Визуализация и кластеризация ключевых слов в области Computer Science

Национальный Исследовательский Университет «МЭИ»

Шарифуллин М.Р

[SharifullinMR@mpei.ru](mailto:SharifullinMR@mpei.ru)

В данном исследовании рассматривается применение методов для извлечения КС, а также последующая иx визуализации и кластеризации извлечённых ключевых слов (КС) в области компьютерных наук (Computer Science). Процесс извлечения ключевых слов играет ключевую роль в анализе текстов, таких как научные статьи, где важно выделить основные идеи и темы. Визуализация ключевых слов позволяет не только оценить их значимость, но и выявить закономерности, что является важным этапом для понимания и интерпретации больших объёмов текстовой информации. Кластеризация ключевых слов, в свою очередь, позволяет группировать схожие термины, облегчая понимание тематических областей и их взаимосвязей. Данные для исследования были взяты из набора под названием ArXiv Dataset (24,000+ Papers)[[1]](#footnote-1), который включает в себя статьи, относящиеся к таким областям, как Machine Learning (ML), Computational Linguistics (CL), Named Entity Recognition (NER), Artificial Intelligence (AI) и Computer Vision (CV). Эти данные охватывают период с 1992 года по февраль 2018 года. Данный текстовый набор предоставляет уникальную возможность для анализа тенденций и развития ключевых тем в указанных научных областях, а также служит источником для задач, связанных с обработкой естественного языка.

Ключевые слова: интеллектуальный анализ данных, визуализация, кластеризация, машинное обучение, визуализация, ключевые слова, компьютерные науки.

This study examines the application of visualization and clustering techniques for extracted keywords (KWs) in Computer Science. The process of keyword extraction plays a key role in analyzing texts, such as scientific articles, where it is important to highlight the main ideas and themes. Visualization of keywords not only allows us to assess their relevance but also to identify hidden patterns, which is an important step for understanding and interpreting large amounts of textual information. Keyword clustering, in turn, allows similar terms to be grouped together, making it easier to understand topic areas and their relationships. The data for the study was taken from a dataset called ArXiv Dataset (24,000+ Papers) , which includes articles related to the fields of Machine Learning (ML), Computational Linguistics (CL), Named Entity Recognition (NER), Artificial Intelligence (AI) and Computer Vision (CV). The data covers the period from 1992 to February 2018. This text set provides a unique opportunity to analyze trends and developments of key topics in the mentioned scientific fields and also serves as a source for tasks related to natural language processing.

Keywords: data mining, visualization, clustering, machine learning, visualization, keywords, computer science.

В данной работе исследуется два метода по извлечению ключевыx слов с разными подxодами работы над подаваемым текстом. BERTopic – извлекающий ключевые слова на уровне кластеров, сформированных на всей подаваемой выборке исследуемых статей, и KeyBert, формирующий индивидуальный набор КС для каждого документа в отдельности. С каждой статьи для анализа взято её название и краткое содержание.

Модель BERTopic представляет собой современный подход к тематическому моделированию текстов. Основой модели служит языковая модель BERT, которая позволяет получать глубокие семантические представления текста. BERT применяется для создания векторных эмбеддингов текстов, учитывающих их контекстное значение, что значительно улучшает качество тематической классификации и выделения ключевых слов. Модель эффективно применяется в задачах анализа текстов, где важно понять их основное содержание или структуру. Наше использованного данной модели будет заключаться в том, что она группирует подаваемые нами содержания статей в тематически схожие кластеры, и позволяет извлечь ключевые слова для каждого кластера.

Ключевой особенностью BERTopic является то, что он изначально ориентирован на извлечение ключевых слов для тем, которые могут быть общими для нескольких статей исследуемого архива, а не строго для каждой строки, то есть КС извлекаются на основе всей сформированной темы, к которой относится строка, а не индивидуально

Алгоритм BERTopic строится как последовательность отдельных шагов по следующему принципу:

1. Выбор встраиваемой модели для преобразования подаваемого текста в числовые представления, а также извлечение смыслового содержания текста. BERTopic предлагает практически любую общедоступную модель встраивания. Это может быть как Sentence-BERT, RoBERTa, Glove, Word2Vec или FastText, модели фреймворка Spacy.
2. Уменьшение размерности полученных эмбеддингов. Поскольку вложения часто имеют высокую размерность, кластеризация становится сложной. Решение состоит в том, чтобы уменьшить размерность вложений до рабочего размерного пространства. BERTopic предоставляет по умолчанию алгоритм UMAP, однако не ограничивает в выборе и позволяет выбрать PCA, Truncated SVD и так далее.
3. Кластеризация. Кластеризация выявляет группы текстов, которые имеют схожее содержание, автоматически определяя количество кластеров(тем). Для этого по умолчанию используется и рекомендуется алгоритм HDBSCAN, который выделяет также "шумовые" данные, не относя иx ни к какому кластеру. Также BERTopic поддерживает K-Means, который будет уже принудительно помещать "шумовые" данные в темы, и множество других методов кластеризации
4. Модели векторизации. В BERTopic векторизатор отвечает за создание мешка слов или n-грамм из исходных текстов, помогая модели выделить комбинации, которые встречаются часто и могут быть важными для выявления тем. Можно использовать как CountVectorizer, так и TF-IDF.
5. C-TF-IDF. Является скорректированным представлением TF-IDF и учитывает то, что отличает документы в одном кластере от документов в другом кластере. Он проводит взвешивание n-грамм, учитывая частоту слов в контексте конкретных тем. То есть в отличии от шага 4. C-TF-IDF помогает взвесить эти n-граммы с учётом важности слов не для всего текста в целом, а в определенном кластере.
6. Дополнительная настройка. Является необязательной и не используются по умолчанию. Данная настройка позволяет настроить модель под конкретные задачи или использовать более специализированные модели для работы над подаваемым текстом. BERTopic не ограничивает в выборе и предлагает встраивание от моделей типа GPT до быстрого извлечения ключевых слов с моделями типа KeyBERT.

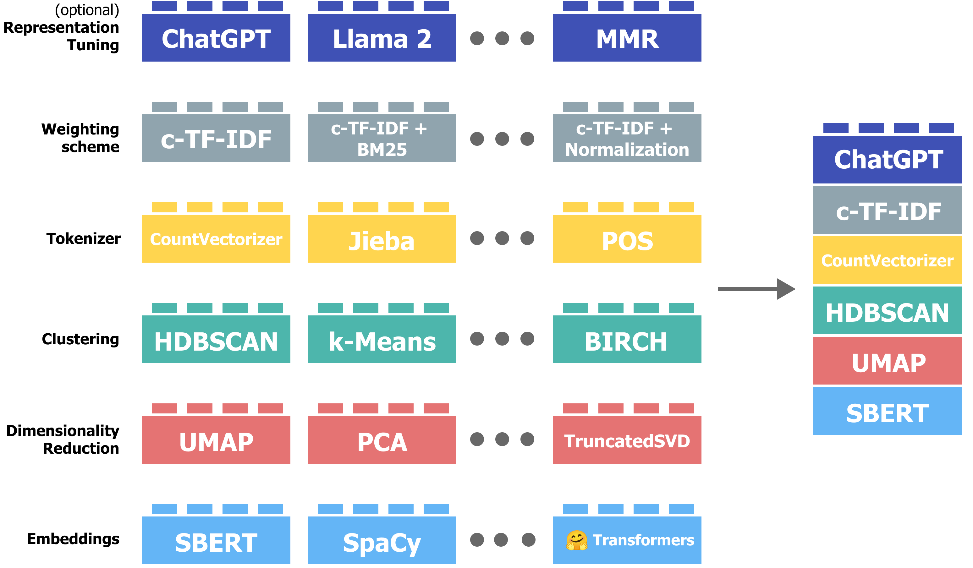


Рис.1. Представление алгоритма BERTopic.

URL: https://newsletter.maartengrootendorst.com/p/bertopic-what-is-so-special-about

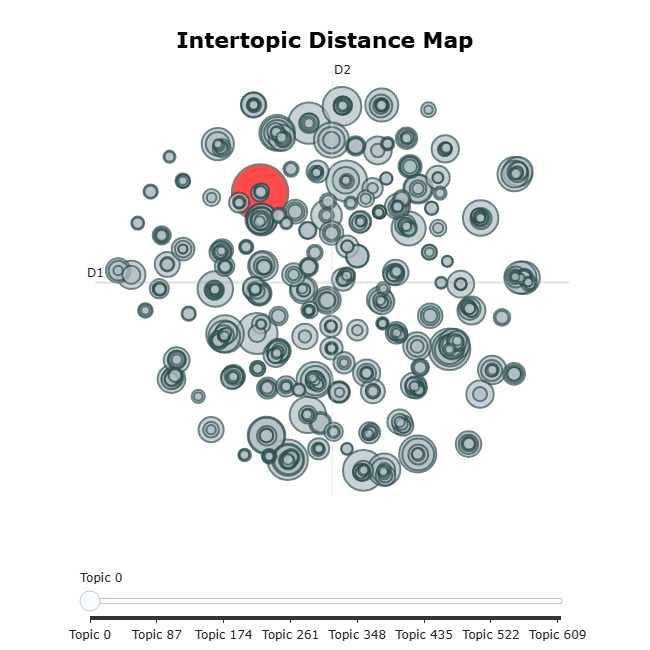


Рис.2. Результат кластеризации на темы

Всего моделью было выделено 612 тем. На графике каждая точка представляет собой тему, а расстояние между точками отражает степень их схожести. Темы, расположенные рядом, имеют схожее семантическое содержание, что может указывать на их общие ключевые слова. Размер кластера также указывает на количество текстов, отнесенных к данной теме, что позволяет определить доминирующие и редкие темы.

Модель KeyBERT представляет собой инструмент для извлечения ключевых слов и фраз из текстовых данных. Основной целью является нахождение тех слов или фраз, которые наиболее точно описывают содержание документа. KeyBERT способен использовать предварительно обученную модель BERT, а также её модификации, что позволяет ему учитывать контекст и значимость слов в предложении. Это позволяет более точно извлекать ключевые слова и фразы, основанные на глубоком понимании текста.

Отличительной особенностью от BERTopic является то, что KeyBERT извлекает ключевые слова, которые строго соответствуют содержанию конкретного текста; то есть в рамках нашего исследования KeyBert способен извлекать ключевые слова из набора данных для каждой отдельной статьи.

С помощью KeyBERT для исследуемой выборки получено

**KeyBERTInspired** будет работать с эмбеддингами, сгенерированными **embedding\_model** (например, с использованием **SentenceTransformer**), чтобы извлечь и предоставить **ключевые слова**, которые наилучшим образом представляют каждую тему. Это помогает понять, какие именно слова или фразы составляют основную концепцию каждой темы.

**representation\_model** (например, KeyBERTInspired):

* Используется для генерации "описания" уже выделенных тем.
* Оно отвечает за выбор ключевых слов, наиболее подходящих для представления конкретной темы.
* **KeyBERTInspired** основывается на эмбеддингах, а не на частотности слов, чтобы сделать представление тем более осмысленным, улучшая выбор слов для их описания.

**Принцип работы BERTTopic**

1. **Представление текста:**
   * Сначала текстовые данные преобразуются в векторные представления с использованием языковой модели BERT или её производных, таких как Sentence Transformers. Эти эмбеддинги отражают контекстное содержание текста.
2. **Кластеризация:**
   * После создания эмбеддингов выполняется кластеризация текстов на основе их векторных представлений. Для этого используется алгоритм **HDBSCAN** (Hierarchical Density-Based Spatial Clustering of Applications with Noise), который автоматически определяет количество кластеров (тем) и выделяет шумовые данные.
3. **Выделение ключевых слов:**
   * Для каждого кластера определяются ключевые слова с помощью анализа частотности слов, моделей TF-IDF или других методов. Эти ключевые слова характеризуют содержание темы и служат основой для генерации ключевых слов документа.
4. **Динамика тем:**
   * Модель может дополнительно анализировать, как распределение тем меняется во времени, что полезно для выявления трендов в текстовых данных.



**Когда используется BERTTopic?**

Модель эффективно применяется в задачах анализа текстов, где важно понять их основное содержание или структуру. Она подходит для обработки больших текстовых коллекций, например:

* анализа отзывов пользователей для выявления основных проблем,
* обработки научных статей для классификации по темам,
* анализа социальных медиа для выявления трендов.

**Преимущества**

BERTTopic выгодно отличается от традиционных методов тематического моделирования (например, LDA) за счёт использования контекстных эмбеддингов. Это позволяет:

* извлекать более точные ключевые слова, отражающие содержание текста,
* находить скрытые темы без необходимости заранее задавать их количество,
* анализировать даже небольшие текстовые наборы, где традиционные методы теряют свою эффективность.

**Заключение**

Применение модели BERTTopic для генерации ключевых слов позволяет выделять темы и ключевые слова, максимально соответствующие контексту. Это делает её ценным инструментом для задач анализа текстов и позволяет извлечь структурированную информацию из больших массивов текстовых данных.

**Sentence-BERT (SBERT)** — это модификация модели BERT, которая была разработана для улучшения качества эмбеддингов предложений, с акцентом на задачи, такие как кластеризация и семантический поиск. В отличие от стандартного BERT, который используется для извлечения эмбеддингов для отдельных слов, **Sentence-BERT** оптимизирован для создания эмбеддингов на уровне предложений. Это делает его более подходящим для задач, где нужно работать с полными предложениями, а не только с отдельными словами.

**Как работает Sentence-BERT**

В основе SBERT лежит концепция **составных эмбеддингов**. Вместо использования обычной BERT, которая генерирует контекстуальные эмбеддинги для каждого токена (слова или части слова), SBERT использует технику **параллельного обучения** через **система с фиксированной длиной эмбеддингов** для предложений. Это достигается с помощью двух основных шагов:

1. **Siamese и Triplet сетки**: SBERT обучает модели с использованием архитектуры, схожей с "сиамскими сетями" (Siamese Networks), где пара предложений обучается на схожести/различиях между ними. Это позволяет эффективно сравнивать и кластеризовать предложения.
2. **Прямое сопоставление предложений**: Для получения эмбеддингов предложений, SBERT генерирует один эмбеддинг для каждого предложения, а затем использует его для выполнения задач, таких как кластеризация или поиск по смыслу.

**Преимущества Sentence-BERT по сравнению с обычным BERT:**

1. **Эффективность**: В отличие от стандартного BERT, который требует подачи каждой пары предложений по отдельности для обработки, SBERT позволяет создавать эмбеддинги для предложений, что значительно ускоряет процессы кластеризации и поиска.
2. **Качество эмбеддингов**: SBERT обеспечивает более устойчивые и информативные эмбеддинги для предложений. Он подходит для задач, связанных с синтаксической и семантической близостью между предложениями, таких как задача нахождения похожих предложений или поиска контекста.
3. **Гибкость**: SBERT может использоваться в таких приложениях, как **поиск по смыслу**, **сравнение предложений**, **кластеризация**, и т.д.

<https://github.com/UKPLab/sentence-transformers> - официальный github sentencebert

**Официальная литература и источники:**

* **Reimers, N., & Gurevych, I. (2019)**. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. arXiv:1908.10084. Это основная работа по SBERT, где авторы подробно описывают архитектуру и принципы работы модели, а также методы обучения. [Ссылка на статью](https://arxiv.org/abs/1908.10084)

https://github.com/MaartenGr/BERTopic?tab=readme-ov-file

 **term**: Это код категории статьи на ArXiv, например:

* cs.AI: Искусственный интеллект в области Computer Science (Artificial Intelligence).
* cs.CL: Компьютерная лингвистика (Computational Linguistics).
* cs.CV: Компьютерное зрение (Computer Vision).
* cs.NE: Сетевые технологии (Networking).
* stat.ML: Машинное обучение (Machine Learning) в статистике.

 **scheme**: URL схемы классификации, которая описывает, как классифицируются термины. В данном случае это схема ArXiv — http://arxiv.org/schemas/atom.

 **label**: Это метка, которая может быть использована для дополнительной информации или описания категории, но в данном случае для всех категорий метка равна None.

.

<https://github.com/MaartenGr/KeyBERT>

Используемая литература:

1.

1. https://www.kaggle.com/datasets/neelshah18/arxivdataset/data [↑](#footnote-ref-1)