Тестовое задание на ставку специалиста по машинному обучению

Цель: Построение анализатора тональности обзоров Steam на основе заданного корпуса текстов: Ссылка

Задачи:

1. Реализовать предобработку датасета:
    a. Сбалансировать классы по числу текстов
    b. Удалить все параметры, за исключением текста (параметр “review”) и класса (параметр “voted_up”))
2. Реализовать алгоритм предобработки текстов:
    a. Токенизация — перевод всех символов в один регистр и удаление всех символов, не являющихся буквами английского алфавита
    b. Удаление стоп-слов (предлоги, союзы, междометия и др.)
3. Реализовать алгоритм построения словаря токенов (за токен принимается слово):
    a. В словаре не должны присутствовать слова крайне низкой и крайне высокой частотности
    b. Словарь должен представлять собою набор пар “токен-частота”, отсортированный по убыванию частоты
4. Реализовать алгоритм векторизации текстов (TF-IDF, Bag-Of-Words)
5. Реализовать алгоритм создания тестовой и обучающей выборок
6. Предобработать и векторизовать обучающую и тестовую выборки
7. Создать модель, классифицирующую полученные вектора (допускается применение готовых реализаций)
8. Произвести обучение модели на обучающей выборке
9. Произвести проверку модели, используя тестовую выборку.
10. Произвести оценку качества классификации на основе точности, полноты и F-меры
11. Разместить код с результатами его выполнения в репозитории Git и подготовить Readme файл по запуску

Дополнительные задачи (не обязательно, но будет плюсом):

1. Лемматизация или стемминг токенов

2. Реализация графического отображения зависимости ошибки от поколения в matplotlib

Примечания:

- Допускается использование spacy, scikit-learn и других пакетов для предобработки текстов и реализации моделей. Также pandas, numpy, scipy и др.