То что желтым это под сомнением.

Сравнение методов тематического моделирования: BERTopic, Top2vec, LDA и (возможно) NMF

# **Структура**

1. Введение

* Важность тематического моделирования в анализе научных текстов.
* Краткий обзор и мотивация для сравнения BERTopic, Top2vec, LDA и (возможно) NMF.

1. Корпус данных и Методы

* Описание корпуса данных из 550 тысяч абстрактов научных статей.
* Процесс подготовки данных
* Реализация моделей
* Методика сравнения: использование метрик coherence, topic diversity, inner recall, joint inner recall.

1. Результаты

* Результаты моделирования для каждого алгоритма.
* Сравнение моделей по каждой из используемых метрик.
* Визуализации результатов для наглядности.

1. Заключение

* Выводы исследования.
* Возможные направления для будущих исследований.
* Итоговые рекомендации по выбору алгоритма тематического моделирования.

1. Список литературы

# Введение #TODO больше референсов

Тематическое моделирование играет ключевую роль в анализе научных текстов, обеспечивая уникальные возможности для извлечения значимой информации из больших объемов текстовых данных. В научной среде, где ежедневно публикуется огромное количество статей, обзоров и отчетов, эффективные инструменты для обработки и анализа текстов становятся не просто полезными, но и необходимыми для поддержания актуальности и продуктивности исследований.

Тематическое моделирование позволяет автоматически выявлять скрытые тематические структуры в текстовых данных, что особенно ценно в научной области. Это не только способствует более глубокому пониманию содержания научных работ, но и облегчает процессы поиска и классификации документов, сегментации текстов по темам и отслеживания развития исследовательских трендов. Кроме того, тематическое моделирование может выявить междисциплинарные связи, что особенно важно в эпоху расширения границ научных дисциплин. Среди множества подходов к тематическому моделированию особое внимание заслуживают такие алгоритмы, как LDA (Latent Dirichlet Allocation), BERTopic, Top2vec и (возможно) NMF (Non-negative Matrix Factorization). Каждый из этих методов имеет свои уникальные особенности и подходы к обработке текста, что делает их сравнительный анализ особенно актуальным для научного сообщества. Особенно важно понимать, как эти модели работают с большими научными датасетами, такими как наборы абстрактов научных статей, и какие преимущества каждый метод может предложить в контексте научных исследований.

Целью данной статьи является всестороннее сравнение этих алгоритмов тематического моделирования с акцентом на их применение к научным текстам. Мы рассмотрим каждый алгоритм в контексте датасета из 550 тысяч абстрактов научных статей, оценивая их способность выявлять значимые и интерпретируемые темы, а также их эффективность и универсальность в различных научных дисциплинах.

## Обзор и мотивация для сравнения BERTopic, Top2vec, LDA и (возможно) NMF

В выборе алгоритмов для тематического моделирования, особенно в контексте научных текстов, ключевыми являются их способность выявлять глубокие, осмысленные темы и обрабатывать большие объемы данных с высокой точностью и эффективностью. Рассмотрение различных подходов к моделированию — LDA, BERTopic, Top2vec и (возможно) NMF — мотивируется их уникальными характеристиками и различиями в принципах работы, что может существенно влиять на результаты анализа научных текстов.

* LDA (Latent Dirichlet Allocation) является одним из наиболее известных и широко используемых алгоритмов тематического моделирования, основан на вероятностном подходе и предполагает, что документы представляют собой смесь тем.
* BERTopic базируется на передовых нейронных сетях и методах обработки естественного языка, что позволяет ему более эффективно работать с контекстуальными особенностями языка и выявлять более сложные тематические структуры.
* Top2vec использует подходы на основе векторных представлений слов и документов, позволяя автоматически обнаруживать темы, которые тесно связаны с содержанием и структурой текста.
* NMF (Non-negative Matrix Factorization) (если будет включен в анализ) представляет собой альтернативный подход, основанный на факторизации матриц, и может предложить интересные варианты выделения тем за счет разложения текстов на компоненты.

Цель сравнения этих методов заключается в том, чтобы определить, какие подходы наиболее эффективны для различных типов научных текстов, а также какие уникальные преимущества и недостатки имеет каждый метод. Особенно важно понять, как эти различные подходы способны справляться с высокой степенью техничности и специфичности научных абстрактов, и какие из них могут предложить наиболее глубокое и полезное понимание содержания научных исследований.

# Корпус данных и Методы

## Корпус Научных Абстрактов: STA (Scientific Thematic Archive)

В рамках нашего исследования мы сформировали уникальный корпус данных, который для удобства дальнейшего анализа и ссылок мы условно обозначаем как 'Научный Тематический Архив' (Scientific Thematic Archive, STA). Корпус данных STA, используемый в нашем исследовании, представляет собой значительную часть метаданных из репозитория ArXiv, широко известного открытого архива, который предоставляет доступ к научным статьям в различных областях науки. Этот архив включает в себя широкий спектр тематик, от физики и информатики до математики, статистики, электротехники, квантовой биологии и экономики. Выбранный нами корпус STA является частью оригинального набора данных ArXiv, доступного на Kaggle, который включает метаданные примерно для 2.2 миллиона статей. Метаданные охватывают различные характеристики, такие как названия статей, авторы, категории, абстракты и полный текст в формате PDF.

Этот богатый репозиторий научных статей является ценным ресурсом для анализа данных, выявления трендов и разработки моделей машинного обучения. Он может способствовать таким приложениям, как анализ трендов, системы рекомендации статей, предсказание категорий, анализ сети со-цитирования, построение знаний графов и семантические поисковые интерфейсы. Данные особенно подходят для тех, кто интересуется обработкой естественного языка и текстовым анализом в академической сфере. В нашем исследовании мы сосредоточились на абстрактах статей, выбрав 550 тысяч из них для создания многоаспектного и мультидисциплинарного корпуса STA. Это обеспечило нам широкий и разнообразный набор данных для анализа и сравнения различных методов тематического моделирования. Основное внимание в процессе подготовки данных было уделено стандартизации и очистке текстов, чтобы обеспечить качественное и однородное исходное содержание для последующего анализа.

## Препроцессинг данных # TODO написать

## Реализация моделей # TODO доработать

### Модель 1: Latent Dirichlet Allocation (LDA)

LDA, наиболее популярная техника тематического моделирования, представляет собой генеративную вероятностную модель для дискретных наборов данных, таких как текстовые корпуса (Сидоров и др., 2024). Эта модель является трехуровневой иерархической байесовской моделью, в которой каждый элемент коллекции представлен в виде конечной смеси на основе набора тем, а каждая тема, в свою очередь, представлена как бесконечная смесь на основе коллекции вероятностей тем. Таким образом, число тем в модели LDA не нужно определять заранее (Иванов и др., 2024), что предоставляет исследователям эффективный ресурс для получения явного представления документа.

В нашем исследовании, для определения оптимальных значений

### Модель 2:

### Модель 3:

### Модель 4:

## Методика сравнения #TODO придумать как правильно начать и потом расписать

* **Coherence** (Согласованность): Эта метрика измеряет степень семантической согласованности слов в теме. Высокая согласованность указывает на то, что слова хорошо сочетаются друг с другом, что делает тему более понятной и значимой.
* **Topic Diversity** (Разнообразие тем): Эта метрика помогает оценить, насколько разнообразны темы, генерируемые моделью. Если модель генерирует темы, которые сильно перекрываются, это может указывать на переобучение или недостаточную способность различать разные темы.
* **Inner Recall** (Внутренний отклик): Эта метрика может использоваться для оценки того, насколько хорошо модель группирует похожие документы в одну и ту же тему. Это помогает понять, насколько эффективно модель справляется с задачей кластеризации текстов.
* **Joint Inner Recall** (Совместный внутренний отклик): Это вариация Inner Recall, которая учитывает не только группировку похожих документов, но и то, насколько хорошо эти группы отделены друг от друга. Это помогает оценить, насколько чётко модель разделяет разные темы.

# Результаты

# Заключение

# Список литературы

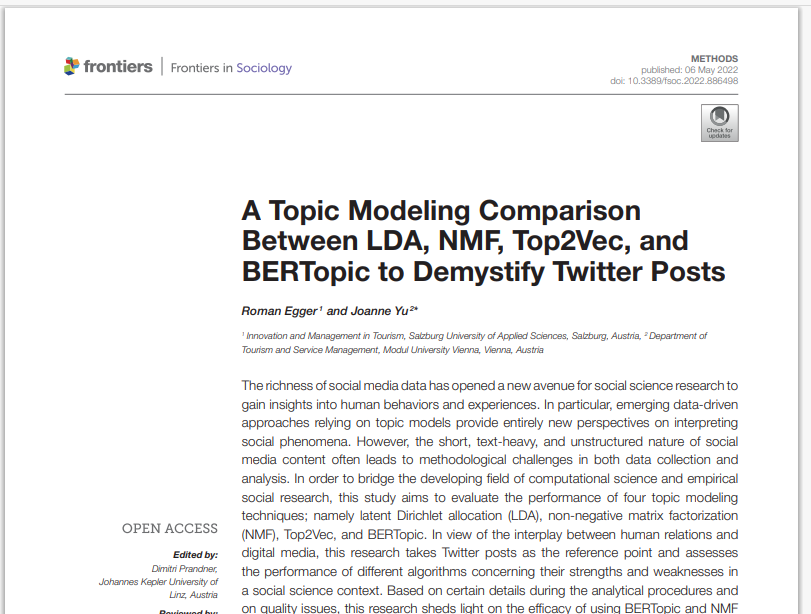
1. ?<https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/68e91a69-b0c5-4b8c-bf1a-f6847d77a8bd/content>
2. ?A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts <https://www.frontiersin.org/articles/10.3389/fsoc.2022.886498/full>
3. Angelov, D. (2020). Top2Vec: Distributed Representations of Topics (arXiv:2008.09470). arXiv. http://arxiv.org/abs/2008.09470

# # TODO

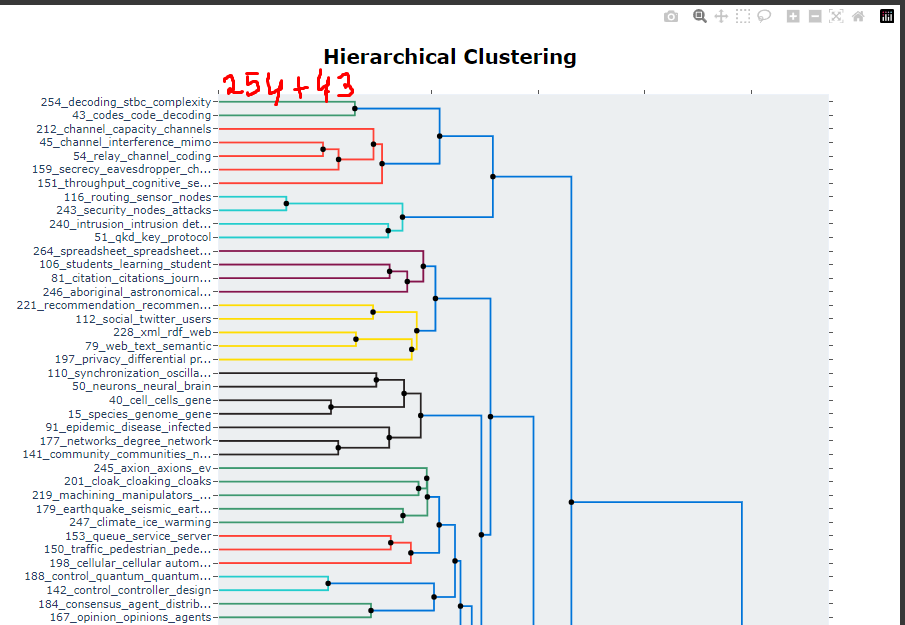
Оформить Hugging face (model, dataset)

# Статья похожая на нашу

2022 года. <https://www.frontiersin.org/articles/10.3389/fsoc.2022.886498/full>



## Иерархическая кластеризация.



\*Note that for legal issues, the complete full-text of journal articles could not be included with the dataset (but can be retrieved by asking JSTOR and the respective publishers).

Т.к корпус данных не полностью состоит из английского языка, придется переводить с помощью машинного перевода и не получится воссоздать тот же датасет, который использовался в статье.