№1

Был дан список статей Хабра, которых надо было спарсить, и был дан json файл уже с данными.

Важно, файлы не читались, поэтому в файлах надо было удалить “info”

Первое чем я занялась, так это импортированием библиотек, дальше я заняласьь парсером данных.(картинка) Я спарсила данные, которые смогла. (картинка)Набор данных включал в себя названия, описание, рейтинг, сферу деятельности компании, дату публикации и непосредственно текст статей.(картинка таблицы полученной)1.2………..

Дальше импортировала следующее библиотеки(фото). Провела предварительную выборку текста с помощью методов NLP: токенизацию, лемматизацию, выделение значимых частей речи, а также удаление стоп-слов, пунктуации, спецсимволов. (фото + краткое объяснение)

На такой таблице закончился модуль\_А(фото)

№2

Продолжаем работать с результатом Модуля\_А, выполнила поиск ключевых слов/биграмм/триграмм в тексте различными способами.(фото) Добавила ключевые слова/биграммы/триграммы, как новые признаки в набор данных. (фото).Занималась тематическим моделированием различными способами.Пробовала выполнить кластеризацию данных по сходству компаний несколькими способами и выбрать метрику оценки качества кластеризации.Провела анализ плотности распределения атрибутов и целевой переменной(фото)итог(фото)

Конец модуля\_Б

№3

Выполнила разбиение полученной выборки на обучающую и тестовую. Провела обучение моделей, основанных на трёх различных алгоритмах(фото их)

Пробовала сделать предсказание номинации организации на тестовой выборке, настраивать гиперпараметры, строить кривые валидации.

Итог(фото)—-----------------------------------------------------------------------------------------------------------Так же был дан target.json файл с информацией по номинациям для представленных организаций. Первое что пришло мне на ум, так это с помощью словаря присвоить каждой номинации целое значение, чтобы в дальнейшем было легко работать.

№4Получила control.xlsx

Прочитала файл и начала работать с ним, пробовала выполнить прогнозирование для контрольной выборки с помощью полученной модели.(фото)

Разработала приложение с графическим интерфейсом с помощью tkinter, которое должно было генерировать прогнозы по новым данным в режиме реального времени.Приложение предоставляет справку по имеющимся командам и их параметрам.

Если кратко то,-----(инфу брать со справки написанной)(ФОТО приложения, что будет при нажатии кнопки справка.

Итог: можно увидеть справку приложения, вписывать данные в ячейки, если они будут неправильно вписанны, то выдаст «неправильный формат», моя модель не до конца обучена предсказывать номинацию, но она обязательна будет-в будущем времени.

Перспективы улучшения Моего решения:

В будущем, у моей модели будет улучшено тематическое моделирование различными способами, будет полноценно сделана кластеризация, полностью будут настроены гиперпараметры, для создания кривых валидаций, обучу модель, чтобы она в конечном итоге предсказывала номинацию!

ВЫВОД:

На соревнованиях было очень интересно, даже порой весело, но трудно. Благодаря им я поняла, что такое машинное обучение, какие бывают виды обучения модели, как строить кривые валидации, делить выборку , много чего…Познакомилась с классными ребятами, с которыми соревнование было веселее. Поэтому просто хочу сказать спасибо всем!