DOI 10.31891/2307-5732-2022-307-2-37-41
УДК 637.5.02

БЕРДНИК Д. А.
https://orcid.org/0000-0002-8092-9228
e-mail: danylo.berdnyk.knm.2018@lpnu.ua
БОЙЧУК А. Б.
Національний університет "Львівська Політехніка"
https://orcid.org/0000-0002-0563-5748
e-mail: andrii.r.boichuk@lpnu.ua

ПОРІВНЯЛЬНИЙ АНАЛІЗ МЕТОДІВ ТЕМАТИЧНОГО МОДЕЛЮВАННЯ ДЛЯ АНАЛІЗУ ВІДГУКІВ В ІНТЕРНЕТ-МАГАЗИНІ ЦИФРОВИХ ТОВАРІВ

В цьому дослідженні проводиться порівняльний аналіз методів тематичного моделювання для використання на текстових документах взятих з відгуків до цифрових товарів у інтернет-магазині. Тематичне моделювання – це техніка машинного навчання без спостерігача, яка дозволяє розкрити, дослідити та анотувати колекцію документів. Три з найбільш популярних моделей тематичного моделювання, які представлені у цій роботі, для дослідження документів є прихований семантичний аналіз LSA, ймовірнісний прихований семантичний аналіз PLSA та приховане розміщення Діріхле LDA. Порівняльний аналіз проводиться за допомогою таких числових метрик як когерентність та перплексія та метрики оцінки "на людське око" за допомогою візуалізації результатів за допомогою хмари слів для різних параметрів цих методів. На додачу було проведено порівняння методів за продуктивністю.

Ключові слова: тематичне моделювання, порівняльний аналіз, та приховане розміщення Діріхле, прихований семантичний аналіз, перплексія, когерентність

DANYLO BERDNYK, ANDRII BOICHUK
Lviv Polytechnic National University

COMPARATIVE ANALYSIS OF THEMATIC MODELING METHODS FOR ANALYSIS OF REVIEWS IN THE ONLINE STORE OF DIGITAL GOODS

Nowadays, people often use online services for their daily tasks. The Internet has increased the demand for applications and services to provide a better customer experience. However, nowadays the Internet is full of information that can make it difficult to understand customer needs and confuse users when searching for the information they need. Therefore, there is a need to use effective methods and tools that can help in identifying and analyzing information from a large number of sources stored as online text. For such tasks, it is convenient to use natural language processing - an industry that combines the capabilities of computational linguistics, computer science and artificial intelligence to allow computer to understand and analyze meaning of human speech. One of the fundamental tasks of natural language processing is the definition of keywords. Identified keywords are used to determine the needs of users of the product when it comes to analyzing product reviews, and quickly find information about the product by the average user. Topic modeling methods are often used to determine keywords in the text

This study provides a comparative analysis of topic modeling methods for use in text documents taken from reviews of digital products in the online store. Topic modeling is an unsupervised machine learning technique that allows you to analyse collection of documents and divide them into different topics. Three of the most popular topic modeling methods presented in this paper for document research are latent semantic analysis LSA, probabilistic latent semantic analysis PLSA, and latent Dirichlet allocation LDA. Comparative analysis is performed using numerical metrics such as coherence, perplexity and "human eye" evaluation metrics using word cloud visualization of results for different parameters of these methods. In addition, a comparison of performance methods was performed.

Keywords: topic modelling, comparative analysis, latent semantic analysis, latent Dirichlet allocation, coherence, perplexity

Постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями

В сучасному світі люди часто використовують інтернет сервіси для своїх щоденних задач. Інтернет збільшив попит на розробку додатків та послуг для забезпечення більш якісного досвіду клієнтів. Проте в наші дні інтернет переповнений інформацією, яка може ускладнити розуміння потреб клієнтів та заплутати користувачів під час пошуку потрібної інформації, яка їх цікавить. Тому існує потреба у використанні ефективних методів та інструментів, які можуть допомогти у виявленні та аналізі інформації з великої кількості джерел, які зберігаються як онлайн-текст. Для таких задач зручно використовувати обробку природної мови – галузь, яка поєднує можливості обчислювальної лінгвістики, інформатики та штучного інтелекту, щоб дозволити комп'ютеру розуміти, аналізувати та генерувати значення природного людського мовлення. Однією з фундаментальних задач обробки природної мови є визначення ключових слів. Визначені ключові слова використовуються для визначення потреб користувачів товару, якщо йдеться мова про аналіз відгуків до товару, та швидкого пошуку інформації про товар звичайним користувачем. Для визначення ключових слів у тексті часто використовують методи тематичного моделювання. Тематичне моделювання є технікою машинного навчання без спостерігача (unsupervised learning), яка використовує статистичні операції для отримання корисного вмісту з неструктурованих даних. Одними з найвідоміших методів тематичного моделювання є латентний семантичний аналіз (LSA), імовірнісний латентний семантичний аналіз (PLSA) та латентний розподіл Діріхле (LDA). Основна ідея LSA полягає у тому, що сукупність всіх контекстів у яких слово з'являється або не з'являється забезпечує набір взаємних обмежень, які значною мірою визначають схожість значень слів і наборів слів один до одного. PLSA, порівняно зі стандартним латентним семантичним аналізом (LSA) що випливає з лінійної алгебри та виконує декомпозицію одиничного значення (SVD) таблиці співучасностей, pLSA заснована на розкладі суміші отриманої з

латентної класової моделі. Це призводить до більш принципового підходу, який має тверде підґрунтя у статистиці. У LDA кожна тема змодельована як безкінечна суміш покладена в основі набору тематичних ймовірностей. В контексті текстового моделювання тематичні ймовірності надають явне представлення документу. Проте існують багато проблем у аналізі текстів взятих з інтернету, такі як помилки в словах, зашумленість текстів беззмістовними словами та вживання сленгових слів.

Отже, актуальність теми дослідження зумовлена зростанням інформаційних потоків та необхідністю зберігання великої кількості інформації призначеної для подальшого видобутку інформації з неї та аналізу та потребою подальшого дослідження використання методів тематичного моделювання для аналізу текстів з інтернету. Результати дослідження можуть бути корисними для отримання нюансів використання методів тематичного моделювання, які використовуються в даній роботі, для аналізу відгуків до товару в інтернет-магазині цифрових товарів.

Аналіз досліджень та публікацій

Дослідниками у праці [1] було представлено та описано алгоритм LDA. У цій статті надається порівняння моделі LDA з іншими відомими тематичними моделями такими як LSA та PLSA. Також в цій статі автори надають приклад використання LDA на прикладі реальних даних новин. У праці [2] автором Thomas Hofmann було описано модель ймовірнісного латентного аналізу PLSA, наведено приклад її застосування. Також у цій роботі наведено порівняння моделі PLSA з моделлю LSA.

У роботі [3] автором Ashish Bindra було досліджено роботу алгоритму LDA для аналізу даних зібраних зі сторінок користувачів у соціальних мережах. Автором було показано різницю в часі роботи різних моделей для різної кількості тем. Авторами роботи [4] Thomas K Landauer, Peter W. Foltz, Darrell Laham був описаний алгоритм LSA, були надані приклади роботи алгоритму на текстових даних та було оцінено як LSA справляється з визначенням синонімів у текстах.

Проте використання та порівняння методів тематичного моделювання для текстових даних з інтернету потребує подальшого дослідження.

Формулювання цілей статті

Метою даного дослідження є провести аналіз відгуків до товару в інтернет-магазині цифрових товарів такими методами тематичного моделювання як метод метод латентного семантичного аналізу LSA імовірнісного латентного семантичного аналізу PLSA та латентне розміщення Діріхле LDA. Також до мети входить продемонструвати роботу кожного з перерахованих методів на прикладі аналізу відгуків до товару та провести порівняльну характеристику використаних методів для даної області дослідження і визначити який з методів має ліпші результаті у проведеному аналізі.

Методи тематичного моделювання

Розглянемо задачу побудови тематичної моделі.

Нехай задана колекція текстових документів D. Кожен документ *d* з колекції D представляє собою послідовність слів $W_d = (w_1,\dots, w_{n_d})$ з словника W, де n_d - довжина документу *d* Передбачається, що кожен документ може відноситись до однієї або декількох тем. Теми відрізняються один від одного різною частотою вживання слів. Потрібно знайти число тем, розподіл частот слів, характерний для кожної з тем та тематику кожного документу, тобто в якій степені він відноситься до кожної з тем.

Перед застосуванням тематичної моделі на даних потрібно здійснити попередню обробку даних. Кроками обробки текстових даних є наступні:

- Застосувати регулярні вирази/нормалізація - замінити великі перші літери на малі, забрати пунктуацію та цифри,
- Токенізація - розділити текст на малі частини, які називаються токенами,
- Забрати стоп слова - забрати слова, які використовуються дуже часто в мові. Наприклад слово “the” в англійській мові,
- Леманізація - привести всі слова до їхніх нормальних форм.

LSA – одна з фундаментальних технік, яка дозволяє вирішити цю проблему. LSA базується на виконанні таких дій, як: обрахувати матрицю частот появи певних слів у певних документах, рядками матриці є документи, а стовпцями - слова; виконати SVD (singular value decomposition) над отриманою матрицею. Обрахування матриці частот виконується без урахування порядку слів. Замість обрахування звичайної частоти слів у документах зазвичай використовують *tf-idf* частоти, які обраховуються за наступною формулою:

$$w_{ij} = tf_{i,j} * log(\frac{N}{df_i}) \quad (1)$$

де $w_{i,j}$ – значення *tf-idf* частоти, $tf_{i,j}$ – частота появи слова у документі, *N* – загальна кількість документів, df_i – документи які містять слова.

Таке представлення частот дозволяє словам, які часто з’являються в одному документі але також часто з’являються в інших документах не мати відносно *tf-idf* показник, а словам, які часто з’являються в одному документі, а в інших з’являється відносно рідко мати відносно високе значення *tf-idf* частоти. Оскільки знайдена матриця частот є надлишкова у багатьох своїх вимірах ми маємо провести зменшення

розмірності матриці частот. Зменшення розмірності може бути виконане розбиттям матриці за допомогою SVD методу. Після розкладання матриці частот A на добуток трьох матриць $A = USV^T$, S є діагональна матриця сингулярних значень. Кожне значення матриці S є потенційною темою знайденою в документах. З цих значень обирається k найбільших значень разом з першими k стовпцями матриць U і V. Оскільки ми маємо k в значенні гіперпараметра алгоритму цей алгоритм називається truncated SVD. Перевагами LSA є: простота реалізації, розуміння та використання; оскільки LSA передбачає лише декомпозицію матриці документів, він працює швидше порівняно з іншими моделями. Недоліками LSA є: відсутність вбудованої інтерпретації, необхідність дуже великого набору документів та словника для отримання точних результатів.

PLSA, на відміну від LSA, використовує замість SVD ймовірнісний метод. Його головна ідея полягає у пошуку імовірнісної моделі із прихованими темами, яка може генерувати дані, які ми спостерігаємо в матриці. PLSA базується на статистичній моделі яка називається Аспектна модель (aspect model) для співпояви даних, яка асоціює приховану класову змінну $z \in Z(z_1,\dots,z_K)$ з кожним спостереженням.

Спільний розподіл наборів документів та слів визначається наступною формулою:

$$P(d,w) = P(d)P(w|d), P(w|d) = \sum_{z\in Z} P(w|z)P(z|d), \quad (2)$$

де $P(d)$ – ймовірність документу, $P(w|z)$ – ймовірність слова при заданій темі, $P(z|d)$ – ймовірність теми при заданому документі

Стандартна процедура оцінки прихованих змінних є expectation maximization (EM) алгоритм. Він пов'язує два кроки: перший - expectation (E) – (3), на ньому обраховуються апостеорні ймовірності для прихованих змінних, та maximization (M) – (4), на якому оновлюються параметри:

$$P(z|d,w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z'\in Z} P(z')P(d|z')P(w|z')} \quad (3)$$

$$P(w|z) \approx \sum_{d\in D}\sum_{w\in W} n(d,w)P(z|d,w) \quad (4)$$

На відміну від LSA, PLSA визначає генеративну модель даних, що дає можливість застосовувати стандартні методи зі статистики для тренування моделі, вибору моделі, контролювання складності моделі. Проте PLSA все ще має деякі недоліки, такі як: кількість параметрів для pLSA лінійно зростає із збільшенням кількості наявних документів, тому вона схильна до перенавчання; оскільки ми не маємо параметрів моделі P(D) незрозуміло, як призначити ймовірність документу поза навчальним набором.

LDA базується на тій самій ймовірнісній моделі що і PLSA, проте в LDA усунуті недоліки які має PLSA завдяки додаткових припущень:

Вектори документів $\theta_d = (p(t|d): t \in T)$ беруться з сімейства розподілів Діріхле $Dir(\theta,\alpha), \alpha \in R^{|T|}$.

Вектори тем $\phi_t = (p(w|t): w \in W)$ беруться з сімейства розподілів Діріхле $Dir(\theta,\beta), \beta \in R^{|W|}$.

Алгоритм LDA полягає в наступному:

1. Генерується розподіл θ_d~ Dir(α), де $d \in D$.
2. Генерується розподіл ϕ_t~ $Dir(\beta)$, де $t \in T$.
3. Для кожного слова документу: обрати тему з розподілу θ_{td}, обрати слово з розподілу ϕ_{wt}.

Один з методів навчання LDA моделі базується на варіаційному висновку. Він є одним з варіантів EM алгоритму, в якому на кожному кроці оцінки(M крок) ϕ_{wt} і θ_{td} згладжуються в чисельнику та знаменнику:

$$\phi_{wt} = \frac{n_{wt} + \alpha_t}{n_t + \alpha_0},$$

$$\theta_{td} = \frac{n_{wt} + \beta_w}{n_d + \beta_0}.$$

Параметри α і β впливають на розподіл Діріхле. Альфа представляє щільність документа-теми – при вищій альфі документи складаються з більшої кількості тем, а при нижчій альфа документи містять менше тем. Бета представляє щільність тематичних слів - при високій бета теми складаються з більшості слів у корпусі, а при низькій бета вони складаються з декількох слів. Параметр α контролює форму розподілу. Якщо α_i <1 для всіх i, ми отримуємо «шипи» на кутах симплексу. Для значень α_i> 1 розподіл прямує до центру симплексу. Зі збільшенням α_0 розподіл стає більш щільно зосередженим навколо центру симплексу. При всіх α_i = 1 розподіл утворюється рівномірний розподіл.

Одним із критеріїв якості моделі є перплексія. Це міра невідповідності моделі p(w|d) значенням w, які спостерігаються в документах *d* колекції D. Вона визначається за наступною формулою:

$$PP = exp\left(\frac{-\sum_d logp(w_d)}{\sum_d N_d}\right), PP' = log(PP) \quad (7)$$

Чим менша перплексія, тим ліпше модель передбачає появу термінів w у документі d. Для логарифму від перплексії, більше значення є ліпшим. Недоліком перплексії є неочевидність її чисельних значень, а також її залежність не тільки від якості моделі, але і від таких факторів як довжина документів, потужність та розрідженість словника.

В якості метрики оцінювання інтерпретованості моделі була також обрана метрика когерентності. Було виявлено, що експертні оцінки добре корелюють з такою мірою якості. Когерентність теми - міра, яка показує, наскільки слова, що зустрічаються поруч в текстах, виявляються в топах одних і тих самих тем. Когерентність теми t - середня точкової взаємна інформація топ-слів теми.

$$PMI_t = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i}^{k} PMI(w_i, w_j), \quad (8)$$

де w_i- i-й термін в порядку спадання ϕ_{wk}. Поточкова взаємна інформація:

$$PMI(u,v) = ln \frac{|D|N_{uv}}{N_u N_v}, \quad (9)$$

де N_{uv} - число документів, в яких терміни u і v хоча б один раз зустрічаються поруч (у вікні в k слів), N_u - число документів, в яких термін u зустрінувся хоча б один раз. Чим вище величина поточкової взаємної інформації, тим вища невипадковість того, що два слова стоять поруч.

Результати експериментів

Для експериментів аналізу було обрано дані, в кількості 10000 записів, з коментарів під сторінками товарів у магазині цифрових товарів Steam. Перед застосування методів до аналізу даних, дані було оброблено наступним чином:

- забрано пунктуацію та цифри, замінено великі перші літери на малі за допомогою регулярних виразів,
- поділено речення на окремі слова,

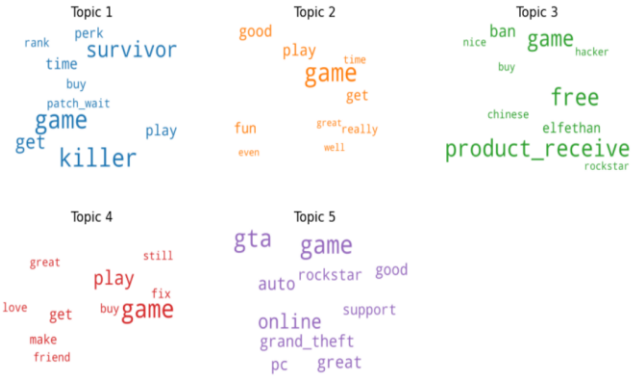

Рис. 2. Розподіл слів по кожній з 5 тем визначених моделлю PLSA

допомогою хмари слів (рис. 1).

Далі, до оброблених попередньо за допомогою нормалізації, токенізації та леманізації даних обраних для експериментів застосуємо модель pLSA. Для цієї моделі було взято для порівняння різні значення параметру кількості моделей k, а саме значення 4, 5, 6, 7, 8. Найліпше значення кількості тем є 5, оскільки для цієї кількості тем модель має найбільше значення когерентності, а саме 0.4212,

Рис 1. Розподіл слів по кожній з 4 тем визначених моделлю LSA

- забрано слова, які використовуються дуже часто в мові, такі як “the”,
- за допомогою сторонньої бібліотеки було застосовано леманізацію.

Спочатку для розділення оброблених попередньо за допомогою нормалізації, токенізації та леманізації даних на теми було застосовано тематичну модель LSA. Для цієї моделі було взято для порівняння різні значення параметру кількості моделей k, а саме значення 4, 5, 6, 7, 8. Найкраще значення кількості тем є 4, оскільки для цієї кількості тем модель має найбільше значення когерентності, а саме 0.480377, найменше - 0.375668 для кількості тем 7. Побудуємо модель з гіперпараметром кількості тем 4 та візуалізуємо розподіл слів по кожній з тем за

Рис 3. Розподіл слів по кожній з тем моделі LDA з гіперпараметром кількості тем 8, альфа = 0.3, бета = 0.9

найменше - 0.367 для кількості тем 8. Побудуємо модель з гіперпараметром кількості тем 5 та візуалізуємо розподіл слів по кожній з тем за допомогою хмари слів (рис. 2).

Далі, до даних застосуємо модель LDA. Для цієї моделі було взято для порівняння різні значення параметру кількості моделей *k*. Для цього параметру були взяті наступні значення параметра: 4, 5, 6, 7, 8. Також було взято різні значення гіперпараметрів альфа та бета такі як 0.01, 0.1, 0.3, 0.9. Найкращі 2 комбінації гіперпараметрів є наступні: кількість тем – 8, альфа – 0.01, бета – 0.9 та кількість тем – 8, альфа – 0.3, бета – 0.9 зі значеннями перплексії 0.552 та 0.537 відповідно. Побудуємо модель з гіперпараметром кількості тем 8, альфа = 0.01, бета = 0.9, та візуалізуємо розподіл слів по кожній з тем (рис. 3).

Зобразимо найкращі числові результати для кожної з моделей LSA, PLSA, LDA разом з оцінкою логарифму перплексії для даних гіперпараметрів. (оцінка перплексії не проводилась для LSA, оскільки у результаті роботи алгоритму можуть з'являтись негативні значення ймовірностей).

Таблиця 1

Найкращі числові результати для кожної з моделей, які були досліджені

	К-сть тем	alpha	betha	Перплексія (логарифм)	Когерентність
LSA	4				0.480377
PLSA	5			-7.414	0.421221
LDA	8	0.01	0.9	-7.281	0.552201

З метрики логарифму перплексії можна побачити, що для обох моделей PLSA та LDA перплексія є досить малою ≈0.006, що свідчить про те, що модель p(w|d) відповідає значенням w, які спостерігаються в документах *d* колекції D. Також порівняємо та зобразимо у вигляді таблиці час роботи кожного з алгоритмів для кількості тем 8.

Таблиця 2

Час роботи кожного з алгоритмів для кількості тем 8

Модель	Час (сек.)
LSA	6.73
PLSA	12.9
LDA	13.6

При порівнянні за часом результати моделі LSA для кількості тем 8 було отримано час виконання 6.73 секунди, що є майже в 2 рази швидше за тренування моделей PLSA та LDA, які показали результати, які не набагато відрізняються один від одного, відповідно 12.9 та 13.6 секунд відповідно. Як правило модель LSA є швидшою, оскільки використовує лише декомпозицію матриці документів, на відміну від PLSA та LDA, які використовують EM алгоритм, який може довго сходитись.

Висновки з даного дослідження і перспективи подальших розвідок у даному напрямі

Під час виконання даного дослідження було проведено аналіз відгуків до товарів у інтернет-магазині цифрових товарів шляхом розбиття їх на теми за допомогою методів тематичного моделювання таких як метод латентного семантичного аналізу LSA імовірнісного латентного семантичного аналізу PLSA та латентне розміщення Діріхле LDA. Було протестовано різні гіперпараметри для цих методів, та було порівняно найкращі з протестованих моделей результати розбиття на теми даних. Оцінка моделі проводилась за допомогою метрики коренетності та оцінка "на око" виведенням "хмари слів" для кожної з тем. Найліпші значення когерентності показала модель LDA з кількістю тем 8, проте при візуалізації деякі створені теми важко зрозуміти, на противагу цьому при візуалізації найбільш інтуїтивно зрозумілі теми було створено за допомогою LSA моделі з кількістю тем 4. Результати отримані в ході виконання даної роботи можна застосувати для оцінки коментарів, наприклад визначення найбільш популярних тем обговорення у коментарів для визначення які недоліки або переваги має той чи інший продукт. Також результати дослідження можуть бути застосовані для визначення до яких продуктів відноситься отриманий набір коментарів. Проте відібраного числа даних, а саме 10000 записів у даних імовірно не є достатньою кількістю для отримання вагомих результатів. Також було проведено порівняння моделей LSA, PLSA та LDA за часом виконання. Модель LSA показала результати вдвічі ліпші за швидкістю виконання ніж PLSA і LDA. Отже можливо зробити висновок, що для оцінки даних, над якими проводився аналіз у даному дослідженні серед досліджених моделей найліпше підходить модель латентного семантичного аналізу оскільки показала найліпші результати по часу виконання, а також хороші результати відносно інших моделей.

References

1. Blei D. M., Ng A. Y., Jordan M. I. Latent dirichlet allocation. Journal of Machine Learning Research. 2003. Vol. 3, No. Jan. P. 993–1022.

2. Hofmann T. Probabilistic latent semantic analysis. arXiv:1301.6705 [cs, stat]. 2013.

3. Bindra A. SocialLDA:scalable topic modeling in social networks. P. 58.

4. Landauer T. K., Foltz P. W., Laham D. An introduction to latent semantic analysis. Discourse Processes. 1998. Vol. 25, No. 2–3. P. 259–284.