Домашнее задание, модуль по NLP:

Задание 1:

В папке с данными (data) расположен датасет **bbc** , который содержит **5 классов** новостных статей (**business, entertainment, politics, sport, tech**). Вам необходимо реализовать следующее:

1. Обучить **любой** (на ваш выбор) алгоритм классификации из **классических алгоритмов** машинного обучения (не нейронная сеть!) и измерить метрику качества работы модели (не забываем также проверять качество на инференсе).
2. Дообучить модель DistillBert для многоклассовой классификации и проверить также качество работы алгоритма (н забываем проверять качество на инференсе). Похожий код был продемонстрирован на лекции, думаем головой, не копипастите, есть нюансы в данной задаче.
3. Оформить оба решения в пайплайны. Отдаем новости – получаем метку класса и вероятность. Соскрапить или собрать вручную по пять свежих новостей с ресурса (https://www.bbc.com/news - ссылки на новости прикрепить в ноутбуке) для каждого из классов и прогнать на них пайплайны. Получить результаты инференса пайплайнов, и оценить какое из решений лучше.

Оценка: 5 баллов

Задание 2:

По каждому из классов (**business, entertainment, politics, sport, tech**) вам необходимо взять каждый класс и **смоделировать распределение топиков (тем)** по каждому из классов и построить визуализации. Необходимо объяснить о чем больше всего говорится в каждом из классов, то есть сделать на основании вашей работы анализ по каждому из классов. **Используем любой подход, который вам известен для задачи Topic Modelling.**

Оценка: 5 баллов

Рекомендации по выполнению домашнего задания:

Можете выполнять его **в Jupyter Notebook** (2 задание обязательно в Jupyter Notebook так как это будет ваш аналитический отчет).

Для тех, у кого **недостаточно ресурсов**, можете выполнить работу в **Google Collab**.

Для формирования датасета, **рекомендую** пользоваться модулем **datasets**, который мы изучили на лекции.

По поводу представления корпуса данных, выбираем любой удобный вам способ, который как вы считаете лучше всего решит вам задачу.

Реализовываем весь пайплайн от загрузки, очистки и NLP пайплайна до подачи данных в датасет, весь тот процесс который был продемонстрирован в первых лекциях.

Удачи.