Отчёт о реализации системы векторизации документов

# 1. Введение

Данный отчёт описывает реализацию системы для создания векторов документов с использованием двух методов обработки текста: выделения словоформ и выделения начальных форм слов. Система разработана для обработки размеченных экспертом наборов данных.

# 2. Реализованные алгоритмы

## 2.1. Выделение словоформ

Алгоритм выделения словоформ включает следующие этапы:

* Токенизация текста с использованием регулярных выражений
* Приведение всех слов к нижнему регистру
* Фильтрация стоп-слов (общеупотребительных слов)
* Удаление слов длиной менее 3 символов
* Сохранение исходных словоформ для построения векторов

## 2.2. Выделение начальных форм слов

Алгоритм лемматизации реализует простые морфологические правила для русского языка:

* Обработка существительных: удаление падежных окончаний
* Обработка прилагательных: приведение к именительному падежу
* Обработка глаголов: удаление личных окончаний
* Обработка окончаний множественного числа
* Сохранение коротких слов (менее 4 символов) без изменений

# 3. Пример обработки DOCX файла

Обработан файл: спортивная\_медицина\_01.docx

Размер текста: 3562 символов

Количество словоформ: 319

Количество начальных форм: 319

Уникальных словоформ: 261

Уникальных лемм: 258

# 4. Формат выходных данных

Система генерирует Excel файл со следующими листами:

* Словоформы - векторы документов на основе словоформ
* Начальные\_формы - векторы на основе лемм
* Статистика - общая статистика по набору данных