

Домашнее задание по теме «Seq2Seq, Encoder-Decoder. Attention»

Формулировка задания

Работа с переводчиком текстов, основанном на архитектуре Seq2Seq и механизме внимания

План работы

- 1) Выбрать и сделать копию одного из ноутбуков с лекции.
- 2) Выбрать и загрузить датасет. Обратить внимание на размер датасета, возможно, потребуется сократить входные данные. Для тренировки можно работать со словарем 1000 - 5000 строк.
- 3) Входную выборку необходимо поделить на тренировочную и тестовую. Рекомендованный размер выборки 80% тренировочной, 20% тестовой.
- 4) Натренировать модель Seq2Seq с механизмом внимания на новых данных. Для тренировки будет достаточно 5-15 эпох. Если обучение идет очень медленно, сократить датасет.
- 5) Оценить результаты модели по критерию `val_loss`;
- 6) Дополнительно под *. Оценить модель с помощью функции `evaluate` на тестовом наборе.
- 7) Дополнительно под *. Выбрать несколько предложений из тестового набора данных `test` и перевести их. Результат отобразить таблицей или на графике.
- 8) Открыть доступ для чтения ноутбука по ссылке.
- 9) Прикрепить ссылку на ноутбук в качестве ответа на домашнее задание на платформе `learn.innopolis.university`

Перечень инструментов, необходимых для реализации деятельности

Google Colab <https://colab.research.google.com/>

Библиотека **keras** фреймворка **tensorflow**

Фреймворк **pyTorch**

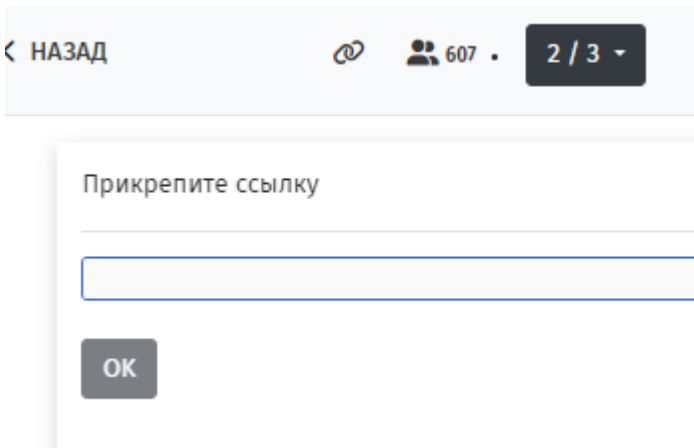
Обратить внимание, что может потребоваться режим GPU в Colab ноутбуке

Датасеты для работы

1. Предложения на русском и английском языке
EnglishRussiansentencepairs
<https://www.kaggle.com/datasets/lgtfeather/englishrussiansentencepairs>
2. Датасет статей из википедии Wikibooks Dataset (два языка на выбор)
<https://www.kaggle.com/datasets/dhruvildave/wikibooks-dataset>
3. Датасет переводчика Яндекс Англо-русский параллельный корпус
<https://translate.yandex.ru/corpus?lang=en>
4. Короткий датасет для перевода с русского на английский
https://drive.google.com/file/d/166Msc4oTDy2wFWd5_VnuZES3d_duvsMD1/view

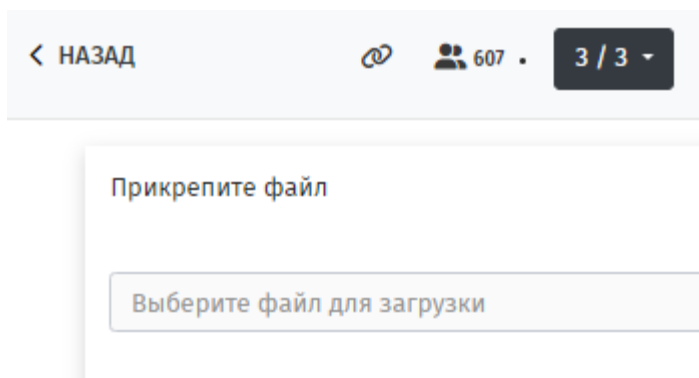
Форма загрузки

- В поле ссылки (2 страница задания) загрузить ссылку на ноутбук google colab или github репозиторий.



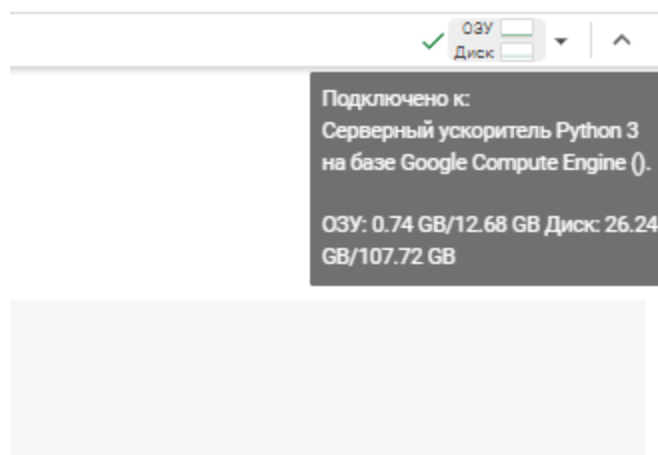
The screenshot shows a Google Colab interface. At the top, there is a navigation bar with a back arrow, the text "НАЗАД", a share icon, a user icon with "607", and a tab indicator "2 / 3". Below this, a modal dialog box is open with the title "Прикрепите ссылку". It contains a single-line text input field and an "ОК" button at the bottom left.

- В поле файла (3 страница задания) загрузить ноутбук с решением (файл с расширением .ipynb).

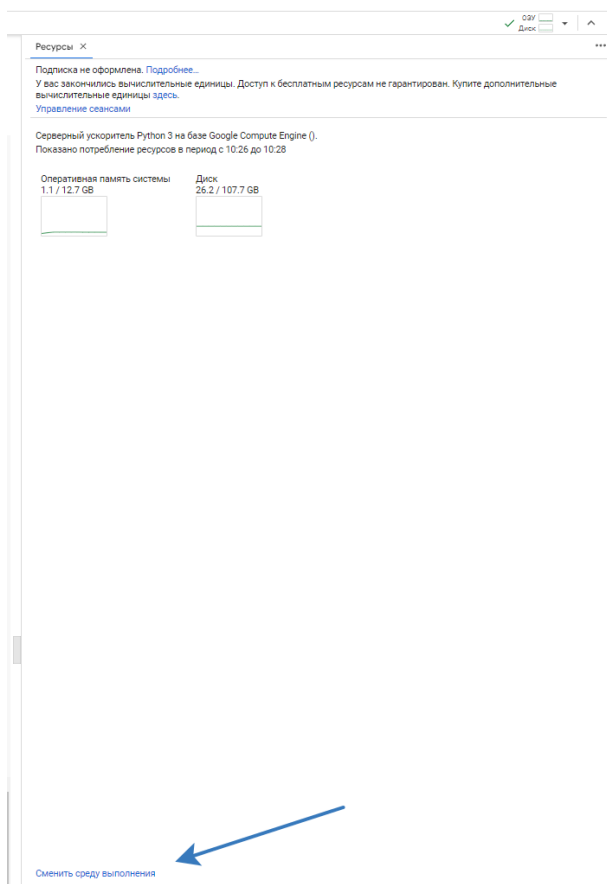


Инструкция по переключению на режим GPU в google colab

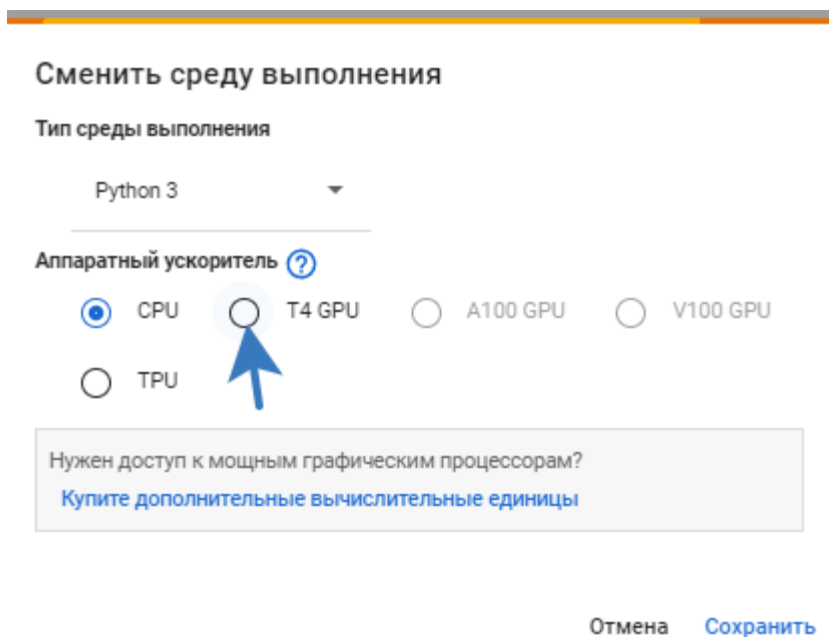
1. Нажмите на панель **“ОЗУ ... Диск”** в правом верхнем углу экрана, рядом с лого вашего google аккаунта



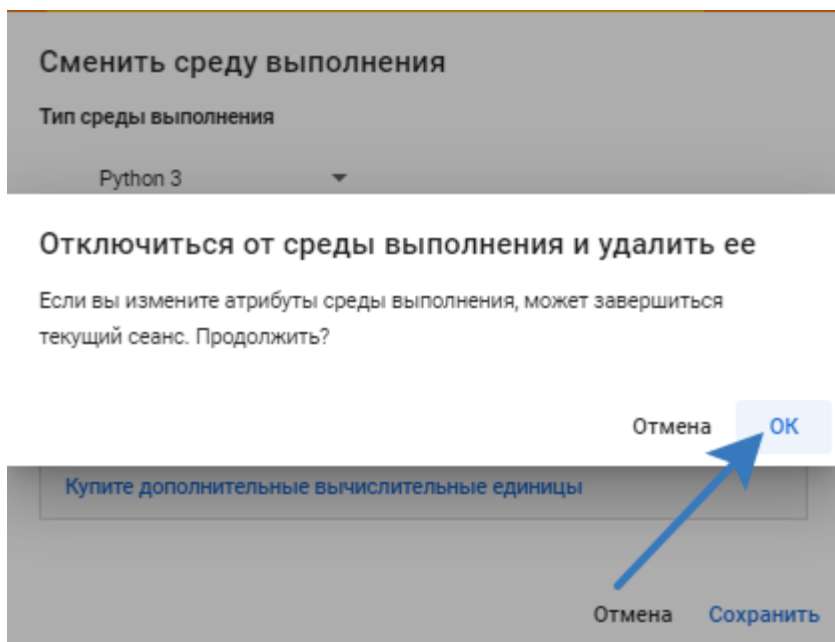
2. Внизу выберите ссылку **“Сменить среду выполнения”**



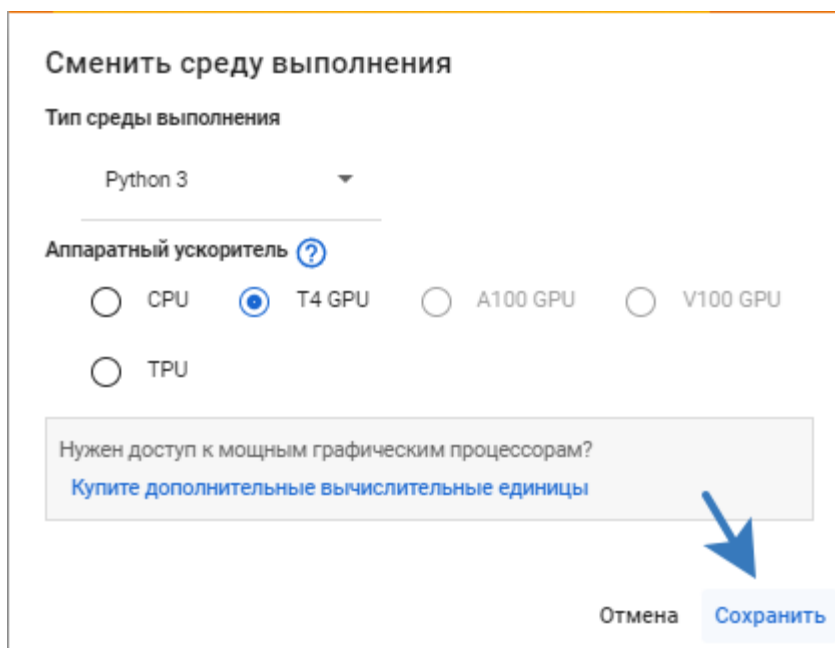
3. В окне со средами выполнения выбрать **“T4 CPU”**



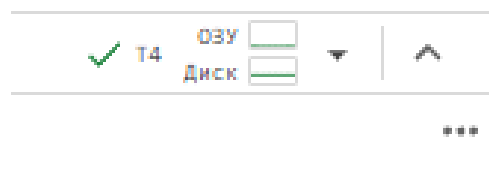
4. Согласиться с предупреждением об окончании сеанса работы



5. Сохранить изменения

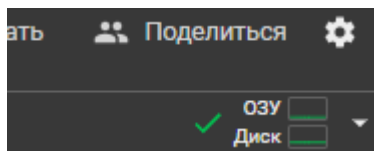


6. Дождаться перезагрузки среды. В верхнем правом углу отобразится панель “ОЗУ ... Диск” и будет указано наименование “T4”

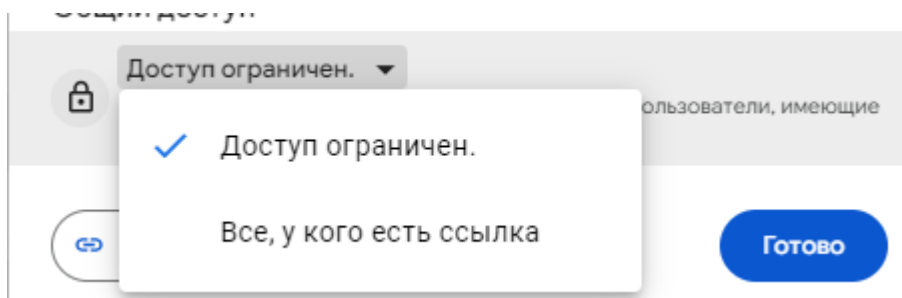


Инструкция по получению ссылки на ноутбук google colab

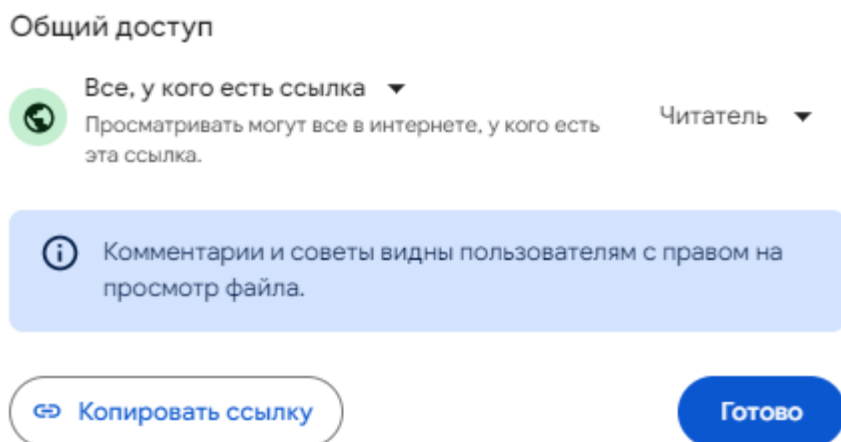
1. Нажмите **“Поделиться”** в правом верхнем углу экрана, рядом с лого вашего google аккаунта



2. В поле **“Общий доступ”** вместо **“Доступ ограничен”** выберите **“Все у кого есть ссылка”**

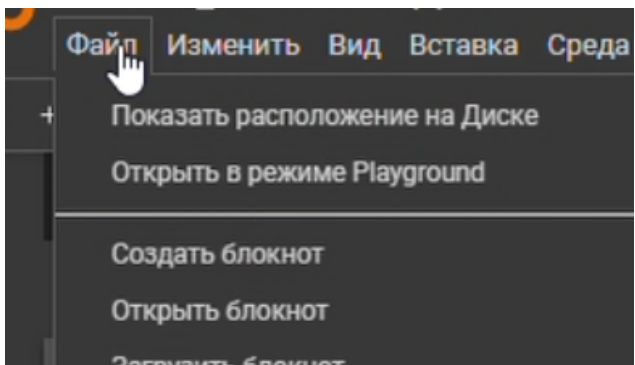


3. Нажмите **“Копировать ссылку”** и вставьте ее в поле ссылки

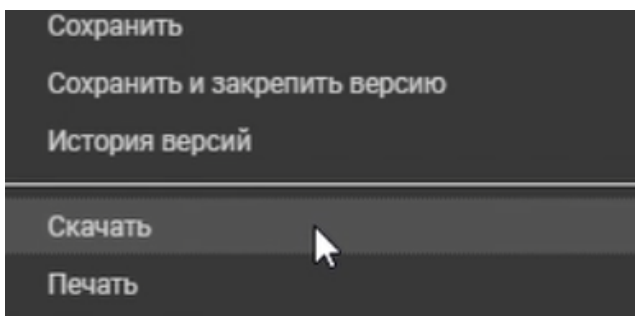


Инструкция по скачиванию файла с google colab

В меню **“Файл”**



Выбрать пункт **“Скачать”**



Выбрать пункт **“IPYNB”**

