Тема 3: «Анализ тональности и внутренней структуры текстов новостей»

Датасет: новостной датасет от Lenta.ru
https://www.kaggle.com/datasets/yutkin/corpus-of-russian-news-articles-from-lenta?resource=download

Параметры:

- Период: сентябрь 1999 - декабрь 2019
- Количество новостей: 800K+

Признаки:

- адрес url
- заголовок новости (title)
- новость (text)
- topic
- tag
- дата

Задачи:

1. Первый этап:
    a. Сбор данных: поиск готового датасета, подходящего для предстоящей задачи
    b. Разведочный анализ данных:
        - кол-во уникальных значений
        - кол-во пропущенных значений
        - создать список наиболее употребляемых слов во всем датасете и в каждом теге
        - создать словарь ключевых слов по каждой теме
        - наиболее популярные темы и теги
        - выявления других особенностей данных
    c. Подготовка данных
        - очистка данных
        - кодировка данных
            - Bag of words
            - TF – IDF
            - N-gram of words
            - WORD2VEC (CBOW, skip-gram)

2. Второй этап (ML модели):
    a. Классификации новостей по темам
    b. Кластеризация новостей: применить алгоритмы и анализ результатов
    c. Доп. задачи:
        - ранжирование новостей
        - изменение тематики новостей во времени

3. Третий этап (DL модели):
    a. Классификация новостей с использованием моделей глубинного обучения
    b. Доп задачи:
        - генерация заголовков для новостей