МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ
Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра дифференциальных уравнений и математической экономики

Построение темпоральной тематической модели коллекции русскоязычных текстов

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студента 4 курса 441 группы

Направления 09.03.03 Прикладная информатика

механико-математического факультета

Павлюковича Владислава Михайловича

Научный руководитель
профессор, д.э.н., профессор _____ В.А. Балаш
дата, подпись

Заведующий кафедрой
зав. каф., д.ф.-м.н., профессор _____ С.И. Дудов
дата, подпись

Саратов 2024 год

Введение. Темпоральные тематические модели играют важнейшую роль в анализе текстовых данных с учётом изменения тематик во времени. Они помогают изучению эволюции тем в текстах, прогнозированию и предсказанию тематических трендов, предсказанию событий в новостной аналитике, а также помогают анализировать контент в медиа-сфере. К тому же, эти модели нашли себе применение и в исследовании исторических документов, где позволяют проводить анализировать развитие культуры, политики и тенденций давно минувших дней.

Актуальность данной дипломной работы заключается в том, что объёмы, обрабатываемые темпоральными тематическими моделями, каждый год растут. К тому же, развитие этих моделей помогают людям во множестве сфер деятельности человека: от научно-исторической и лингвистической сфер до рекламной аналитики и сфер медиа-контента. Это мощный инструмент для анализа текстовых данных на разных временных отрезках, помогающие увидеть изменения и тенденции, которые скрыты в динамике информации.

Целью данной бакалаврской работы является изучение темпоральных тематических моделей, их применение в ранее перечисленных областях, рассмотрение этих моделей с математической точки зрения и реализация этих моделей в программном коде с помощью коллекции русскоязычных текстов. Для достижения поставленной цели необходимо:

- Изучить существующие темпоральные тематические модели;

- Создать программу для анализа текстов, взяв за основу одну из темпоральных тематических моделей;

- Проанализировать полученные данные и эффективность программного кода.

Объект исследования - темпоральные тематические модели, использующиеся для анализов текстов. Используются как разработанные под это модели, так и существенно модернизированные. Предмет исследования - коллекция русскоязычных текстов.

Бакалаврская работа состоит из введения, трёх основных разделов, заключения, списка литературы и приложения. В первом разделе дается краткий обзор основных известных темпоральных тематических моделей, которые уже используются в нишевых задачах, а также описываются основные понятия, использующиеся в тематическом моделировании. Во втором разделе подробно описывается каждая модель, её принцип работы, основные математические формулы, а также достоинства и недостатки. В третьем разделе описывается работа написанной программы и построение графиков кластеров на основе DTM-модели. Тестирование модели проводилось на коллекции русскоязычных текстов от 1983 года до 2019 года.

Основное содержание работы. В разделе 1 описываются наиболее известные существующие темпоральные тематические модели. Благодаря ним известна динамика тем в различных категориях статей.

Для начала рассмотрим полностью оригинальную модель, которая специально разрабатывалась под данную цель - Dynamic Topic Model. Dynamic Topic Model - это модель, позволяющая моделировать изменение тематик в текстах с течением времени. Она учитывает распределение тем в документах на различных временных срезах, что помогает анализировать эволюцию тематик в больших коллекциях текстовых документов с временными метками.

Чтобы создать DTM-модель, используем BERTopic. Эта библиотека представляет собой комбинацию методов, в которых используются

преобразователи и класс TF-IDF для создания плотных кластеров, которые легко понять, сохраняя при этом важные слова в описании темы.

Метод тематического моделирования состоит из трёх этапов:

1) Встраивание документов.
2) Кластеризация документов.
3) Документ TF-IDF.

BERTopic позволяет использовать DTM, вычисляя представление темы на каждом временном шаге без необходимости запускать всю модель несколько раз. Чтобы сделать это, нужно подогнать BERTopic так, как если бы в данных не было временного аспекта. Далее используется глобальное представление основных тем. Коллекцию документов, для начала, разделяют на документы по темам. Далее, разделенные документы делятся по временным отрезкам. Для каждой темы и временного интервала вычисляется представление c-TF-IDF. Это приведет к представлению определенной темы на каждом временном шаге без необходимости создавать кластеры из вложений. Предварительно настроенный c-TF-IDF применяется к каждому подмножеству документов, после чего настраиваются параметры c-TF-IDF для каждого временного интервала или усредняя представление с глобальным представлением.

Вероятность появления слова в теме в определенный момент времени:

$$P(w(t,d,n) \mid z(t,d,n)) = \frac{n(t,d,k)^{w(t,d,n)}}{n(t,d,k) + V},$$

где:

- $n(t,d,k)^{w(t,d,n)}$ - количество раз, когда слово w(t,d,n) встречается в теме k в документе d в момент времени t.

- n(t,d,k) - общее количество слов в теме k в документе d в момент времени t.
- V - общее количество уникальных слов в корпусе.

Вероятность эволюции темы из предыдущего момента времени:

$$P(z(t,d,n) \mid z(t\text{-}1,d,n)) = \frac{n(k't-1,kt)^d}{n(kt-1)^d + K},$$

где:

- $n(k't - 1, kt)^d$ - количество переходов от темы k' в момент времени t-1 к теме k в момент времени t в документе d.
- $n(kt - 1)^d$ - общее количество переходов в момент времени t-1 в документе d.
- K - общее количество тем.

Следующая рассматриваемая модель - Author-Topic Model (ATM). Эта модель тематического моделирования, которая расширяет классическую модель LDA (Latent Dirichlet Allocation), чтобы учитывать влияние авторства на содержание текста. Эта модель помогает выявить предпочтения авторов в тематиках и выделять особенности конкретных авторов в контексте тем. Затем эти данные можно использовать в динамическом анализе тематик. К примеру, провести анализ изменения тем в документах за определённым автором, запоминая даты написания документов. Далее, с помощью визуализации можно определить, каким темам автор отдавал предпочтения за определённый период.

Чтобы добиваться хороших результатов, в модели сделали привязку предпочтений к авторам. То есть выявляется связь между авторами и

темами его текстов. Модель стремится выделить эти предпочтения и оценить влияние авторства на тематику текста. Таким образом, в коллекции документов создаётся связь между документами и их авторами. Модель стремится определить, какие темы наиболее характерны для каждого автора и как авторы влияют на содержание текстов.

- Пользователь задает количество тем, которые требуется идентифицировать в текстовой коллекции.
- Для каждого документа, ATM определяет веса тем, с которыми он связан, в соответствии с распределением вероятностей.
- Для каждого документа, ATM определяет временной период, с которыми он связан, в соответствии с указанной в документе датой.
- Для каждой темы, ATM определяет веса слов, которые ей присущи, в соответствии с распределением вероятностей.

$$P(w, d, t) = P(w, T, t) \times P(T, t, d),$$

где:

w - последовательность слов $<w_1, .., w_n>$,

T - тема,

t - временной период,

d - документ.

Следующей моделью описывается TLSA. Она является доработанным вариантом классической модели Латентно-Семантического Анализа, с той разницей, что, в случае TLSA, анализ учитывает эволюцию тем во времени, что позволяет учитывать временной аспект в моделировании. Классическая модель LSA - это метод обработки

информации на естественном языке, анализирующий взаимосвязь между библиотекой документов и терминами, в них встречающимися, и выявляющий характерные факторы, присущие всем документам и терминам. А также выявляет зависимость между темами и авторами.

В основе метода латентно-семантического анализа лежат принципы факторного анализа, в частности, выявление латентных связей изучаемых явлений или объектов. При классификации документов этот метод используется для извлечения контекстно-зависимых значений лексических единиц при помощи статистической обработки больших корпусов текстов.

Алгоритм LSA лучше всего работает с матрицами TF-IDF для сбора слов по темам. LSA также оптимизирует темы, чтобы сохранить разнообразие этих измерений. Число тем, необходимых модели для захвата смысла документов, намного меньше количества слов в словаре векторов TF-IDF. Поэтому LSA часто называется методикой понижения размерности. Она уменьшает необходимое для захвата смысла документов число измерений.

Вероятность появления слова в теме:

$$\boldsymbol{P(w(d,n) \mid z(d,n)) = \frac{n(d,k)^{w(d,n)} + \beta}{n(d,k) + V\beta}},$$

где:

- $n(d,k)^{w(d,n)}$ - количество раз, когда слово w(d,n) встречается в теме k в документе d.
- n(d,k) - общее количество слов в теме k в документе d.

- β - параметр сглаживания.
- V - общее количество уникальных слов в корпусе.

Вероятность автора в теме:

$$P(a(d)|z(d,n)) = \frac{c(d,k)^{a(d)} + \gamma}{c(d,k) + A\gamma},$$

где:

- $c(d,k)^{a(d)}$ - количество раз, когда автор встречается в теме k в документе d.
- c(d,k) - общее количество авторов в теме k в документе d.
- γ - параметр сглаживания.
- A - общее количество авторов в корпусе.

Общая формула для Time-aware Latent Semantic Analysis (TSLA) объединяет данные вероятности и учитывает временной аспект при моделировании тем в текстовых данных. В общем случае, формула TSLA может быть представлена как комбинация вероятностей появления слова в теме и вероятности авторства в теме с учетом временных меток. Таким образом, матрицы классической модели LSA становятся трёхмерными, поскольку появляется новая ось координат - время. В этом случае формулы LSA принимают вид $P_t(w(d,n)|z(d,n))$ и $P_t(a(d)|z(d,n))$ соответственно, где t - временной период.

Четвёртым описываемым методом в работе является Time Series Topic Modeling (TSTM). Это метод, который объединяет принципы тематического моделирования с временными рядами для анализа изменения тем в текстовых данных во времени. В принципах работы Time Series Topic Modeling лежат несколько последовательных этапов:

1. Представление текстов:
   - Исходные тексты разбиваются на документы, где каждый документ содержит текст, относящийся к определенному временному отрезку (например, год, месяц, неделя).
2. Построение базовой тематической модели:
   - Начальная тематическая модель анализирует структуру тем в текстовых данных без учета времени. Это позволяет определить основные темы и ключевые слова в документах.
3. Учет временных факторов:
   - Вводятся временные ряды, которые связывают темы с определенными временными точками. Это позволяет отслеживать эволюцию тем во времени.
4. Моделирование временной динамики тем:
   - TSTM анализирует, как темы меняются и эволюционируют во времени, учитывая временные зависимости и динамику изменения тематик в текстах.
5. Прогнозирование будущих тем:
   - На основе модели временной динамики тем возможно прогнозировать будущие изменения в тематиках текстовых данных и идентифицировать вероятные тренды.
6. Интерпретация результатов:
   - Результаты TSTM могут быть интерпретированы для понимания эволюции тем в текстах, выявления ключевых изменений в тематиках и прогнозирования будущих тематических трендов.

Последней описываемой моделью является Structured Topic Models (STM). STM - это метод тематического моделирования, который позволяет учитывать структуру документов и взаимосвязи между темами. Модель ведёт учёт структуры документов. STM учитывает не только

содержание текстов, но и их структуру, такую как заголовки, подзаголовки, списки, абзацы и другие элементы, что позволяет учитывать контекст и организацию информации в документах.

После этого модель назначает каждому документу набор тем с учетом его содержания и структуры. Темы могут быть связаны с различными частями документа в зависимости от их распределения в тексте. Затем метод ведёт определение взаимосвязей тем и моделирует взаимосвязи между темами в документах, позволяя выявлять связанные и семантически близкие темы, которые встречаются вместе в текстовых данных. К тому же, модель учитывает контекст структуры документов при анализе тем, что может помочь в понимании особенностей тематики в зависимости от их места в тексте.

Результаты STM позволяют интерпретировать темы не только по содержанию, но и по их распределению в структуре документов. Это помогает лучше понять организацию информации и тематические связи в текстах.

В разделе 3 описывает применение написанной программы, разработанной на основе DTM-модели. Программа написана на языке программирования Python 3.9 с использованием библиотеки BERTopic. Для тестирования работы программы использовалась коллекция русскоязычных документов от 1983 по 2019 гг. Для простоты вычислений данный набор был сокращён до 1999 года.

Для проведения тестирования программой выполняются следующие действия:

1) Чтение всех файлов с документами. В одном файле содержатся документы за один год.

2) Проведение токенизации и лемматизации текстов для создания категорий тем.
3) Кластеризация слов по отдельным категориям.
4) Фильтрация кластеров для сравнения определенных категорий.
5) Выявление наиболее встречаемых слов внутри кластеров.
6) Создание облака слов за определённый год.
7) Повторение этих процедур для остальных документов.
8) Визуализация графика динамики тем на всём временном периоде и вывод наиболее популярных слов.

Заключение. В данной работе была рассмотрена большая часть темпоральных тематических моделей. На основе наиболее эффективной модели была разработана и протестирована программа. На полученных данных от программы был проведен анализ динамики тем за выбранный период. Программа для вычисления тем, популярных слов и визуализации была написана на языке программирования Python 3.9, библиотека BERTopic и описана в разделе приложения. Использование темпоральной тематической модели на практике очень простое. Все что нужно пользователю– изучить набор данных и смотреть результаты на полученных графиках. Такое графическое представление обеспечивает наглядность и удобство использования.