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

Выпуская квалификационная работа бакалавра

Метод классификации новостных текстов по тематикам с использованием опорных векторов

Студент: **Динь Вьет Ань ИУ7И-84Б**

Руководитель: **Кострицкий Александр Сергеевич**

2024г

Цель и задачи

Цель: разработка метода классификации новостных текстов по тематикам с использованием опорных векторов.

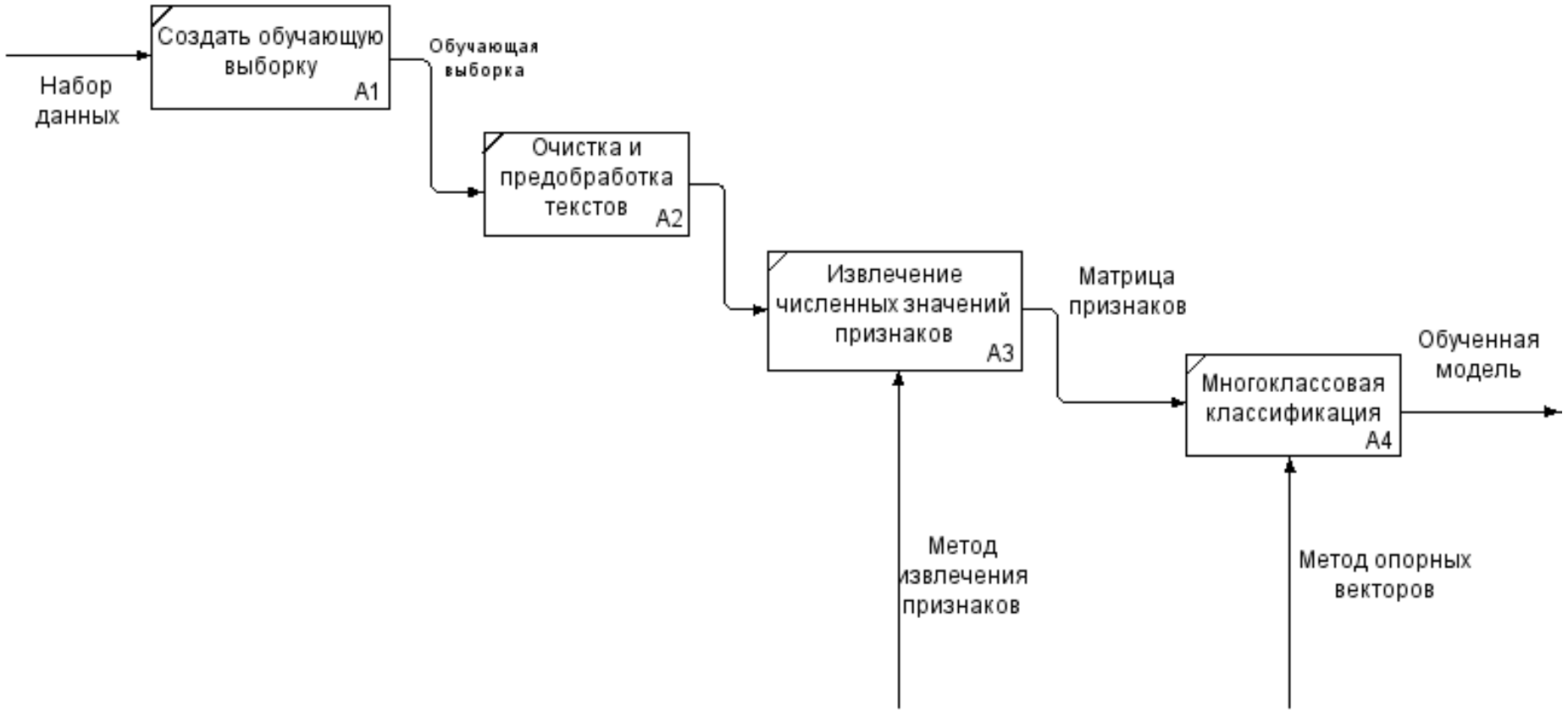
Задачи:

- провести анализ предметной области и основных методов классификации текстов;
- спроектировать метод классификации новостных текстов по тематикам с использованием опорных векторов;
- разработать программное обеспечение, реализующее данный метод;
- провести оценку качества классификации текстов.

Сравнительный анализ методов классификации текстов

Метод	Быстрота построения (обучения) классификатора	Требуется большой размер выборки	Возможность применять к многоклассовым задачам
Наивный байесовский классификатор	Да	Да	Да
Метод опорных векторов	Нет	Нет	Да
Дерево решений	Да	Да	Нет
Метод К-ближайших соседей	Нет	Нет	Да
Нейронные сети	Нет	Да	Да

Функциональная схема обучения классификатора



Этап очистки и предобработки текстов

- **Очистка текстов**: включает в себя преобразование текста в нижний регистр, удаление лишних пробелов и символов, не являющихся буквенно-цифровыми

- **Предобработка текстов**:
  + Токенизация – разбиение непрерывной строки на отдельные токены.
  + Удаление стоп-слов
  + Лемматизация – приведение слова к его начальной форме

Извлечение численных значений признаков из текста

- Для извлечения признаков из текста использоваться мера TF-IDF.
- TF-IDF – статистическая мера, используемая для оценки важности слова в контексте текста.

TF-IDF(word) = TF(word) · IDF(word)

TF(word) = $\frac{n_{word}}{A}$,

где n_{word} – количество вхождений слова word в текст,
A – количество всех слов в тексте

IDF(word) = $\log\left(\frac{D}{DW(word)}\right)$,

где D – общее количество документов,
DW(word) – количество документов, которые содержат слово word.

Набор данных

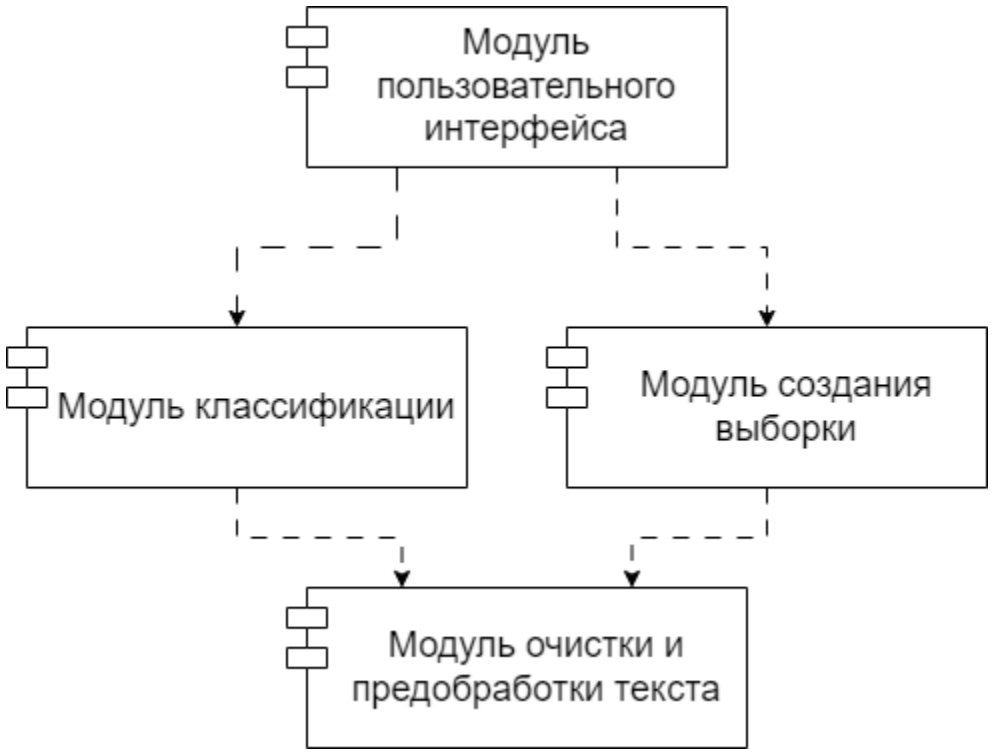
- Набор данных “Dataset from Lenta.Ru”.
- Содержит более 800 тысяч новостей на русском языке, соответствующих более чем 20 тематикам.
- Источник новостей — сайт lenta.ru, российское новостное интернет-издание, основанное в 1999 году.
- Выборка, созданная из этого набора

данных, состоит из 5 тематик:

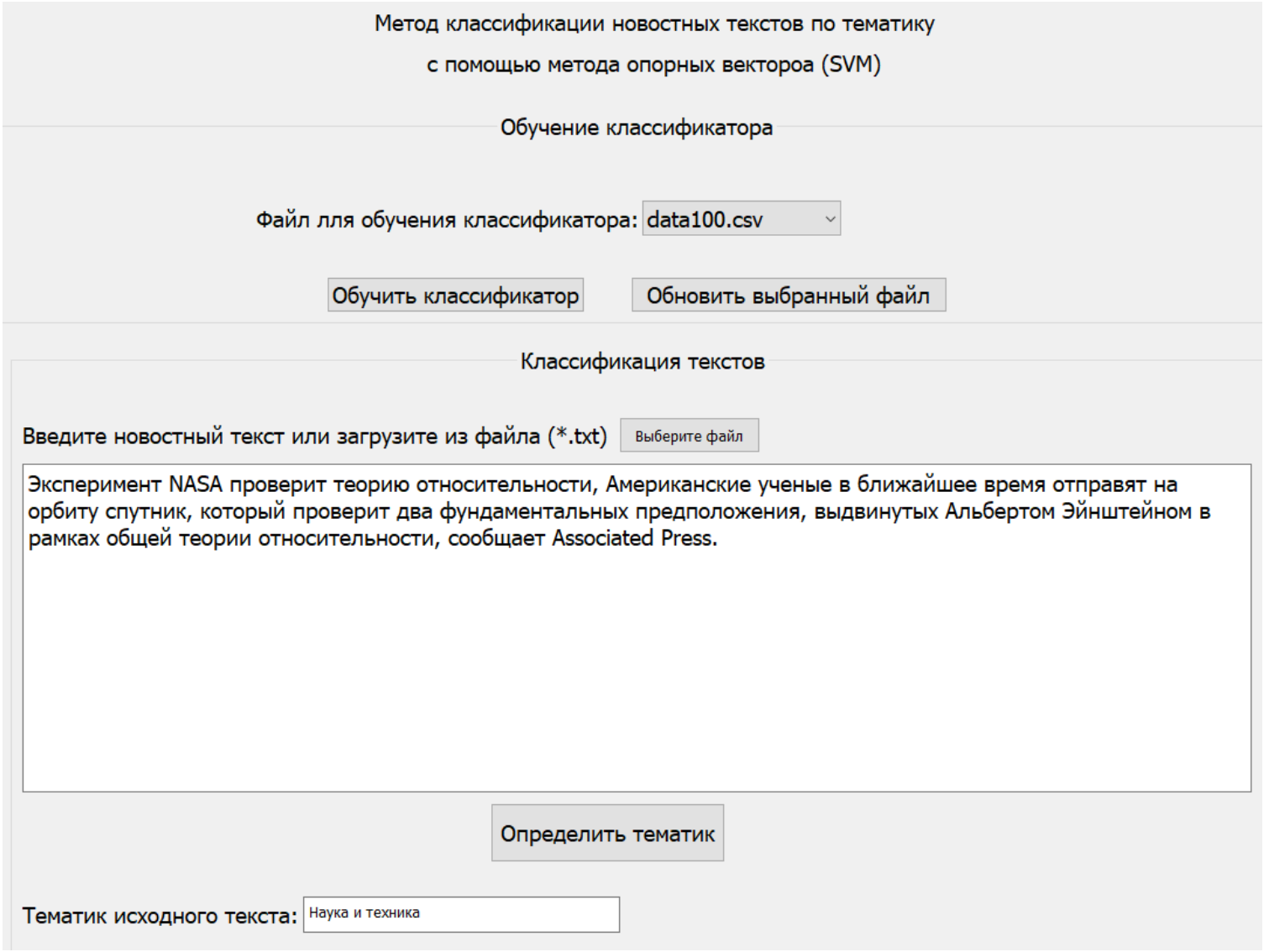
+ Наука и технологии
+ Спорт
+ Экономика
+ Культура
+ Мир.

Тематика	Количество текстов в наборе данных
Наука и технологии	53136
Спорт	64413
Экономика	79528
Культура	53797
Мир	136621

Схема разработанного ПО



Интерфейс пользователя программы



Метрики оценки качества классификации текстов

- Метрики аккуратности (Accuracy) – количество правильно проставленных меток класса от общего количества данных.

$$\text{Accuracy} = \frac{\text{TP+ FP}}{\text{TP+FP+TN+FN}}$$

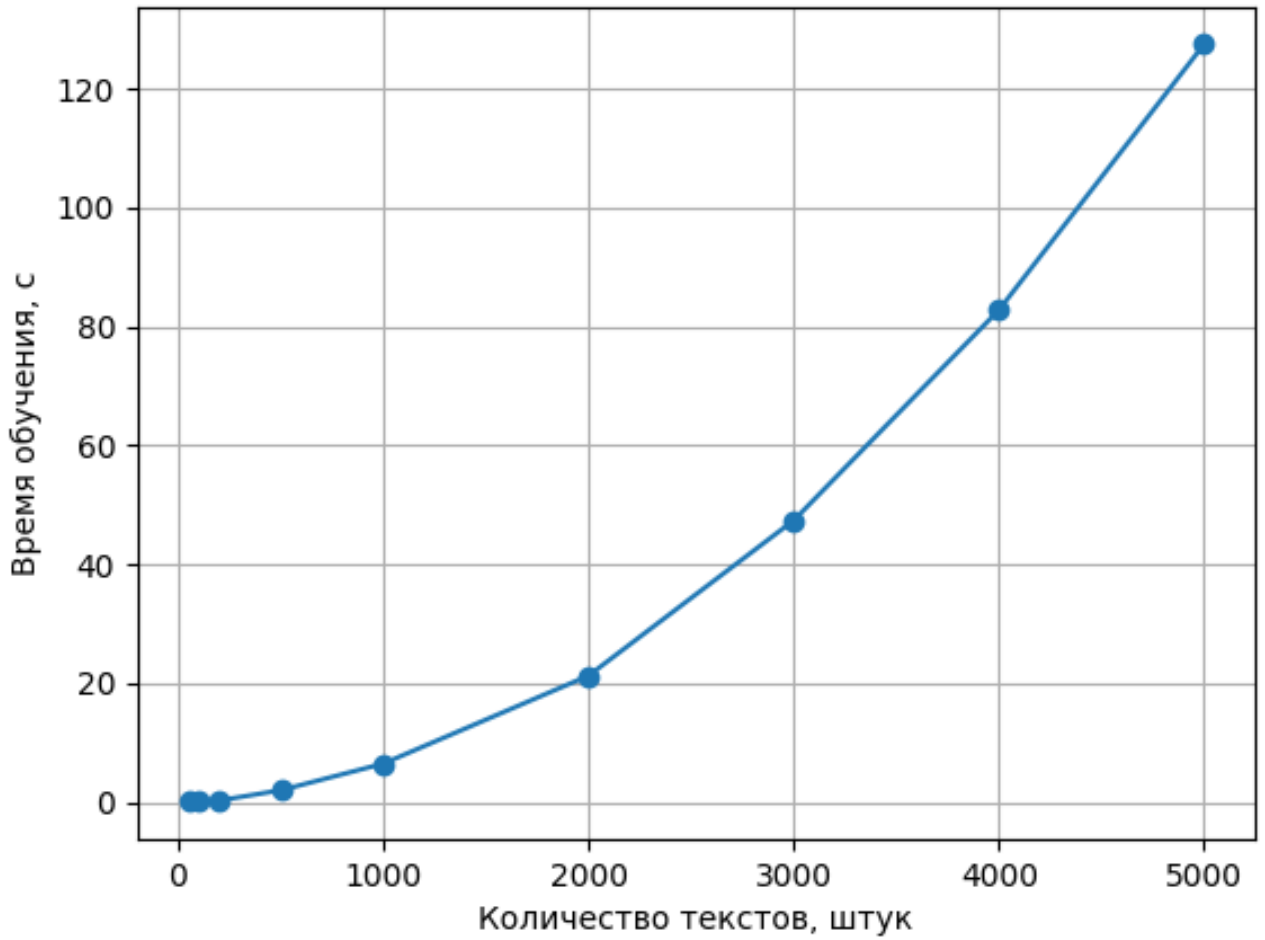
- F1-мера:

$$\text{Precision} = \frac{TP}{TP+FP} \qquad \text{Recall} = \frac{TP}{TP+FN} \qquad \text{F1} = 2\,\frac{Precision*Recall}{Precision+Recall}$$

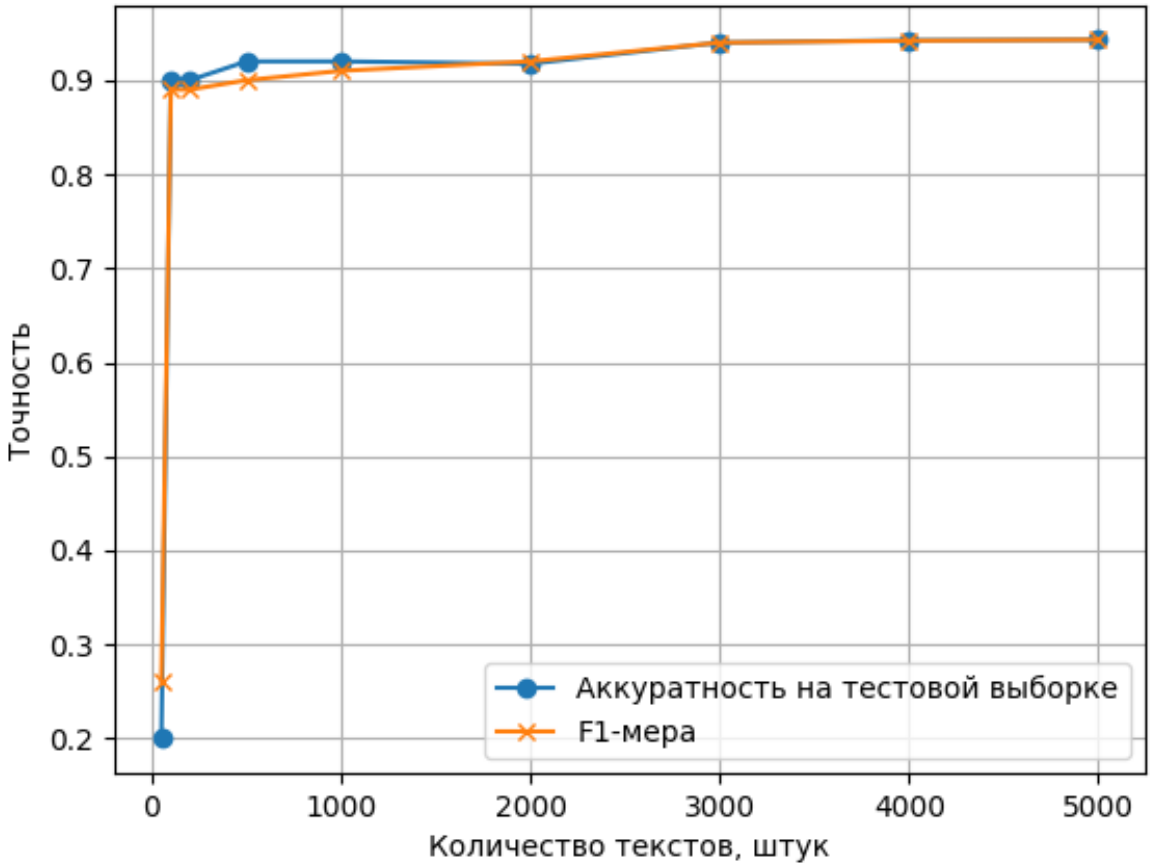
где:

- TP – истинный положительный результат
- TN – истинный отрицательный результат
- FP– ложный положительный результат
- FN – ложный отрицательный результат

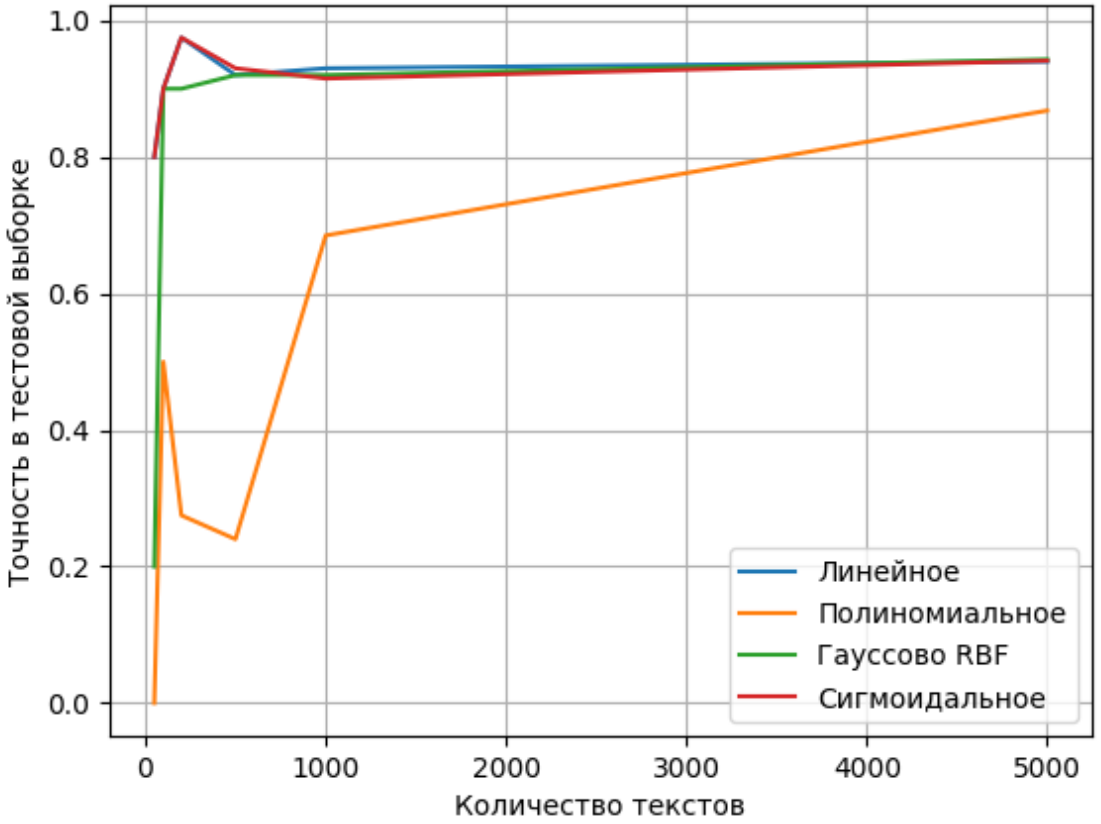
Зависимость времени обучения классификатора от количества текстов в выборке

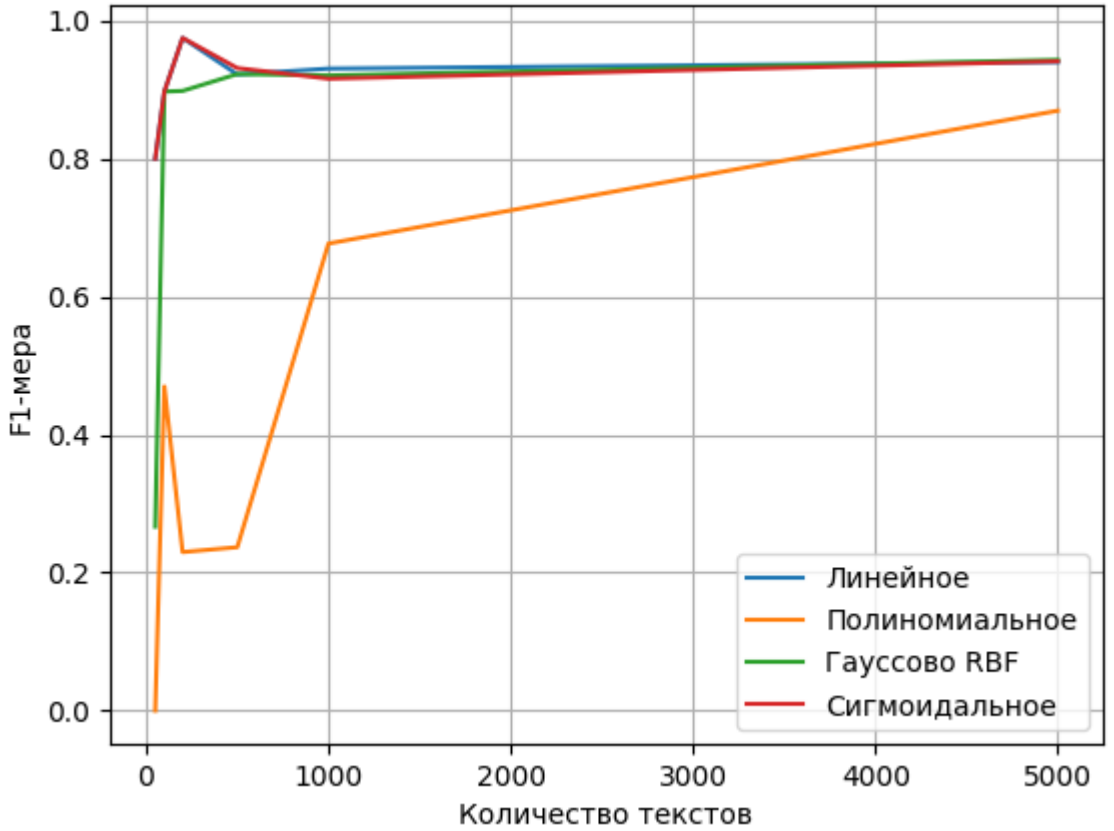


Зависимость качества классификатора от количества текстов в выборке



Зависимость качества классификатора от ядер метода опорных векторов





Заключение

Разработан и реализован метод классификации новостных текстов по тематикам с использованием опорных векторов.

Все задачи решены. Цель достигнута.

Дальнейшее развитие:

- ускорение работы метода.
- добавление возможности работы с различными языками одновременно.