HW: Text classification using Spark

Описание работы и критериев оценивания

Домашнее задание основано на соревновании
https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/data

Цель: обучить текстовый классификатор простых фичах с помощью Spark ML

1) HashingTF и IDF - 70 баллов
2) Word2Vec - 30 баллов

ДЗ предлагается сделать в ноутбуке и в формате ipynb загрузить в репозиторий на Github и прислать ссылку на него в интерфейсе сдачи

Бонусы и штрафы:

- **100%** за плагиат
- **30%** за посылку решения в течение недели после deadline

Блок 1. HashingTF и IDF

1) Подготовить фичи комментариев с помощью HashingTF и IDF
2) Обучить линейные классификаторы и сравнить метрики качества моделей
3) Сделать выводы о влиянии параметра numFeatures в HashingTF на качество метрик

Блок 2. Word2Vec

1) Подготовить фичи комментариев с помощью w2v
2) Обучить линейный классификатор и сравнить метрики качества с предыдущими подходами