Информационно-контекстный поиск в больших текстовых данных со слабой разметкой

Комплексный подход: интерпретируемые представления • эвристическое извлечение контекста • аналитика эмбеддингов

Научный руководитель: Красовицкий А.М., PhD

Цель проекта

Создание комплексного подхода к информационно-контекстному поиску, сочетающего интерпретируемые представления текстов, эвристические механизмы извлечения контекста и средства визуализации для анализа и оптимизации распределений векторных признаков. Итог — программные инструменты, повышающие эффективность, прозрачность и объяснимость результатов поиска.

Объект исследования

Методы, алгоритмы и вычислительные модели машинного обучения для извлечения интерпретируемых признаков, построения эвристических алгоритмов и разработки средств визуализации. Особое внимание — свойствам векторных представлений текстов и возможностям их смысловой интерпретации.

Сравнение с аналогами

Проект используя передовые алгоритмы ML и методы NLP. Неотрицательная матричная факторизация (NMF), кластеризация (k-means, иерархическая), анализ внимания в трансформерах и графовые представления данных. Эвристические алгоритмы на основе LLM (BERT, GPT, LLaMA) с промпт-инженерией. Визуализация эмбеддингов методами снижения размерности UMAP, PCA, t-SNE и оптимизации топологических отображений.

Эффективность проекта

Комплексное сочетание интерпретируемости векторных представлений, эвристической адаптации к слабой разметке и визуальной аналитики. Сформированы осмысленные оси смыслового пространства текста и контекстно-ориентированное извлечение фрагментов без точной аннотации.

Область применения

Интеллектуальный анализ текстовых данных, тематический и контекстный поиск в научных, технических и информационных средах, разработка инструментов объяснимого машинного обучения и семантической аналитики текстов.

Публикации

Mathematics (Web of Science; Scopus, 65-й проц., Q2)
Optimizing the Mean Shift Algorithm for Efficient Clustering, 2025, 13, 3408
Computación y Sistemas (Scopus, 14-й проц., Q4)
Entropy–Distance Approach to Evaluating Diversity and Robustness in Organizational Information Retrieval. Vol. 29, No. 4, 2025.
Applied Sciences (Web of Science; Scopus 69-й проц., Q2)
LLM-Enhanced Semantic Text Segmentation. 2025, 15, 10849.
Informatics (Web of Science; Scopus 94-й проц., Q1)
Cross-Lingual Transfer of Named Entity Markup with Large Language Models, Vol. 13, no. 5: 70
Computación y Sistemas (Scopus, 14-й проц., Q4)
Revisiting Arabic Morphology: A Machine-Learning-Based Approach to Stemming and Root Character Permutations with a Publicly Available Open-Source Implementation, Vol. 30, No. 1, 2026
Computación y Sistemas (Scopus, 14-й проц., Q4)
Comparing Sparse and Dense Information Retrieval Methods on a Wikipedia-Derived NLP Dataset, Vol. 30, No. 1, 2026

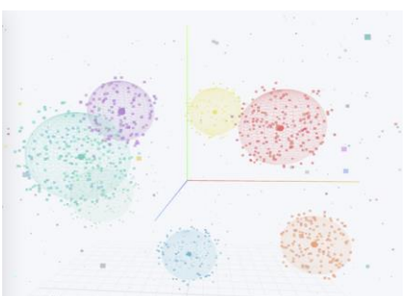
Зарубежные партнеры (вузы и ученые)

нет

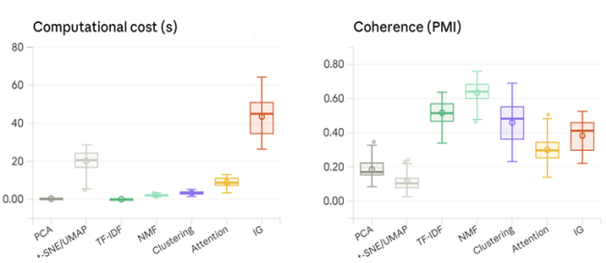
Бизнес партнер

не предусмотрено

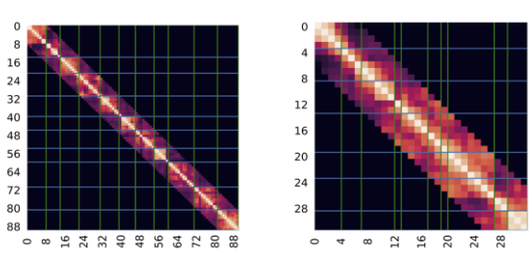
Двухуровневая система извлечения информации на основе LLM и промпт-инженерии



Семантические кластеры текстов в UMAP-based проекции эмбеддингового пространства на топики

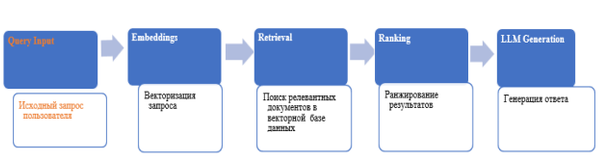


Оценки извлечения интерпретируемых признаков: разложение векторов на смысловые оси



Сравнительный анализ моделей сегментации на эмбеддинговых моделях

$$\text{Prompt}_{i+1} = f(\text{LLM}(\text{Prompt}_i), \text{Context})$$



Многоуровневая обработка запросов в RAG-pipeline