КРОСС-МОДЕЛЬНОЕ СРАВНЕНИЕ NNMF, LDA И LSI НА РАЗНОМ ОБЪЕМЕ ТЕКСТОВЫХ ДАННЫХ

Семченко О.П.[1], Кайда А.Ю.[2]
[1] ТПУ, ИШИТР, гр. 8К13, e-mail: semchenko@tpu.ru
[2] ТПУ, ОИТ, ст. преподаватель, e-mail: ayk13@tpu.ru

Аннотация

Исследование сравнивает эффективность трех моделей тематического моделирования: LDA, LSA и NNMF, на корпусе из 1000 документов, представляющих несколько тем с разной лексикой. Также использовались такие наборы тем, чтобы были такие их сочетания, в которых есть схожая лексика, так как это должно было усложнить работу модели. Результаты показали превосходство LDA при различных объемах данных.

Ключевые слова: тематическое моделирование, LDA, LSI, NNMF

Введение

Тематическое моделирование — это метод анализа текстовых данных, направленный на раскрытие скрытых тематических паттернов в собрании документов. Он автоматически определяет темы, заложенные в текстах, и выявляет слова, связанные с каждой из этих тем. Это применяется для адекватного анализа и понимания обширной текстовой информации.

Тематическое моделирование относится к направлению NLP (*Natural Language Processing*) в ML (*Machine Learning*). Natural Language Processing - это область искусственного интеллекта, которая занимается разработкой методов и алгоритмов для обработки и анализа естественного языка, такого как тексты и речь [1].

Целью исследования является определение эффективности и сравнение моделей тематического моделирования в зависимости от объема текстовой информации. Основная задача заключается в выявлении преимуществ и ограничений каждой модели при работе с различными объемами текста.

Новизна данного исследования проявляется в уникальном подходе к анализу трех основных моделей тематического моделирования на различных объемах текстовых данных. Это позволяет понять, как каждая модель работает в различных сценариях, что имеет важное значение для практического применения в области анализа текстов.

Подготовка датасета (корпуса)

Так как цель исследования – сравнить результат работы моделей, то следует на вход подавать им одинаковые данные. Вручную был собран корпус из 1000 документов приблизительно одинакового размера (числа слов), состоящий из статей различных новостных источников. Были определены следующие тематики для текстов:

1. «Музыка» – набор статей, посвященных мировой музыкальной индустрии. Сюда входят статьи, посвященные выходам нового музыкального материала, концертов и т. д.
2. «Мода» – набор статей, посвященных мировой индустрии моды. Материал в статьях был о модных показах, новых тенденциях в моде, стилю определенных людей и т. д.
3. «Кулинария» – набор статей, посвященных рецептам разных блюд и описаниям традиционных блюд разных стран.
4. «Спорт» – набор статей, посвященных соревнованиям (чемпионатам, кубкам и т. д.) по футболу, баскетболу, хоккею и другим видам спорта.
5. «Киберспорт» – набор статей, посвященных обзорам киберспортивных мероприятий.

Выбор таких тем был не случайным: у тем «Музыка», «Мода», «Кулинария» и «Спорт» разная лексика, что должно отразиться на чистоте результата, в то время как у тем «Спорт» и «Киберспорт» схожая лексика, и могли возникнуть трудности при определении тем в корпусе.

Выбор моделей для обработки текстовых данных

Тематическое моделирование осуществляется посредством применения следующих методов для анализа текстовых данных:

1. Латентное размещение Дирихле (LDA) – это статистическая модель, которая используется для анализа текстовых данных и выявления скрытых тематик в них [105]. В контексте исследования, LDA представляется как вероятностная модель, основанная на предположении, что каждый документ представляет собой смесь нескольких тем, а каждая тема представляет собой смесь нескольких слов.

2. Латентно-семантический анализ (LSA) – это метод анализа текста, который основан на сингулярном разложении матрицы терминов-документов. LSA рассматривается как техника для уменьшения размерности пространства признаков и выявления скрытых семантических структур в текстах. Путем анализа матрицы терминов-документов LSA позволяет находить семантические аналогии и связи между словами и документами [2].

3. Неотрицательное матричное разложение (NNMF) – это метод разложения неотрицательной матрицы на две или более факторных матриц. NNMF рассматривается как метод, который может быть применен для анализа неотрицательных данных, таких как тексты или изображения. Путем разложения матрицы терминов-документов на неотрицательные факторные матрицы NNMF позволяет выявлять скрытые семантические структуры в текстах и извлекать семантические признаки из данных [3].

Для исследования возможностей анализа текстовых данных были выбраны модели из библиотеки Gensim. Gensim обеспечивает простоту использования, эффективность работы с большими объемами данных и масштабируемость алгоритмов. Использование моделей из Gensim позволяет проводить комплексный анализ текстовых корпусов, выявлять скрытые тематики и исследовать семантические связи между словами и документами [1].

Описание процесса исследования

После подготовки корпуса и выбора модели, был этап проведения эксперимента. Перед использованием моделей, текстовые данные были токенизированы и нормализованы с помощью модуля Pymorphy3, позволяющего приводить в нормальную форму слова русского языка. Также был сделан словарь – список уникальных слов из всего корпуса, так как именно его следует подавать на вход моделям из библиотеки Gensim.

Эксперимент проводился в 3 этапа:

1. Применение для корпуса с 5 темами и различными размерами корпуса.

Сначала на вход моделям подавались маленькие размеры документов (от 10 до 50 по каждой теме), а затем подавались корпусы размерами 150 и 200 документов по каждой теме.

Было зафиксировано, что результаты становились лучше при увеличении корпуса. На маленьких размерах все модели выводили некорректные результаты. Лучше всего на всех итерациях отрабатывало размещение Дирихле, результаты работы которого приведены ниже.

```
Тема 1 : 0.012 * бренд + 0.012 * коллекция + 0.008 * дизайнер + 0.006 * новый + 0.006 * показ + 0.006 * мода
Тема 2 : 0.022 * кухня + 0.019 * блюдо + 0.009 * рецепт + 0.008 * мясо + 0.007 * приготовление + 0.007 * соус
Тема 3 : 0.012 * команда + 0.009 * игрок + 0.007 * матч + 0.007 * первый + 0.006 * сезон + 0.006 * игра
Тема 4 : 0.019 * альбом + 0.016 * трек + 0.009 * новый + 0.008 * песня + 0.005 * музыка + 0.005 * группа
Тема 5 : 0.021 * команда + 0.017 * team + 0.010 * турнир + 0.009 * игра + 0.008 * esports + 0.008 * место
```

Рис. 1. Результаты работы LDA на корпусе из 250 документов

```
Тема 1 : 0.011 * масло + 0.009 * блюдо + 0.008 * вода + 0.007 * добавить + 0.007 * вкус + 0.007 * сахар
Тема 2 : 0.023 * команда + 0.013 * team + 0.011 * турнир + 0.010 * игра + 0.009 * игрок + 0.006 * место
Тема 3 : 0.011 * бренд + 0.009 * коллекция + 0.007 * дизайнер + 0.006 * мода + 0.006 * вещь + 0.006 * кухня
Тема 4 : 0.015 * матч + 0.009 * команда + 0.008 * лига + 0.007 * чемпион + 0.006 * первый + 0.006 * игра
Тема 5 : 0.011 * альбом + 0.008 * новый + 0.008 * песня + 0.007 * трек + 0.005 * музыка + 0.004 * группа
```

Рис. 2. Результаты работы LDA на корпусе из 1000 документов

На каждом из рисунков выше приведено вероятностное распределение токенов по всем темам: числа – веса токенов в теме, а слова – токены. Оба результата показали почти идеальное распределение слов по темам – все темы были определены правильно.

Также было выявлена получена диаграмма соотношения словарей тем (рис. 3), с помощью которой была обнаружена погрешность в работе модели LDA: на диаграмме она определяла темы «Мода» и «Музыка» как схожие по лексике, хотя лексика не является таковой.

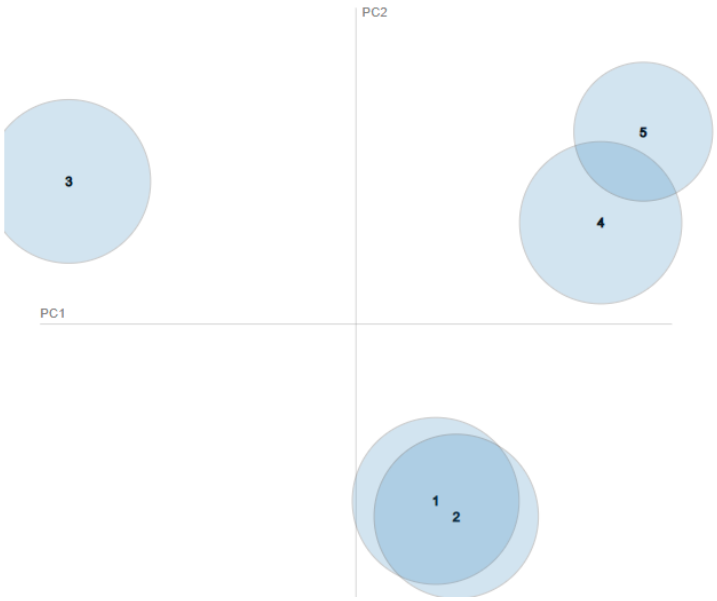

Рис. 3. Диаграмма соотношения словарей тем (1 – «Мода», 2 – «Музыка», 3 – «Кулинария», 4 – «Спорт», 5 – «Киберспорт»)

2. Применение для корпуса с 2 темами со схожей лексикой.

Размеры корпуса, исследуемые на этом этапе, были такими же, как и в предыдущем: подавались сначала от 10 до 50 документов по каждой теме, а затем давались по 200 документов на тему.

Была выявлена такая же тенденция, как и в предыдущем этапе – чем больше корпус, тем больше результаты. На маленьких размерах лучше отрабатывала LSA, но на больших лучшие результаты были получены от модели LDA.

```
Тема 0 : 0.467 * команда + 0.369 * team + 0.202 * турнир + 0.173 * игра + 0.166 * игрок + 0.159 * esports
Тема 1 : 0.424 * team + -0.343 * тренер + -0.317 * команда + -0.296 * игрок + -0.184 * cs + -0.163 * go
```

Рис. 4. Результаты работы LSA на маленьких размерах корпуса

```
Тема 1 : 0.013 * матч + 0.009 * команда + 0.007 * лига + 0.007 * первый + 0.006 * игра + 0.006 * чемпион
Тема 2 : 0.025 * команда + 0.015 * team + 0.011 * турнир + 0.010 * игра + 0.009 * игрок + 0.007 * international
```

Рис. 5. Результаты работы LDA на большом размере корпуса

На каждом из рисунков выше приведено вероятностное распределение токенов по всем темам: числа – веса токенов в теме, а слова – токены. В работе LSI была погрешность – в первой теме, выявленной моделью, присутствуют токены «esports» и «team», которые относятся только ко второй теме, в том время как модель LDA отработала отлично на большом корпусе – темы были выявлены верно.

3. Применение для корпуса с 3 темами с разной лексикой

На данном этапе во всех экспериментах лучше всего отрабатывали модели NNMF и LDA, в то время как LSA выводило некорректные результаты.

```
Тема 1: бренд, коллекция, дизайнер, платье, вещий, мода, показ, байер, работа, рид
Тема 2: альбом, музыка, музыкант, свифт, the, музыкальный, стать, выпустить, новый, тейлор
Тема 3: масло, тесто, пирог, творог, минута, рецепт, добавить, приготовление, соль, сливка
```

Рис. 6. Результат работы NNMF на маленьком размере корпуса

```
Тема 1: бренд, коллекция, дизайнер, мода, вещь, показ, новый, платье, модный, стать
Тема 2: масло, сахар, вода, добавить, блюдо, соль, нарезать, минута, вкус, ингредиент
Тема 3: альбом, песня, трек, новый, музыка, группа, певица, музыкант, выпустить, клип
```

Рис. 7. Результат работы NNMF на большом количестве документов

На каждом из рисунков выше приведено вероятностное распределение токенов по всем темам: числа – веса токенов в теме, а слова – токены. В работе NNMF не было погрешностей и модель правильно выделила тематические словари, в отличие от остальных моделей.

Заключение

В заключении нашего эксперимента, проведенного для сравнения работы моделей LDA, LSA и NNMF на различных объемах текстового корпуса, мы обнаружили, что модель размещения Дирихле (LDA) проявила наилучшую эффективность по сравнению с другими моделями.

Несмотря на то, что модель LSA также показала хорошие результаты, она оказалась немного менее точной по сравнению с LDA. В то же время, неотрицательное матричное разложение (NNMF) проявило себя весьма эффективно только при работе с тремя темами, имеющими разную лексику. Однако при анализе двух тем с схожей лексикой и пяти тем с разной лексикой NNMF демонстрировало результаты, которые можно считать удовлетворительными, хотя и не столь выдающимися, как в случае с LDA и LSA.

Список использованных источников

1. Бенгфорт Б., Билбро Р., Охеда Т. Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка. – СПб.: Питер, 2019. – 368 с.: ил. – (Серия «Бестселлеры O’Reilly»)

2. Траск Э. Грокаем глубокое обучение. – СПб.: Питер, 2024. – 352 с.

3. Хобсон Л., Ханнес Х., Коул Х. Обработка естественного языка в действии. – СПб.: Питер, 2020. – 576 с.

4. Основы Natural Language Processing для текста. [Электронный ресурс]. – URL: https://habr.com/ru/company/Voximplant/blog/446738/ (дата обращения: 18.03.2024).