**Вопросы и ответы:**

1) Эссе, которое содержит фактическую ошибку имеет ответ True или False?

* Для нас это без разницы.

2) В датасете есть дисбаланс по кол. эссе. Какой он?

В каком колабе он был выявлен? (Я увидел дисбаланс только в кол. слов, что содержатся в выборках).

3) Где указано кол. эссе по каждому предмету

Количество эссе по каждому предмету - указано графой text\_count в файле.json.

4) Как организовать хранение наших выводов и предположений по обучению нейросети? Как это делают датасентисты?

Используйте Excel файлы.

Учите нейросети до того момента, пока они не начнут переобучаться.

Т.к. нейросети с разными структурами могут нуждаться в разном кол. эпох и learning\_rate.

Учить нейросети разных структур нужно на одинаковых данных, сформированных вручную.

Тестируйте нейросети на тестовых выборках – это наилучший способ проверить точность нейросети разной структуры.

5) Удачный ли мы выбрали размер словаря (кол. слов, что учитываются при обучении), при формировании датасета?

Чтобы сделать такой вывод нужно проверить с разными размерами словаря.

6) Правильно ли я сделал экспорт датасета Мелании (Словарь частотности)?

Да, экспорт и импорт сделаны верно.

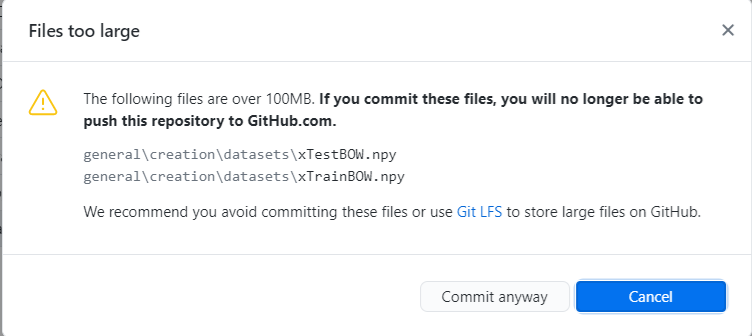
Но это не словари частотности, а закодированные тексты.

Словарь частотности – это токенайзер.

7) Как сохранить сформированный датасет, если он весит очень много?

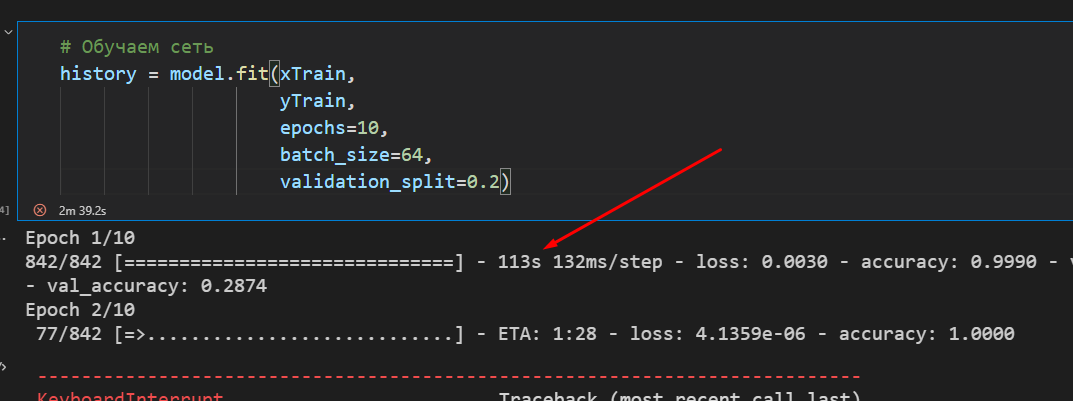
Есть возможность сохранения .npy файлов с архивацией.

Это позволит решить проблему.



Почему у меня уходит 2 минуты на 1 эпоху?

Как быстрее учить нейросети на пк.



Нужно мощное железо.

Советы

1. Сформировать несколько .npy датасеты, которые будут содержать закодированные тексты сформированные с разным словарем.

Неограниченный словарь, 10000, 20000, 50000

Используй перевод слов к нормальной форме. Убирает склонения.

Выкусить ненужные символы.

Pymurphy.

Сократит кол. слов. Меньше слов может помочь лучше обучить нейросеть. Но может быть и хуже.

**Дальнейшие шаги:**

Датасеты готовьте на лок. Пк. Это быстрее, чем в колабе.

Обучайте нейросети в колабе. Это быстрее, чем на пк.

Сделать датасеты с и без них паймерфи (2 датасета).

Сделать датасеты с разными размерами словарей (3 размера словаря).

= 3\*2\*3\*2\*3= 18\*3 = **54 датасетов (файлов)**

Доп. Параметры, что меняются:

xLen (100), step (50, 100),

xLen (300), step (150, 300),

xLen (500), step (250, 500),

Деление на выборки (2)

Тестовая не нужна

ААААА потом – ЕКСЕЛЬ!

**Самое важное**

Сформировать датасеты

Получить доступ к платформе тестирования от организаторов конкурса.

Начать оформлять исследование (если будет доступ к их платформе, тогда нам не придется формировать тестовые выборки)

Только Мелания может загружать в тестовую систему.

Мелания – формирование датасетов.

Глеб – Подготовка таблиц для анализа.

Марат – **Итоговый проект по ML Эхо.**

Саша -> Обработка тестового датасета с платформы.

Игорь = мощный пк –> Игорь учит нейросети