ЛЕСОВЕДЕНИЕ, ЛЕСОВОДСТВО, ЛЕСНЫЕ КУЛЬТУРЫ, АГРОЛЕСОМЕЛИОРАЦИЯ, ОЗЕЛЕНЕНИЕ, ЛЕСНАЯ ПИРОЛОГИЯ И ТАКСАЦИЯ/FORESTRY, FORESTRY, FOREST CROPS, AGROFORESTRY, LANDSCAPING, FOREST PYROLOGY AND TAXATION

DOI: https://doi.org/10.60797/IRJ.2025.158.33

ПОСТРОЕНИЕ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ ДЛЯ АНАЛИЗА БОЛЬШИХ ТЕКСТОВЫХ МАССИВОВ В ЛЕСОМЕЛИОРАТИВНЫХ ИССЛЕДОВАНИЯХ

Научная статья

Танкова Ж.В.[1,] *, Танков А.А.[2]

[1] Оренбургский государственный аграрный университет, Оренбург, Российская Федерация

[2] Федеральный научный центр биологических систем и агротехнологий Российской академии наук, Оренбург, Российская Федерация

* Корреспондирующий автор (tankovazhv[at]yandex.ru)

Аннотация

Апробирована методика анализа информационного потенциала корпуса научных публикаций с применением методов тематического моделирования, в частности алгоритма Латентного распределения Дирихле (LDA). Исследование охватило 533 научные статьи и материалы конференций по тематике защитного лесоразведения, опубликованные с 2000 по 2024 год. Анализ проводился с использованием платформы Orange 3.38.0 с надстройкой для интеллектуального анализа текстов.

Результаты показали эффективность LDA для выявления скрытых тематических паттернов в области защитного лесоразведения и агролесомелиорации. Оптимальное количество тем (10) было определено на основе показателей лог-перплексии (12772) и тематической согласованности (0,54). Визуализация результатов осуществлялась с помощью облака слов и многомерного шкалирования (MDS), что обеспечило наглядное представление ключевых тем и их взаимосвязей.

Исследование демонстрирует потенциал тематического моделирования как инструмента для автоматизации анализа научной литературы, выявления трендов и пробелов в исследованиях, а также для поддержки принятия решений в области экологического управления и устойчивого развития лесных экосистем.

Ключевые слова: тематическое моделирование, защитное лесоразведение, интеллектуальный анализ текстов, латентное распределение Дирихле, облака слов, Text Mining.

CONSTRUCTING THEMATIC MODELS FOR ANALYSING LARGE TEXT CORPORA IN FOREST MELIORATION STUDIES

Research article

Tankova Z.V.[1,] *, Tankov A.A.[2]

[1] Orenburg State Agrarian University, Orenburg, Russian Federation

[2] Federal Scientific Center for Biological Systems and Agrotechnologies of the Russian Academy of Sciences, Orenburg, Russian Federation

* Corresponding author (tankovazhv[at]yandex.ru)

Abstract

The methodology for analysing the information potential of the corpus of scientific publications was tested using thematic modelling methods, in particular the Latent Dirichlet Distribution Algorithm (LDA). The study covered 533 research articles and conference proceedings on the subject of protective forestry published from 2000 to 2024. The analysis was conducted using the Orange 3.38.0 platform with an extension for text mining.

The results showed the effectiveness of LDA for identifying latent thematic patterns in the field of protective forestry and agroforestry. The optimal number of themes (10) was determined based on log-perplexity (12772) and thematic consistency (0.54). The results were visualised using word cloud and multidimensional scaling (MDS), which provided a visual representation of key themes and their relationships.

The research demonstrates the potential of topic modelling as a tool to automate the analysis of scientific literature, identify trends and research gaps, and support decision-making in environmental management and sustainable development of forest ecosystems.

Keywords: thematic modelling, protective forestry, latent Dirichlet distribution, word clouds, Text Mining.

Введение

Современные исследования в области защитного лесоразведения и агролесомелиорации демонстрируют растущую междисциплинарность, объединяя экологию, агрономию, гидрологию и геоинформационные технологии [1], [2]. Однако стремительный рост объема научных публикаций, посвященных противоэрозионным мероприятиям, устойчивому землепользованию и климатической адаптации лесных экосистем, порождает методологические сложности. Традиционные подходы к анализу литературы, основанные на ручной классификации данных, теряют эффективность в условиях информационной перегрузки. Существующие обзоры, как правило, фокусируются на узких темах — будь то природные пожары, почвенные условия или лесозащитные технологии, — но не позволяют выявить скрытые тематические паттерны, которые могут трансформировать понимание глобальных трендов.

В этой связи тематическое моделирование, включая алгоритмы LDA (Latent Dirichlet Allocation), предлагает инструментарий для автоматизации анализа неструктурированных текстовых данных. Этот метод способен идентифицировать латентные темы, такие как роль защитных лесонасаждений в борьбе с эрозией, оценка гидрологического эффекта лесополос или применение геномных технологий в лесовосстановлении. Например, моделирование тем с помощью латентного распределения Дирихле (LDA) было признано эффективным методом алгоритмического и автоматического выявления абстрактных тем, присутствующих в большом и неструктурированном наборе статей [3], [4], [5]. Моделирование тем с помощью LDA основано на строгих статистических принципах, которые позволяют генерировать темы с минимальным вмешательством человека и/или ручной обработкой [4], [5]. Такой автоматический метод позволяет создавать более содержательные и реалистичные темы и обеспечивает надёжность и достоверность результатов в отличие от методов, используемых вручную [6]. LDA успешно применяется для тематического моделирования в области информационных наук [7], маркетинга [8] статистики [9], туризма [10], принятия решений [11], компьютерных наук [10] и исследований в области гидроэнергетики [12].

Между тем в российских научных изданиях недостаточно освещены вопросы интеллектуального анализа текста для автоматизации структурирования информации, выявления скрытых тематических структур, анализа трендов и динамики исследований, идентификации пробелов в литературе, поддержке систематических обзоров, визуализации данных. Литературные данные свидетельствуют, что в последние годы инструменты анализа текстовых данных начинают использовать применительно к оценке состояния зелёных насаждений города [13], [14].

Цель настоящего исследования — продемонстрировать эффективность LDA-моделирования для структурирования знаний в области защитных лесонасаждений и агролесомелиорации. Актуальность работы обусловлена необходимостью преодоления методологических разрывов между традиционными и алгоритмическими подходами к анализу литературы, а также потребностью в инструментах поддержки принятия решений для устойчивого управления лесными экосистемами.

Структура статьи включает описание корпуса данных, этапы предобработки текстов, обучение моделей, а также прикладные примеры, демонстрирующие потенциал метода для российских условий. Исследование призвано стать мостом между компьютерными науками и лесомелиоративной практикой, открывая новые горизонты для анализа глобальных экологических вызовов.

Материал и методы исследования

Для анализа были выбраны тексты 533 научных статей и материалов конференций, опубликованных в период с 2000 по 2024 год. Источниками данных стали научные электронные библиотеки (CyberLeninka, eLibrary) и сайт конференций БГИТУ (cyberleninka.ru, elibrary.ru, science-bsea.bgita.ru). Основным критерием отбора публикаций было наличие термина «защитные насаждения» в тексте. Все документы были опубликованы на русском языке и соответствовали тематике защитного лесоразведения.

Анализ текстового корпуса проводился с использованием платформы Orange 3.38.0, которая предоставляет инструменты для интеллектуального анализа данных на основе Python [15], [16]. Для работы с текстами была установлена надстройка Text Mining, обеспечивающая функционал для обработки, моделирования и визуализации текстовых данных. Реализация модели анализа текста представлена на рис. 1.

Предварительная обработка текста:

1. Трансформация: приведение текста к нижнему регистру, удаление диакритических знаков, HTML-тегов, URL-адресов.
2. Токенизация: разделение текста на слова и фразы с использованием регулярных выражений.
3. Нормализация: применение алгоритма Porter Stemmer для сокращения слов до их базовой формы.
4. Фильтрация: удаление стоп-слов (предлогов, местоимений, союзов), чисел и нерелевантных терминов.
5. Визуализация частотности слов.

Для отображения наиболее упоминаемых токенов использовался виджет «Облако слов» (Word Cloud). Размер слова на графике пропорционален его частоте упоминания в тексте, что позволяет выделить ключевые термины корпуса.

2.1. Тематическое моделирование

Алгоритм Латентного распределения Дирихле (LDA) был применён для выявления скрытых тематических паттернов. Оптимальное количество тем (10) определено на основе лог-перплексии (12772) и тематической согласованности (0,54). Виджет LDAvis использовался для визуализации тем с учётом их распространённости и семантической значимости.

2.2. Оценка значимости терминов

Метод TF-IDF применялся для определения веса слов в рамках документов и всего корпуса. Это позволило идентифицировать наиболее значимые термины для каждой темы.

2.3. Интерпретация результатов

Визуализация тем осуществлялась с помощью многомерного шкалирования (MDS), что позволило отобразить темы в виде кругов на двумерной плоскости с учётом их взаимосвязей и распространённости.

Таким образом, предложенная методика обеспечивает комплексный подход к обработке и анализу текстовых данных, позволяя автоматизировать процесс выделения тем и структурирования информации в области защитного лесоразведения.

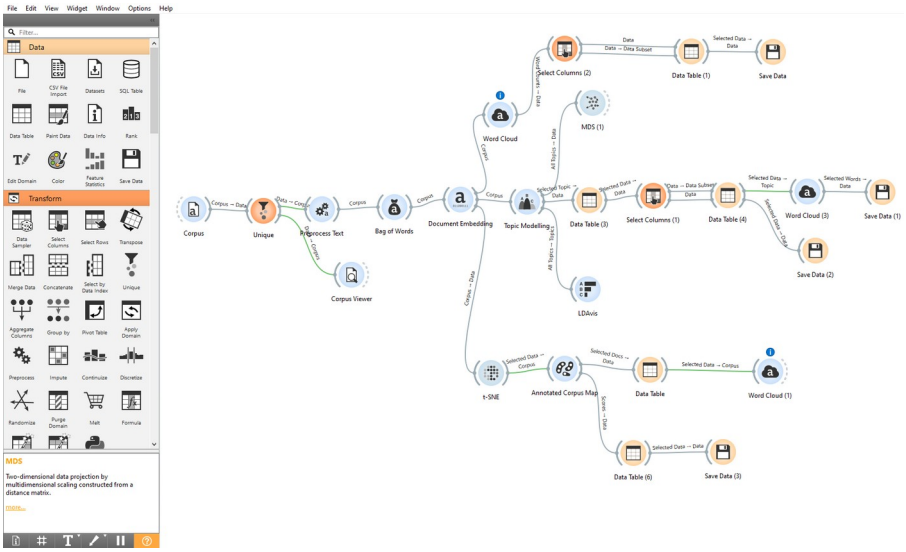

Рисунок 1 - Схема реализации модели анализа текста в Orange Data Mining
DOI: https://doi.org/10.60797/IRJ.2025.158.33.1

2.4. Виджеты для определения количества слов, их частоты и значимости

Для визуального отображения наиболее упоминаемых слов (токенов) в тексте использовался виджет «Облако слов (Word Cloud)». Размер слова на графике пропорционален частоте его упоминания в тексте. Этот виджет выдает список слов, отсортированных по убыванию частоты. Статистически значимые слова для всего корпуса и для каждого отдельного тематического блока были идентифицированы при помощи интеграции виджетов «Мешок слов (Bag of Words)», «Векторное представление документов (Document Embedding)» и инструмента «Просмотр корпуса (Corpus Viewer)».

2.5. Тематическое моделирование (Topic Modeling)

Тематическое моделирование выявляет абстрактные темы в текстовом корпусе на основе кластеров слов и фраз, встречающихся в каждом документе, а также их частоты. Обычно один документ содержит несколько тем в различных пропорциях, поэтому виджет также предоставляет информацию о весе темы (вклада) в каждом документе. В данном анализе для тематического моделирования был выбран метод Латентного размещения Дирихле (Latent Dirichlet Allocation, LDA) [4]. Один из ключевых параметров модели, который необходимо определить заранее, — количество тем [17]. На основе нашего анализа и с помощью виджета Topic modelling были выбраны первые десять тем, наиболее полно описывающие содержание корпуса, обеспечивающих оптимальное сочетание низкой лог-перплексии и высокой тематической согласованности.

Для изучения связи между частотными и специфичными словами в конкретной теме виджет LDA-based visualization (LDA-vis) был подключен к виджету тематического моделирования. Это позволило выявить топ-слова для каждой темы, взвешенные по критерию релевантности. Для LDAvis использовалась оптимальная, по мнению Зиверта и Ширли ($\lambda = 0{,}6$), настройка релевантности [18].

Результаты тематического моделирования визуализировались для интерпретации тем с использованием виджета MDS (многомерное шкалирование). Виджет многомерного шкалирования создает проект визуализации данных, который содержит следующее:

- отображает темы в виде кругов на двумерной плоскости, центры которых определяются путем вычисления расстояния между темами;
- кодирует общую распространенность каждой темы с использованием площадей кругов [18].

Результаты и их обсуждение

3.1. Предварительная обработка текста и генерация облака слов

В рамках исследования, с помощью виджета предварительной обработки текста был инициирован процесс преобразования 533 текстовых документов в 619 559 токенов 45808 типов.

Для оценки значимости терминов применялся метод TF-IDF, который определяет вес слов на основе их частоты в рамках документа (term frequency, TF) и обратной частоты встречаемости в корпусе документов (inverse document frequency, IDF).

С использованием виджета «Облако слов» были визуально отображены 200 наиболее часто встречающихся слов (рис. 2).

Рисунок 2 - Облако слов для всего корпуса документов
DOI: https://doi.org/10.60797/IRJ.2025.158.33.2

Облако слов: отображает токены в корпусе, где их размер обозначает «вес» или частоту слова в корпусе, т. е. токены, которые встречаются чаще, отображаются в большем формате, как показано на рис. 2.

Порядок самых рейтинговых слов по их весу: «лесных», «насаждений», «деревьев», «дуба», «защитных», «насаждения», «территории», «полосы», «почвы», «состояния». Самым распространённым словом в анализируемых документах и первым в списке оказалось «лесных», что очевидно.

Используя описанный выше метод визуализации, можно было только определить, как часто встречается определённое слово и/или фраза, не объясняя контекст, в котором они использовались.

Поэтому для определения конкретных тем в полученных документах использовался виджет тематического моделирования «Topic modelling» и виджет «Annotated Corpus Map» (аннотированная карта корпуса).

3.2. Выбор количества тем

Тематические модели – это модели скрытых переменных документов, которые используют корреляции между словами и скрытыми семантическими темами в коллекции документов [19]. Важным следствием этого определения является то, что ожидаемое количество тем (т. е. скрытых переменных) должно быть установлено до вычисления самой модели. Таким образом, поскольку количество тем должно быть установлено априори, выбор наилучшего количества тем для данной коллекции документов не является тривиальной задачей. В литературе [4], [20] эта проблема решалась разными способами, но всегда находился компромисс между необходимостью большого количества тем для охвата всех тем в коллекции документов и необходимостью ограниченного числа тем, которые могут быть легче поняты и проверены экспертами в области собранных данных.

Для того чтобы помочь в выборе количества интересующих тем, была введена мера — перплексия [21]. Идея заключается в том, чтобы выбор модели в отношении количества тем возможен путём разделения данных на обучающие и тестовые наборы данных. Затем вероятность для тестовых данных аппроксимируется с использованием нижней границы, полученной в рамках оценки вариационным EM-алгоритмом (VEM).

В частности, перплексия является мерой способности модели обобщать скрытые данные. Она определяется как обратное геометрическое среднее вероятности тестового корпуса с учётом модели, как указано ниже:

$$\text{Perplexity}(w) = \exp\left\{-\frac{\log(p(w))}{\sum_{d=1}^{D}\sum_{j=1}^{V} n^{(jd)}}\right\}$$

где обозначает частоту встречаемости *j*-го термина в *d*-м документе.

Более высокие значения перплексии указывают на худшую аппроксимацию слов документов темами, изученными моделью. Перплексия служит мерой качества модели LDA в прогнозировании новых данных из того же распределения, что и обучающая выборка. Таким образом, она оценивает ключевую характеристику алгоритма вывода: при условии одинаковой структуры модели наилучший алгоритм (по качеству обучения) покажет более низкую перплексию [22].

В виджете Тематическое моделирование применяется логарифмическая форма этой величины, используемая для удобства интерпретации — Log perplexity (логарифмическая перплексия или лог-перплексия).

Перплексия традиционно считается одной из основных метрик для оценки статистических моделей (наряду с правдоподобием тестовой выборки или предельной вероятностью данных при заданной модели). Однако она слишком груба, поэтому в последние годы сообщество тематического моделирования переходит к более точным метрикам [23].

Topic coherence (тематическая согласованность) — метрика, оценивающая семантическую связность слов в тематическом кластере. Она используется для проверки качества тематического моделирования (например, LDA) [24], [25], [26], [27]. Высокое значение показателя Topic coherence для темы подтверждает её семантическую значимость в кластеризации научных публикаций.

Первые десять тем, заданных алгоритмом LDA, были признаны адекватными представителями всего корпуса текстов на основании хороших значений лог-перплексии (12772) и Topic coherence (0,54) (рис.3).

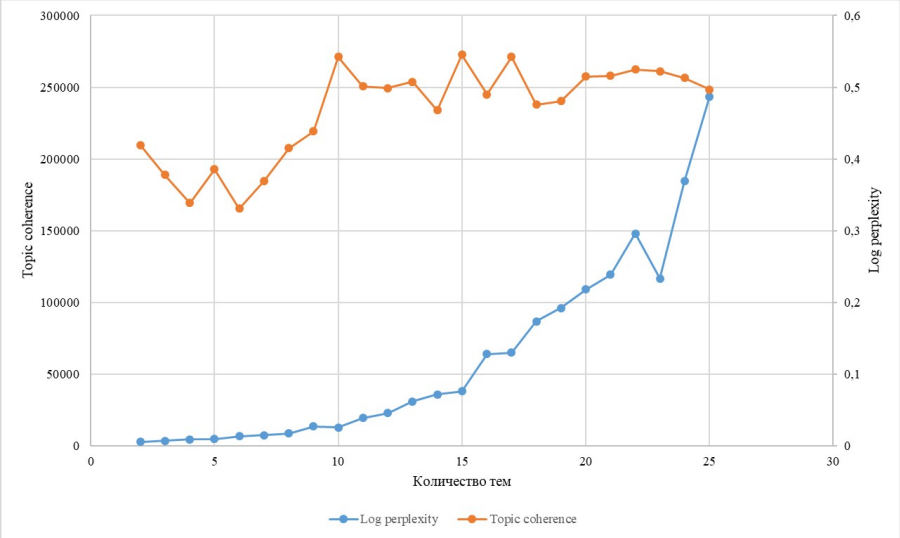

Рисунок 3 - График показателей Log perplexity и Topic coherence для количества тем в промежутке от 2 до 25
DOI: https://doi.org/10.60797/IRJ.2025.158.33.3

Десять наиболее часто встречающихся слов, которые лучше всего описывают каждую тему, приведены в таблице 1.

Таблица 1 - Наиболее распространённые слова в каждой теме

DOI: https://doi.org/10.60797/IRJ.2025.158.33.4

Тема	Наиболее распространённые слова
Тема 1	Жизнеустойчивости, лесных, деревьев, лжеакации, горизонт, дубняк, ГЗЛП, дуба, контрольном, агролесного
Тема 2	Лесных, агроклиматической, гниль, крымского, Крым, декабрь, колок, защитных, жуки
Тема 3	Дятлов, агрегатов, асимметрии, заветренном, лесных, дубняк, звука, деревьев, вспомогательных, асимметрия
Тема 4	Деревьев, лесных, лесополос, возможное, изгородей, влагоперенос, железнодорожного, групповых, групповым, дуба
Тема 5	Западное, лесопастбище, лесных, генетических, интегрального, дуба, возобновлением, интегральный, волоках, геосистемы
Тема 6	Залежи, залежью, горизонтах, выщелоченного, кальция, карбонатов, залежь, берёзой, Вейбулла
Тема 7	Здоровый, захламлённости, белого, балла, ключевом, водоохранного, лесных, газоочистки, карбонатная, геосистемы
Тема 8	Лесных, деревьев, дуба, защитных, лесной, лесные, земель, исследования, лесополос, дуб
Тема 9	Лесных, выщелоченном, асимметрии, агролесоводство, защитных, инфляции, лесонепригодные, лесистости, изолинейная, безлистного
Тема 10	Каштановая, горизонтальной, ланцетным,

Тема	Наиболее распространённые слова
	залежь, дубрав, гниль, глееватой, завода, затенение

Примечание: перечислены по убыванию частоты

3.3. Оценка и визуализация

Для визуализации и интерпретации тем был использован инструмент LDAvis [28], интегрированный в программу Orange Data Mining. Он позволяет получить общее представление о выделенных темах, оценить их различия и детально проанализировать их ключевые слова. В рамках виджета LDAvis в Orange используются два ключевых показателя:

1. Overall term frequency (Общая частота термина) — отражает, как часто термин встречается во всём корпусе документов. Этот параметр помогает выделять слова, которые значимы для различения тем между собой.

2. Term frequency within topic (Частота термина в теме) — показывает, насколько термин специфичен для конкретной темы, ранжируя слова по их релевантности внутри неё.

LDAvis в Orange позволяет динамически изменять ранжирование слов, комбинируя эти метрики. Например, термины с высокой релевантностью (Term frequency within topic) отображаются как ключевые для темы, а их размер на визуализации зависит от значимости (Overall term frequency). Это обеспечивает интуитивное понимание структуры тем и упрощает их интерпретацию.

В случаях, когда темы включают термины из одной области, и в связи с этим, было бы довольно сложно различать на основе наиболее вероятных слов темы или их частоты в корпусе. В общем, на самом деле, темы часто имеют тенденцию отображать общие термины среди первых слов, появляющихся в списке, слова, которые впоследствии повторяются в нескольких темах. Чтобы обойти эту трудность, применяют LDAvis — веб-интерактивную визуализацию тем, разработанную Sievert и Shirley [28].

Для оптимальной интерпретации тем Sievert и Shirley предложена мера Relevance (релевантности) термина теме.

Перечень наиболее распространённых слов, характерных для тем № 1 и 8 при показателях релевантности равным 0,6 и 1,0, приведён в таблице 2.

Таблица 2 - Наиболее распространённые слова в темах № 1 и 8

DOI: https://doi.org/10.60797/IRJ.2025.158.33.5

Тема 1		Тема 8	
Relevance=0,6	Relevance=1,0	Relevance=0,6	Relevance=1,0
пастбище	насаждений	лесных	лесных
дубняк	пастбище	насаждений	насаждений
агролесного	жизнеустойчивости	деревьев	деревьев
секции	лесных	полос	полос
толщине	деревьев	дуба	дуба
садозащитной	лжеакации	защитных	защитных
масличных	горизонт	насаждения	насаждения
окс	дубняк	территории	территории
секциях	гзлп	полосы	полосы
тамарикса	секции	почвы	почвы
жизнеустойчивости	состояния	состояния	состояния
горизонт	дуба	пород	пород
кжу	контрольном	лесной	лесной
гумусовые	насаждения	лесные	лесные
нарушена	окрестностях	почв	почв
поврежденное	толщине	земель	земель
пробе	агролесного	площади	площади
окрестностях	тамарикса	условиях	условиях
кистей	окс	степи	степи
прореживание	лиственницей	пп	пп

Примечание: при показателе Relevance=0,6

На рис. 4 слова, связанные с «Темой 1» и «Темой 8» при показателе релевантности равном 0,6, ранжированы по их частоте внутри темы (красные столбцы), а серые столбцы отражают общую частоту терминов в корпусе. Оба показателя важны: чёткость различения тем повышается, если учитывать, как частоту терминов, так и их «исключительность» (степень специфичности термина для темы).

Рассмотрим рисунок. Абсолютная ширина красной полосы делает слово «жизнеустойчивости» одним из самых важных слов в определении Темы 1. Тем не менее это довольно распространенный термин (как показывает его серая полоса). А вот слово «пастбище» также одно из самых определяющих слов в теме, но, практически в два раза менее распространённое слово в корпусе.

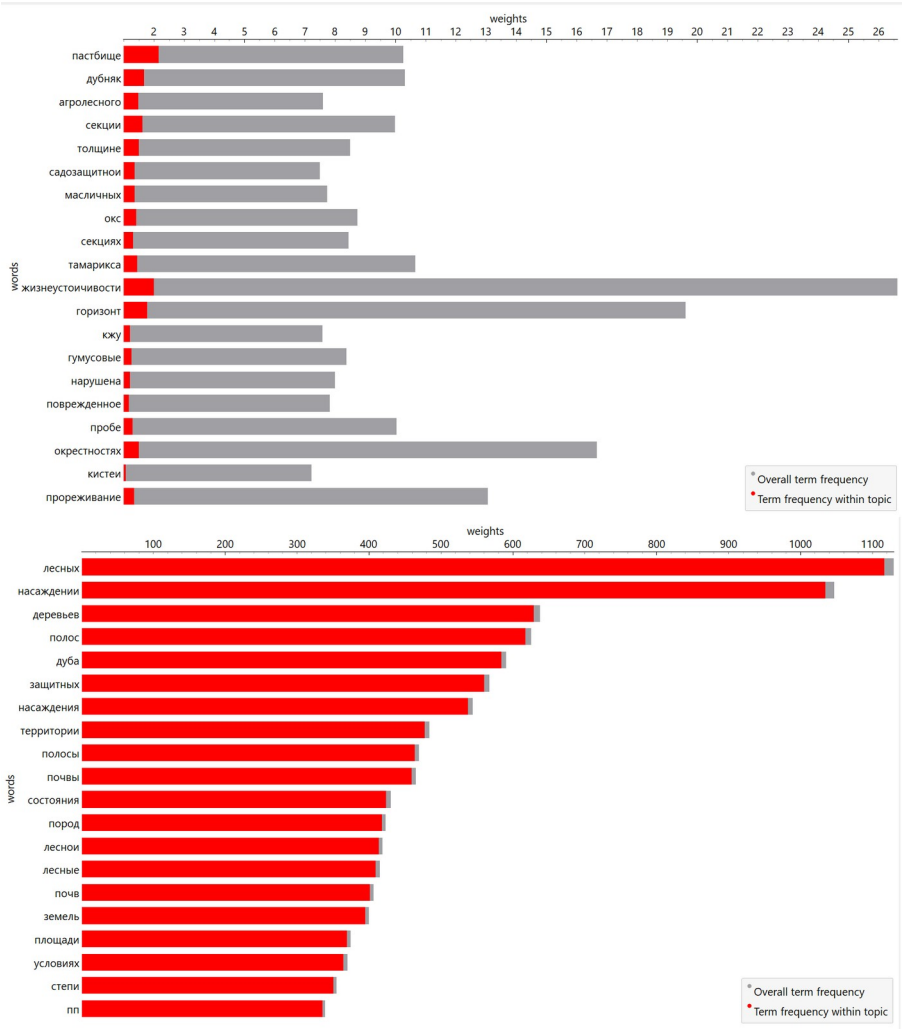

Рисунок 4 - Интерпретация тем № 1 и 8
DOI: https://doi.org/10.60797/IRJ.2025.158.33.6

Примечание: при показателе релевантности 0,6

Анализ показывает, что при снижении показателя релевантности с 1,0 до 0,6 список слов Темы 1 существенно меняется, и из 20 слов остаётся в списке лишь 8: «пастбище», «жизнеустойчивости», «дубняк», «секции», «толщине», «агролесного», «тамарикса», «ОКС».

В то время как для Темы 8 при аналогичном снижении показателя релевантности список слов не меняется — все 20 слов имеют широкое распространение в корпусе и в то же время обладают высокой эксклюзивностью для данной темы.

Снижение показателя релевантности с 1,0 до 0,6 помогает избежать «шума» от слишком частых, но неинформативных слов и выделить ключевые термины, которые действительно объясняют суть темы.

Рисунок 5 - Взаимосвязи между выделенными темами LDA посредством многомерного шкалирования
DOI: https://doi.org/10.60797/IRJ.2025.158.33.7

В программе Orange Data Mining для анализа корпуса научных текстов использовался метод многомерного шкалирования (MDS), где визуальные параметры отражают маргинальную вероятность тем (MTP):

- цветовая шкала: интенсивность цвета соответствует значению MTP (например, светло-жёлтый — доминирующая Тема 8 с MTP = 0,802, тёмно-синий — остальные, малозначимые темы.
- размер точек: пропорционален MTP (крупные точки — высокая вероятность, мелкие — низкая).
- связи между темами (настройка «Show similar pairs», SSP).

3.4. Семантическая интерпретация тематических кластеров

Результаты тематического моделирования выявили 10 кластеров, отражающих ключевые направления исследований в области защитного лесоразведения. Каждая тема была проинтерпретирована на основе анализа наиболее релевантных терминов, их контекстной значимости и междисциплинарных связей.

Тема 1: Агролесомелиорация пастбищ.

Ключевые термины: пастбище, жизнеустойчивость, дубняк, агролесные технологии, лжеакация.

Кластер фокусируется на интеграции лесных насаждений в пастбищные экосистемы. Основные аспекты включают:

- роль дубовых и акациевых насаждений в предотвращении деградации почв;
- оценку жизнеустойчивости агролесных систем в условиях антропогенных нагрузок;
- методы оптимизации пастбищного хозяйства через лесомелиоративные мероприятия.

Тема 2: Агроклиматические аспекты лесополос.

Ключевые термины: агроклиматические условия, территориальные исследования, лесополосы, ветровая эрозия.

Тема объединяет исследования влияния климатических факторов на эффективность защитных лесополос:

- анализ пространственного распределения лесных полос в зонах рискованного земледелия;
- взаимосвязь между структурой насаждений и их защитной функцией (снижение скорости ветра, сохранение влаги);
- региональные особенности агроклиматического районирования.

Тема 3: Биоразнообразие лесных экосистем.

Ключевые термины: дятлы, энтомофауна, медоносные виды, фитомасса.

Кластер посвящён изучению биотических взаимодействий в лесных экосистемах:

- роль птиц (дятлов) и насекомых в поддержании экологического баланса;
- влияние медоносных растений на продуктивность лесных сообществ;
- оценка флуктуирующей асимметрии как индикатора антропогенного стресса.

Тема 4: Восстановление нарушенных экосистем.

Ключевые термины: разрушенные почвы, сосна, фундук, интродукция.

Тема акцентирует методы рекультивации деградированных территорий:

- использование сосны и фундука для фиторемедиации почв;
- технологии восстановления растительного покрова в зонах горных выработок и карьеров;
- роль микоризных симбиозов в ускорении сукцессионных процессов.

Тема 5: Генетические ресурсы и резерваты.

Ключевые термины: генетические исследования, интегральная мелиорация, резерваты.

Кластер объединяет работы по сохранению биоразнообразия:

- создание генетических резерватов для защиты редких видов (например, крымской сосны);
- интеграция генетических и мелиоративных подходов в лесовосстановлении;
- оценка адаптивного потенциала древесных пород к климатическим изменениям.

Тема 6: Почвы и растительность залежных земель в контексте лесоразведения.

Ключевые термины: горизонты почв, карбонаты, чернозем, солонцы.

Тема исследует почвенные процессы:
- влияние карбонатных горизонтов на продуктивность лесных насаждений;
- связь между гранулометрическим составом почв и их мелиоративной ценностью;
- роль солонцов в формировании лесопастбищных ландшафтов.

Тема 7: Интродукция древесных видов.
Ключевые термины: орех, тамарикс, скрещивание, интродукция.
Кластер посвящён адаптации нетрадиционных видов в защитном лесоразведении:
- перспективы культивирования грецкого ореха и тамарикса в аридных регионах;
- гибридизация видов для повышения устойчивости к засухе и засолению;
- оценка инвазивного потенциала интродуцентов.

Тема 8: Основы защитного лесоразведения.
Ключевые термины: лесные полосы, дуб, защитные насаждения, лесоразведение.
Доминирующая тема (МТР = 0,802) охватывает базовые принципы дисциплины:
- принципы проектирования лесополос для защиты сельхозугодий;
- роль дуба как ключевой породы в мелиоративных насаждениях;
- методы оценки состояния и эффективности защитных лесонасаждений.

Тема 9: Деградация почв и морфологический стресс растений.
Ключевые термины: выщелоченные почвы, асимметрия, псевдотсуга, опустынивание.
Тема анализирует последствия деградации земель:
- влияние опустынивания на морфологию древесных пород;
- роль псевдотсуги в восстановлении экосистем;
- стратегии борьбы с деградацией чернозёмов в степных регионах.

Тема 10: Полифункциональные мелиоративные системы.
Ключевые термины: снегозадержание, горизонтальная планировка, полифункциональные системы.
Кластер объединяет инженерно-экологические подходы:
- технологии снегораспределения для повышения урожайности полей;
- интеграция лесных полос с гидротехническими сооружениями;
- оптимизация ландшафта для многоцелевого использования (защита почв, аккумуляция воды, биоразнообразие).

3.5. Интерпретация структуры корпуса

Доминирование Темы 8 подтверждает её роль системообразующего элемента в корпусе, объединяющего общеотраслевые аспекты. Периферийные темы (1–7, 9–10) специализируются на узких прикладных задачах, что соответствует междисциплинарной природе защитного лесоразведения. Визуализация MDS выявила изоляцию Темы 8 и кластеризацию остальных, что указывает на их семантическую уникальность, но ограниченную интеграцию с базовыми концепциями. Это подчёркивает необходимость расширения корпуса данных для уточнения межтематических связей.

Заключение

Проведённое исследование демонстрирует значительный потенциал методов тематического моделирования, в частности алгоритма латентного распределения Дирихле (LDA), для анализа и систематизации больших массивов научных текстов в области защитного лесоразведения. На основе корпуса из 533 публикаций (2000–2024 гг.) выделены 10 ключевых тем, отражающих междисциплинарный характер дисциплины: от агролесомелиорации пастбищ до химических процессов в карбонатных почвах. Качество модели подтверждено метриками лог-перплексии (12772) и тематической согласованности (0,54), что свидетельствует о её статистической устойчивости и семантической интерпретируемости.

Визуализация данных с помощью многомерного шкалирования (MDS) и инструмента LDAvis выявила доминирование Темы 8 («Основы защитного лесоразведения»), интегрирующей базовые принципы проектирования лесополос и оценки их эффективности. Периферийные темы, такие как «Восстановление нарушенных экосистем» (Тема 4) и «Деградация почв и морфологический стресс растений» (Тема 9), акцентируют узкоспециализированные аспекты, требующие углублённого изучения. Полученные результаты позволяют:

1. Структурировать знания в области лесомелиорации, выявляя связи между агроклиматическими факторами, биоразнообразием и почвенными процессами.

2. Идентифицировать пробелы в исследованиях, например, недостаточную изученность генетических ресурсов (Тема 5) и полифункциональных мелиоративных систем (Тема 10).

3. Способствовать принятию решений в агролесомелиорации и защитном лесоразведении, предоставляя данные для оптимизации технологий восстановления деградированных территорий.

Однако результаты также указывают на ограничения метода. Семантическая неоднозначность терминов (например, «защитные насаждения» в контексте эрозии и биоразнообразия) и неравномерное распределение тем в корпусе (доминирование общеотраслевых аспектов) требуют расширения выборки и интеграции с традиционными методами анализа.

Таким образом, тематическое моделирование выступает не только инструментом анализа, но и катализатором междисциплинарных исследований, способствуя переходу от фрагментарных данных к системному пониманию механизмов устойчивого развития лесных экосистем. Дальнейшие работы должны быть направлены на синтез алгоритмических и эмпирических подходов для преодоления методологических разрывов.

Конфликт интересов

Не указан.

Рецензия

Эминов Б.Ф., Казанский национальный исследовательский технический университет им. А.Н. Туполева – КАИ, Казань Российская Федерация
DOI: https://doi.org/10.60797/IRJ.2025.158.33.8

Conflict of Interest

None declared.

Review

Eminov B.F., Kazan National Research Technical University named after A.N. Tupolev – KAI, Kazan Russian Federation
DOI: https://doi.org/10.60797/IRJ.2025.158.33.8

Список литературы / References

1. Кулик К.Н. Лесомелиорация — основа создания устойчивых агроландшафтов в условиях недостаточного увлажнения / К.Н. Кулик, А.М. Пугачева // Лесотехнический журнал. — 2016. — Т. 6. — № 3 (23). — С. 29–40. — EDN: WMUWWL.

2. Кулик К.Н. Методическая основа агролесомелиоративной оценки защитных лесных насаждений по данным дистанционного мониторинга / К.Н. Кулик, А.В. Кошелев // Лесотехнический журнал. — 2017. — Т. 7. — № 3 (27). — С. 107–114. — DOI: 10.12737/article_59c22527885b57.91268039. — EDN: XCBVCL.

3. Antons D. Mapping the topic landscape of JPIM, 1984–2013: In search of hidden structures and development trajectories / D. Antons, R. Kleer, T.O. Salge // Journal of Product Innovation Management. — 2015. — Vol. 33. — № 6. — P. 726–749. — DOI: 10.1111/jpim.12300.

4. Blei D.M. Probabilistic topic models / D.M. Blei // Communications of the ACM. — 2012. — Vol. 55. — № 4. — P. 77–84. — DOI: 10.1145/2133806.2133826.

5. Zhang L. Aspect and entity extraction for opinion mining / L. Zhang, B. Liu // Data mining and knowledge discovery for big data: Methodologies, challenge and opportunities. Berlin, Heidelberg: Springer Berlin Heidelberg. — 2014. — P. 1–40. — DOI: 10.1007/978-3-642-40837-3_1.

6. Griffiths T.L. Finding scientific topics / T.L. Griffiths, M. Steyvers // Proceedings of the National Academy of Sciences. — 2004. — Vol. 101. — № suppl_1. — P. 5228–5235. — DOI: 10.1073/pnas.0307752101.

7. Yan E. Research dynamics: Measuring the continuity and popularity of research topics / E. Yan // Journal of Informetrics. — 2014. — Vol. 8. — № 1. — P. 98–110. — DOI: 10.1016/j.joi.2013.10.010.

8. Tirunillai S. Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation / S. Tirunillai, G.J. Tellis // Journal of Marketing Research. — 2014. — Vol. 51. — № 4. — P. 463–479. — DOI: 10.1509/jmr.12.0106.

9. De Battisti F. A decade of research in statistics: A topic model approach / F. De Battisti, A. Ferrara, S. Salini // Scientometrics. — 2015. — Vol. 103. — № 2. — P. 413–433. — DOI: 10.1007/s11192-015-1554-1.

10. Guo Y. Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent Dirichlet allocation / Y. Guo, S.J. Barnes, Q. Jia // Tourism management. — 2017. — Vol. 59. — P. 467–483. — DOI: 10.1016/j.tourman.2016.09.009.

11. Chae B. A topical exploration of the intellectual development of decision sciences 1975-2016: intellectual development of decision sciences 1975-2016 / B. Chae, D. Olson // Decision Sciences. — 2018. — Vol. 52. — № 3. — P. 543–566. — DOI: 10.1111/deci.12326.

12. Jiang H. A topic modeling based bibliometric exploration of hydropower research / H. Jiang, M. Qiang, P. Lin // Renewable and Sustainable Energy Reviews. — 2016. — Vol. 57. — P. 226–237. — DOI: 10.1016/j.rser.2015.12.194.

13. Танков А.А. Опыт применения методов технологии text mining для анализа экспертных описаний городских насаждений / А.А. Танков, Д.А. Танков, Ж.В. Танкова [и др.] // Национальные приоритеты развития агропромышленного комплекса: Материалы национальной научно-практической конференции с международным участием. — Оренбург: Агентство Пресса, 2023. — С. 1204–1210. — EDN: MYULTA

14. Танков А.А. Опыт применения методов технологии Text Mining для анализа экспертных описаний городских зелёных насаждений / А.А. Танков, Ж.В. Танкова, Д.А. Танков // Теория и практика инновационных исследований в области естественных наук: сборник материалов II Всероссийской научно-практической конференции с международным участием. Оренбург: Оренбургский государственный университет. — 2023. — С. 82–85. — EDN: GLGTTS.

15. Demšar J. Orange: data mining toolbox in Python / J. Demšar [et al.] // Journal of Machine Learning Research. — 2013. — Vol. 14. — № 1. — P. 2349–2353. — DOI:10.5555/2567709.2567736.

16. Orange Data Mining. — URL: https://orangedatamining.com (accessed: 01.11.2024).

17. Daud A. Knowledge discovery through directed probabilistic topic models: a survey / A. Daud, J. Li, F. Muhammad // Frontiers of Computer Science in China. — 2009. — № 4. — P. 280–301. — DOI: 10.1007/s11704-009-0062-y.

18. Sievert C. LDAvis / C. Sievert K. Shirley // CRAN R Repository. — 2015. — DOI: 10.32614/CRAN.package.LDAvis.

19. Blei D.M. A correlated topic model of science / D.M. Blei, J.D. Lafferty // The annals of applied statistics. — 2007. — № 1 (1). — P. 17–35. — DOI: 10.1214/07-aoas114.

20. Hall D. Studying the history of ideas using topic models / D. Hall, D. Jurafsky, C. Manning // Conference on Empirical Methods in Natural Language Processing. — 2008. — P. 363–371. — DOI: 10.3115/1613715.1613763.

21. Grün B. Topicmodels: an R package for fitting topic models / B. Grün, K. Hornik // Journal of Statistical Software. — 2011. — № 13. — P. 1–30. — DOI: 10.18637/jss.v040.i13.

22. Kapadia S. Evaluate topic models: Latent Dirichlet Allocation (LDA) / S. Kapadia // Towards Data Science. — 2019. — URL: https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0 (accessed: 11.02.2025).

23. Mimno D. Bayesian Checking for Topic Models / D. Mimno, D.M. Blei // Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference. — 2011. — P. 227–237. — URL: https://aclanthology.org/D11-1021/ (accessed: 01.12.2024).

24. Liu S. Analysis and prospect of clinical psychology based on topic models: hot research topics and scientific trends in the latest decades / S. Liu, R.-Y. Zhang, T. Kishimoto // Psychology, Health & Medicine. — 2020. — № 26 (4). — P. 1–13. — DOI: 10.1080/13548506.2020.1738019.

25. Röder M. Exploring the Space of Topic Coherence Measures / M. Röder M., A. Both A., A. Hinneburg // Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM'15. — 2015. — P. 399–408. — DOI: 10.1145/2684822.2685324.

26. Kumar K. Evaluation of Topic Modeling: Topic Coherence / K. Kumar // Data Science+. — 2018. — URL: https://datascienceplus.com/evaluation-of-topic-modeling-topic-coherence/ (accessed: 06.10.2022).

27. Chang J. Reading Tea Leaves: How Humans Interpret Topic Models / J. Chang, J. Boyd-Graber, S. Gerrish [et al.] // Neural Information Processing Systems. — 2009. — № 32. — P. 288–296.

28. Sievert C. LDAvis: A method for visualizing and interpreting topics / C. Sievert, K. Shirley. — 2014. — DOI: 10.3115/v1/W14-3110.

Список литературы на английском языке / References in English

1. Kulik K.N. Lesomelioracija – osnova sozdanija ustojchivyh agrolandshaftov v uslovijah nedostatochnogo uvlazhnenija [Forest melioration — the basis for the creation of sustainable agrolandscapes in conditions of insufficient moisture] / K.N. Kulik, A.M. Pugacheva // Lesotehnicheskij zhurnal [Forestry Journal]. — 2016. — Vol. 6. — № 3 (23). — P. 29–40. — EDN: WMUWWL. [in Russian]

2. Kulik K.N. Metodicheskaja osnova agrolesomeliorativnoj ocenki zashhitnyh lesnyh nasazhdenij po dannym distancionnogo monitoringa [Methodological basis for agroforestry assessment of protective forest plantations based on remote monitoring data] / K.N. Kulik, A.V. Koshelev // Lesotehnicheskij zhurnal [Forestry Journal]. — 2017. — Vol. 7. — № 3 (27). — P. 107–114. — DOI: 10.12737/article_59c22527885b57.91268039. — EDN: XCBVCL. [in Russian]

3. Antons D. Mapping the topic landscape of JPIM, 1984–2013: In search of hidden structures and development trajectories / D. Antons, R. Kleer, T.O. Salge // Journal of Product Innovation Management. — 2015. — Vol. 33. — № 6. — P. 726–749. — DOI: 10.1111/jpim.12300.

4. Blei D.M. Probabilistic topic models / D.M. Blei // Communications of the ACM. — 2012. — Vol. 55. — № 4. — P. 77–84. — DOI: 10.1145/2133806.2133826.

5. Zhang L. Aspect and entity extraction for opinion mining / L. Zhang, B. Liu // Data mining and knowledge discovery for big data: Methodologies, challenge and opportunities. Berlin, Heidelberg: Springer Berlin Heidelberg. — 2014. — P. 1–40. — DOI: 10.1007/978-3-642-40837-3_1.

6. Griffiths T.L. Finding scientific topics / T.L. Griffiths, M. Steyvers // Proceedings of the National Academy of Sciences. — 2004. — Vol. 101. — № suppl_1. — P. 5228–5235. — DOI: 10.1073/pnas.0307752101.

7. Yan E. Research dynamics: Measuring the continuity and popularity of research topics / E. Yan // Journal of Informetrics. — 2014. — Vol. 8. — № 1. — P. 98–110. — DOI: 10.1016/j.joi.2013.10.010.

8. Tirunillai S. Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation / S. Tirunillai, G.J. Tellis // Journal of Marketing Research. — 2014. — Vol. 51. — № 4. — P. 463–479. — DOI: 10.1509/jmr.12.0106.

9. De Battisti F. A decade of research in statistics: A topic model approach / F. De Battisti, A. Ferrara, S. Salini // Scientometrics. — 2015. — Vol. 103. — № 2. — P. 413–433. — DOI: 10.1007/s11192-015-1554-1.

10. Guo Y. Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent Dirichlet allocation / Y. Guo, S.J. Barnes, Q. Jia // Tourism management. — 2017. — Vol. 59. — P. 467–483. — DOI: 10.1016/j.tourman.2016.09.009.

11. Chae B. A topical exploration of the intellectual development of decision sciences 1975-2016: intellectual development of decision sciences 1975-2016 / B. Chae, D. Olson // Decision Sciences. — 2018. — Vol. 52. — № 3. — P. 543–566. — DOI: 10.1111/deci.12326.

12. Jiang H. A topic modeling based bibliometric exploration of hydropower research / H. Jiang, M. Qiang, P. Lin // Renewable and Sustainable Energy Reviews. — 2016. — Vol. 57. — P. 226–237. — DOI: 10.1016/j.rser.2015.12.194.

13. Tankov A.A. Opyt primenenija metodov tehnologii text mining dlja analiza jekspertnyh opisanij gorodskih nasazhdenij [Experience in the application of text mining methods to analyse expert descriptions of urban plantations] / A.A. Tankov, D.A. Tankov, Zh.V. Tankova [et al.] // Nacional'nye prioritety razvitija agropromyshlennogo kompleksa: Materialy nacional'noj nauchno-prakticheskoj konferencii s mezhdunarodnym uchastiem [National priorities of agro-industrial complex development: Proceedings of the National Scientific and Practical Conference with international participation]. — Orenburg: Agentstvo Pressa, 2023. — P. 1204–1210. — EDN: MYULTA [in Russian]

14. Tankov A.A. Opyt primenenija metodov tehnologii Text Mining dlja analiza jekspertnyh opisanij gorodskih zeljonyh nasazhdenij [Experience of application of Text Mining technology methods to analyse expert descriptions of urban green spaces] / A.A. Tankov, Zh.V. Tankova, D.A. Tankov // Teorija i praktika innovacionnyh issledovanij v oblasti estestvennyh nauk: sbornik materialov II Vserossijskoj nauchno-prakticheskoj konferencii s mezhdunarodnym uchastiem [Theory and practice of innovative research in the field of natural sciences: Proceedings of the II All-Russian Scientific and Practical

Conference with international participation]. Orenburg: Orenburg State University. — 2023. — P. 82–85. — EDN: GLGTTS. [in Russian]

15. Demšar J. Orange: data mining toolbox in Python / J. Demšar [et al.] // Journal of Machine Learning Research. — 2013. — Vol. 14. — № 1. — P. 2349–2353. — DOI:10.5555/2567709.2567736.
16. Orange Data Mining. — URL: https://orangedatamining.com (accessed: 01.11.2024).
17. Daud A. Knowledge discovery through directed probabilistic topic models: a survey / A. Daud, J. Li, F. Muhammad // Frontiers of Computer Science in China. — 2009. — № 4. — P. 280–301. — DOI: 10.1007/s11704-009-0062-y.
18. Sievert C. LDAvis / C. Sievert K. Shirley // CRAN R Repository. — 2015. — DOI: 10.32614/CRAN.package.LDAvis.
19. Blei D.M. A correlated topic model of science / D.M. Blei, J.D. Lafferty // The annals of applied statistics. — 2007. — № 1 (1). — P. 17–35. — DOI: 10.1214/07-aoas114.
20. Hall D. Studying the history of ideas using topic models / D. Hall, D. Jurafsky, C. Manning // Conference on Empirical Methods in Natural Language Processing. — 2008. — P. 363–371. — DOI: 10.3115/1613715.1613763.
21. Grün B. Topicmodels: an R package for fitting topic models / B. Grün, K. Hornik // Journal of Statistical Software. — 2011. — № 13. — P. 1–30. — DOI: 10.18637/jss.v040.i13.
22. Kapadia S. Evaluate topic models: Latent Dirichlet Allocation (LDA) / S. Kapadia // Towards Data Science. — 2019. — URL: https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0 (accessed: 11.02.2025).
23. Mimno D. Bayesian Checking for Topic Models / D. Mimno, D.M. Blei // Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference. — 2011. — P. 227–237. — URL: https://aclanthology.org/D11-1021/ (accessed: 01.12.2024).
24. Liu S. Analysis and prospect of clinical psychology based on topic models: hot research topics and scientific trends in the latest decades / S. Liu, R.-Y. Zhang, T. Kishimoto // Psychology, Health & Medicine. — 2020. — № 26 (4). — P. 1–13. — DOI: 10.1080/13548506.2020.1738019.
25. Röder M. Exploring the Space of Topic Coherence Measures / M. Röder M., A. Both A., A. Hinneburg // Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM'15. — 2015. — P. 399–408. — DOI: 10.1145/2684822.2685324.
26. Kumar K. Evaluation of Topic Modeling: Topic Coherence / K. Kumar // Data Science+. — 2018. — URL: https://datascienceplus.com/evaluation-of-topic-modeling-topic-coherence/ (accessed: 06.10.2022).
27. Chang J. Reading Tea Leaves: How Humans Interpret Topic Models / J. Chang, J. Boyd-Graber, S. Gerrish [et al.] // Neural Information Processing Systems. — 2009. — № 32. — P. 288–296.
28. Sievert C. LDAvis: A method for visualizing and interpreting topics / C. Sievert, K. Shirley. — 2014. — DOI: 10.3115/v1/W14-3110.