Лабораторная работа №4

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.

Требования по отправке и оформлению лабораторной работы аналогичны требованию лабораторной работы №1, №2 и №3.

Основные этапы выполнения работы:

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.