**Инструкция к сдаче**

1. Настоятельно рекомендуем сдавать практическое задание в виде ссылки на личный репозиторий на github.
2. Рекомендуемый способ организации данных в репозитории: создать отдельные папки по темам и помещать в них отдельные файлы для каждой задачи с правильным расширением.

Ссылка на инструкцию по работе с git и сдачу практики:

<https://docs.google.com/document/d/1RAT_ukE39iOfbz1xa39QXae2hBUEZ4U6Fko_wFDdrsM/edit>

Ссылка на видеокурс по Git:

<https://geekbrains.ru/courses/66>

Если остались сложности с системой git, то обратитесь к преподавателю или наставнику.

## Тема “Улучшение качества POS-taggera”

на вебинаре был рассмотрен пример тегирования с использованием корпуса на Русском языке, вам необходимо улучшить модель что бы качество классификации было выше чем с лог регрессией

## Тема “Создание признакового пространства”

Используем предобработанные в рамках 1-ого домашнего задания датасет combine\_df\_prepocessed.pkl. Используем столбец 'clean\_tweet'.

##### ***Задание 1.***

Используя библиотеку Spacy, вывести ТОП-20 популярных NER в combine\_df датасете. Какой тип NER (ORG, GPE, PERSON и тд) оказался самым популярным? (Учтите, что max\_word\_limit\_spacy для Spacy = 1000000)

С помощью Spacy выяснить: какие персоны и организации самые обсуждаемые в train и test датасетах? вывести ТОП-20 самых популярных. Действительно ли в топ вошли только персоны и организации или есть мусор?

Повторим шаги из заданий 1 и 2, используя библиотеку nltk.

##### ***Задание 2.***

Используя библиотеку nltk, вывести ТОП-20 популярных NER в combine\_df датасете. Какой тип NER (ORG, GPE, PERSON и тд) оказался самым популярным? Для данного задания используем ограничение на количество символов во входном датасете (max\_word\_limit\_spacy = 1000000), чтобы иметь возможность сравнить результаты работы Spacy и nltk. Обратите внимание, что nltk чувствителен к регистру.

С помощью nltk выяснить: какие персоны и организации самые обсуждаемые в train и test датасетах? вывести ТОП-20 самых популярных. Действительно ли в топ вошли только персоны и организации или есть мусор?

##### ***Задание 3.***

Какая из библиотек по вашему лучше отработала? Сравните качество полученных most\_common NER и количество распознаных NER.