Файлы в папке:

1. dataset - папка с обучащей и тестовой выборкой
2. country.csv - файл с названием всех стран
3. Task.ipynb - ноутбук с решением
4. kontur.jpg - картинка для первой ячейки
5. navec_hudlit_v1_12B_500K_300d_100q.tar - файл с эмбедингами русских слов (вектора размерностью 300)
6. task.md - описание задания
7. utils.py - python файл с функциями для предобработки текстовых признаков

Версии основных используемых библиотек:

- sklearn = 0.23.2
- pandas = 1.1.2
- tensorflow = 2.7.0
- catboost = 1.0.5
- xgboost = 1.2.0
- pymorphy = 0.9.1 (не оптимизированный)
- navec = 0.10.0

Ноутбук с решением Task.ipynb условно можно разделить 2 части:
В первой части предобрабатывается текст заголовков новостей. Затем создаются, извлекаются из текста новые признаки. И анализируются полученные переменные с небольшой визуализацией.
Вторая часть - моделирование. Сравниваются различные модели: LinearRegressor, Catboost, Xgboost, NN+RNN.

Лучший результат f1 = 0.89 был получен с помощью CatBoostClassifier, обученный на признаках, извлеченных из текста + эмбеддинги для слов в заголовках + сам текстовый признак title. Catboost использует подход bag of words для обработки текстовых признаков.