|  |  |
| --- | --- |
| **Gerb-BMSTU_01** | ***«*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)** |

ФАКУЛЬТЕТ \_\_\_\_\_\_\_ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ\_\_\_\_\_\_\_\_

КАФЕДРА \_\_\_\_\_\_\_\_\_КОМПЬЮТЕРНЫЕ СИСТЕМЫ И СЕТИ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

НАПРАВЛЕНИЕ ПОДГОТОВКИ: 09.03.03 ПРИКЛАДНАЯ ИНФОРМАТИКА\_\_\_

**ОТЧЕТ**

**по лабораторной работе №** \_\_4**\_\_**

**Дисциплина: Прикладной анализ данных**

**Название лабораторной работы:** Прогнозирование моделью линейной регрессии

Студент гр. ИУ6-55Б **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**  А.М. Латышев

(Подпись, дата) (И.О. Фамилия)

Преподаватель  **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(Подпись, дата) (И.О. Фамилия)

Москва, 2024

**Цель работы:**

**Задание:**

Вариант 12

Вы используете данные из набора твитов для обучения моделей определения семантической окраски. Обучающую и тестовую выборки вы можете найти в файлах с соответствующими названиями по ссылке -<https://github.com/sismetanin/rusentitweet>

Вы решаете задачу обучения модели определения семантической окраски твита.

Этапы:

1. Оставьте в выборках только строки с классами positive и negative.
2. Определите и реализуйте креативные методы очистки набора данных. Например, в твитах часто встречаются ссылки на аккаунты других пользователей, оформленные однотипным образом – кажется, что это лишняя информация.
3. Осуществите стемминг подготовленного набора данных и преобразуйте каждый твит в мешок слов. Помните, что кастомные преобразования обучаются только на train выборке. Если они необучаемые, то нужно взять один и тот же тип преобразования для обеих выборок (один и тот же метод из одной библиотеки).
4. Составьте Count-матрицу и рассчитайте на ней tf-idf. Обратите внимание, что tf-idf – это обучаемое преобразование, которое нужно зафитить на обучающих данных и применить затем к тестовым.
5. Обучите модели логистической регрессии и случайного леса на обучающей выборке, примените их к тестовым данным. Посчитайте качество на обучающих и тестовых данных, сравните результаты. Определите наиболее важные признаки (слова).
6. В пункте 3 вместо стемминга осуществите лемматизацию и проделайте пункты 3-5 с учетом другого типа подготовки данных.
7. Сравните результаты по качеству и по наиболее важным признакам (словам) между 2 обученными вариантами

**Ход работы:**

1. Из выборок были удалены строки с классами, отличными от "positive" и "negative".
2. Были проведены креативные методы очистки данных, такие как удаление ссылок на аккаунты других пользователей, удаление символов пунктуации и т.д.

  
Рисунок 1 – Загрузка набора данных и очистка

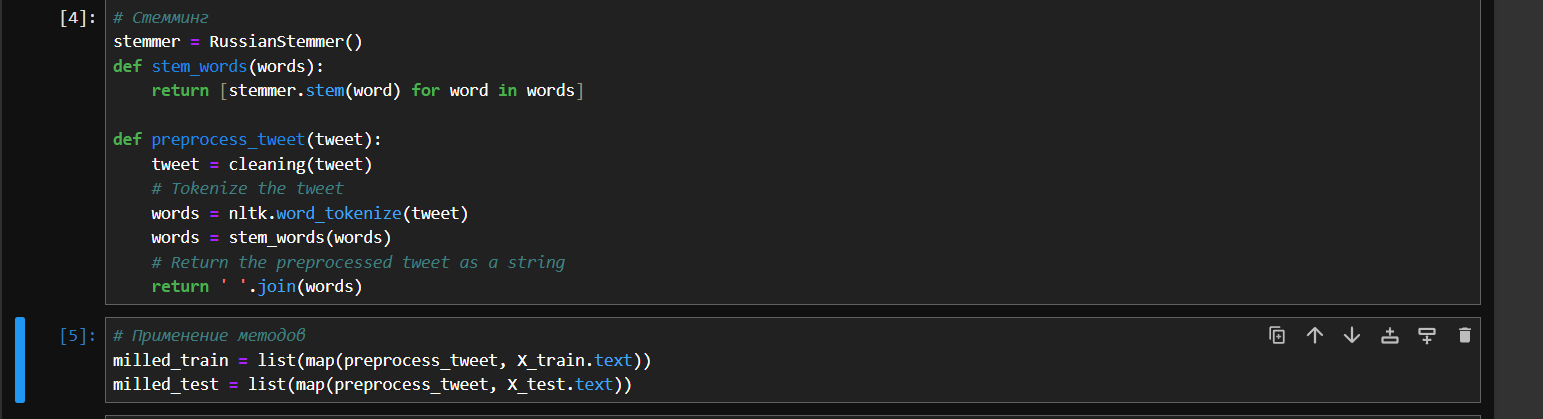
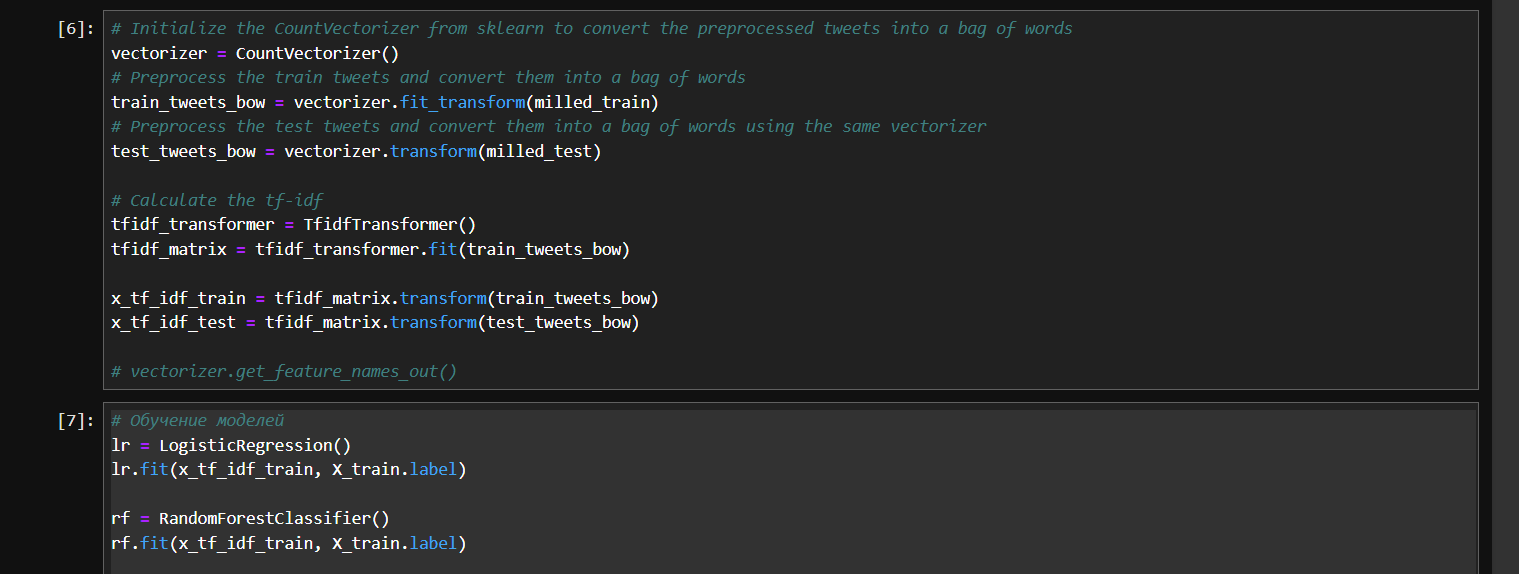
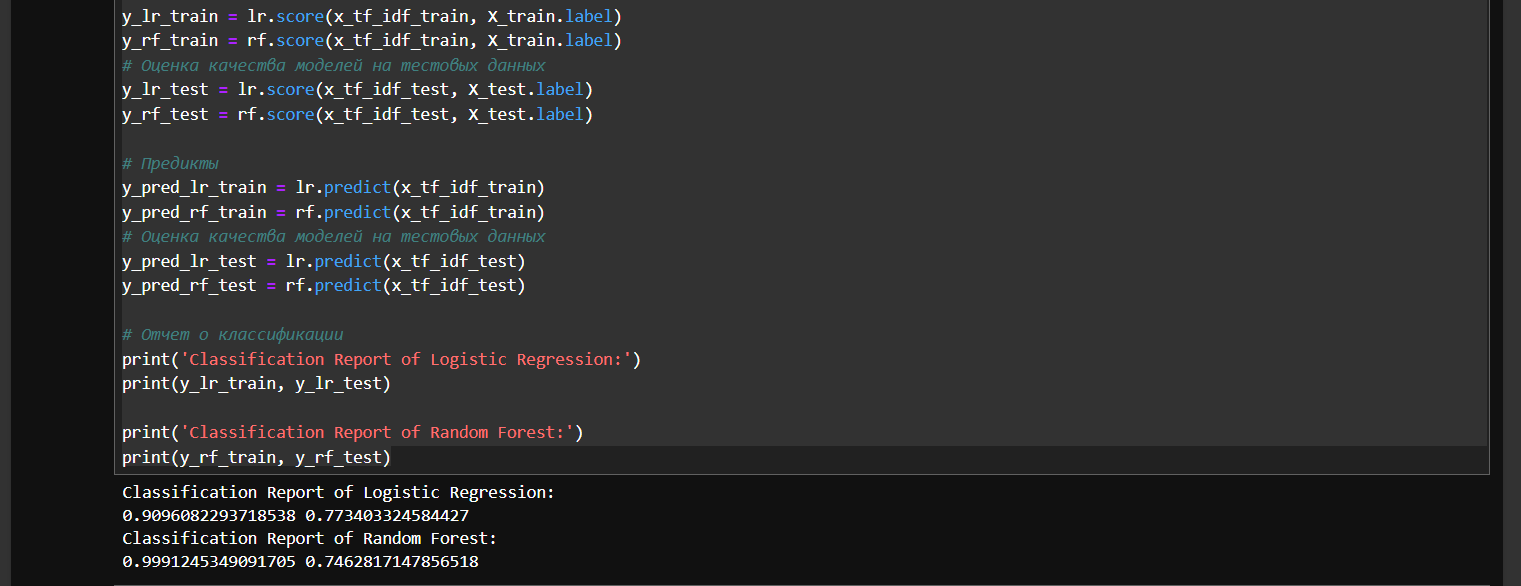
1. Был осуществлен стемминг и преобразование каждого твита в мешок слов.  
   

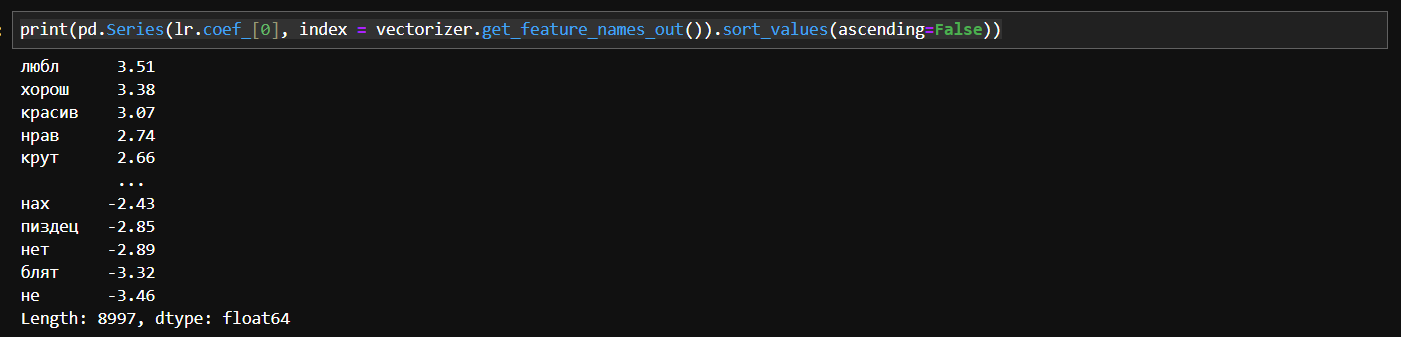
Рисунок 2 – Применение стемминга и очистки

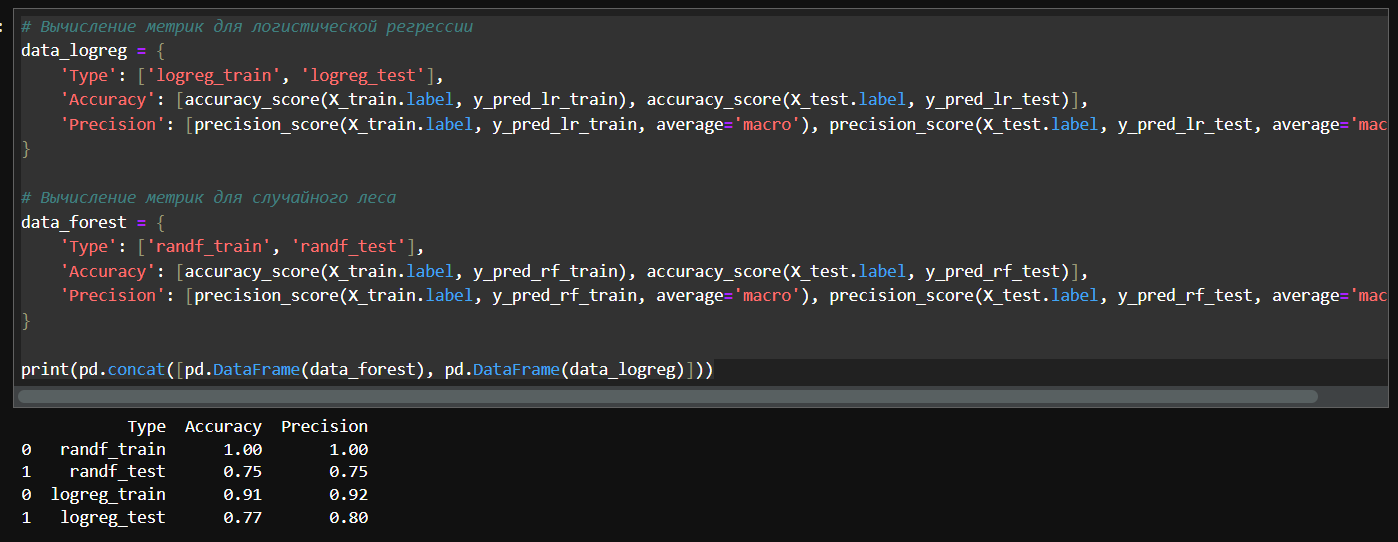
1. Была составлена Count-матрица и рассчитан tf-idf.

  
Рисунок 3 – Составление Count-матрицы и подсчет tf-idf

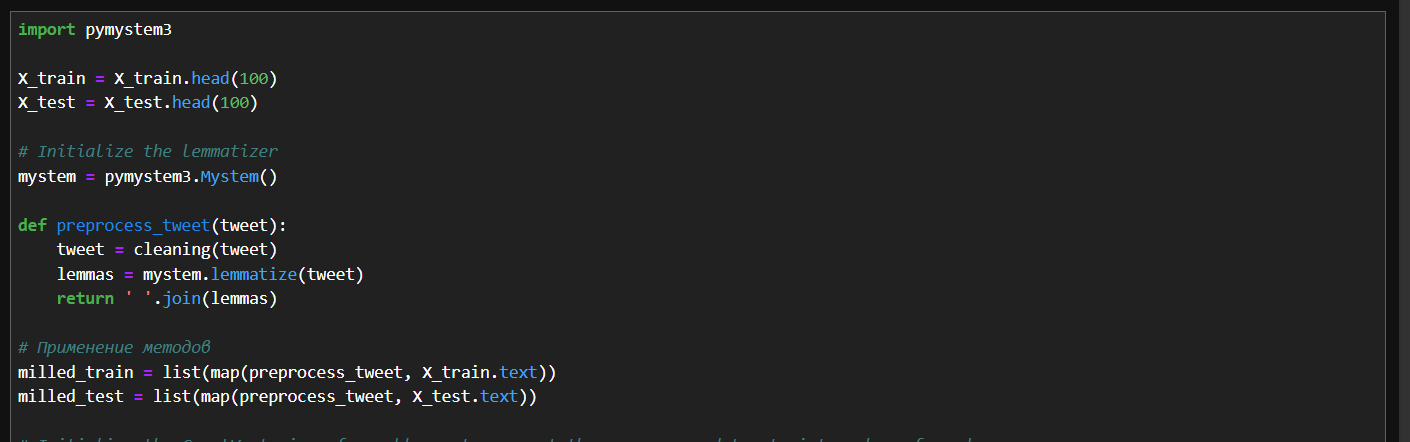
1. Были обучены модели логистической регрессии и случайного леса на обучающей выборке, применены к тестовым данным и оценены по качеству на обучающих и тестовых данных. Также были определены наиболее важные признаки (слова).

  
Рисунок 4 – Модели логистической регрессии и случайного леса

  
Рисунок 5 – Наиболее важные признаки

  
Рисунок 6 – Сравнение результатов

1. Вместо стемминга была проведена лемматизация и проделаны шаги 3-5 с учетом этого типа подготовки данных.

  
Рисунок 7 – Лемматизация

1. Были сравнены результаты по качеству и по наиболее важным признакам (словам) между двумя обученными вариантами.

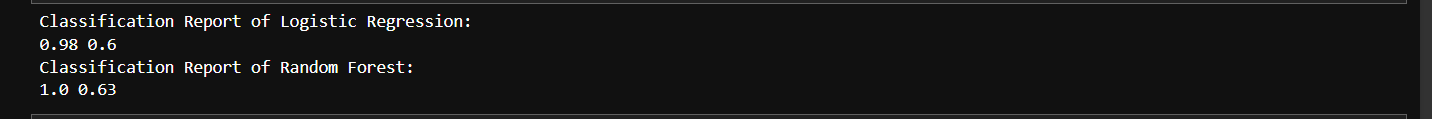
  
Рисунок 8 – Качество моделей с лемматизацией

  
Рисунок 9 – Наиболее важные признаки.

**Вывод:**

В ходе выполнения лабораторной работы были изучены методы предобработки текста, такие как лемматизация и стемминг, чистка пунктуации и спец.символов. С использованием подхода “мешок слов” были обучены модели логистической регрессии и случайного леса. Также был проведён анализ моделей на основе метрик Accuracy и Precision и оценены наиболее важные признаки, выделенные моделями.