**Домашнее задание № 13**

**Лабораторная работа по теме «Решение задач Обработки естественного языка»**

**Автор: Сорокин Максим Евгеньевич**

**Формулировка задания:**

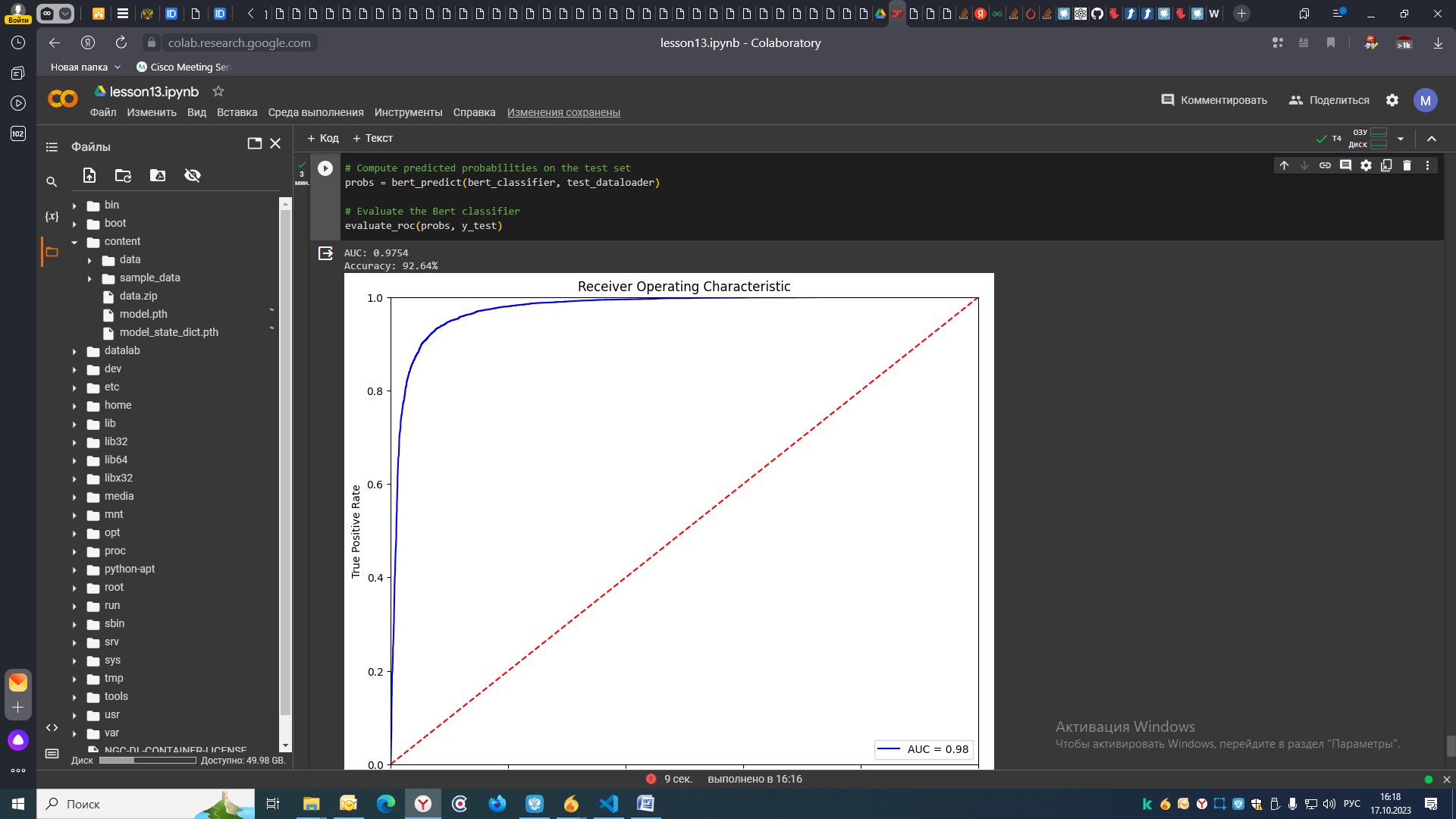
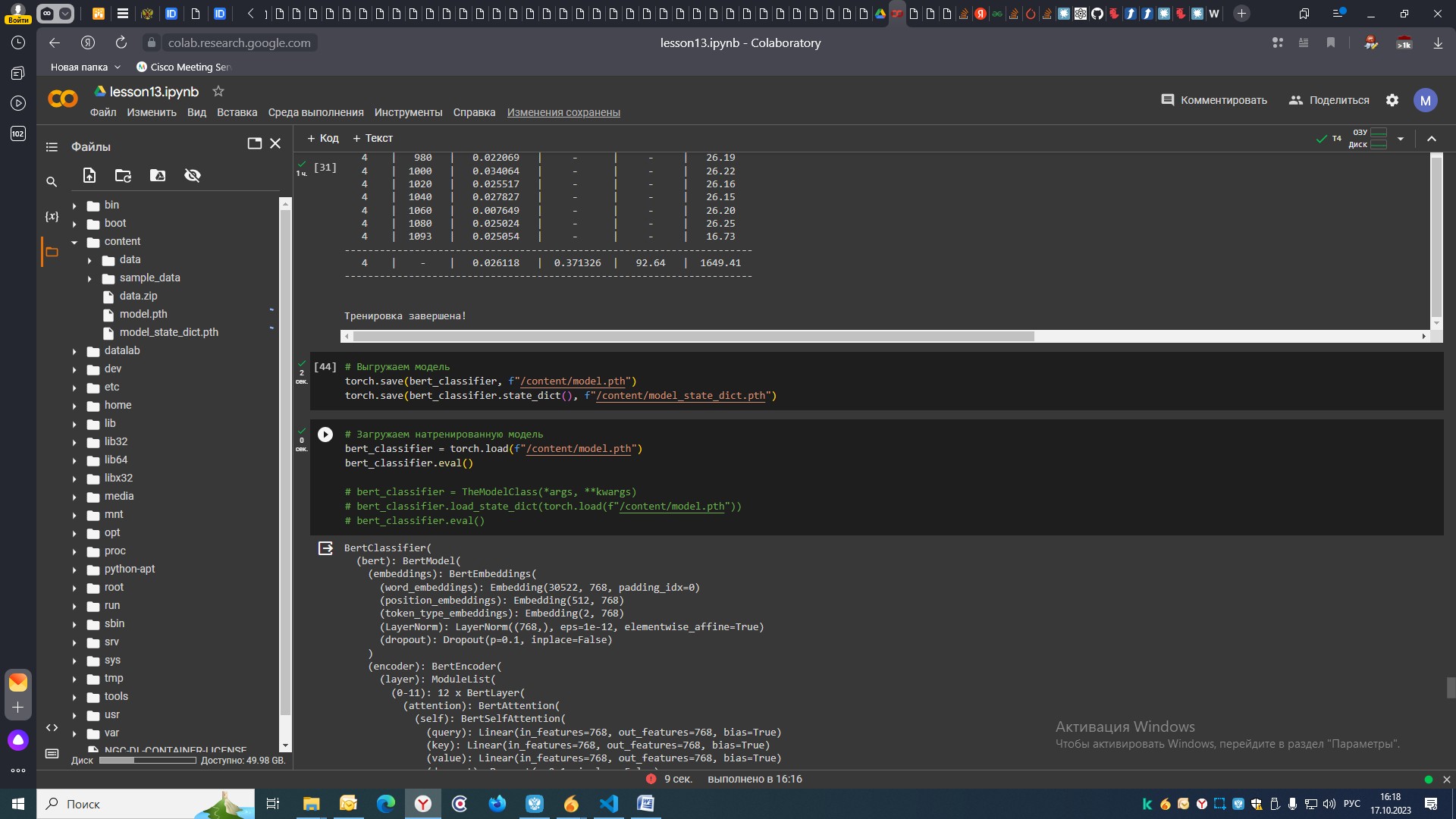
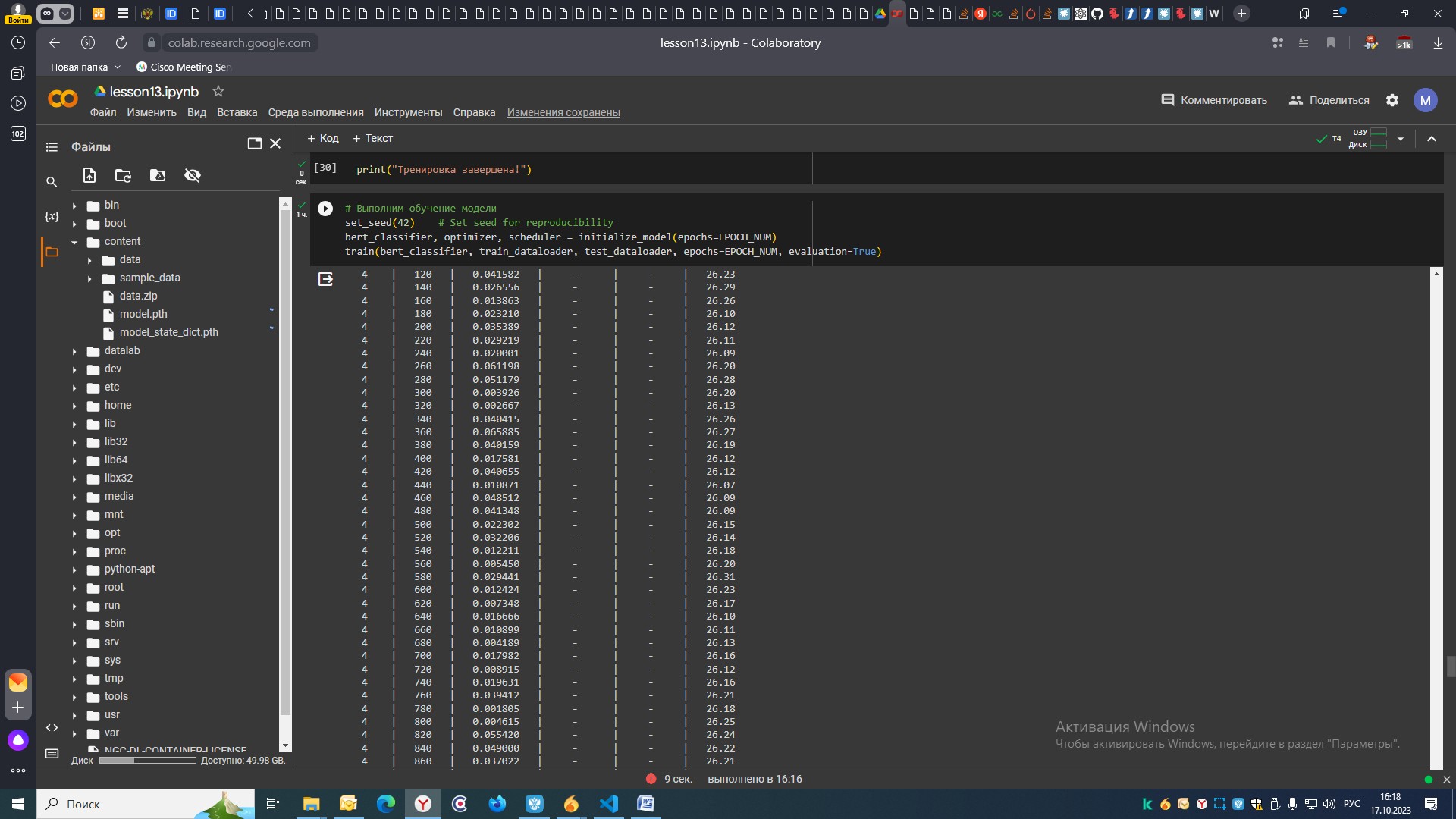
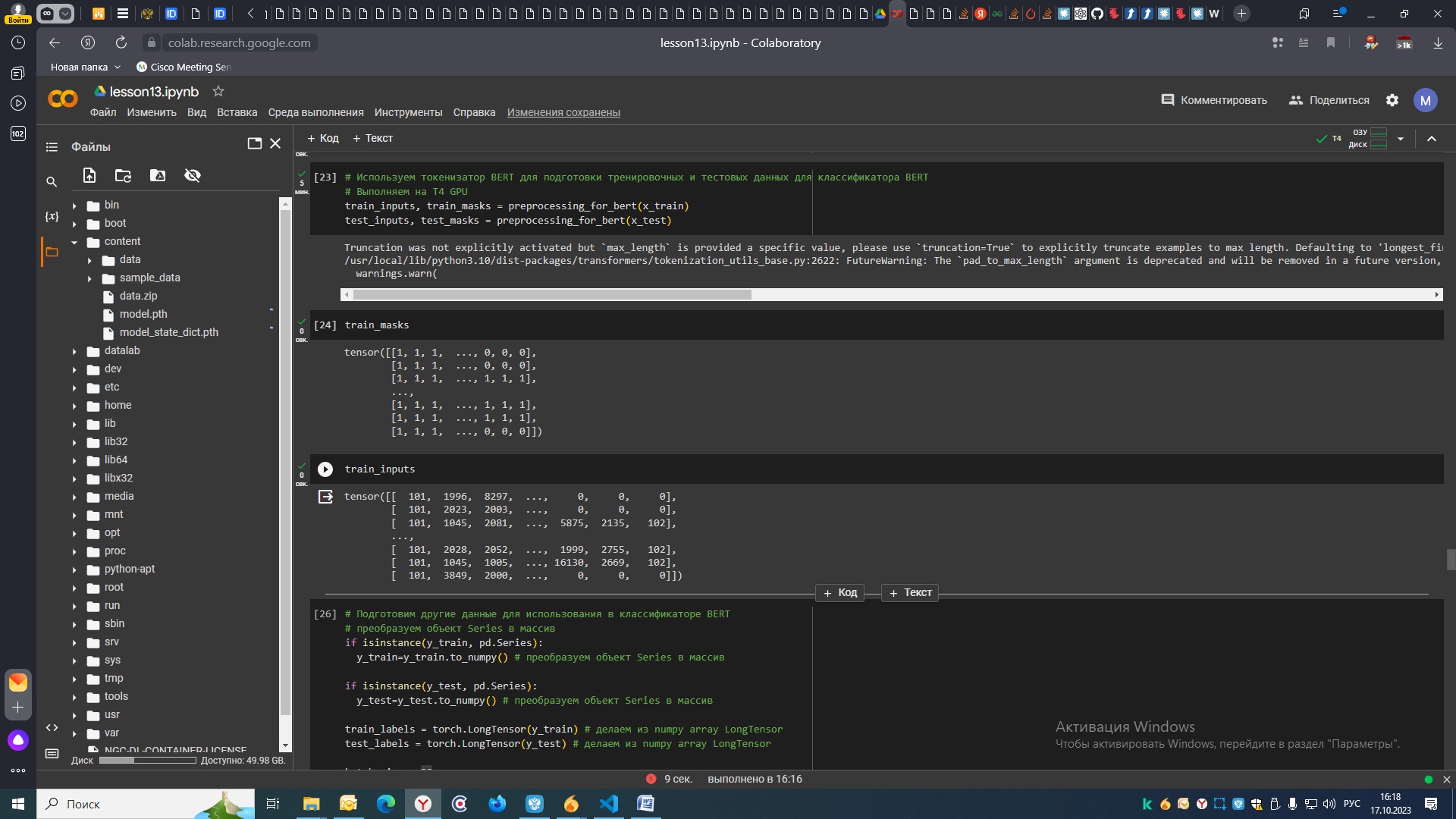
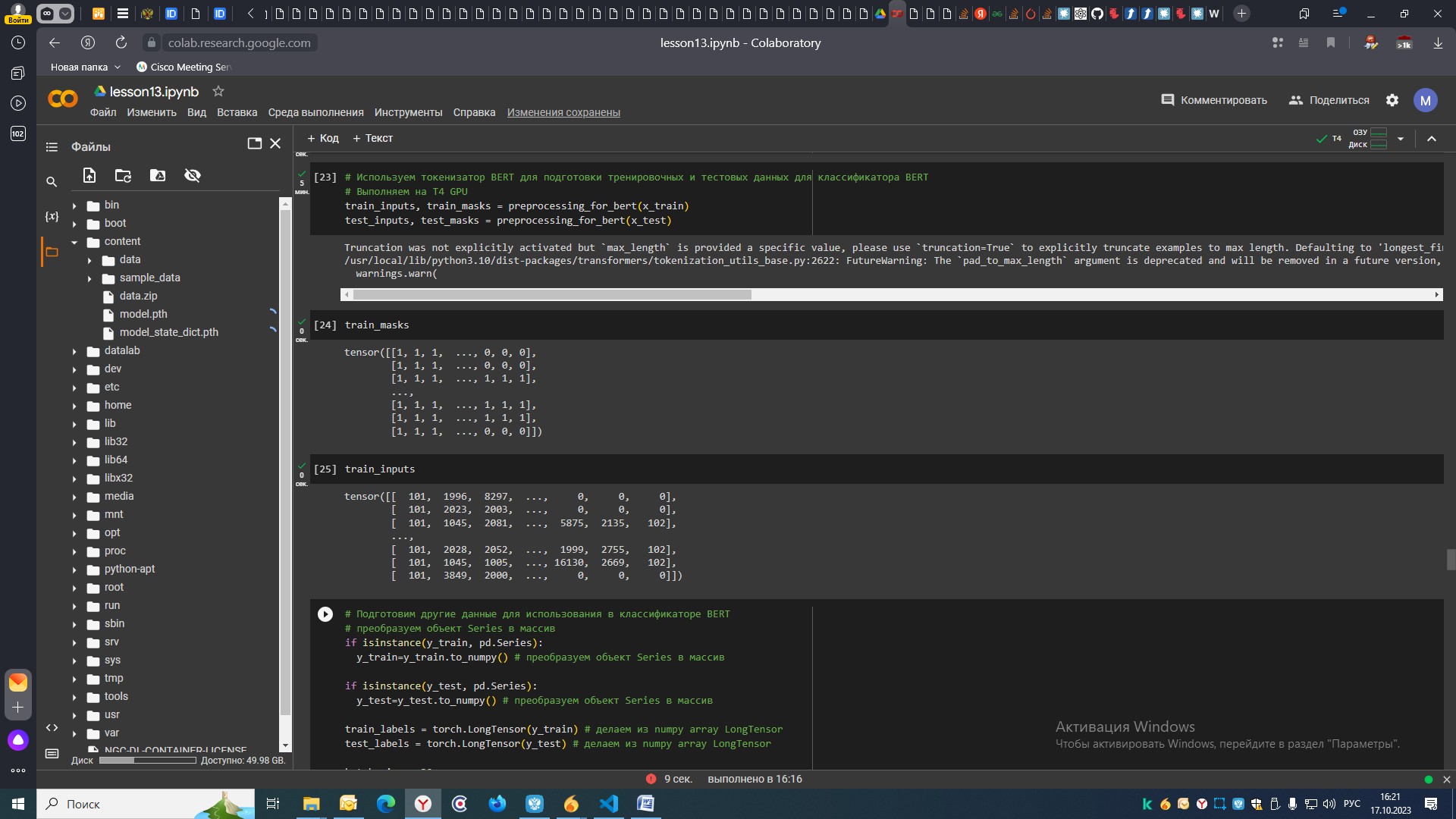
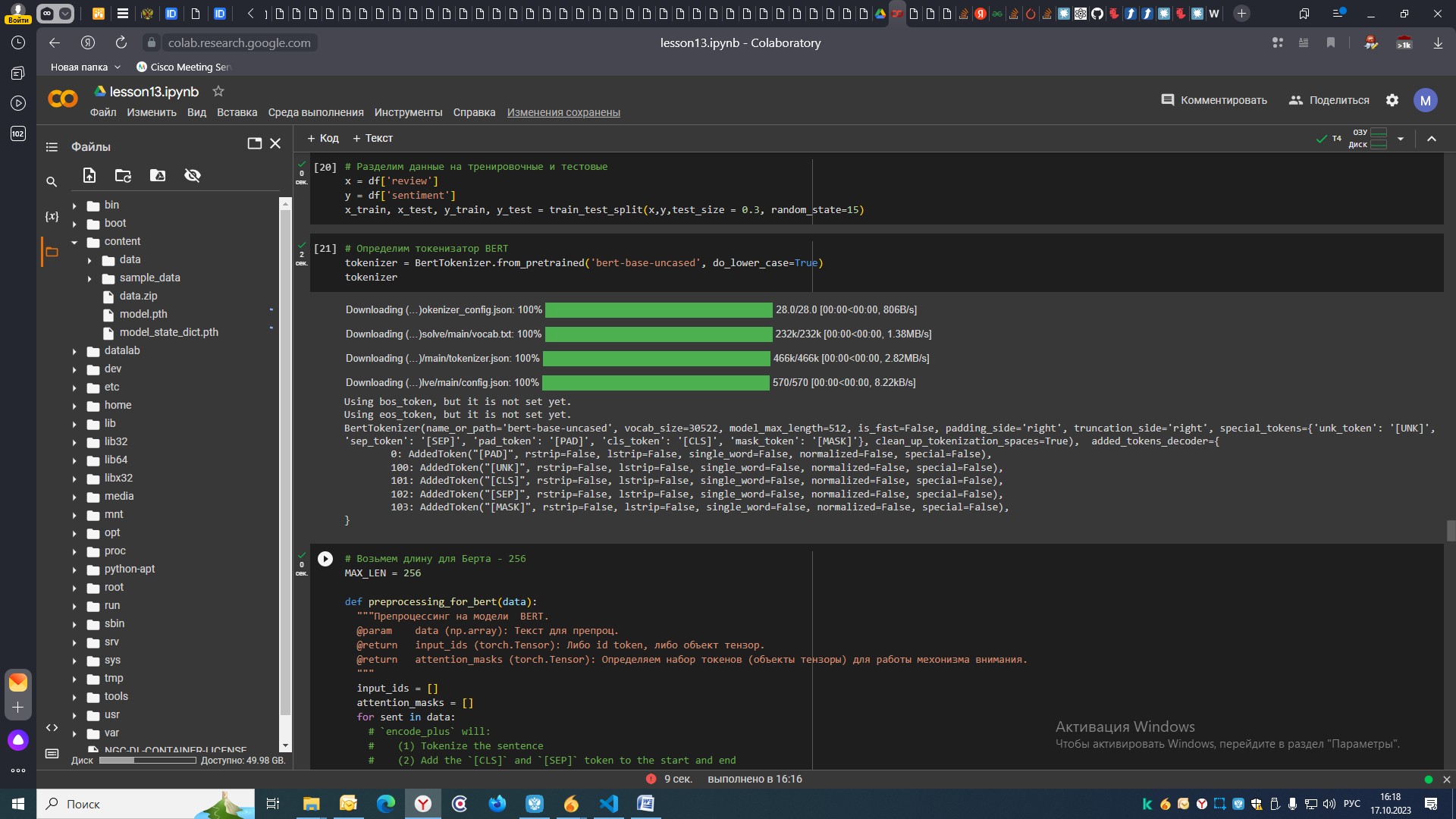
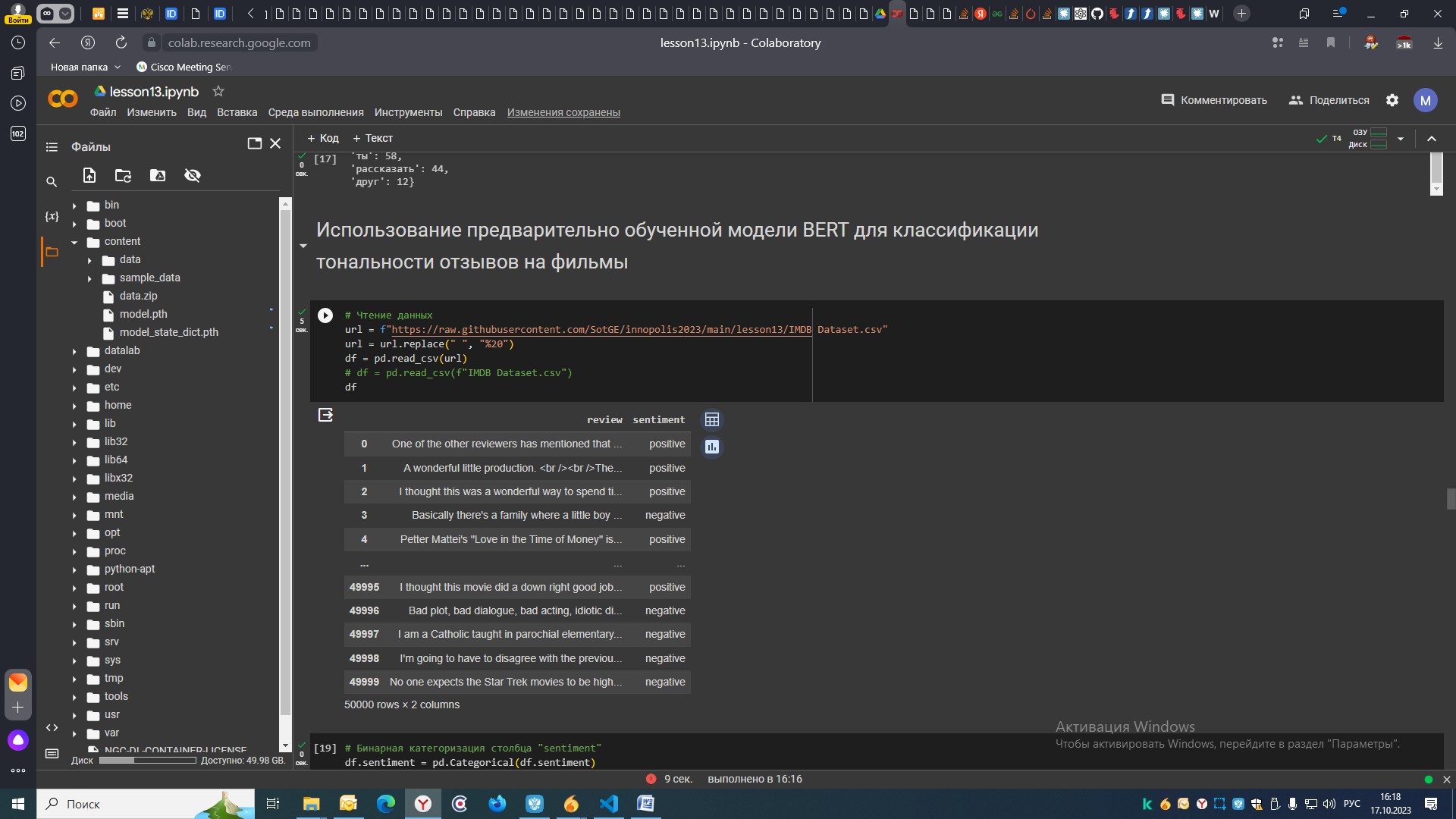
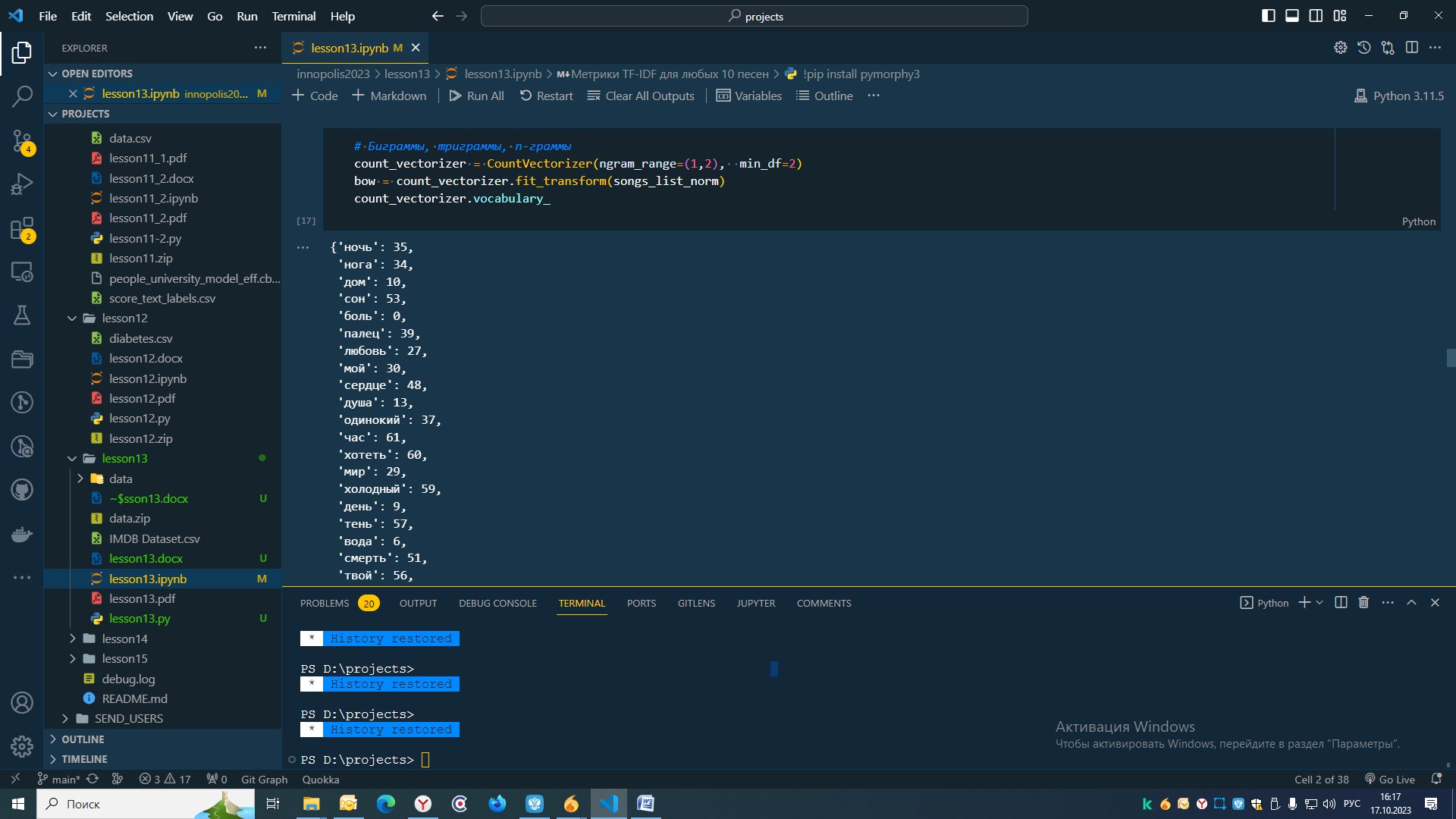
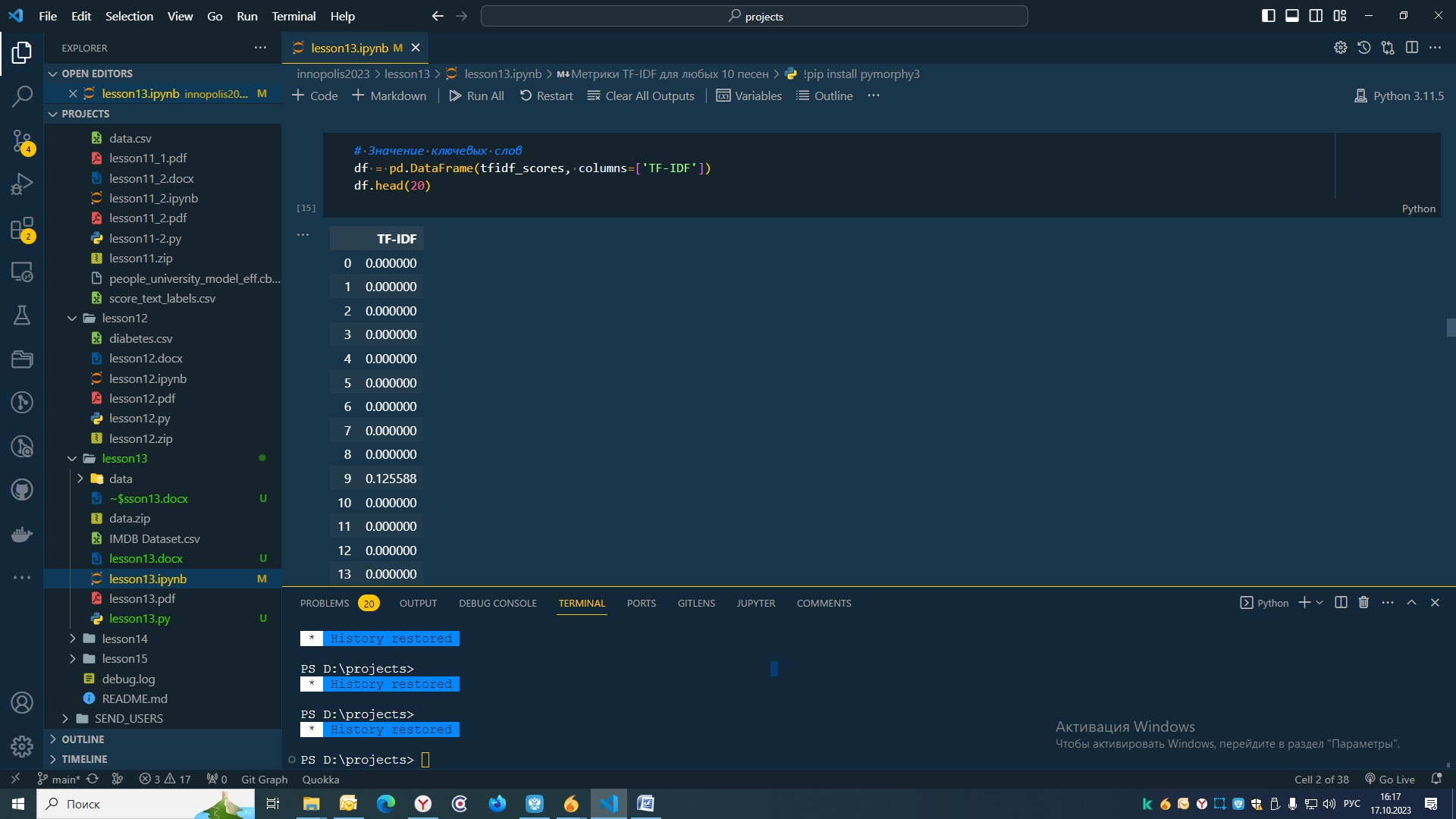
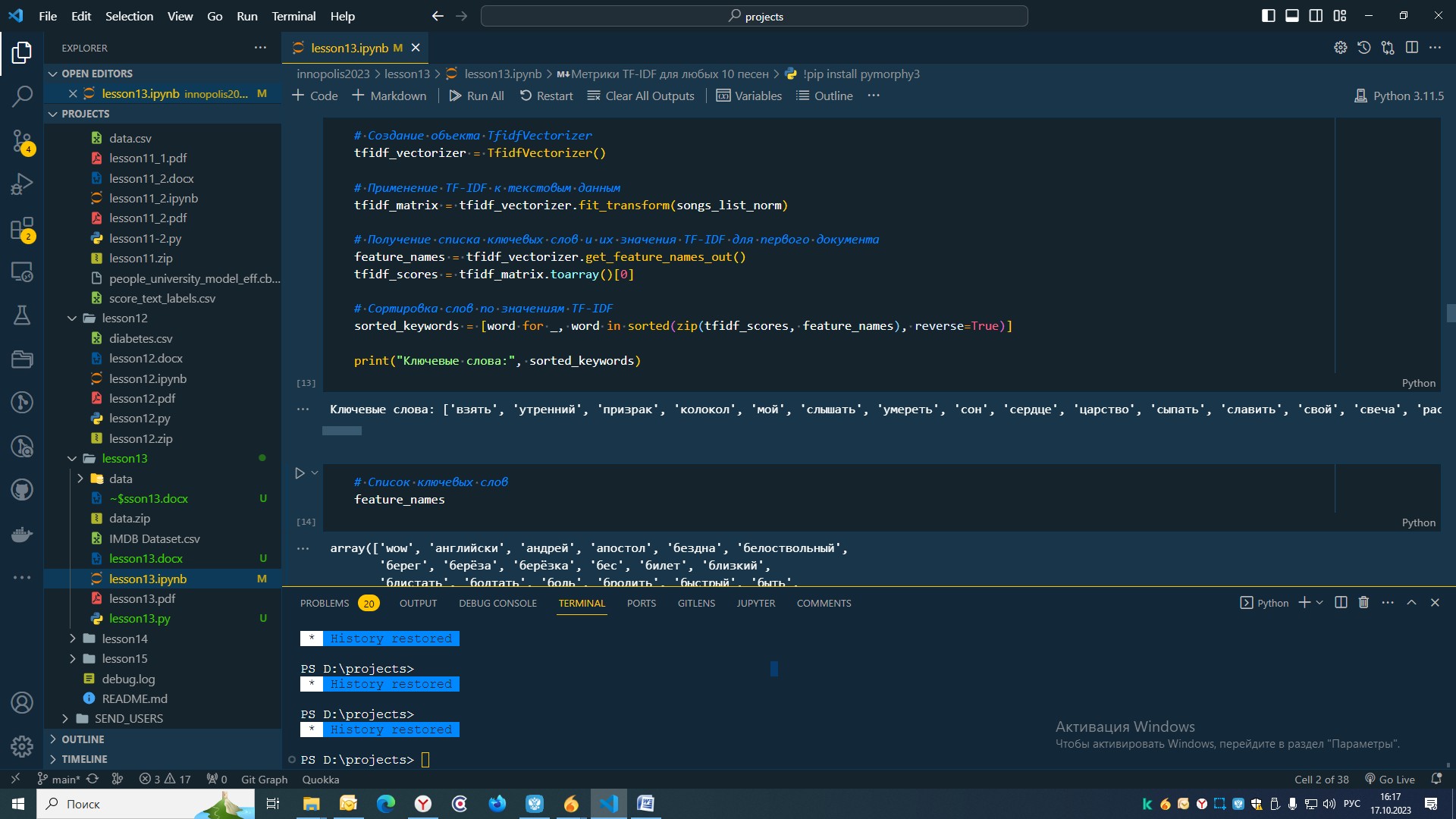
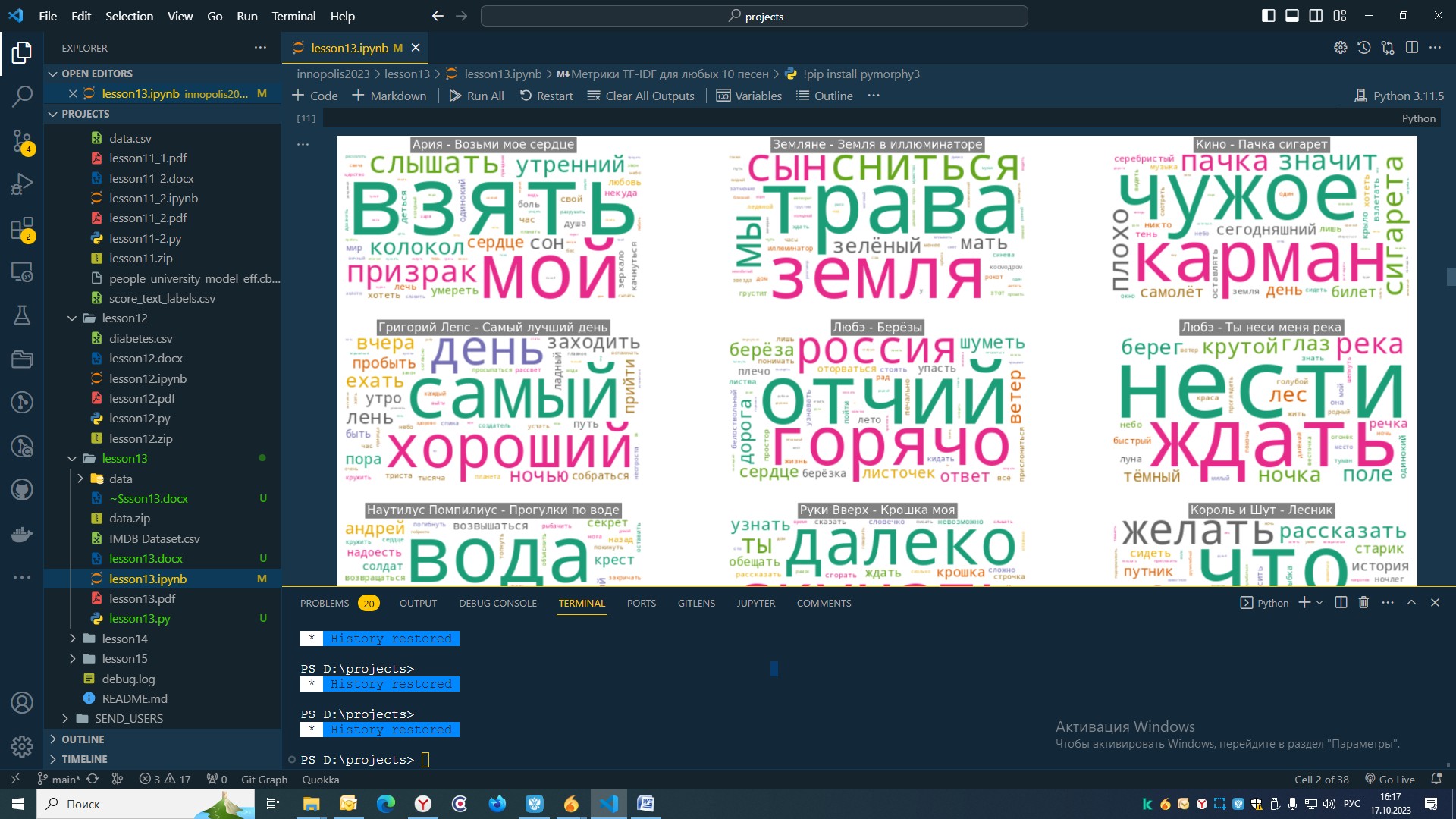
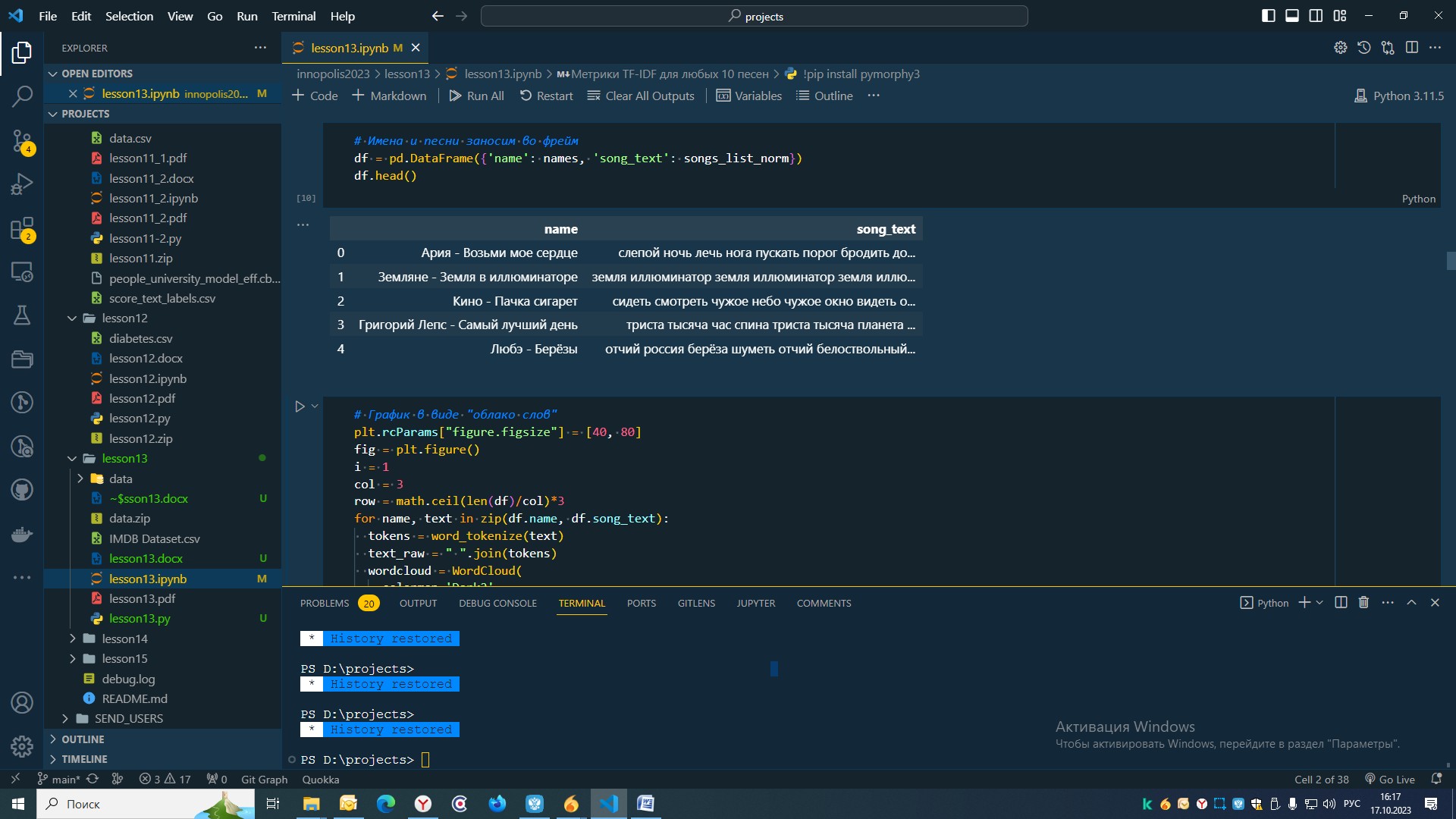
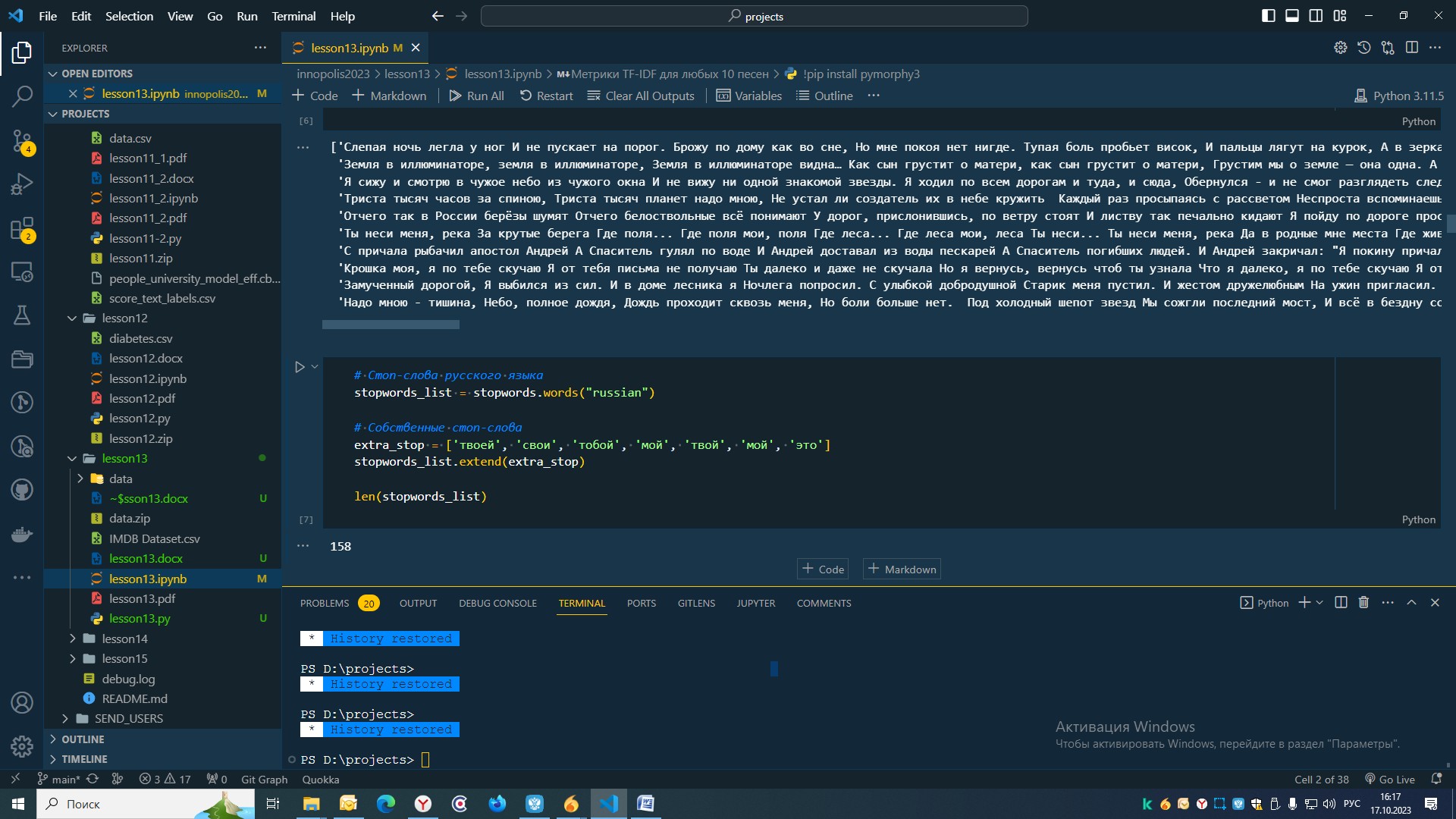
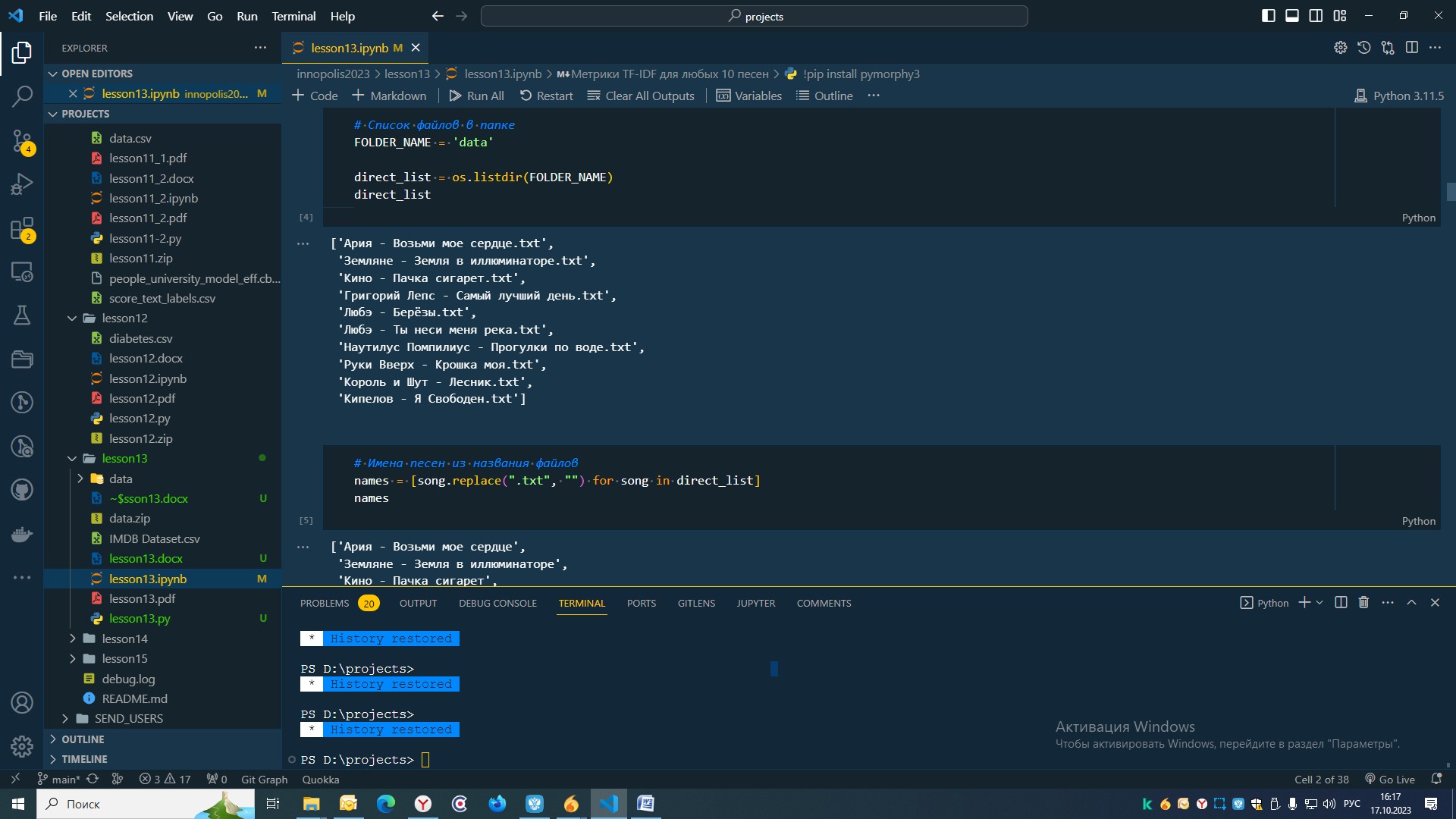
1. Рассчитайте метрики TF-IDF для любых 10 песен на одном языке, которые вы сами выберите.

Не забудьте, что нужно привести слова к начальной форме, убрать стоп-слова.

1. Цель этого задания - использовать предварительно обученную модель BERT для классификации тональности отзывов на фильмы.

* Скачайте датасет отзывов на фильмы. Датасет содержит текст отзыва и бинарную метку тональности (положительный/отрицательный).
* Используйте библиотеку Hugging Face для загрузки предварительно обученной модели BERT и токенизатора.
* Подготовьте данные: используйте токенизатор BERT для преобразования текстовых данных в формат, который можно подать на вход модели BERT.
* Создайте классификатор на основе BERT: это может быть модель BERT с одним линейным слоем для классификации на вершине.
* Обучите классификатор на данных обучения и оцените его производительность на данных для тестирования.

**Решение:**



**К решению приложены файлы:**

«lesson13.ipynb» – для Google Colab

«lesson13.py» – исходный код для Python

«data» – набор из файлов с песнями

«IMDB Dataset.csv» –датасет с отзывами к фильмам

Ссылка на Notebook Colab:

<https://colab.research.google.com/drive/1O_158GPJzEk6eE0KF0KtQa1Kjb_yKt2s?usp=sharing>

Ссылка на GitHub:

<https://github.com/SotGE/innopolis2023/tree/main/lesson13>

Обученная модель:

<https://disk.yandex.ru/d/78FKBHTYEsnizA>