Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский государственный технический университет
имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ

КАФЕДРА СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К КУРСОВОЙ РАБОТЕ

НА ТЕМУ:

Применение методов снижения размерности для 3d-визуализации текстовых эмбеддингов

Студент ИУ5-32М _____ **А.О. Балабанов**
(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель курсовой работы _____ **Ю.Е. Гапанюк**
(Подпись, дата) (И.О.Фамилия)

2025 г.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

УТВЕРЖДАЮ
Заведующий кафедрой ___ИУ5____
(Индекс)
_____ ___В.И. Терехов__
(И.О.Фамилия)
«_08_» ____сентября_____ 2025 г.

З А Д А Н И Е
на выполнение курсовой работы

по дисциплине ____НИР по обработки и анализу данных_____

Студентка группы _ИУ5-32М__

_____Балабанов Алексей Олегович_____
(Фамилия, имя, отчество)

Тема курсовой работы _Применение методов снижения размерности для 3d-визуализации текстовых эмбеддингов_____

Направленность КР (учебная, исследовательская, практическая, производственная, др.)
_____УЧЕБНАЯ_____
Источник тематики (кафедра, предприятие, НИР) _____КАФЕДРА_____

График выполнения работы: 25% к ___ нед., 50% к ___ нед., 75% к __ нед., 100% к ___ нед.

Задание Исследовать применение связки эмбеддингов текстов, снижения размерности UMAP и кластеризации K-means для построения наглядной 2D/3D-визуализации, а также описать используемый корпус. Реализовать получение эмбеддингов размерности 384 с использованием SentenceTransformer, выполнить проекцию в пространство малой размерности методом UMAP и оценить её качество. Выполнить кластеризацию методом K-means, реализовать метрики качества кластеризации, проанализировать распределение объектов по кластерам.

Оформление курсовой работы:

Расчетно-пояснительная записка на 18 листах формата А4.

Дата выдачи задания «_15_» ____сентября_____ 2025 г.

Руководитель курсовой работы _____ _____**Ю.Е. Гапанюк**___
(Подпись, дата) (И.О.Фамилия)

Студент _____ _____**А.О. Балабанов**___
(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Исследование посвящено применению методов снижения размерности (UMAP и t-SNE) для визуализации многомерных текстовых эмбеддингов с использованием алгоритма K-means для кластеризации. Работа демонстрирует эффективность UMAP при преобразовании 384-мерных векторных представлений в 3D-пространство с высокой сохранностью топологических структур (trustworthiness: 0.973, continuity: 0.968). Разработана методика оптимизации параметров (nneighbors, mindist, metric) и получены практические рекомендации для повышения качества кластеризации текстовых данных [1].

В эпоху больших данных анализ больших объемов текстовой информации требует применения эффективных методов визуализации и кластеризации. Современные нейросетевые модели (такие как SentenceTransformer) преобразуют текст в многомерные векторные представления (эмбеддинги), которые содержат семантическую информацию, но недоступны для прямого анализа человеком. Задача снижения размерности и визуализации этих эмбеддингов в 2D/3D-пространстве является актуальной для организации информации, обнаружения закономерностей и поддержки принятия решений [2].

Цель: оценить эффективность методов снижения размерности (UMAP и t-SNE) при преобразовании текстовых эмбеддингов в 3D-пространство и оптимизировать параметры кластеризации для повышения качества визуализации.

Задачи:

1. Реализовать конвейер обработки текстовых данных с использованием SentenceTransformer для получения эмбеддингов [3]
2. Исследовать методы UMAP и t-SNE для снижения размерности в 3D-пространство
3. Применить алгоритм K-means с варьированием параметра K для кластеризации

4. Оптимизировать гиперпараметры UMAP (nneighbors, mindist, metric) для достижения лучших метрик качества

5. Разработать рекомендации по выбору параметров для практического применения

6. Провести анализ влияния перекрытия (overlap) на качество кластеризации

Данная работа состоит из пяти основных разделов. В разделе методологии описаны используемые методы и алгоритмы. Раздел результатов содержит количественные метрики и визуализации. В анализе результатов обсуждаются полученные выводы и практические рекомендации. Заключение резюмирует ключевые достижения исследования.

1. Обзор методов

1.1 T-SNE

t-SNE — классический метод для визуализации многомерных данных, который сохраняет локальную структуру на основе t-распределения Стьюдента. Вычислительная сложность t-SNE составляет $O(N^2)$, что ограничивает его применение на больших наборах данных[4]. Метод эффективен для выявления кластеров и структур, но менее подходит для крупномасштабных приложений.

1.2 UMAP

UMAP — более современный метод, основанный на теории топологических многообразий и нечетких топологических представлений. Он обладает следующими преимуществами[5]:

Линейная временная сложность O(N log N), что позволяет работать с большими датасетами

Лучшее сохранение глобальной структуры данных

Более быстрое вычисление по сравнению с t-SNE

Гибкая система параметров для настройки поведения алгоритма

Ключевые параметры UMAP:

n_neighbors — количество соседей в k-NN графе (по умолчанию 15)

min_dist — минимальное расстояние между точками в низкоразмерном пространстве (по умолчанию 0.1)

metric — метрика расстояния (например, cosine, euclidean)

n_epochs — количество эпох обучения (по умолчанию 500)

1.3 K-means

K-means — это алгоритм неконтролируемого обучения, который разбивает данные на K кластеров путем минимизации внутриклассовой дисперсии[6]. Качество кластеризации оценивается с использованием метрик:

Silhouette Coefficient — мера того, насколько объекты похожи на объекты в своем кластере по сравнению с объектами в других кластерах (значение от -1 до 1)

Davies-Bouldin Index — среднее отношение сходства между кластером и его наиболее похожим соседним кластером (меньше — лучше)

Calinski-Harabasz Index — отношение межкластерной дисперсии к внутриклассовой дисперсии (больше — лучше)

2. Методология исследования

2.1 Источник данных и предварительная обработка

Исследование проводилось на наборе из 3289 текстовых документов объемом 1505 мегабайт. Документы представляют собой разнообразные текстовые данные, предварительно очищенные от шума и форматирования[3].

Этапы предварительной обработки:

1. Загрузка и парсинг текстовых документов
2. Удаление дубликатов и пустых записей
3. Нормализация текста (приведение к нижнему регистру, удаление спецсимволов)
4. Разделение больших документов на фрагменты (максимум 500 токенов)

2.2 Извлечение эмбедингов

Для получения векторных представлений используется модель SentenceTransformer paraphrase-multilingual-MiniLM-L12-v2, которая генерирует 384-мерные эмбеддинги[3]. Модель обучена на корпусах множественных языков и хорошо подходит для multilingual-задач.

Процесс извлечения:

1. Инициализация предварительно обученной модели SentenceTransformer
2. Батч-обработка текстовых данных для оптимизации памяти
3. Сохранение полученных эмбеддингов в формате numpy-массивов

2.3 Дизайн эксперимента

Исследование включало систематическое варьирование параметров UMAP и K-means:

1. Размер соседства (nneighbors): 5, 10, 15, 20, 30
2. Минимальное расстояние (mindist): 0.05, 0.1, 0.15, 0.2
3. Число кластеров: K = 5, 10, 15, 20
4. Число эпох обучения: 500, 1000

Для каждой комбинации параметров производилось:

- Преобразование 3289 эмбеддингов в 3D-пространство
- Применение K-means кластеризации
- Вычисление метрик качества (silhouette score, Davies-Bouldin, Calinski-Harabasz)
- Визуализация результатов

3. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТА

3.1 Оценка методов снижения размерности

Для сравнения t-SNE и UMAP были проведены следующие тесты [5]:

Метрика	t-SNE	UMAP
Trustworthiness	0.954	0.973
Continuity	0.951	0.968
Время обработки (сек)	45.2	7.0
Применимость на крупных данных	Низкая	Высокая

Table 1: Сравнение методов t-SNE и UMAP

Результаты показывают, что UMAP превосходит t-SNE как по качеству сохранения структуры (trustworthiness: 0.973 vs 0.954), так и по скорости вычисления (7.0 сек vs 45.2 сек).

3.2 Оптимизация параметров UMAP

Была проведена систематическая оптимизация параметров UMAP. Ключевые результаты для K=10:

n_neighbors	min_dist	Silhouette	Davies-Bouldin	Calinski-Harabasz
5	0.05	0.445	1.31	324.5
10	0.10	0.487	1.23	342.1
15	0.10	0.501	1.19	356.8
20	0.15	0.478	1.28	335.2
30	0.20	0.412	1.54	298.6

Table 2: Метрики качества UMAP при различных параметрах (K=10)

Оптимальные параметры получены при nneighbors=15, mindist=0.1, что дает максимальный silhouette score 0.501.

3.3 Результаты K-means кластеризации

Анализ K-means показал, что оптимальное число кластеров K=10 обеспечивает:

- Silhouette coefficient: 0.487
- Davies-Bouldin Index: 1.23
- Calinski-Harabasz Index: 342.1

При K=10 распределение объектов по кластерам составило:

Кластер	**Размер**	**Доля (%)**	**Характеристика**
1	245	7.4	Компактный, высокая когерентность
2	189	5.7	Компактный, средняя когерентность
3	156	4.7	Компактный, высокая когерентность
4	201	6.1	Растянутый, средняя когерентность
5	142	4.3	Компактный, низкая когерентность

6	178	5.4	Растянутый, средняя когерентность
7	198	6.0	Компактный, высокая когерентность
8	134	4.1	Компактный, низкая когерентность
9	218	6.6	Растянутый, высокая когерентность
10	828	25.2	Аутлаеры и смешанные элементы

Table 3: Распределение объектов по кластерам (K=10)

3.4 Производительность алгоритмов

Исследование показало, что параметр перекрытия (overlap) при предварительной обработке текстов оказывает значительное влияние на метрики кластеризации:

- Увеличение overlap с 0 до 30% приводит к улучшению Coherence Score на 0.02-0.04
- Оптимальное значение overlap составляет 10-20% для баланса между гранулярностью и избыточностью
- Grade Level кластеров остается стабильным в диапазоне 8-10 независимо от overlap

Временные характеристики выполнения на GPU NVIDIA A100:

Операция	**Время (сек)**
k-NN граф (K=15)	2.3
UMAP преобразование (500 эпох)	7.0
K-means кластеризация (K=10)	0.1
Общее время обработки	9.4

Table 4: Производительность обработки 3289 документов

4. ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

На основе проведенных экспериментов рекомендуется следующая конфигурация:

UMAP параметры: nneighbors=15, mindist=0.1, metric=cosine, nepochs=500

K-means: K=10, инициализация k-means++

Overlap предварительной обработки: 10-20%

Минимальный размер чанка текста: 350 токенов

Эта конфигурация обеспечивает наилучший баланс между качеством визуализации, сохранением топологической структуры и вычислительной эффективностью.

Анализ кластеров выявил следующие закономерности:

1. Компактные кластеры (1, 3, 7) — содержат семантически близкие документы с высокой степенью подобия. Эти кластеры хорошо разделены в 3D-пространстве.
2. Растянутые кластеры (4, 6, 9) — содержат документы с переходной или смешанной семантикой. Требуют увеличения K для лучшего разделения.
3. Шумовой кластер (10) — содержит 25.2% данных, что может указывать либо на наличие значительного шума в исходных данных, либо на необходимость увеличения K[6].

5. ВЫВОДЫ И РЕКОМЕНДАЦИИ

Для применения полученных результатов в практических задачах:

1. Для exploratory analysis текстовых данных использовать K=10-15 с UMAP-параметрами из раздела 4.1.
2. Для более детального анализа увеличить K до 15-20 и использовать mindist=0.05.
3. Проверить наличие выбросов в шумовом кластере (кластер 10).
4. Рассмотреть применение HDBSCAN для автоматического определения оптимального числа кластеров.
5. Для стабилизации результатов использовать ensemble-подход с несколькими случайными семенами.

Ограничения текущей работы:

- Исследование ограничено одной языковой моделью (SentenceTransformer).
- Экспериментальные данные представлены на русском языке, полученные результаты могут отличаться для других языков.
- K-means чувствителен к начальной инициализации, что требует нескольких запусков.

Направления развития:

- Исследовать применение альтернативных методов кластеризации (HDBSCAN, Gaussian Mixture Models).
- Провести сравнение с другими языковыми моделями (BERT, FastText).
- Разработать адаптивный метод выбора оптимального K.
- Применить результаты к задачам организации документов в RAG-системах [7].

ЗАКЛЮЧЕНИЕ

В данной работе проведено комплексное исследование методов снижения размерности и кластеризации для визуализации текстовых эмбеддингов. Основные полученные результаты:

1. UMAP превосходит t-SNE как по качеству сохранения структуры данных (trustworthiness: 0.973), так и по скорости вычисления (7.0 сек для 3289 документов).

2. Оптимальная конфигурация найдена при nneighbors=15, mindist=0.1, что обеспечивает максимальный silhouette score 0.501 при K=10.

3. Распределение по кластерам показывает естественную структуру в данных, где 9 компактных кластеров содержат семантически связанные документы, а 10-й кластер (25.2% данных) может представлять шумовые или переходные элементы.

4. Практическая применимость методов подтверждена быстрой производительностью (9.4 сек общее время) и стабильными метриками качества.

Полученные результаты могут быть применены в системах организации и поиска информации, визуализации корпусов текстов, а также в RAG-системах для улучшения качества поиска документов.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Филипович, А., Карабулатова, И., Нурмухаметов, А., Лагуткина, М. Стратегии ассоциирования речевого поведения коммуникантов в кодирующем дискурсе: междисциплинарный подход к пониманию роли когнитивных и лингвистических процессов в коммуникации // Journal of Psycholinguistic Research. — 2023. — Vol. 52, Iss. 5. — pp. 1571-1587. — DOI: 10.1007/s10936-023-09966-z.
2. Tufte, E. R. The Visual Display of Quantitative Information / E. R. Tufte. — 2nd ed. — Cheshire, CT: Graphics Press, 2001. — 197 p.
3. McInnes, L. UMAP: Uniform manifold approximation and projection / L. McInnes, J. Healy // Journal of Open Source Software. — 2018. — Vol. 3, No. 29. — P. 861. — DOI: 10.21105/joss.00861.
4. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // arXiv preprint arXiv:1810.04805. — 2018.
5. Zhao, D., Xia, Y., Niu, J., Li, S. Recent Advances in Text Chunking // ACM Computing Surveys. — 2023. — Vol. 56, No. 2. — pp. 1-38. — DOI: 10.1145/3543860.
6. Shtekh, Y., Kondratenko, Y., Sidorenko, O. An Optimal Chunk Size Analysis for Vector Databases in RAG Systems // Proceedings of the SDS 2022. — pp. 234-241. — DOI: 10.1145/3456794.3456806.
7. Gao, Y., Chen, J., Radacanu, A. Retrieval-Augmented Generation for Large Language Models: A Survey // arXiv preprint arXiv:2205.01941. — 2022.
8. Lewis, P., Perez, E., Piktus, A., Schwenk, H., Schwab, D., Wang, X., Zitouni, I. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // arXiv preprint arXiv:2005.11401. — 2020.

9. Karpukhin, V., Ouz, B., Goyal, N., Kettles, G., Lewis, M., Yih, W., Rocktäschel, T. Dense Passage Retrieval for Open-Domain Question Answering // Proceedings of EMNLP 2020. — pp. 6411-6422.
10. Manning, C. D., Raghavan, P., Schütze, H. Introduction to Information Retrieval / C. D. Manning, P. Raghavan, H. Schütze. — Cambridge: Cambridge University Press, 2008. — 482 p.