**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ**

**РОССИЙСКОЙ ФЕДЕРАЦИИ**

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ**

**ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**

**«НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

**ФАКУЛЬТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ**

Кафедра Систем информатики

Направление подготовки 09.06.01 – Информатика и вычислительная техника

**ОТЧЕТ**

**Обучающегося** Козубенко Андрея Алексеевича **группы № 22215 курса 4**

(Ф.И.О. полностью)

**Тема задания**: Построение эмбеддингов фрагментов текста

Новосибирск 2025

Оглавление

[Введение 3](#_8xl0k4o7g53y)

[Реализация 4](#_18lqawz87498)

[Листинги](#_18lqawz87498) 5

[Заключение](#_ftfmarjidp06) 6

[Список литературы](#_8i44esq2u94t) 7

# Введение

Целью лабораторной работы является **разработка модуля для построения эмбеддингов фрагментов текста** в рамках существующего Django-проекта.  
Эмбеддинг — это числовое представление текста, позволяющее оценивать смысловую близость между предложениями или документами.

Для достижения цели были поставлены следующие задачи:

* Реализовать функцию **разбиения текста на фрагменты** (get\_chunks);
* Реализовать функцию **генерации эмбеддингов** (get\_embeddings), использующую предобученную модель sentence-transformers/paraphrase-multilingual-mpnet-base-v2;
* Реализовать функцию **сравнения эмбеддингов** по **косинусному сходству** (cos\_compare);

# Реализация

Модуль реализован в составе Django-приложения в файле  
db/api/embedding\_utils.py.

### 1. Основные функции

**1.1. get\_chunks(text, chunk\_size=200)**  
Разбивает текст на последовательные части (чанки) длиной примерно по 200 слов для дальнейшего анализа.

**1.2. get\_embeddings(texts)**  
Использует модель paraphrase-multilingual-mpnet-base-v2 для получения векторных представлений каждого текста или чанка.

**1.3. cos\_compare(emb1, emb2)**  
Вычисляет **косинусное сходство** между двумя эмбеддингами:

Результат принимает значения от −1 до 1 (где 1 — максимальное сходство).

### 2. REST-эндпоинты

Для интеграции функций с Django REST Framework созданы API-методы:

| **Операция** | **Метод** | **URL** | **Описание** |
| --- | --- | --- | --- |
| Разбить текст на чанки | POST | /api/embeddings/chunk/ | Возвращает список фрагментов текста |
| Получить эмбеддинги | POST | /api/embeddings/build/ | Возвращает эмбеддинги для переданных текстов |
| Сравнить эмбеддинги | POST | /api/embeddings/compare/ | Вычисляет косинусное сходство между двумя векторами |

# Листинги

**Листинг 1 – Функции модуля embedding\_utils.py**



**Листинг 2 – Пример REST-эндпоинта**



# Заключение

В ходе выполнения лабораторной работы была реализована функциональность по **генерации и сравнению эмбеддингов фрагментов текста** в рамках Django-проекта.  
Были разработаны и протестированы:

1. Функция разбиения текста на фрагменты (get\_chunks);
2. Механизм получения эмбеддингов с использованием модели Sentence-Transformers;
3. Метрика косинусного сходства (cos\_compare);
4. REST-эндпоинты для интеграции с системой и тестирования в Postman.

Полученный модуль может быть использован для анализа смысловой близости текстов в корпусе — например, при поиске похожих фраз или автоматическом сопоставлении переводов.

# Список литературы

1. Django documentation: <https://docs.djangoproject.com/>
2. Sentence Transformers Documentation — <https://www.sbert.net/>
3. REST framework: <https://www.django-rest-framework.org/>