СПРИНТ №6

ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА

В наших данных отсутствуют признаки, представленные текстом.

Использовалась библиотека `fetch_20newsgroups` из встроенных в sklearn датасетов:

Проведенная работа

1. Выбраны новостные категории из датасета и проанализирован текст;
2. Проведена очистка текста от метаданных, символов и пунктуации;
3. Текст токенизирован. Проведено удаление стоп-слов. Текст лемматизирован;
4. Проведена векторизация и определена точность используемых моделей

Векторизация	Test accuracy	Train time
По униграммам	0.9480	4.73
По униграммам и биграммам	0.9556	20.87
По униграммам и биграммам, параметр кросс-валидации увеличен до cv=5	0.9547	33.63
По униграммам и триграммам	0.8618	30.34
Дополнительный метод случайного леса	0.8816	10.09
Сверточная нейронная сеть	0.9192	Не оценивалось