



ПӘНДЕР АРАЛЫҚ ҒЫЛЫМДАР  
МЕЖДИСЦИПЛИНАРНЫЕ НАУКИ  
INTERDISCIPLINARY SCIENCES

CENTRAL ASIAN SCIENTIFIC JOURNAL  
ISSN 3135-3061  
2026. №4 (32), Том 2. С. 79–86  
DOI: <https://doi.org/10.5281/zenodo.23771572>  
Сайт журнала: <https://cajournal.kz>

УДК 004.912:821.512.122

МОРФОЛОГИЧЕСКИ ОРИЕНТИРОВАННАЯ МЕТОДИКА  
ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ КАЗАХОЯЗЫЧНОЙ ПОЭЗИИ:  
ПРОЕКТИРОВАНИЕ ИССЛЕДОВАНИЯ НА МАТЕРИАЛЕ  
ПРОИЗВЕДЕНИЙ М. МАКАТАЕВА

Қ.Д. Сәрсембай

Международный университет Астана, Астана, Казахстан

**Сәрсембай Қуаныш Даниярұлы** — магистрант ОП «7М06102 IT менеджмент», Высшая школа информационных технологий и инженерии, Международный университет Астана, Астана, Казахстан.

**Научный руководитель:** Байгожанова Даметкен Сагидуллаевна — к.п.н., ассоциированный профессор, Высшая школа информационных технологий и инженерии, Международный университет Астана, Астана, Казахстан.

© Қ.Д. Сәрсембай, 2026

**Аннотация:** В статье предложена методика тематического моделирования поэтических текстов на казахском языке, учитывающая агглютинативную морфологию языка и специфику стихотворной формы. На материале произведений Мукагали Макатаева описаны принципы формирования и верификации корпуса, схема разбиения текстов на документы, пять сценариев лингвистической предобработки (от базовой токенизации до морфологической лемматизации с выделением устойчивых словосочетаний), а также способы векторизации для моделей LDA, NMF и BERTopic. Обосновано, почему перплексия не может служить основным критерием сравнения моделей, построенных на разных словарях, и предложен интегральный критерий выбора числа тем, объединяющий когерентность  $C_v$ , NPMI, разнообразие тем и устойчивость решения. Представлена модульная архитектура программного решения для анализа и визуализации результатов. Методика может применяться к корпусам других казахских авторов.

**Ключевые слова:** тематическое моделирование, казахский язык, М. Макатаев, лемматизация, LDA, BERTopic, когерентность тем, перплексия, цифровые гуманитарные исследования

**Введение**

Творчество Мукагали Макатаева (1931–1976) – одно из центральных явлений казахской поэзии XX века. Его наследие хорошо изучено литературоведами: исследуются поэтика, образная система, лирический герой, отдельные тематические пласты, например произведения для детей [1]. Однако такие работы опираются на ручной анализ выборки текстов, и выводы о том, какие темы преобладают во всём наследии поэта и как они соотносятся между собой, остаются экспертными оценками, которые трудно проверить количественно.

Методы тематического моделирования позволяют автоматически выявлять скрытые темы в больших коллекциях текстов. Базовой моделью остаётся латентное размещение Дирихле (LDA) [2]. В цифровых гуманитарных исследованиях она применялась к корпусам художественной прозы [3] и поэзии [4]. Для поэзии задача сложнее, чем для новостей или



научных статей: тексты короткие, лексика метафорична, частотны повторы и рефрены, а одна и та же тема выражается разными образами.

Для казахского языка добавляется ещё одна трудность – агглютинативная морфология. От одного корня образуются десятки словоформ, поэтому без морфологической нормализации словарь «раздувается», матрица «документ–термин» становится сверхразреженной, и модель распределяет одно понятие по разным темам. Работы по тематическому моделированию казахских текстов немногочисленны и посвящены в основном публицистике и учебным текстам [5; 6]; казахская поэзия как объект тематического моделирования практически не рассматривалась.

Цель статьи – спроектировать воспроизводимую методику тематического моделирования казахоязычной поэзии на материале произведений М. Макатаева, включающую формирование корпуса, морфологически ориентированную предобработку, обоснованный выбор метрик качества и архитектуру программного решения. Новизна работы состоит в следующем: (1) предложена схема разбиения поэтических текстов на документы, учитывающая жанр; (2) построена система сравнимых сценариев предобработки, позволяющая измерить вклад морфологической нормализации; (3) введён интегральный критерий выбора модели, в котором перплексия используется только как ограничение сходимости, а не как критерий сравнения; (4) предложено вычислять когерентность по внешнему референсному корпусу казахского языка, лемматизированному тем же конвейером.

### **Сбор данных и формирование корпуса произведений М. Макатаева**

**Источники.** Основными источниками служат печатные собрания сочинений и сборники произведений поэта; электронные публикации используются как вспомогательные – для сверки и восполнения пропусков. Для каждого текста фиксируется источник (издание, том, страница), что делает корпус проверяемым. Тексты, доступные только в виде сканов, распознаются системой OCR Tesseract с моделью казахского языка, после чего проходят ручную вычитку.

**Верификация.** Поскольку ошибки распознавания напрямую искажают словарь модели, вводится процедура двойной сверки: каждый текст сопоставляется как минимум с двумя независимыми источниками, а на контрольной выборке (не менее 5% строк) вычисляется доля ошибочных символов (CER). Текст включается в корпус, если CER после вычитки не превышает 0,5%. Расхождения между редакциями фиксируются в метаданных, а в корпус включается текст по основному источнику.

**Нормализация кодировки.** Для казахских текстов характерны специфические ошибки: специфические буквы алфавита (ә, ғ, қ, ң, ө, ұ, ү, һ, і) при распознавании или наборе заменяются похожими символами других алфавитов – латинской «i» вместо казахской «і», латинским «ə» (U+0259) вместо кириллической «ә» (U+04D9), латинскими «a», «e», «o», «c» внутри кириллических слов. Внешне слово выглядит правильно, но для модели это другой токен. Поэтому на этапе загрузки текст приводится к форме Unicode NFC, а смешанные по алфавиту слова автоматически выявляются и исправляются по таблице соответствий.

**Структура корпуса и единица документа.** Корпус делится на подкорпусы по жанру (таблица 1). Ключевое проектное решение – выбор единицы документа. Если считать документом целую поэму, она «растворит» в себе несколько тем; если считать документом отдельную строфу лирического стихотворения, текст окажется слишком коротким для вероятностной модели. Поэтому применяется жанрово-зависимое разбиение: лирическое стихотворение – один документ; крупные поэтические произведения делятся на псевдодокументы по главам или фрагментам 40–60 строк с сохранением границ строф. Переводы из мировой поэзии выделяются в отдельный подкорпус и исключаются из основной модели, чтобы темы переводимых авторов не смешивались с собственными темами поэта.

**Таблица 1 – Структура корпуса и правила разбиения на документы**

| Подкорпус            | Состав                                                 | Единица документа                                | Роль в анализе                           |
|----------------------|--------------------------------------------------------|--------------------------------------------------|------------------------------------------|
| Лирика               | Стихотворения                                          | Стихотворение целиком                            | Основной                                 |
| Поэмы                | Крупные поэтические произведения                       | Глава или фрагмент 40–60 строк по границам строф | Основной                                 |
| Проза и публицистика | Прозаические тексты, статьи (при наличии в источниках) | Блок 150–300 слов по границам абзацев            | Дополнительный, для сопоставления        |
| Переводы             | Переводы из мировой поэзии                             | Стихотворение или песнь                          | Контрольный, в основную модель не входит |

**Метаданные.** Для каждого документа хранятся: идентификатор, название (или первая строка), жанр, подкорпус, год написания или первой публикации (если известен), принадлежность к циклу, источник, статус верификации и номер версии очистки. Датировка позволяет на следующем этапе исследовать динамику тем по периодам творчества; тексты без надёжной даты в динамический анализ не включаются.

#### **Препроцессинг данных: стемминг, лемматизация и векторизация текста**

**Базовые операции.** Конвейер начинается с приведения к нижнему регистру, удаления пунктуации и чисел, токенизации с сохранением составных слов через дефис (например, «ата-ана») и удаления стоп-слов. Список стоп-слов строится из служебных слов казахского языка (және, мен, да/де, та/те, ғой, ғана, бұл, сол, екен, еді и др.) и дополняется по частотному анализу корпуса. Для поэзии важно не удалять автоматически самые частотные знаменательные слова: в лирике частота слов «жүрек» (сердце), «өлең» (стихотворение) или «тау» (гора) отражает тематику, а не «шум».

**Почему нужна морфологическая нормализация.** В казахском языке аффиксы присоединяются к корню последовательно и подчиняются закону сингармонизма: аффикс множественного числа имеет варианты -лар/-лер/-дар/-дер/-тар/-тер (таулар, елдер, тастар). Так, от корня «жыр» (песнь, стих) образуются формы жырлар, жырларым, жырларымда, жырларымдағы, жырымнан и т. д. Без нормализации каждая из них становится отдельным признаком, и понятие «стих» дробится. Для сравнения вводятся пять сценариев предобработки:

- **P0** (базовый): токенизация, нижний регистр, стоп-слова;
- **P1** (стемминг): отсечение аффиксов по принципу наибольшего совпадения, по аналогии с алгоритмом Портера [7], с учётом вариантов по сингармонизму и минимальной длиной основы;
- **P2** (лемматизация): морфологический анализ на основе конечных автоматов (Apertium-kaz [8]; см. также правилый анализатор [9]) со снятием неоднозначности с помощью нейросетевой модели Stanza, обученной на казахском трибанке Universal Dependencies [10];
- **P3**: P2 с фильтрацией по частям речи (сохраняются существительные, прилагательные и глаголы);
- **P4**: P3 с выделением устойчивых биграмм по NPMI (например, «туған жер» – родная земля).

Различия между сценариями наглядно видны на отдельных словоформах (таблица 2). Простой стемминг сокращает словарь, но в поэтическом тексте даёт опасные ошибки «пересечения»: слово «ақын» (поэт) после отсечения «-ын» превращается в «ақ» (белый), а «қалам» (перо) – в «қала» (город). Для корпуса, где «ақын» – одно из ключевых слов, такая ошибка искажает тему. Лемматизация с морфологическим анализом этого избегает, но требует снятия неоднозначности: «қалам» действительно может означать и «перо», и «мой город».

**Таблица 2 – Результат сценариев предобработки на примерах словоформ**

| Словоформа (перевод) | P0: токен | P1: стемминг | P2: лемматизация |
|----------------------|-----------|--------------|------------------|
|----------------------|-----------|--------------|------------------|



Для каждого сценария фиксируются объём словаря, доля гапаксов (слов, встречающихся один раз) и разреженность матрицы «документ–термин». Эти показатели вместе с метриками качества (раздел 3) позволяют количественно оценить, насколько морфологическая нормализация улучшает темы именно на поэтическом материале.

**Векторизация.** Способ представления текста зависит от модели. Для LDA используется модель «мешка слов» с отсечением редких (встречающихся менее чем в 3 документах) и слишком частотных (более чем в 50% документов) терминов. Для неотрицательной матричной факторизации (NMF) [11] применяются веса TF-IDF. Для BERTopic [12] документы представляются векторами многоязычной модели LaBSE [13], поддерживающей казахский язык. Важная деталь: эмбединги BERTopic вычисляются по исходному, нелемматизированному тексту, поскольку модель учитывает контекст и морфологию сама, а итоговые ключевые слова тем (с-TF-IDF) строятся по лемматизированной версии, чтобы они были сопоставимы с результатами LDA и NMF. В качестве дополнительного варианта рассматривается аддитивная регуляризация тематических моделей (ARTM) [14], позволяющая явно задавать требования к разреженности и различимости тем.

**План эксперимента.** Полный план включает 5 сценариев предобработки  $\times$  3 модели  $\times$  значения числа тем  $K$  от 5 до 40 с шагом 5  $\times$  5 запусков с разными начальными значениями генератора случайных чисел. Повторные запуски нужны для оценки устойчивости тем.

#### Обоснование выбора метрик качества моделей

**Перплексия** – традиционная внутренняя метрика вероятностных моделей [15]. Она показывает, насколько хорошо модель предсказывает слова отложенной выборки:

$$\text{Perplexity}(D_{\text{test}}) = \exp(-\sum_d \log p(w_d) / \sum_d N_d), \quad (1)$$

где  $w_d$  – слова документа  $d$ ,  $N_d$  – их число. Чем ниже перплексия, тем лучше модель описывает данные. Однако у этой метрики три ограничения, принципиальные для нашей задачи. Во-первых, эксперимент Chang и соавторов показал, что модели с лучшей перплексией нередко дают менее интерпретируемые для человека темы [16]. Во-вторых, перплексия зависит от словаря: сценарии P0–P4 порождают словари разного размера, поэтому сравнивать по перплексии модели на разных сценариях некорректно – меньший словарь «автоматически» даёт меньшую перплексию. В-третьих, для NMF и BERTopic, не являющихся вероятностными генеративными моделями в смысле LDA, она не определена. Поэтому в предлагаемой методике перплексия используется только внутри одной модели при фиксированном словаре: для контроля сходимости и подбора гиперпараметров  $\alpha$  и  $\eta$ .

**Когерентность** оценивает смысловую согласованность наиболее вероятных слов темы и лучше согласуется с оценками людей [17; 18]. Основой служит нормированная точечная взаимная информация:

$$\text{NPMI}(w_i, w_j) = \ln[P(w_i, w_j) / (P(w_i) \cdot P(w_j))] / (-\ln P(w_i, w_j)), \quad (2)$$

где вероятности оцениваются по совместной встречаемости слов в скользящем окне ( $k$   $P(w_i, w_j)$  добавляется малая константа  $\epsilon$ , чтобы избежать логарифма нуля). Метрика  $C_v$  [19] объединяет NPMI, косвенное косинусное сходство и скользящее окно и в сравнительных исследованиях показала наилучшее согласие с экспертными оценками. Для поэтического корпуса принципиален выбор корпуса, по которому оцениваются вероятности. Если считать когерентность по тому же небольшому корпусу, на котором обучена модель, метрика «поощряет» случайные совпадения. Поэтому предлагается рассчитывать её по внешнему референсному корпусу казахского языка [20], прошедшему тот же конвейер предобработки, а для сопоставления – и по внутреннему корпусу.

Автоматическая когерентность не гарантирует интерпретируемости: Ноуле и соавторы показали, что для нейросетевых моделей автоматические метрики и оценки людей расходятся [21]. Поэтому в методику включены ещё две автоматические метрики и экспертная проверка. **Разнообразие тем** (topic diversity) [22] – доля уникальных слов среди топ- $N$  слов всех тем:

$$\text{TD} = |\cup_k \text{Top}_N(k)| / (K \cdot N), \quad (3)$$



низкое значение TD означает, что темы дублируют друг друга. **Устойчивость S** – средняя мера Жаккара между топ-N словами сопоставленных тем при разных запусках; она показывает, воспроизводится ли тема или является артефактом случайной инициализации.

Выбор числа тем K, сценария предобработки и модели выполняется по **интегральному критерию**:

$$Q(K) = \alpha \cdot C_v(K) + \beta \cdot \text{NPMI}(K) + \gamma \cdot \text{TD}(K) + \lambda \cdot S(K), \quad \alpha + \beta + \gamma + \lambda = 1, \quad (4)$$

где все слагаемые нормированы на отрезок [0; 1] по методу min-max. Начальные веса:  $\alpha = 0,4$ ,  $\beta = 0,2$ ,  $\gamma = 0,2$ ,  $\lambda = 0,2$ ; они калибруются по экспертной оценке. Для LDA действует дополнительное ограничение: модель участвует в сравнении только при стабилизации перплексии (относительное изменение между итерациями менее 1%). Итоговая модель проверяется экспертами-филологами с помощью теста «лишнего слова» (word intrusion) [16]: в список из пяти топ-слов темы добавляется слово из другой темы, и если эксперты устойчиво его находят, тема считается интерпретируемой. Роли метрик сведены в таблицу 3.

**Таблица 3 – Метрики качества и их роль в методике**

| Метрика        | Что измеряет                                      | Ограничения                                                                                           | Роль в методике                                                          |
|----------------|---------------------------------------------------|-------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------|
| Perplexity     | Предсказательную способность вероятностной модели | Несравнима при разных словарях; не определена для NMF и BERTopic; слабо связана с интерпретируемостью | Контроль сходимости и подбор $\alpha$ , $\eta$ при фиксированном словаре |
| $C_v$          | Смысловую согласованность топ-слов темы           | Зависит от референсного корпуса                                                                       | Основной критерий ( $\alpha$ )                                           |
| NPMI           | Попарную совместную встречаемость слов            | Смещение на редких словах                                                                             | Дополнительный критерий ( $\beta$ )                                      |
| TD             | Разнообразие тем                                  | Не оценивает смысл                                                                                    | Штраф за дублирование тем ( $\gamma$ )                                   |
| S              | Воспроизводимость тем при разных запусках         | Требуется повторных запусков                                                                          | Критерий надёжности ( $\lambda$ )                                        |
| Word intrusion | Интерпретируемость для человека                   | Трудоёмкость                                                                                          | Финальная проверка и калибровка весов                                    |

#### **Архитектура программного решения для анализа и визуализации**

Программное решение построено по модульному принципу (рисунок 1): каждый этап конвейера – отдельный модуль с фиксированным входом и выходом, что позволяет заменять компоненты (например, анализатор морфологии или модель эмбедингов) без изменения остальной системы. Все параметры экспериментов задаются в конфигурационных файлах YAML, а результаты каждого запуска (параметры, метрики, артефакты) регистрируются в журнале экспериментов MLflow. Это обеспечивает воспроизводимость: любую тему из итогового отчёта можно проследить до конкретной версии корпуса, сценария предобработки и начального значения генератора.

Назначение модулей и используемые инструменты приведены в таблице 4. Реализация выполняется на языке Python. Для LDA используется библиотека gensim [23], для NMF – scikit-learn, для BERTopic – одноимённая библиотека с моделями sentence-transformers.

**Таблица 4 – Модули программного решения**

| Модуль      | Функции                                                         | Инструменты             |
|-------------|-----------------------------------------------------------------|-------------------------|
| Сбор данных | Загрузка текстов, OCR, сверка источников, заполнение метаданных | Tesseract (kaz), Python |
| Хранилище   | Хранение текстов, метаданных, версий очистки                    | SQLite / PostgreSQL     |



| Модуль               | Функции                                                                                                  | Инструменты                               |
|----------------------|----------------------------------------------------------------------------------------------------------|-------------------------------------------|
| Предобработка        | Нормализация Unicode, токенизация, стоп-слова, сценарии P0–P4                                            | Apertium-kaz, Stanza, собственные правила |
| Векторизация         | BoW, TF-IDF, эмбединги                                                                                   | gensim, scikit-learn, LaBSE               |
| Моделирование        | Обучение LDA, NMF, BERTopic; перебор K и начальных значений                                              | gensim, scikit-learn, BERTopic            |
| Оценка качества      | Perplexity, C_v, NPMI, TD, S, расчёт Q(K)                                                                | gensim CoherenceModel, собственный модуль |
| Визуализация         | Интерактивная карта тем, тепловые карты «документ–тема», динамика тем по периодам, поиск текстов по теме | pyLDavis, Plotly, Streamlit               |
| Журнал экспериментов | Регистрация параметров, метрик и артефактов                                                              | MLflow, YAML                              |



Рисунок 1 – Архитектура программного решения для тематического анализа корпуса

Модуль визуализации ориентирован на исследователя-филолога, а не на программиста. Интерактивная карта тем на основе LDavis [24] показывает взаимное расположение тем и позволяет регулировать параметр релевантности слов. Тепловая карта «документ–тема» позволяет увидеть, какие стихотворения являются «чистыми» представителями темы, а какие сочетают несколько тем. Для датированных текстов строится динамика долей тем по периодам творчества. Через веб-панель эксперт может открыть любой документ, просмотреть его тематический профиль и оставить разметку для теста



«лишнего слова»; эта разметка возвращается в модуль оценки качества и используется для калибровки весов критерия (4).

### **Заключение**

В статье спроектирована методика тематического моделирования казахоязычной поэзии на материале произведений М. Макатаева. Её основные элементы: верифицируемый корпус с жанрово-зависимым разбиением на документы и нормализацией кодировки; пять сравнимых сценариев предобработки, позволяющих количественно оценить вклад морфологической нормализации для агглютинативного языка; система метрик, в которой перплексия ограничена ролью контроля сходимости, а выбор модели выполняется по интегральному критерию на основе когерентности, разнообразия и устойчивости тем с экспертной проверкой; модульная архитектура программного решения, обеспечивающая воспроизводимость экспериментов и удобство интерпретации результатов.

Следующий этап исследования – формирование корпуса, проведение экспериментов по описанному плану и содержательная интерпретация выявленных тем совместно со специалистами по казахской литературе. Методика не привязана к творчеству одного автора и может применяться для сравнительного тематического анализа произведений других казахских поэтов.

### **Список литературы:**

1. Абишева С.Д., Саметова Ж.Ш. Мұқағали Мақатаевтың балаларға арналған шығармаларының поэтикалық ерекшеліктері // Ясауи университетінің хабаршысы. – 2023. – № 1 (127). – С. 22–34. – DOI: 10.47526/2023-1/2664-0686.02.
2. Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet Allocation // Journal of Machine Learning Research. – 2003. – Vol. 3. – P. 993–1022.
3. Jockers M.L., Mimno D. Significant themes in 19th-century literature // Poetics. – 2013. – Vol. 41, No. 6. – P. 750–769.
4. Navarro-Colorado B. On poetic topic modeling: extracting themes and motifs from a corpus of Spanish poetry // Frontiers in Digital Humanities. – 2018. – Vol. 5. – Art. 15.
5. Akanova A. [et al.] Development of a thematic and neural network model for data learning // Eastern-European Journal of Enterprise Technologies. – 2022. – DOI: 10.15587/1729-4061.2022.263421.
6. Akanova A. [et al.] Neurocomputer system of semantic analysis of the text in the Kazakh language // ACM Transactions on Asian and Low-Resource Language Information Processing. – 2024. – DOI: 10.1145/3652159.
7. Porter M.F. An algorithm for suffix stripping // Program. – 1980. – Vol. 14, No. 3. – P. 130–137.
8. Washington J., Salimzyanov I., Tyers F. Finite-state morphological transducers for three Kypchak languages // Proceedings of LREC 2014. – Reykjavik, 2014. – P. 3378–3385.
9. Kessikbayeva G., Cicekli I. Rule based morphological analyzer of Kazakh language // Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM. – Baltimore, 2014. – P. 46–54.
10. Qi P., Zhang Y., Zhang Y., Bolton J., Manning C.D. Stanza: a Python natural language processing toolkit for many human languages // Proceedings of ACL 2020: System Demonstrations. – 2020. – P. 101–108.
11. Lee D.D., Seung H.S. Learning the parts of objects by non-negative matrix factorization // Nature. – 1999. – Vol. 401. – P. 788–791.
12. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure // arXiv preprint. – 2022. – arXiv:2203.05794.
13. Feng F., Yang Y., Cer D., Arivazhagan N., Wang W. Language-agnostic BERT sentence embedding // Proceedings of ACL 2022. – 2022. – P. 878–891.
14. Vorontsov K., Potapenko A. Additive regularization of topic models // Machine Learning. – 2015. – Vol. 101. – P. 303–323.
15. Wallach H.M., Murray I., Salakhutdinov R., Mimno D. Evaluation methods for topic models // Proceedings of ICML 2009. – 2009. – P. 1105–1112.
16. Chang J., Gerrish S., Wang C., Boyd-Graber J., Blei D.M. Reading tea leaves: how humans interpret topic models // Advances in Neural Information Processing Systems 22. – 2009. – P. 288–296.
17. Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Proceedings of NAACL HLT 2010. – 2010. – P. 100–108.



18. Mimno D., Wallach H., Talley E., Leenders M., McCallum A. Optimizing semantic coherence in topic models // Proceedings of EMNLP 2011. – 2011. – P. 262–272.
19. Röder M., Both A., Hinneburg A. Exploring the space of topic coherence measures // Proceedings of WSDM 2015. – 2015. – P. 399–408.
20. Makhambetov O., Makazhanov A., Yessenbayev Zh., Matkarimov B., Sabyrgaliyev I., Sharafudinov A. Assembling the Kazakh Language Corpus // Proceedings of EMNLP 2013. – 2013. – P. 1022–1031.
21. Hoyle A., Goel P., Hian-Cheong A., Peskov D., Boyd-Graber J., Resnik P. Is automated topic model evaluation broken? The incoherence of coherence // Advances in Neural Information Processing Systems 34. – 2021.
22. Dieng A.B., Ruiz F.J.R., Blei D.M. Topic modeling in embedding spaces // Transactions of the Association for Computational Linguistics. – 2020. – Vol. 8. – P. 439–453.
23. Řehůřek R., Sojka P. Software framework for topic modelling with large corpora // Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. – Valletta, 2010. – P. 45–50.
24. Sievert C., Shirley K. LDavis: a method for visualizing and interpreting topics // Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. – 2014. – P. 63–70.