

Выпускная квалификационная работа бакалавра



Метод определения признаков авторского стиля для текстов на русском языке

Студент:

Кондрашова Ольга Павловна

Группа:

ИУ7-85Б

Научный руководитель:

доцент ИУ-7, кандидат физико-математических наук
Ковтущенко Александр Петрович

Консультант:

старший преподаватель ИУ-7 Волкова Лилия Леонидовна

Цель и задачи

Цель работы – разработка метода определения признаков авторского стиля для текстов на русском языке

Задачи:

- Проанализировать предметную область и существующие методы
- Сформулировать основные положения разрабатываемого метода определения признаков авторского стиля
- Описать использующий его метод классификации
- Разработать ПО, реализующее оба метода
- Провести оценку точности классификации текстов по предложенным признакам

Авторский стиль

- К особенностям авторского стиля можно отнести стилистические приемы, грамматические конструкции, способы построения фраз и абзацев
- Стилеметрия – раздел лингвистики, занимающийся измерением стилевых характеристик
- Авторский инвариант – набор стилеметрических характеристик, присущих данному автору

Существующие методы анализа текста

Метод полного синтаксического анализа – устанавливает связи между структурными единицами текста, требует вмешательства эксперта

- Дерево составляющих
- Дерево зависимостей

Метод энтропийной классификации – дает разные результаты на текстах разной длины с разными алгоритмами сжатия

- Характеристика сжатия

Метод анализа длин слов – различает художественный, научный, публицистический, разговорный стили

- Частотное распределение длин слов

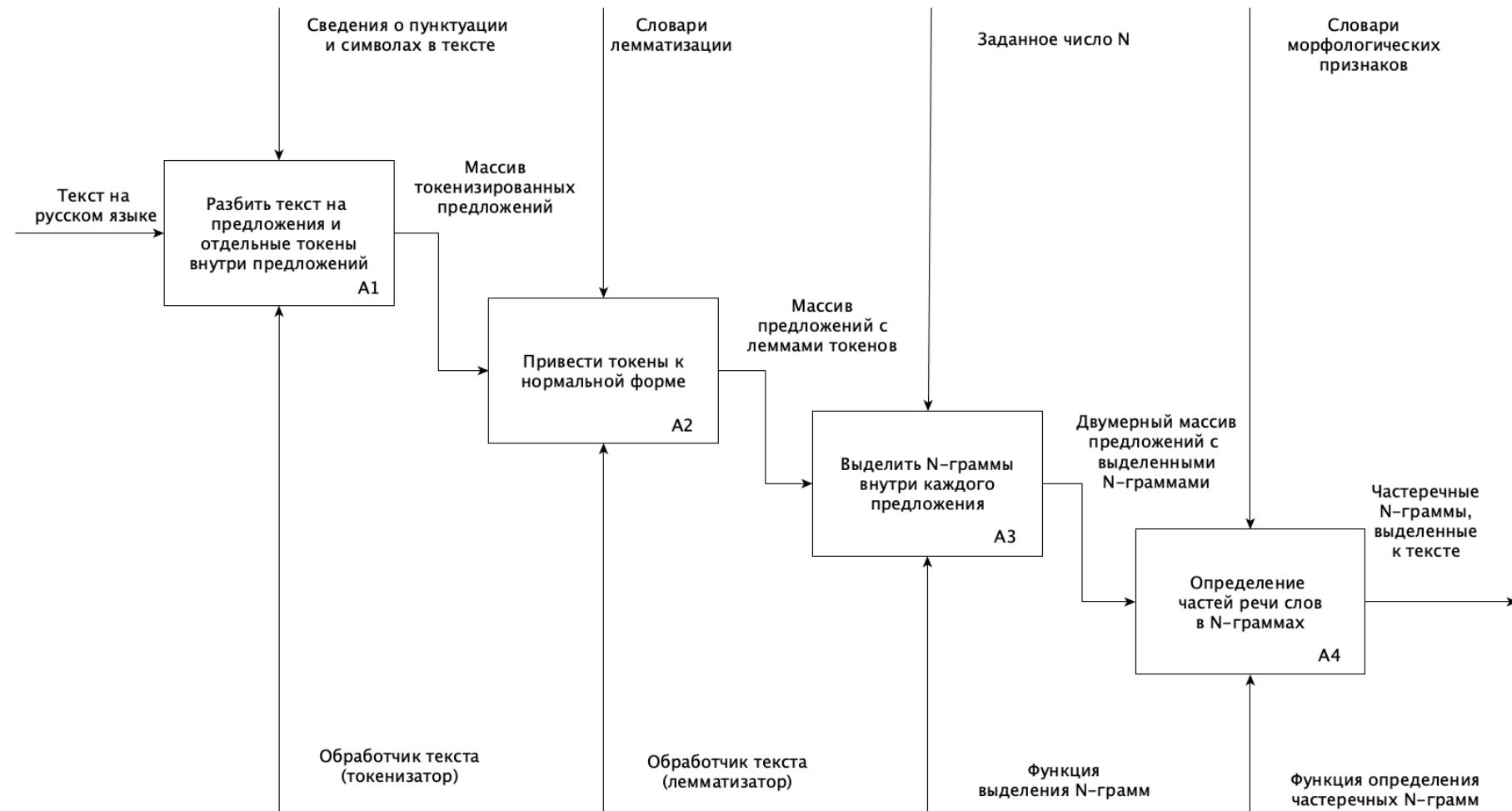
Рекуррентные нейронные сети – работает с последовательностями разной длины в зависимости от функций активации

- Значение на выходном узле

Метод выделения N-грамм – позволяет учитывать порядок слов

- Частотное распределение N-грамм

Функциональная схема метода выделения признаков: образования частеречных N-грамм



Этапы обработки текста

- Токенизация – разбиение непрерывной строки на отдельные токены
- Лемматизация – приведение каждого слова к его начальной форме, основанное на использовании словарей
 - используется библиотека pymorphy2 на основе словарей OpenCorpora

Частеречные N-граммы

N-грамма – N подряд идущих слов, выбранные из предложения

Частеречная N-грамма – кортеж из частей речи слов, вошедших в N-грамму

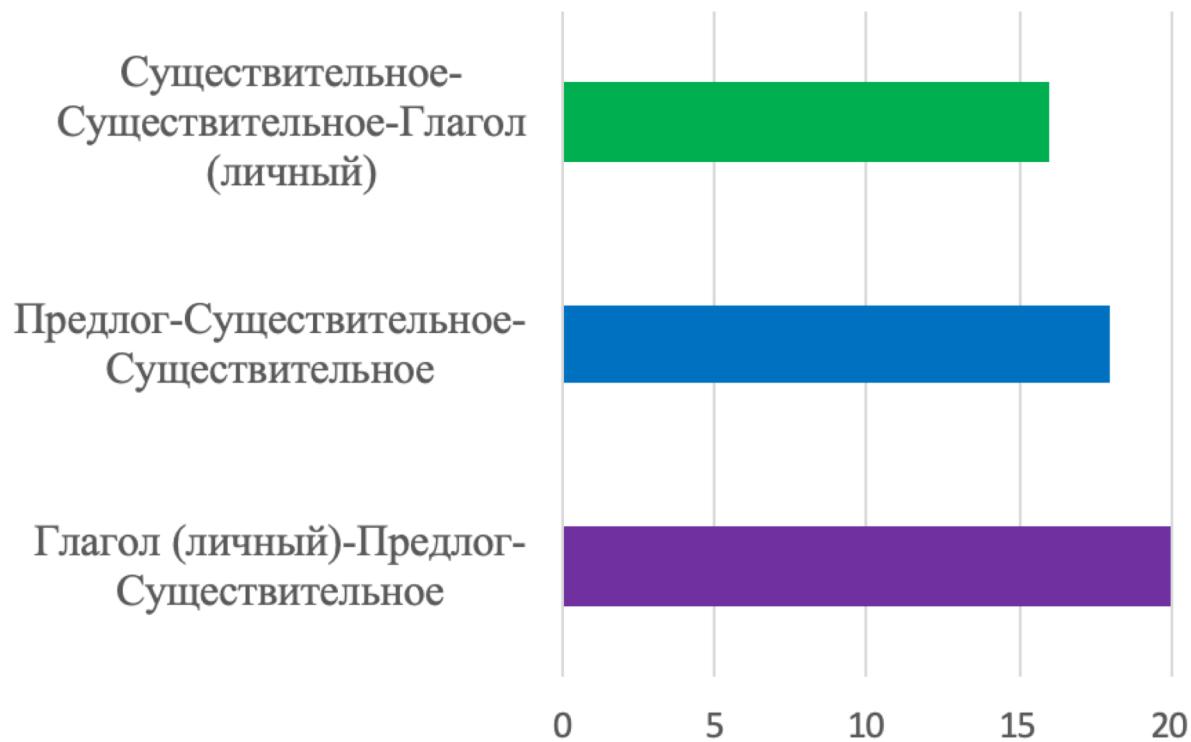
Для предложения «Последовательность подряд идущих слов»

Лемматизация: «Последовательность подряд идущий слово»

	N-граммы из предложения	Частеречные N-граммы
Биграммы (N=2)	Последовательность подряд; подряд идущий; идущий слово.	Существительное – Наречие; Наречие – Причастие; Причастие – Существительное.
Триграммы (N=3)	Последовательность подряд идущий; подряд идущий слово.	Существительное – Наречие – Причастие Наречие – Причастие – Существительное.

Авторский инвариант Чехова

Наиболее часто встречающиеся частеречные триграммы



«Заказчики Луки Александрыча жили ужасно далеко, так что, прежде чем дойти до каждого из них, столяр должен был по несколько раз заходить в трактир и подкрепляться. Каштанка помнила, что по дороге она вела себя крайне неприлично. От радости, что ее взяли гулять, она прыгала, бросалась с лаем на вагоны конножелезки, забегала во дворы и гонялась за собаками».

А.П. Чехов
«Каштанка»

Извлечение численных значений признаков из текста

Документ представляется в виде вектора признаков, каждая N-грамма получает численную оценку по частотной мере TF-IDF

TF-IDF – статистическая мера, используемая для оценки важности N-граммы

$$TF - IDF(t, D) = TF(t, d) \times IDF(t_i, D)$$

где $|D|$ – количество документов;
 $|D_i \in D|$ – число документов, где t_i встретилось хотя бы один раз.

$$\text{TF} — \text{частотность N-граммы: } TF(t, d) = \frac{n_t}{\sum_k n_k}$$

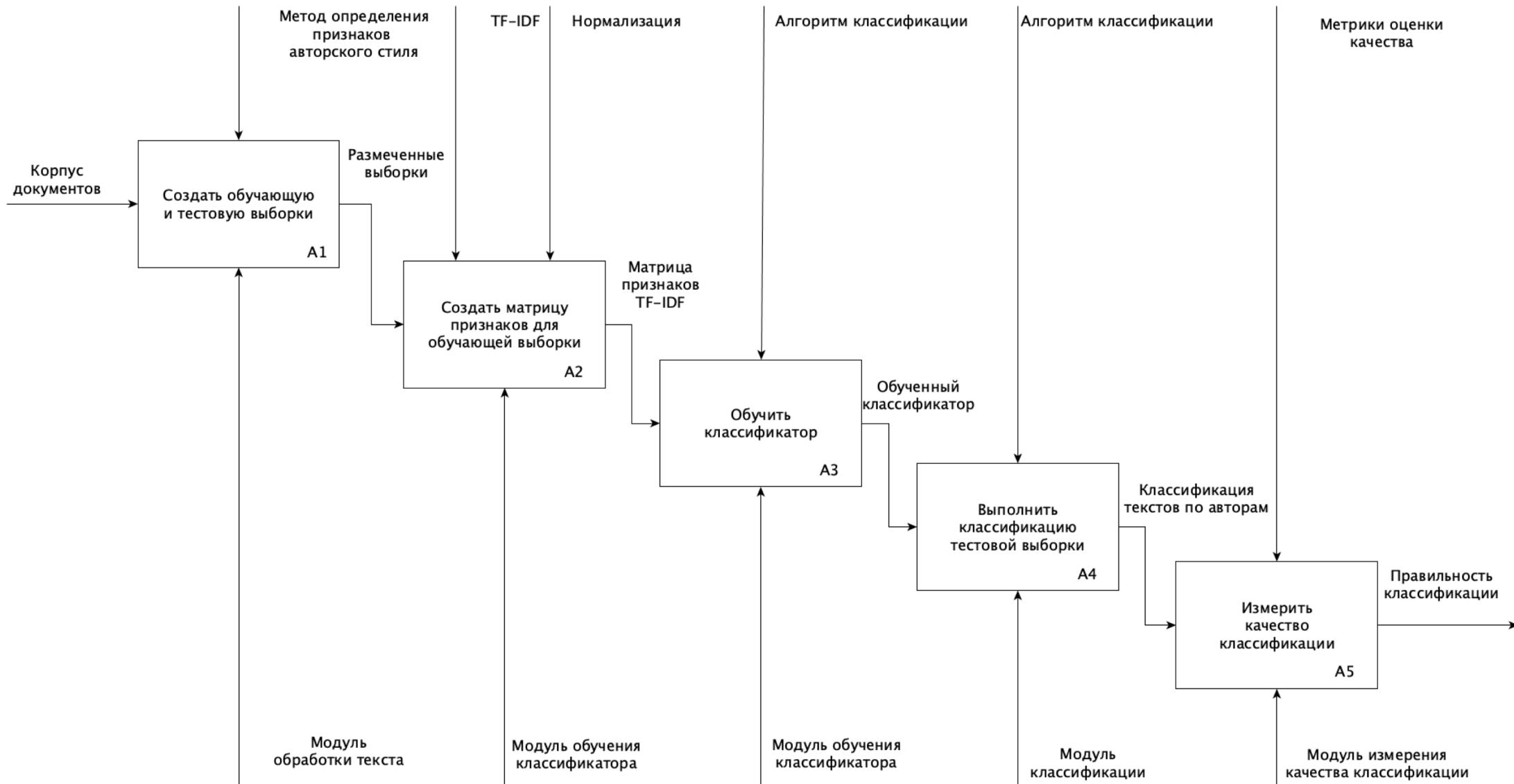
где n_t – количество вхождений N-граммы t в документ;
 $\sum_k n_k$ – общее количество слов в документе.

IDF – инверсия частоты, с которой некоторая N-грамма встречается в документах коллекции

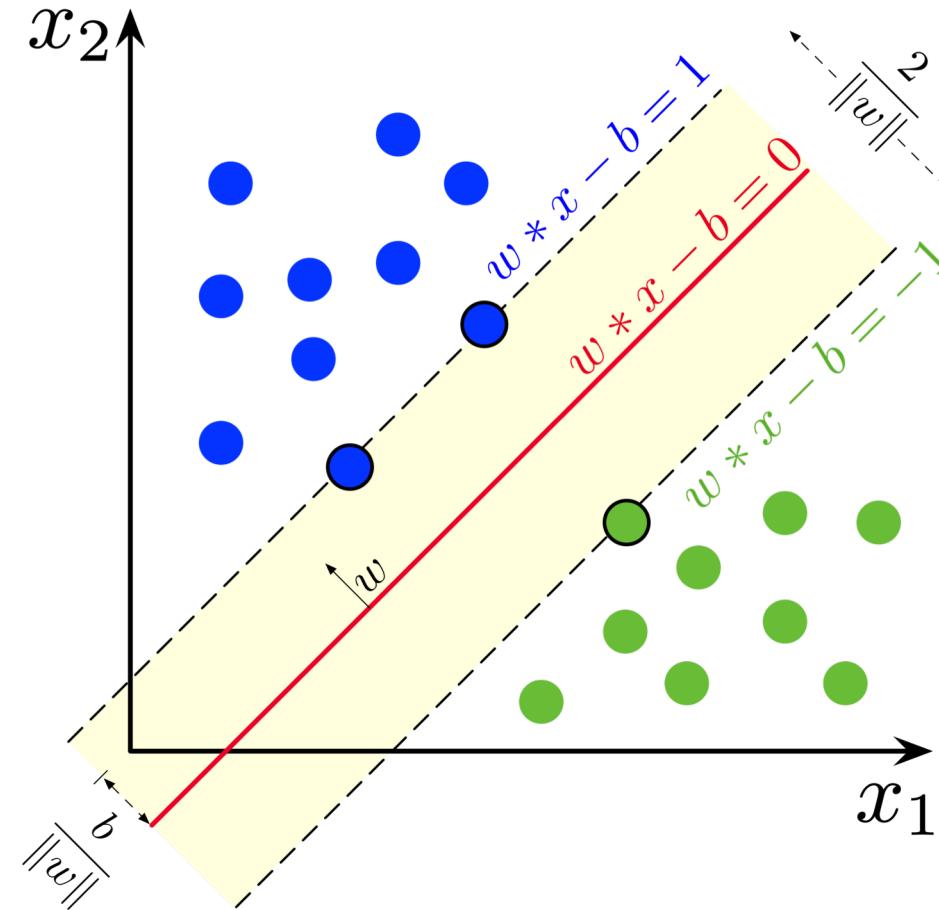
$$IDF(t_i, D) = \log\left(\frac{|D|}{|D_i|}\right)$$

где t – N-грамма;
 D – коллекция документов;
 d – документ из коллекции D .

Функциональная схема обучения и применения метода классификации



Метод классификации опорных векторов



Метод максимизирует ширину полосы зазора между двумя классами

Обучающая выборка

Размеченный корпус документов состоит из произведений 5 авторов:

- Достоевский, Горький, Толстой, Тургенев, Чехов

По 30 произведений каждого автора

$$N \in [2, 6]$$

Итог: 5 выборок из 150 текстов, представленных в виде частеречных N-грамм

Значение N для N-грамм	Объём корпуса (количество N-грамм)
2	2 972 956
3	2 789 053
4	2 606 454
5	2 429 961
6	2 264 315

Используемые метрики качества

- **Точность (accuracy)** – доля правильно предсказанных классификатором классов среди всех предсказанных значений.
- **F-мера**

Категория i		Экспертная оценка	
		Положительная	Отрицательная
Оценка системы	Положительная	TP	FP
	Отрицательная	FN	TN

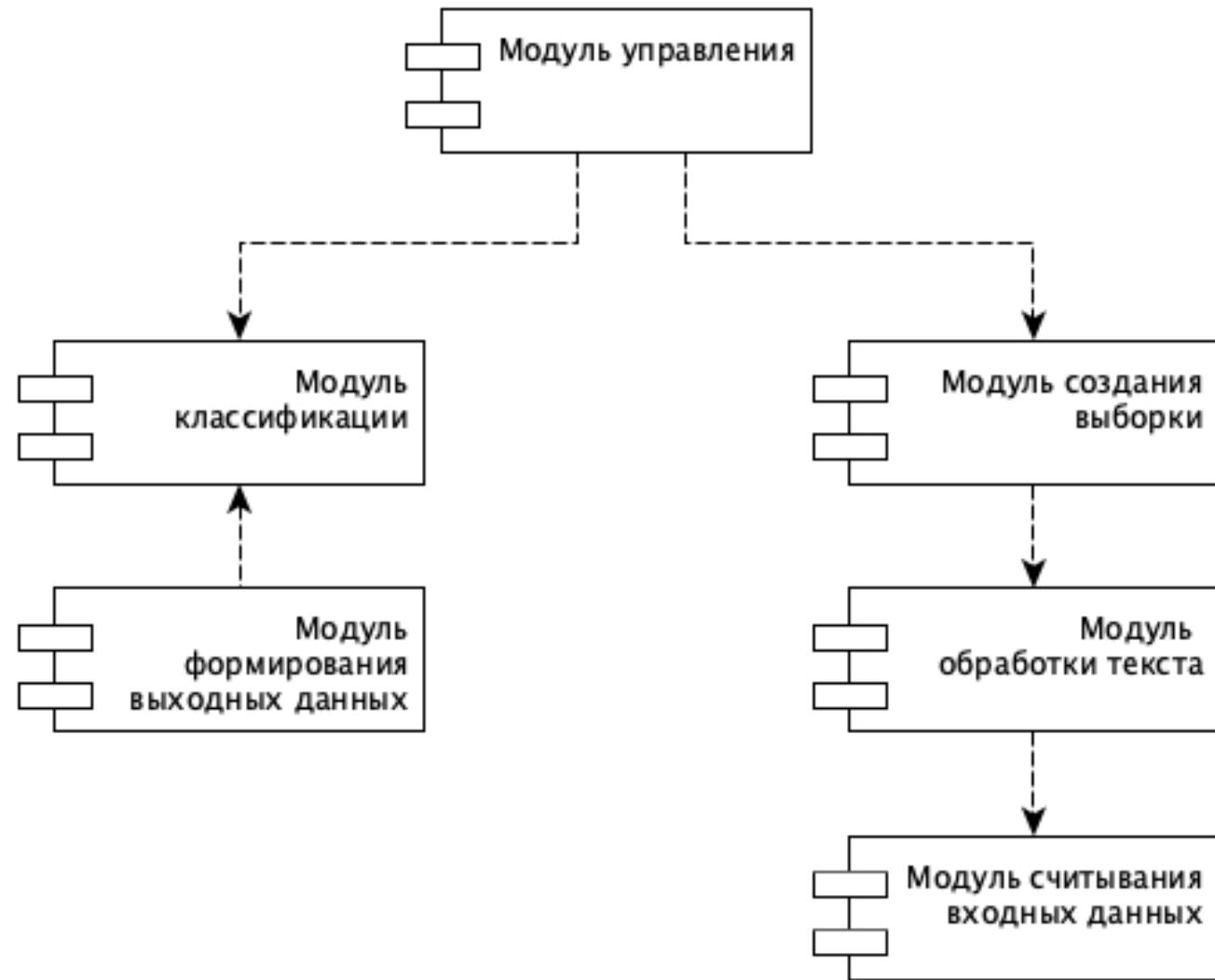
13

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

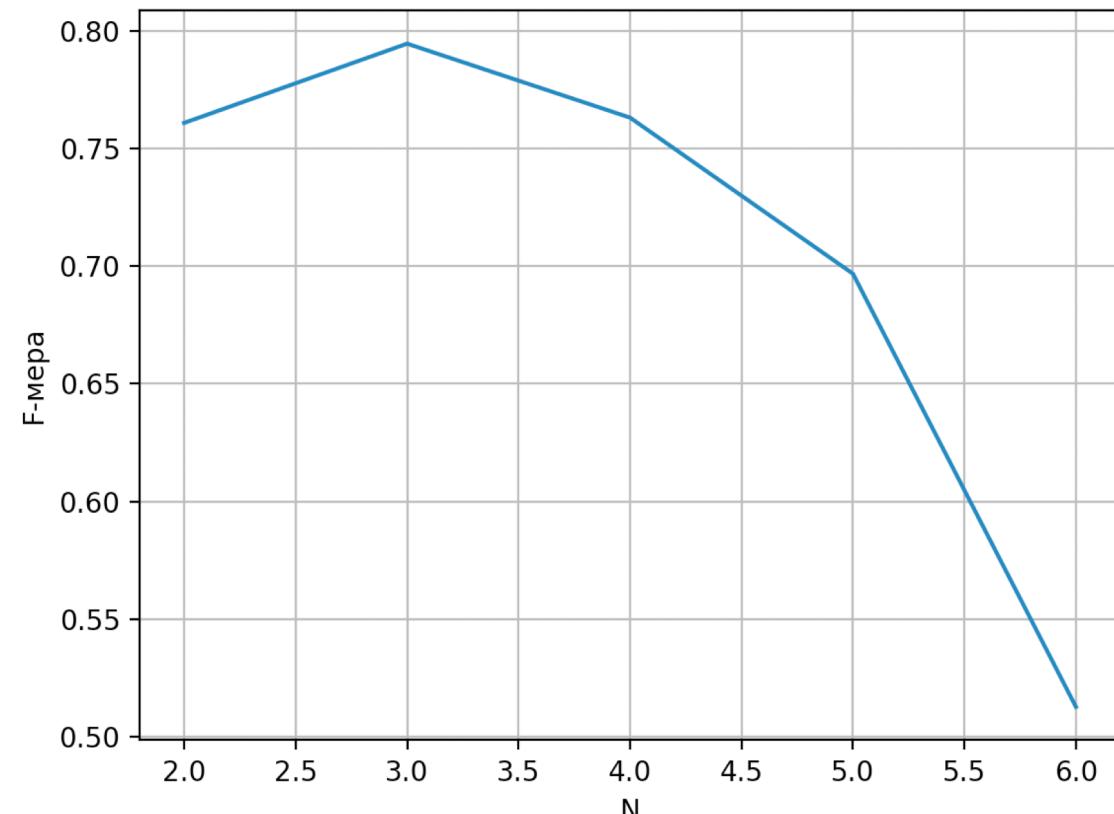
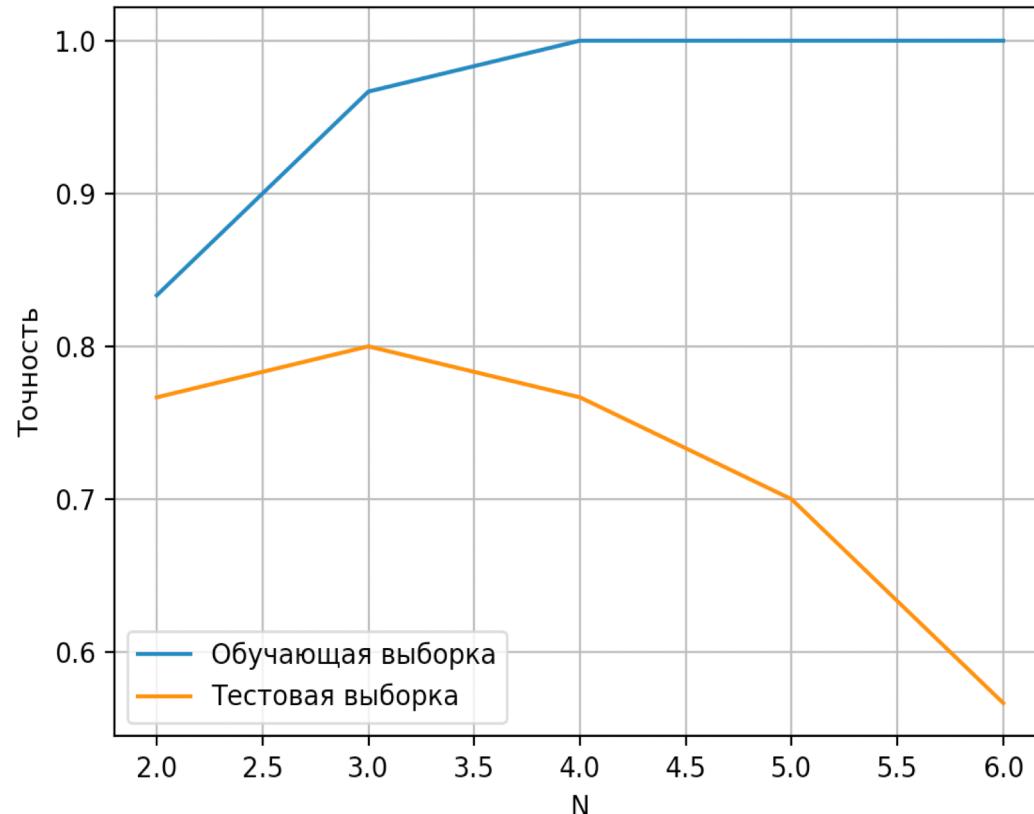
$$F = 2 \frac{Precision \times Recall}{Precision + Recall}$$

Структура ПО



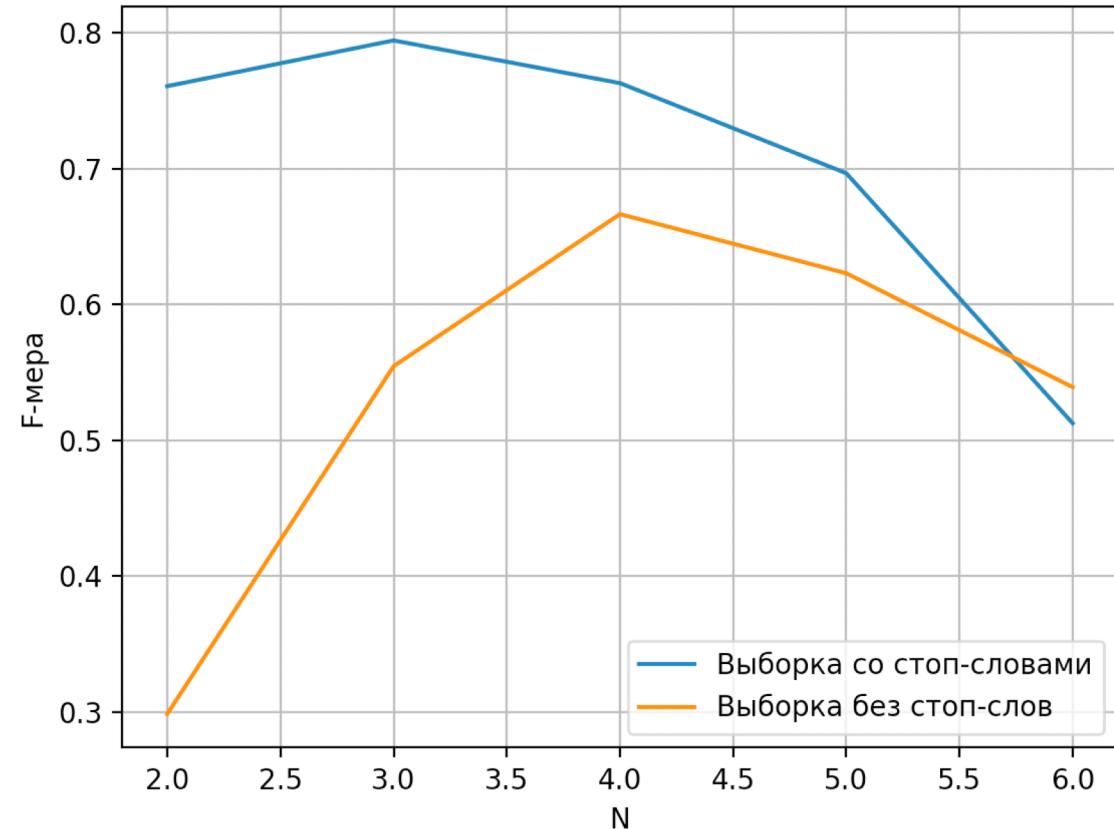
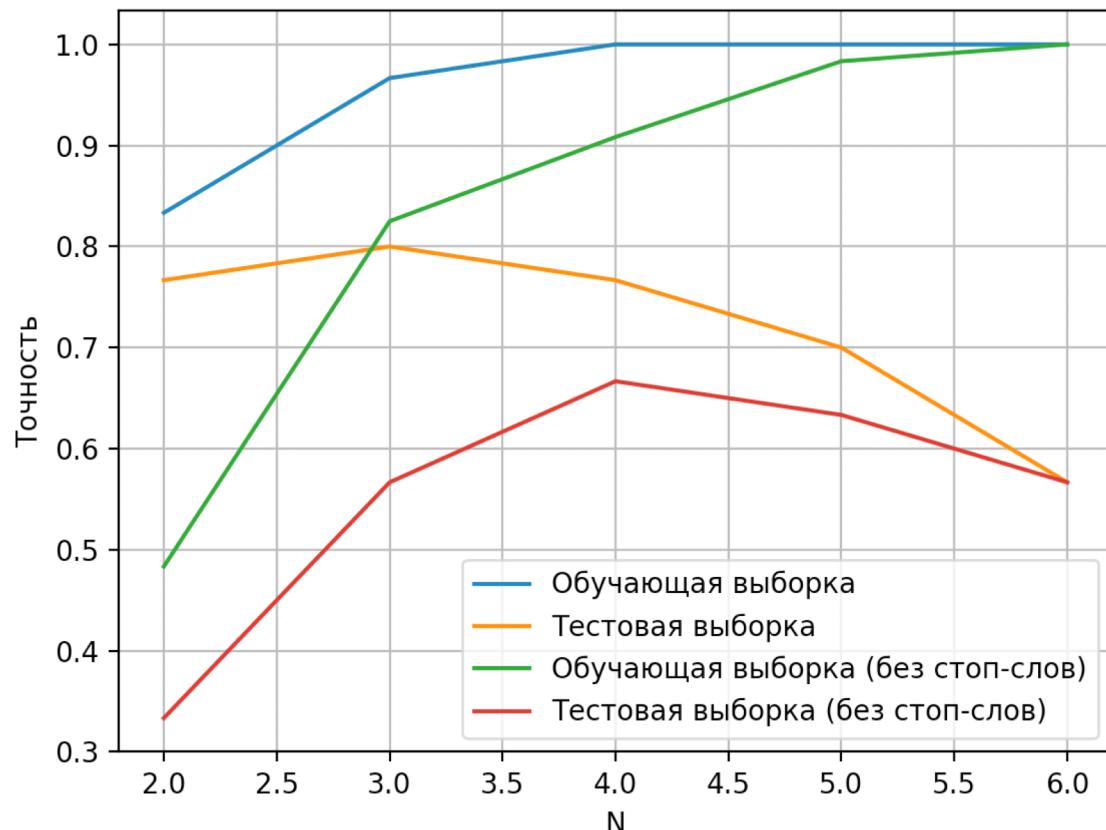
Зависимость метрик качества классификации от значения N

- 5 выборок из частеречных N-грамм: N от 2 до 6
- 150 текстов (5 авторов, по 30 текстов от каждого)
- Достигнута точность 80%



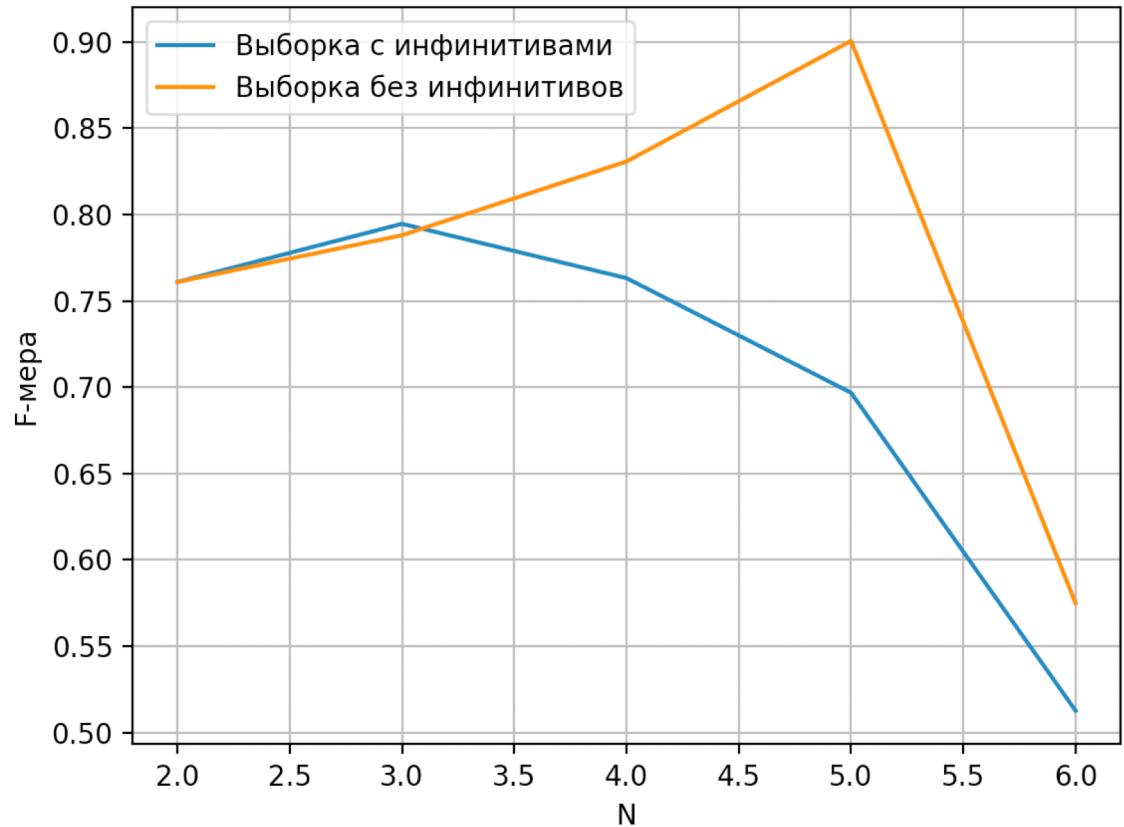
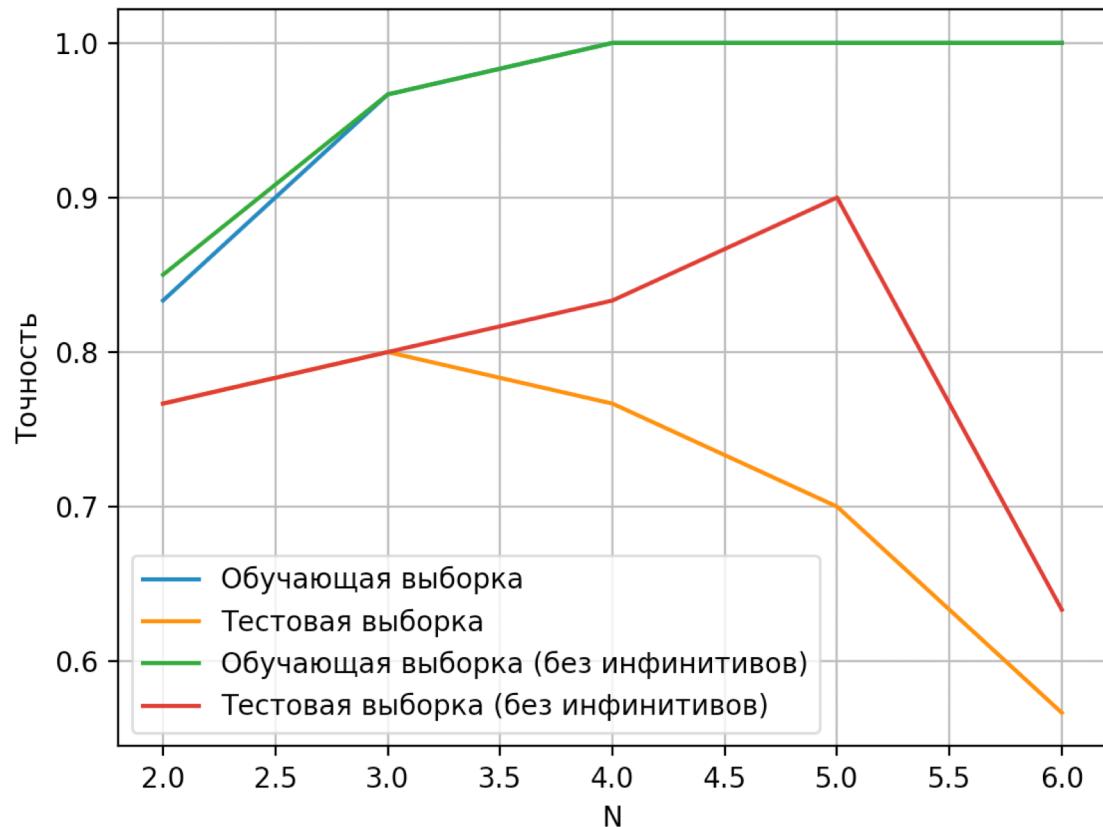
Зависимость метрик качества классификации от удаления стоп-слов

- Из исходных текстов были удалены стоп-слова



Зависимость метрик качества классификации от морфологического анализа

- Был произведен морфологический анализ, не различающий личные глаголы и инфинитивы



Заключение

- Разработан метод определения признаков авторского стиля
- Проанализирована предметная область
- Сформулированы основные положения разрабатываемого метода определения признаков авторского стиля и использующего его метода классификации
- Разработано ПО, реализующее предложенные методы
- Проведена оценка точности классификации прозаических текстов по выделяемым признакам, достигнута точность 80%

Дальнейшее развитие:

- Улучшение качества работы морфологического анализа текста
- Ускорение работы метода

Научные работы

- Принята в печать публикация «**TOWARDS A METHOD FOR DETERMINING SIGNS OF AUTHOR STYLE FOR TEXTS IN RUSSIAN**» в сборнике материалов Международной научно-практической конференции «Information Innovative Technologies» 2021 г., Прага (принято к публикации, РИНЦ)
- Выступление на тему «**К методу определения признаков авторского стиля текстов на естественном языке**» на студенческой конференции «Студенческая научная весна» 2021 г.