Задача: Классификация текста на 3 класса. признак target (0, 1, 2) его и нужно предсказать

1) EDA - анализа данных.

- показать разные статистики для ваших данных

- визуализировать данные

- очистить от мусора

- новые фичи, которые вы придумаете

- какие-нибудь инсайты

2) embeddings:

- one hot encoding

- word2vec обученную (ссылку скину где достать обученную модель) / самим обучить

- fasttext обученную / самим обучить

- tf-idf

3) ML-algorithms:

- 2 простых алгоритма (SVM, LogReg, NaiveBayes)

- random forest

- decision tree

- ансамбли мл алгоритмов (2)

- кросс-валидация

- GridSearch для подбора гиперпараметров

4) Нейронные сети:

- Попробовать любую архитектуру из тех, что прошли на паре. (полносвязанная сетка, сверточная сетка или рекурентная)

P.S.: Постарайтесь сделать подбор количества слоев(желательно сделать бы) и некоторых параметров(если захотите)

Оценка: 8

Опционально:

- Визуализация при помощи bokeh ваших эмбеддингов и попытка пояснить почему они именно такие

- +1 алгоритм нейросетевой

Оценка: +2 балла

Конечные результаты:

- test.csv - требуется предсказать метки классов.

Файл должен состоять из двух колонок Id,Predicted

Id - id записи

Predicted - метка предсказанного класса

Метки классов:

0, 1, 2

P.S.: Не ленитесь пробуйте что-то интересное и новое.

Может быть кто-то попробуем "POS-теги" или еще какие-либо интересные фишки и сможет их объяснить