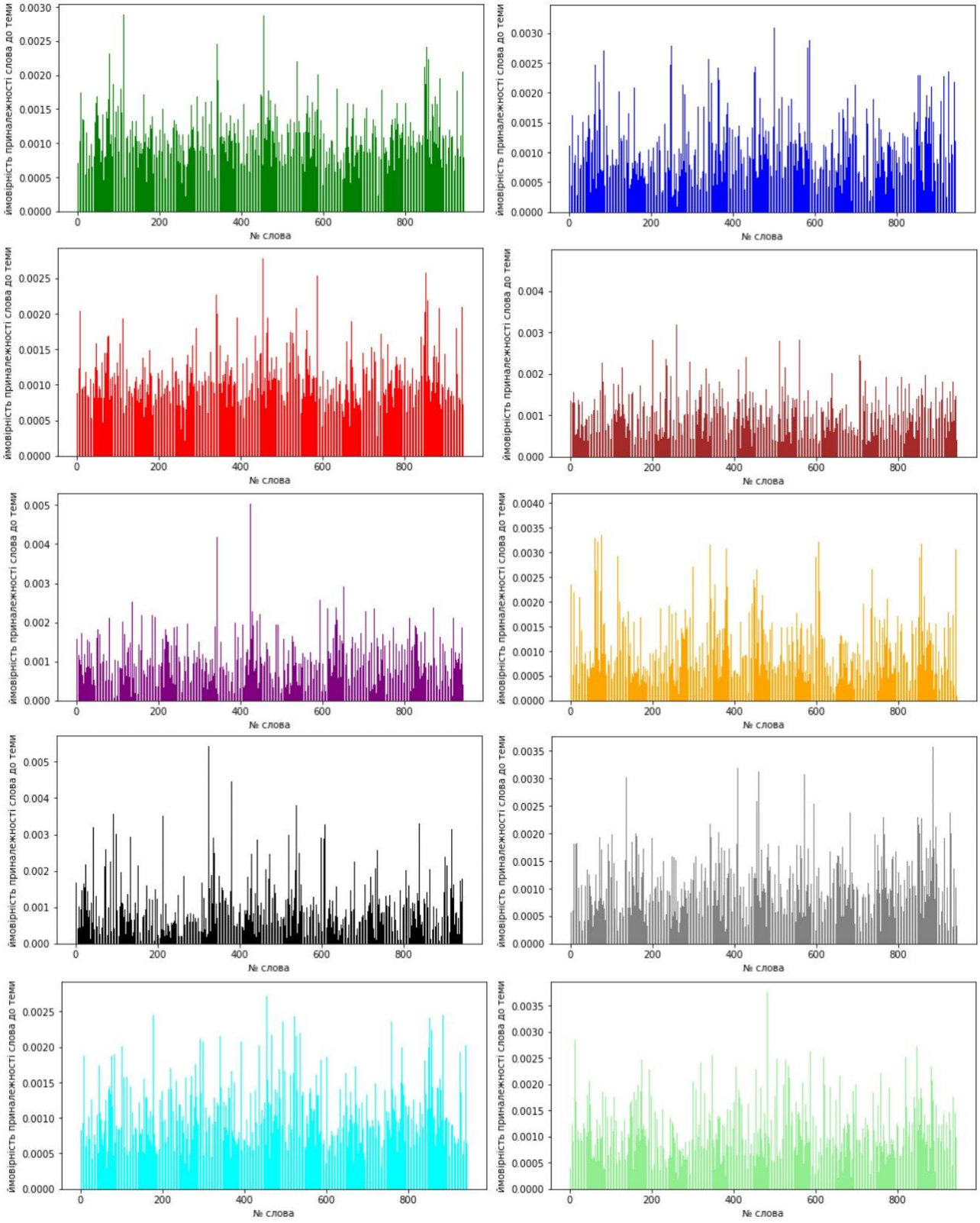

Рисунок 4.3 – Розподіл ймовірностей слів по темам № 1–10 відповідно

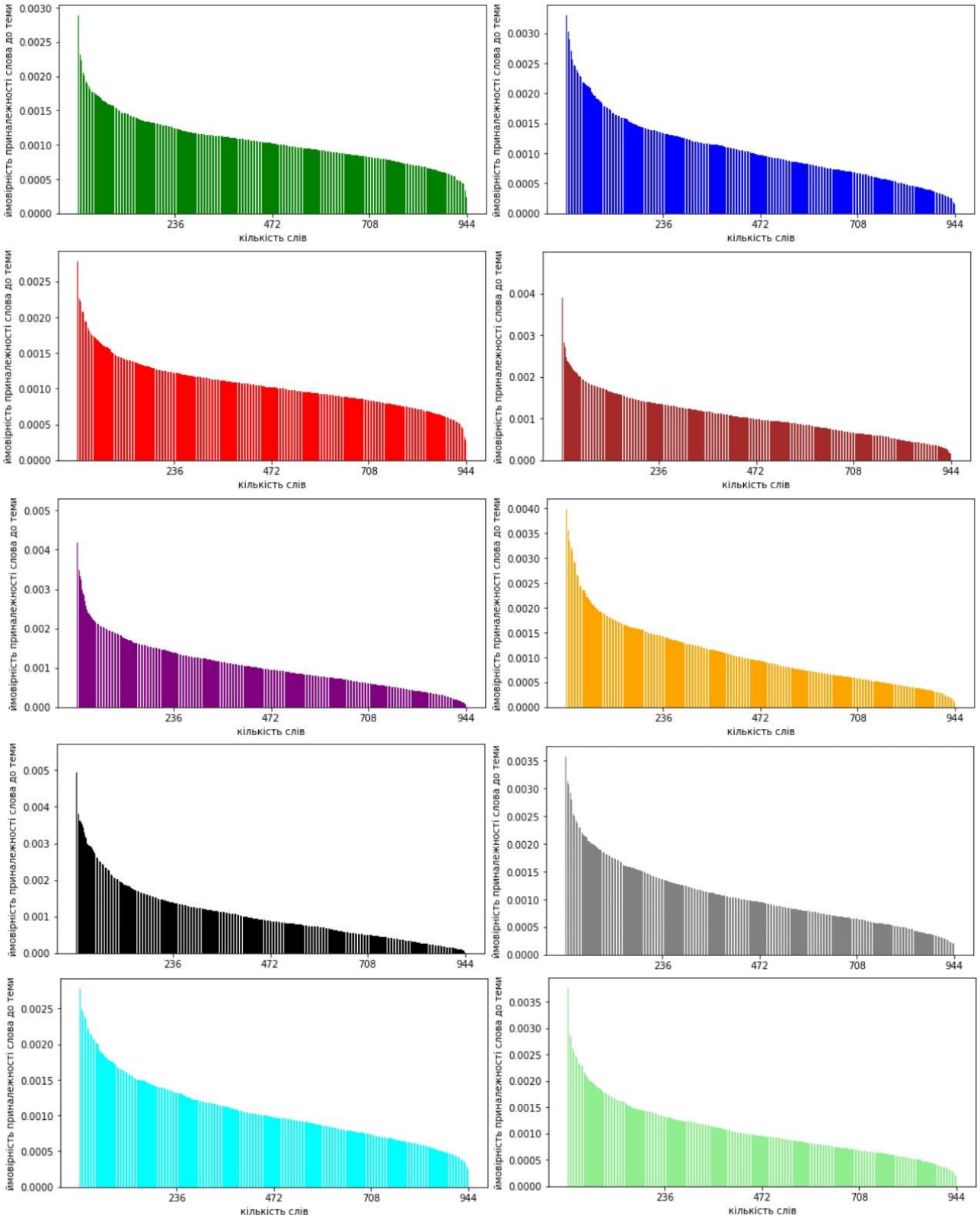

Рисунок 4.4 – Впорядковані за спаданням ймовірності слів по темам № 1–10 відповідно

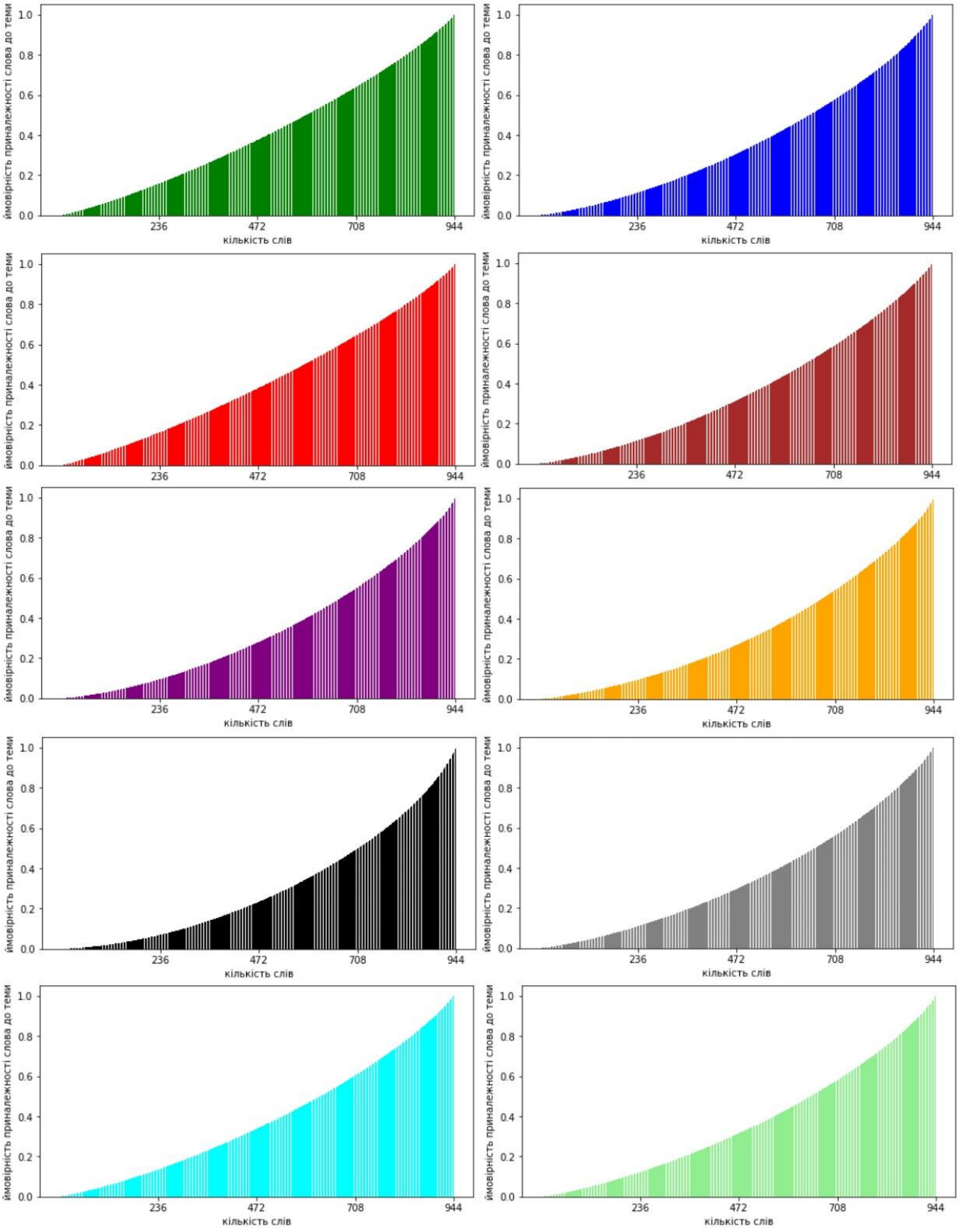

Рисунок 4.5 – Накопичені ймовірності слів у порядку спадання їх ймовірностей

Для тем № 1 та № 3 поставити у відповідність вихідні теми важче, але це не означає, що модель є поганою, бо в даній моделі приховані теми можуть відноситися не до однієї сфери, а можуть описуватися декількома або багатьма сферами.

Розглянемо розподіл тем по документам з вихідного датасету. В таблиці 4.2 представлені номери документів з відповідними номерами тем, які мають найбільшу ймовірність для документа. На рисунку 4.6 представлені графіки розподілів ймовірностей тем по документам. Нижче представлені тексти 10 випадкових документів з датасету:

– документ № 1: «Just in case the original poster was looking for a serious answer. I'll supply one. Yes, even when steering no hands you do something quite similar to countersteering. Basically to turn left, you to a quick wiggle of the bike to the right first, causing a counteracting lean to occur to the left. It is a lot more difficult to do on a motorcycle than a bicycle though, because of the extra weight. (Ok, so my motorcycle is heavy. Maybe yous isn't.)»;

– документ № 2: «Detroit is a very disciplined team. There's a lot of Europeans in Detroit which would make the game fast, so Toronto would have to slow the game down, which means drawing penalties, as a last resort anyway. Toronto will be a good team as soon as they get more good players. Toronto is just an average team, Detroit isn't Ballard screwed Toronto when he was owner. Everyone knows that. and it's going to take time for Toronto to become a real force. I expect Gilmour to be burnt out next year. He can't pull the whole team forever.»;

– документ № 3: «You obviously haven't read the information about the system. The chips put out serial number infomation into the cypher stream to allow themselves to be identified. The system does not rely on registering people as owning particular phone units. And probably as a back door to allow re-generation of the secret key. Have we determined yet that S1 and S2 don't ever change?»;

– документ № 4: «Last year Brein Taylor was in A ball, probably at Tampa in the Florida State League. I believe he began this year in AA which is Albany. Hopefully George won't rush him and he'll be allowed to progress at his own rate to AAA and then to the Bronx. This guy is the real thing.»;

– документ № 5: «For a good discussion of cryptographically "good" random number generators, check out the draft-ietf-security-randomness-00.txt Internet Draft, available at your local friendly internet drafts repository. A reasonably source of randomness is the output of a cryptographic hash function (e.g., MD5), when fed with a large amount of more-or-less random data.»;

– документ № 6: «An interesting note ... I have absolutely no recollection who was on my team. I picked all my players about 2 weeks before the start of the season, and then never touched the roster again. I got wrapped up in my own "money" pool and decided not to get involved at all with the USENET pool (sorry Andrew btw). The only thing I remember about my team is that I had Joe Sacco and maybe John MacLean. Maybe Francis and Kevin Stevens as well.»;

– документ № 7: «This is true, but the main thing the commish i.e. Selig needs to do is to suspend Bobby Cox. You *cannot* allow a team to come out at the ump as the Braves did. I usually rip umps, but in this case, the players were dead wrong. Cox should go for 5 games. If I mhad ever umped a game where that happened, I'd have ejected every player that came out.»;

– документ № 8: «How about rec.radio.amateur.packet? At least at my site, there is no general packet radio (i.e. non-amateur) newsgroup. That said, I would definately subscribe to r.r.a.packet if you want to learn about all aspects of amateur packet radio, at both the high and low ends. Also, I would get the FAQ from the group, and then post any specific questions to that group»;

– документ № 9: «Unfortunately, you're wrong on both counts. The most common method of implementing a tunable receiver is to have a local oscillator. The local oscillator's frequency can be radiated out of the receiver via the antenna unless the circuit is designed and constructed with great care.»;

– документ № 10: «Herman, I would think you of all people would/could distinguish between "health" and "treatment of disease." All the prevention medicine people preach this all the time. You cannot buy health. You can buy treatment of disease, assuming you are lucky enough to have a disease which can be treated. A rich person with a terminal disease is a bit out of luck.».

Таблиця 4.2 – Розподіл тем по документам

Документ	Тема 1	Тема 2	Тема 3	Тема 4	Тема 5	Тема 6	Тема 7	Тема 8	Тема 9	Тема 10
1	0,081	0,022	0,057	0,034	0,080	0,027	0,018	0,012	0,013	0,655
2	$7,48 \cdot 10^{-4}$	$6,68 \cdot 10^{-4}$	$6,24 \cdot 10^{-4}$	$6,58 \cdot 10^{-4}$	$7,34 \cdot 10^{-4}$	$1,85 \cdot 10^{-3}$	$5,97 \cdot 10^{-4}$	0,992	$2,41 \cdot 10^{-4}$	$7,28 \cdot 10^{-4}$
3	0,002	0,003	0,027	0,003	0,003	0,002	0,003	0,001	0,977	0,002
4	0,007	0,005	0,005	0,005	0,005	0,941	0,005	0,005	0,006	0,015
5	0,001	0,002	0,002	0,619	0,001	0,001	0,003	0,001	0,368	0,001
6	0,021	0,015	0,023	0,031	0,014	0,065	0,077	0,701	0,039	0,011
7	$9,82 \cdot 10^{-4}$	$8,51 \cdot 10^{-4}$	$1,01 \cdot 10^{-3}$	$1,06 \cdot 10^{-3}$	$9,21 \cdot 10^{-4}$	0,986	$9,73 \cdot 10^{-4}$	$6,20 \cdot 10^{-3}$	$9,47 \cdot 10^{-4}$	$8,38 \cdot 10^{-4}$
8	0,049	0,029	0,205	0,021	0,027	0,015	0,571	0,024	0,041	0,017
9	0,096	0,066	0,106	0,295	0,147	0,082	0,044	0,041	0,063	0,086
10	$3,48 \cdot 10^{-4}$	0,997	$3,51 \cdot 10^{-4}$	$2,22 \cdot 10^{-4}$	$3,00 \cdot 10^{-4}$	$3,05 \cdot 10^{-4}$	$2,32 \cdot 10^{-4}$	$2,69 \cdot 10^{-3}$	$3,55 \cdot 10^{-4}$	$3,19 \cdot 10^{-4}$

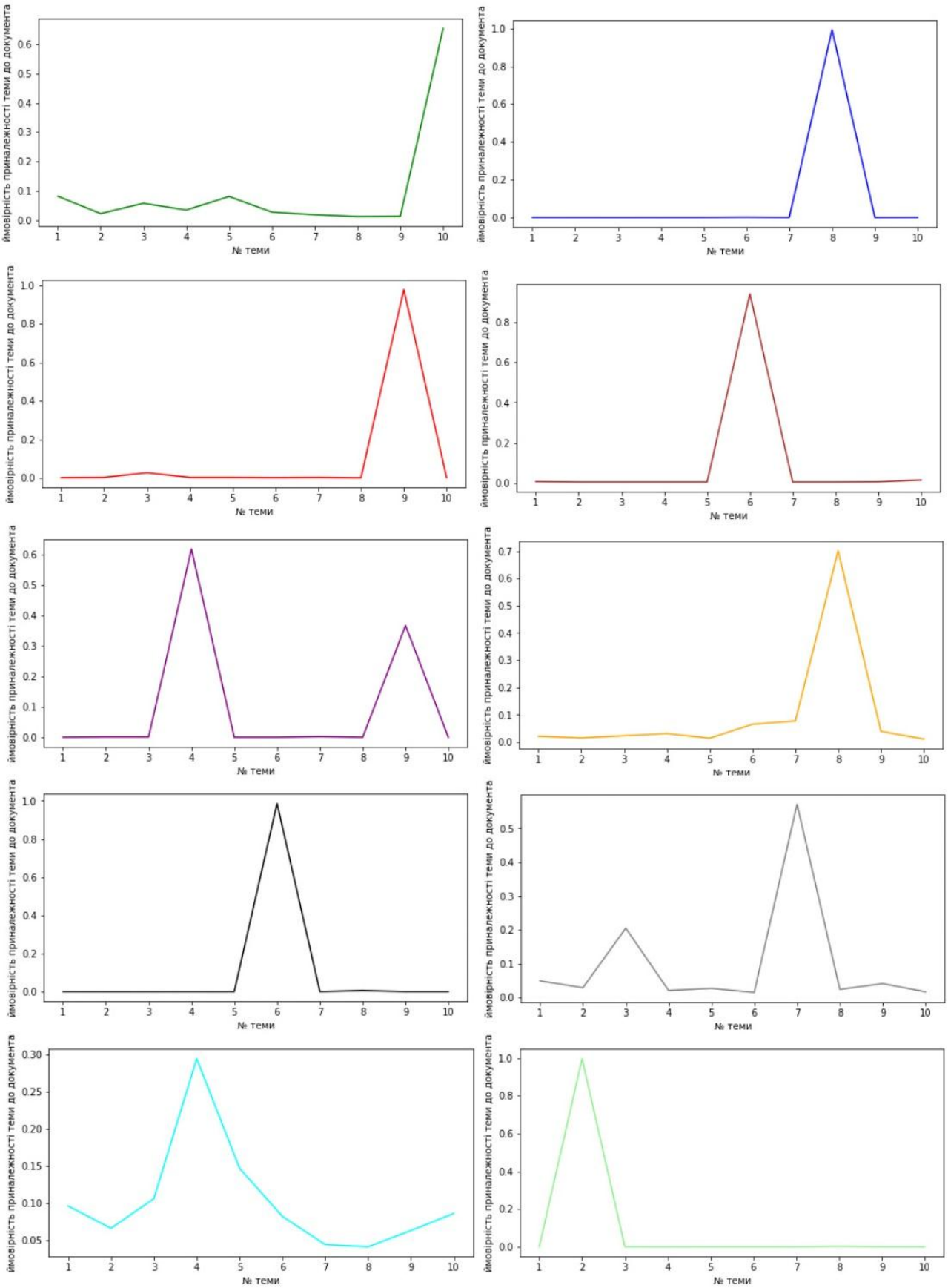

Рисунок 4.6 – Графіки розподілів тем по документам № 1– 10

Отже, аналізуючи отримані результати, можна зробити наступні висновки:

– документ № 1 найімовірніше відноситься до теми № 10, тобто в ньому основною темою є мотоцикли, також цей документ має відношення до тем № 1 та № 5, що говорить нам про відношення до теми авто;

– документ № 2 відноситься до теми № 8 з найбільшою ймовірністю, тобто в ньому йде мова про хокей, при цьому цей документ майже не має відношення до інших тем, що говорить нам строгу тематичну односпрямованість документа;

– для документа № 3 тема № 9 має найбільшу ймовірність, тобто можна зробити висновок, що тема документа – криптологія, до інших тем документ відношення майже не має;

– документ № 4 має відношення до теми № 6 з ймовірністю 0,941, тобто в документі описується тема бейсболу, до інших тем документ відношення майже не має;

– документ № 5 має більш розріджений розподіл тем, теми № 4 та № 9 мають найбільші ймовірності, отже, в документі йде мова про криптологію та електроніку;

– документ № 6 має декілька тематичних направлень: найбільшу ймовірність має для теми № 8, тобто до хокею, а також він відноситься до тем № 7 та № 6, тобто до космосу та бейсболу. Хоча космос має не багато спільного зі спортом, але ми розуміємо, що в цьому документі є слова, які лише відносяться до цієї тематики, і не є домінуючим фактором у визначені тематики документа, про що говорить нам ймовірність цієї теми для документа;

– тема № 6 має велику ймовірність для документа № 7, це означає, що в даному документі розповідається про бейсбол, а також документ зачіпляє деякі інші сфери, виходячи з розподілу ймовірностей тем;

– документ № 8 має розріджений розподіл тем, тобто в ньому йде мова про декілька тем: тему № 7, яка пов'язана з космосом, а також тему № 3, яку важко віднести до конкретної сфери, але остання тема не є домінуючою, тому можна зробити висновок, що документ розповідає саме про космос;

– для документа № 9 можна зробити висновок, що він відноситься до багатьох тем одразу, оскільки бачимо майже рівномірний розподіл тем по цьому документу. Це може означати, що даний документ характеризується великим спектром тематичних направлень або є дуже коротким для визначення домінуючої теми;

– тема № 2 має велику ймовірність документа № 10, це означає, що цей документ розповідає про медицину.

Результати, наведені у таблиці 4.2, графічно проілюстровано на рисунку 4.6, де відображено ймовірності розподілу тем у кожному з проаналізованих вище документів (№ 1– 10).

За допомогою перплексії (2.12) була оцінена якість моделі. Значення перплексії дорівнює 64,7401, довжина словника дорівнює 944. Даний результат дозволяє нам зробити висновок, що отримана модель є гарною, оскільки значення перплексії набагато менше, ніж довжина словника, тобто генеративна модель генерує терми в документах не з рівними ймовірностями, що для моделі є ключовим показником якості.

Проаналізувавши кожен документ та результати, отримані алгоритмом, можна зробити висновок, що розглянутий метод доволі точно зміг визначити тематичне направлення кожного документа, це означає, що реалізований продукт можна використовувати в різних застосунках.

Зауважимо, що для деяких документів визначити домінуючу тему важко, зокрема, через такі причини: документ може бути дуже коротким, тому для нього розподіл тем є рівномірним, або в ньому насправді перетинається велика кількість тем.

ВИСНОВКИ

У ході роботи був проведений аналіз моделей та методів проблеми визначення тем документів. До них, зокрема, належить модель автоенкодера. Ця модель зараз є однією з найпопулярніших серед дослідників і гарно описує процес тематичного оцінювання.

Використовуючи її можна доволі точно оцінювати теми документів, бо модель побудована на нейронних мережах, які стрімко набирають популярність останніми роками, а одним з перспективних напрямків їх застосувань є обробка природньої мови, де вони показують гарні результати. Окрім високої точності перевагою цієї моделі є швидкість алгоритму.

У роботі була розглянута тематична модель на основі автоенкодера, за допомогою якої було розв’язано задачу тематичного моделювання. Порівнюючи результати, отримані за допомогою автоенкодера, з результатами інших методів, зокрема, використаних у бакалаврській роботі, можна з повною впевненістю сказати, що автоенкодер є одним з найкращих методів розв’язання подібних задач обробки текстів природньої мови.

У результаті виконання роботи був реалізований програмний продукт мовою програмування Python, завдяки якому була розв’язана поставлена задача на великому наборі документів. Програмний продукт можна активно використовувати в різних застосунках, які потребують розв’язання задачі тематичного моделювання. Додатково можна сказати, що програмний продукт доволі зрозумілий для інших розробників, що дозволяє покращувати його колективно.

Зауважимо, що в роботі розглянуті не всі можливі форми автоенкодера, а також не всі можливі тематичні моделі. Це означає, що дослідження, розпочаті в даній роботі, можуть бути продовжені, а сама тема є актуальною і дуже перспективною.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Diederik P. Kingma, Max Welling An Introduction to Variational Autoencoders // Foundations and Trends in Machine Learning. 2019. Vol. 12, no 4. Pp. 6–24.

2. Bishop C. Pattern Recognition and Machine Learning. New York : Springer-Verlag, 2006. Pp. 439–450.

3. Воронцов К. В., Потапенко А. А. Модификации EM-алгоритма для вероятностного тематического моделирования // Машинное обучение и анализ данных. 2013. Т.1, № 6. С. 657–686.

4. Stigler S. M. Thomas Bayes's inference // Journal of the Royal Statistical Society: Series A. 1982. Vol. 145. Pp. 250–258.

5. Jordan M., Gharamani Z., Jaakkola T. An introduction to variational methods for graphical models // Learning in Graphical Models. MIT Press, 1998. Pp. 105–162.

6. Corduneanu A., Bishop C. Variational Bayesian model selection for mixture distributions // Proc. AI and Statistics Conf. 2001. Pp. 27–34.

7. Baldi P., Vershynin R. The capacity of feedforward neural networks // Neural networks. 2019. Pp. 288–311.

8. Ganguly A., Earp S. W. F. An Introduction to Variational Inference // Sertis Vision Lab. 2021. Pp. 1–13.

9. Doersch C. Tutorial on Variational Autoencoders. Carnegie Mellon, 2016. Pp. 4–18.

10. 20 Newsgroups. URL : http://qwone.com/~jason/20Newsgroups/ (дата звернення: 05.12.2021).

11. Деркач О. С. Використання варіаційного виведення для латентного розміщення Діріхле в задачі тематичного моделювання // 25-й Міжнародний молодіжний форум «Радіоелектроніка та молодь у XXI столітті» : зб. матеріалів форуму (м. Харків, 20-22 квітня 2021 р.). Т. 7. Харків : ХНУРЕ, 2021. С. 71–72.