Используя библиотеки sklearn и torch, решите следующие задачи классификации (мультиклассификации). Графики рисуйте с помощью библиотеки matplotlib или любой другой удобной библиотеки (например, seaborn)

1. Нормализуйте текст
2. Разбейте его на трейн/тест. Если датасет слишком большой, то оставьте не больше 10к - 100к записей.
3. Обучите логистическую регрессию стохастическим градиентным спуском, используя tf-idf в качестве факторов. Редкие слова удалите. Проанализируйте, какие слова получили наибольший вес. Попробуйте лемматизировать слова. Нарисуйте loss на графике на трейне и на тесте (например, с помощью библиотеки matplotlib). Посчитайте метрику accuracy, F-1 (macro/micro для задачи мультиклассификации)
4. Поэкспериментируйте с весом в L1 регуляризации. Сравните обученную новым способ лог. Регрессию с ранее обученным вариантом. Что лучше? Какие веса занулились? Нарисуйте графики. Посчитайте те же метрики.
5. Обучите нейронную сеть с помощью библиотеки torch с одним скрытым слоем, используя tf-idf над лемматизированными словами. Редкие слова удалите. Нарисуйте loss на трейн и тесте на графике. Сравните лоссы обученные с разными инициализациями: нулевая, xavier, he. В качестве функции активации используйте ReLU. Посчитайте те же метрики.
6. Зафиксируйте лучшее решение по ранее упомянутым метрикам.

Формат сдачи: jupyter notebook (или ссылка на google collab)

Дедлайн сдачи: 13 октября 23:59

Почта для отправления: nlp_bmstu_fn12@mail.ru

Описание датасетов

Natural Language Processing with Disaster Tweets

https://disk.yandex.ru/d/9IIgepBWWNBKDw

Нужно предсказать какие твиттеры написаны про реальные происшествия, а какие — нет

Twitter sentiment analysis

https://disk.yandex.ru/d/-53L5fKvKIJoqw

Негативный/позитивный/нейтральный твит. Мультиклассификация

News category dataset

https://disk.yandex.ru/d/p2HZcK1Lov3B2g

Предсказать рубрику новости (англ). Мультиклассификация

News rubric dataset

https://disk.yandex.ru/d/Uvv4gykwQJGuWw

Предсказать рубрику новости (ру). Мультиклассификация

Toxic russian comment

https://disk.yandex.ru/d/mbrnfedvmEIFsA

Оставьте нетоксичные комментарии, а все остальные пометьте как токсичные. В таком случае останется всего 2 класса. Задача классификации

Распределение датасетов по студентам

Злобнов Даниил Алексеевич — russian news dataset

Иванова Юлия Витальевна — toxic russian_comments

Ириневич Сергей Георгиевич — toxic russian_comments

Касьянова Кристина Александровна — news category dataset

Кононенко Артём Александрович — disaster tweets

Лосев Владислав Александрович — tweeter sentyment analysis

Мужецкий Антон Андреевич — russian news dataset

Работяжева Дарья Михайловна — russian news dataset

Середа Максим Андреевич — tweeter sentyment analysis

Сытник Вероника Александровна — news category dataset

Усольцева Валерия Денисовна — toxic russian_comments

Шабашов Иван Александрович — disaster tweets

Янук Андрей Владимирович — tweeter sentyment analysis