МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»**

Кафедра социальной информатики

ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ НАБОРА ДАННЫХ НАУЧНЫХ ПУБЛИКАЦИЙ

(автореферат бакалаврской работы)

Студента 4 курса 452 группы
направления 09.03.03 – «Прикладная информатика»
профиль «Прикладная информатика в социологии»
Социологического факультета
Дилмурадова Довлетмурада

Научный руководитель
профессор, доктор социологических наук, _____ Н. И. Мельникова
профессор
подпись, дата

Зав. кафедрой
кандидат социологических наук, доцент _____ И. Г. Малинский
подпись, дата

Саратов 2023

ВВЕДЕНИЕ

Актуальность проблемы. В условиях постоянно увеличивающегося объёма информации, в частности информации текстовой и так называемого «информационного бума» появляется необходимость анализировать данные такого объема, который человек не в состоянии в одиночку обработать самостоятельно. Следовательно, необходимы методы и средства для автоматического извлечения необходимой информации из большого набора данных. Теоретически обоснованным и активно развивающимся направлением в анализе текстов на естественном языке является тематическое моделирование коллекций текстовых документов. Именно по этой причине данная тема выпускной квалификационной работы является актуальной.

Построение тематической модели можно рассматривать как задачу одновременно кластеризации документов и слов в одном и том же наборе кластеров, называемых темами. Процесс нахождения тем по кластеризации называется тематическим моделированием.

Целью данной работы является разработка Jupyter Notebook по определению тематик документов.

Для данной поставленной цели необходимо решить следующие **задачи:**

1. Рассмотреть алгоритмы математического моделирования.

2. Рассмотреть библиотеки на языке Python по тематическому моделированию.

3. Разработать ноутбук с использованием библиотеки тематического моделирования.

Объектом является набор данных по тематическому моделированию.

Предметом исследования является использование тематического моделирования для Jupyter Notebook по определению тематик документов.

Теоретико-методическая основа. При решении поставленных задач использовались методы системного анализа, математического и компьютерного моделирования, автоматической обработки естественного языка, алгоритмов, разработки информационных систем и программирования.

Структура работы. Выпускная квалификационная работа состоит из введения, трёх глав, поделённых на параграфы, заключения, списка использованных источников.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе «Тематическое моделирование» рассматриваются основные понятия тематического моделирования, латентно-семантический анализ LSA, латентное размещение Дирихле LDA, тематическая классификация, векторная модель текста и вероятностный латентно-семантический анализ pLSA.

Способом построения модели коллекции текстовых документов, которая определяет, к каким темам относится конкретный документ, и какие слова образуют каждую тему, называется тематическим моделированием.

Тематическое моделирование активно развивается на протяжении последних тридцати лет. За это время разработано множество различных моделей для решения различных зада по обработке текстов на естественных языках, а также графической и видеоинформации.

Методы тематического моделирования позволяют выявить темы, рассматриваемые в текстах. Методы тематического моделирования позволяют выявить темы в не структурируемых данных. При обнаружении таких паттернов, как частота и расстояние между словами, тематическая модель объединяет информацию, которая схожа, и слова, и выражения, которые появляются чаще всего. С помощью этой информации можно определить, о чем идет речь в каждом наборе текстов.

Тематическое моделирования используется для анализа и формирования информационных ресурсов, в анализе экономической деятельности, в социальном прогнозировании п проектировании. Основой для всех этих исследований являются, прежде всего, неструктурированные тактовые наборы данных.

Также отмечено, что для тематического моделирования текстовой информации важно учитывать строение и синтаксическую структуру

предложений для отслеживания изменение тематики во временном отрезке или внутри отдельных документов. Изменения же выстраивают определённые иерархические отношения между темами, которые, в свою очередь, учитывают взаимосвязь авторства, ссылок и других составляющих.

Вероятностные тематические модели применяются для анализа документов и определения смысла слов. Вероятностные тематические модели делают возможным «мягкую» кластеризацию терминов и необходимых документов по кластерам. Так как эти термины и документы в одно и то же время могут относиться к нескольким многим темам с разными вероятностями. Именно благодаря этому становится возможным решение «синонимии» и «омонимии» определённых терминов, которые появляются при уже «жёсткой» кластеризации. Известно, часто употребляемые синонимы в похожих контекстах находят своё применение с гораздо большей вероятностью в одной и той же теме. Тогда как омонимы, исходя из их частоты использования, распределяются между темами и употребляются в значимо различных контекстах.

Пусть D - множество (коллекция) текстовых документов, W - множество (словарь) всех используемых в текстовых документах терминов (слов или словосочетаний). Каждый документ d ∈ D представляет собой последовательность n_d терминов (w_1, . . .,w_{n_d}) из словаря W. Термин может повторяться в документе большое количество раз.

Латентно семантический анализ (ЛСА, LSA) — это статистический метод обработки текстовой информации на естественном языке, позволяющий определить взаимосвязь между коллекциями документов и терминами, которые в них встречаются. В основе данного метода лежит принцип факторного анализа, в частности, выявление латентных связей изучаемых явлений и объектов. При классификации и кластеризации документов данный метод позволяет извлечь контекстно-зависимые значения лексических единиц.

Возможно разделение основного алгоритма вышеупомянутого метода на четыре шага:

1. Предобработка,
2. Нахождение весов слов любым методом, например, с помощью алгоритма tf-idf,
3. Построение весовой матрицы,
4. Разложение матрицы методом сингулярного разложения (англ. Singular value decomposition, SVD).

Результатом работы алгоритма будет являться матрица, визуализация которой позволит отразить общую семантическую близость документов друг к другу.

Вероятностный латентно-семантический анализ — это статистический метод анализа корреляций двух типов данных. В общем смысле, данный метод является развитием уже упомянутого латентно-семантического анализа, однако в отличие от своего предшественника, который по своей сути является алгоритмом построения векторного представления с последующим снижением его размерности, ВЛСА основан на смешанном разложении и использовании вероятностной модели. Именно это позволяет более точно и качественно определять возможные тематики документов.

Рассмотрено латентное размещение Дирихле. Оно применяется в информационном поиске, её еще называют «порождающая модель», позволяющая объяснить результаты наблюдений с помощью скрытых (латентных) групп. Данная модель представляет собой расширение модели pLSA и устраняет основные её недостатки путем использования распределения Дирихле в качестве основного распределения. По итогу получается, что набор тематик выводится более конкретный и четкий.

Вследствие описания и рассмотрения основных методов тематического моделирования можно сказать, что методы, которые основаны на вероятностных моделях наиболее пригодны для решения поставленной задачи, но в то же время требуют высоких вычислительных затрат при реализации в исходном виде. Рассмотренный метод LDA, который позволяет избежать

основных недостатков pLSA, является наиболее сложным, и при этом позволяет достичь наилучших результатов.

С вышеупомянутыми моделями известны и другие тематические модели, которые связаны в той или иной мере. Выделены следующие из них:

- совместная вероятностная модель JPM (англ. Joint Probabilistic Model),
- скрытая тематическая марковская модель AHMM (англ. Aspect Hidden Markov Model),
- автор-тематическая модель ATM (англ. Author-Topic Model),
- модель автор-получатель ARTM (англ. Author-Recipient Topic Model),
- корреляционная тематическая модель CTM (англ. Correlated Topic Model).

Наконец, тематическое моделирование представляет собой технику машинного обучения. Эта техника (то есть способ) автоматически анализирует текстовые данные для определения кластеров слов для набора документов. Также это называется «неконтролируемое» машинное обучение, так как не требует определенного списка тегов или набора обучающих данных, которые ранее были классифицированы людьми.

Тематическая классификация требует дополнительной работы, этот метод анализа тем дает более точные результаты, чем неконтролируемые методы, что означает, что вы получите более ценные сведения, которые помогут вам принимать более эффективные решения на основе данных.

Во второй главе «Командная оболочка Jupyter Notebook и библиотеки тематического моделирования» рассматриваются: командная оболочка Jupyter Notebook и библиотеки BigARTM и Gensim.

Project Jupyter – это некоммерческая организация, созданная для разработки программного обеспечения с открытым исходным кодом, открытых стандартов и услуг для интерактивных вычислений на десятках языков программирования. Project Jupyter, выделенный из IPython в 2014 году

Фернандо Пересом, поддерживает среды исполнения на нескольких десятках языков. Название Project J\upyter является ссылкой на три основных языка программирования, поддерживаемых Jupyter, а именно Julia, Python и R, а также дань уважения к записным книжкам Галилея, в которых записано открытие спутников Юпитера. Project Jupyter разработал и поддерживал продукты для интерактивных вычислений Jupyter Notebook, JupyterHub и JupyterLab, версию Jupyter Notebook следующего поколения.

BigARTM — это свободно распространяемая библиотека, которая предназначена для тематического моделирования больших наборов текстовых документов и массивов транзакционных данных. Она обеспечивает эффективную параллельную обработку данных и основана на вероятностном подходе с использованием аддитивной регуляризации

ARTM (англ. Additive Regularization for Topic Modeling);

APTM (англ. Аддитивная Регуляризация Тематических Моделей).

Были определены функциональные возможности:

BigARTM представляет собой модульную реализацию технологии ARTM. Основное ядро библиотеки написано на языке C++ в соответствии с промышленными стандартами программирования. Благодаря поддержке распараллеливания на многопроцессорных системах, BigARTM позволяет эффективно обрабатывать большие объемы данных без необходимости загружать все данные в оперативную память одновременно. Он демонстрирует линейную вычислительную сложность, зависящую от размера коллекции и количества тем. BigARTM превосходит другие доступные бесплатные библиотеки по скорости вычислений. Также он включает в себя встроенную библиотеку регуляризаторов и метрик качества и предоставляет возможность добавления пользовательских регуляризаторов. В целом, BigARTM иногда называется, как «конструктор LEGO» для создания тематических моделей.

BigARTM представляет собой библиотеку, которая включает в себя несколько механизмов, которые устраняют ограничения простых моделей,

таких как PLSA или LDA, и расширяют возможности в области тематического моделирования. Среди них:

- Regularization. Регуляризаторы, которые можно комбинировать в любых сочетаниях.
- Modality. Модальности, которыми можно описывать нетекстовые объекты внутри документов.
- Hierarchy. Тематические иерархии, в которых темы разделяются на подтемы.
- Intratext. Обработка текста как последовательности тематических векторов слов.
- Co-occurrence. Использование данных о совместной встречаемости слов.
- Hypergraph. Тематизация сложно структурированных транзакционных данных.

Была рассмотрена библиотека Gensim и её основные характеристики.

Так, определено, что Gensim — это библиотека Python для тематического моделирования, индексации документов и поиска сходства в больших корпусах текстов. Данная библиотека преимущественно используется в области обработки естественных языка и в поиске информации. Gensim разработана на основе модели VSM. Алгоритмы, реализованные в Gensim, такие как Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA) позволяют выявить семантическую структуру документа, основываясь на статистических связях между словами в тексте.

Gensim обладает способностью работать с обширными коллекциями текстовых данных, что является отличительной чертой данной библиотеки по сравнению с другими программными библиотеками машинного обучения, которые ориентированы на обработку данных в оперативной памяти. Также Gensim предлагает эффективные многопоточные реализации различных алгоритмов, что способствует повышению скорости обработки данных.

Среди свойств genism: следует отметить отсутствие необходимости загружать весь входной набор данных в память одновременно, что позволяет обрабатывать большие объемы корпусов без ограничений памяти и то. что в библиотеке эффективно реализованы большая часть известных алгоритмов тематического моделирования.

В третьей главе «Разработка ноутбука по тематического моделирование в среде Jupyter Nootebook» продемонстрирован и описан процесс разработки в Jupyter Notebook с использованием алгоритма Дирихле.

Были представлены исходные данные и описан способ их обработки. С помощью рисунков автор показал результат выполнения программного кода.

JupyterNotebook (ранее IPythonNotebooks) - это интерактивная вычислительная среда на основе веб-интерфейса для создания документов Jupyternotebooks. Она часто используется для работы с данными, статистическим моделированием и машинным обучением. Термин «notebook» в разговорной речи может относиться ко многим различным сущностям, в основном к веб-приложению Jupyter, веб-серверу JupyterPython или формату документа Jupyter, в зависимости от контекста. Документ Jupyternotebook - это документ JSON, содержащий упорядоченный список ячеек ввода / вывода, \который может содержат\ь код, текст (с использованием Markdown), математику, графики и мультимедиа, обычно заканчивающиеся на «.ipynb» расширение.

Jupyter Notebook предоставляет REPL (read–eval–printloop/ цикл чтения – оценки – печати) на основе браузера, построенный на ряде популярных библиотек с открытым исходным кодом:

- IPython
- ØMQ
- Торнадо (веб-сервер)
- JQuery
- Bootstrap (front-end framework)
- MathJax

Меню в Jupyter Notebook представляет ряд возможностей для работы с ноутбуками и включает следующие пункты:

- File
- Edit
- View
- Insert
- Cell
- Kernel
- Widgets
- Help

Выполнение практической части заключалось в том, чтобы представить публикации научных сотрудников СГУ, которые поступили в научную библиотеку университета в 2022 году. Для разработки ноутбука использовалась открытая БД «Публикации ученых СГУ», содержащая содержит библиографические описания печатных и электронных публикаций ученых университета, изданных, начиная с 2008 года по настоящее время. Данные по поступлениям в библиотеку за 2022 год следующие:

Таблица 1 – Количество публикаций ученых СГУ

Месяц	Количество-публикаций
январь	147
февраль	92
март	79
апрель	113
май	92
Июнь	199
Июль	154
Август	124
сентябрь	289
Октябрь	259
Ноябрь	167
декабрь	211

Первоначально по каждому месяцу в диапазоне январь-декабрь был создан файл с кратким описанием результатов поиска. Такая возможность

представляется на сайте библиотеки. Далее все данные по месяцам были объединены в единый файл. Затем была выполнена первичная обработка данных по их обезличиванию, т.е. удалению ФИО авторов, рецензентов и пр.. а также составлен список стоп-слов. Файл с исходными данными для ноутбука представлен в формате txt.

ЗАКЛЮЧЕНИЕ

Основные результаты и выводы, полученные при выполнении данной работы. Для достижения поставленной цели были выполнены поставленные задачи данной работы:

1. Рассмотрены алгоритмы тематического моделирования.
2. Проведен сравнительный анализ и выбор библиотек Python, необходимых для работы с документами на естественном языке.
3. Разработан Jupyter Notebook по определению тематик документов на естественном языке с использованием латентного размещения Дирихле.

Латентное размещение Дирихле является базовой вероятностной тематической моделью и из-за большого количества приложений и обобщений является самой распространенной вероятностной тематической моделью. Базовые вероятностные тематические модели позволяют выявлять скрытую тематику документов на основе модели документа как мешка слов.

Полученные в ВКР результаты можно использовать в целях поиска и обнаружения трендов в публикуемых работах, обобщения и актуализации научной работы. Однако использование методов тематического моделирования выдвигает важнейшую проблему, связанную с интерпретацией результатов, что фактически означает семантическое описание выявленных тем. На основе семантического описания тем, возможно продуктивное использование результатов тематического моделирование, например, в социальном прогнозировании и проектировании.