Державний торговельно-економічний університет  
Факультет інформаційних технологій  
Кафедра інженерії програмного забезпечення та кібербезпеки

**ЛАБОРАТОРНА РОБОТА №5**

**З ДИСЦИПЛІНИ «Моделі і структури даних»**

**НА ТЕМУ: «Обробка природної мови. Великі мовні моделі»**

**Виконавець**:

Студентка групи 4-4

Авєріна Наталія Ігорівна

Київ 2024

**Лабораторна робота №5**

**Тема**: «**Обробка природної мови. Великі мовні моделі**»

**Цілі**: ознайомитися з великими мовними моделями (LLM), навчитися використовувати їх для генерації тексту, отримати практичний досвід роботи з платформою Hugging Face.

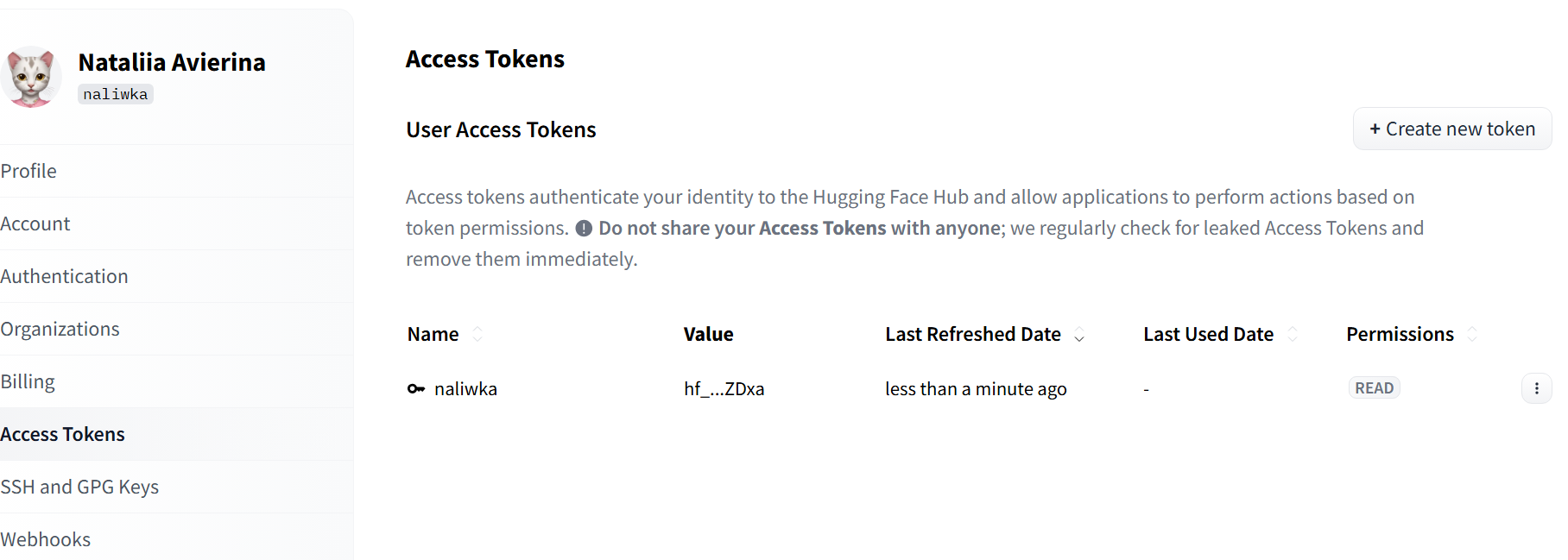
**Хід виконання:**

**Робота в Google Colab:**

<https://colab.research.google.com/drive/1KolSQUUplALkRLVjoLuJB12-OobwH40X?usp=sharing>

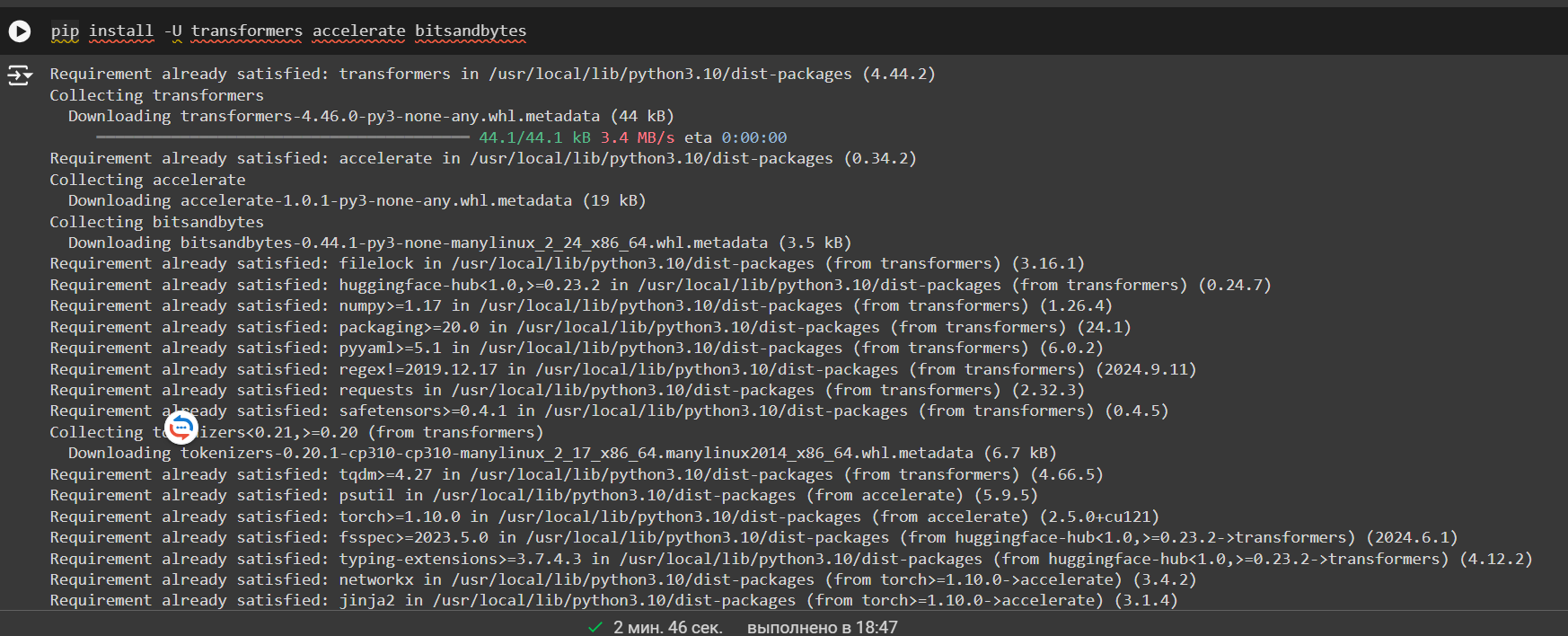
**Завдання 1**

Створюємо обліковий запис на сайті <https://huggingface.co/> та створюємо токен з опцією Read. Зберігаємо токен.



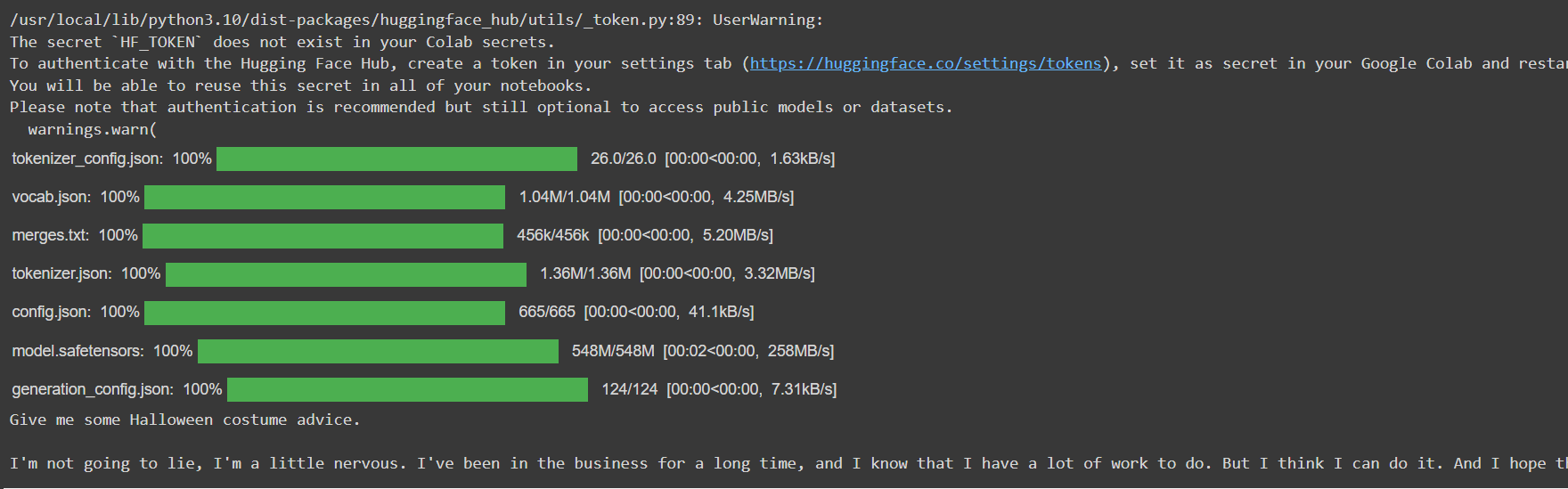
**Завдання 2**

Встановлюємо бібліотеки transformers, accelerate, bitsandbytes.



**Завдання 3**

Використаємо велику мовну модель GPT-2. Попросимо в нього пораду щодо костюму на Хеловін.

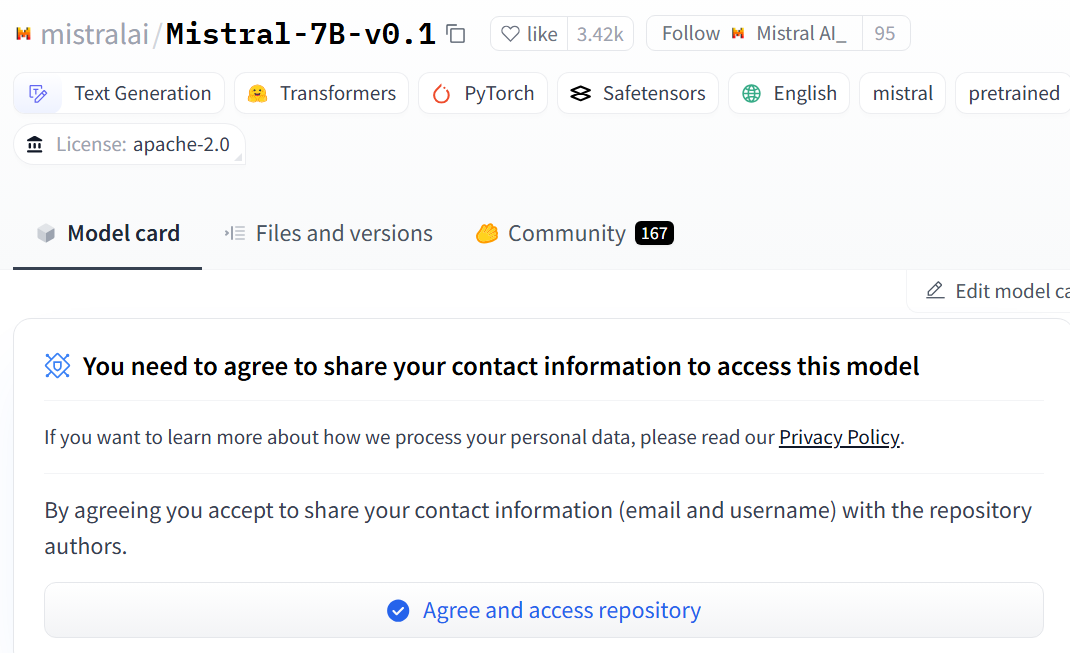


Відповідь GPT-2:

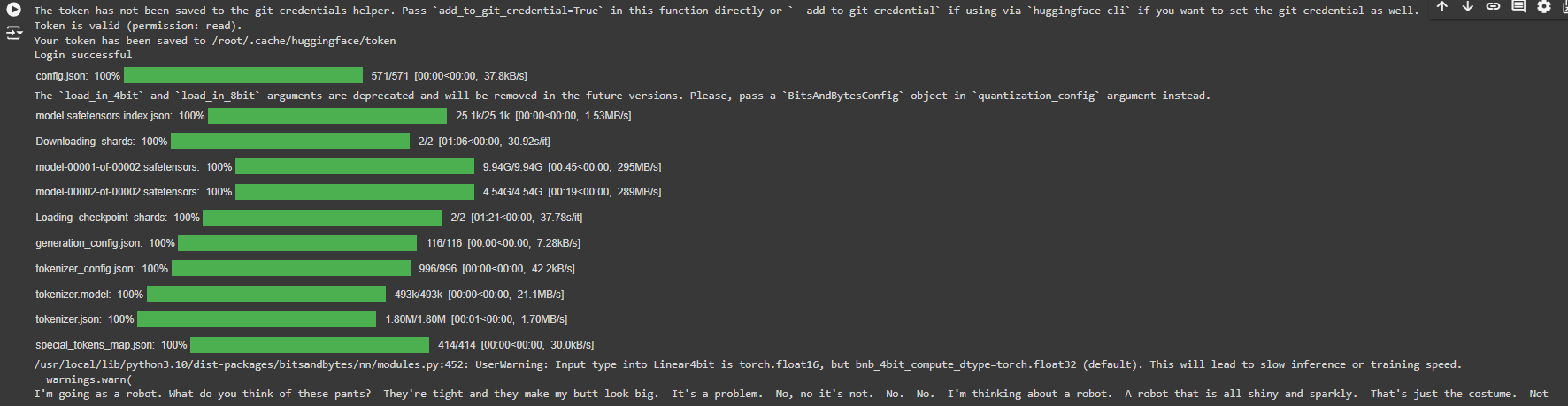
*"I'm not going to lie, I'm a little nervous. I've been in the business for a long time, and I know that I have a lot of work to do. But I think I can do it. And I hope that you'll join me in that effort."*

**Завдання 4**

Спробуємо велику мовну модель Mistral для такої ж цілі. На сторінці <https://huggingface.co/mistralai/Mistral-7B-v0.1> отримаємо доступ до моделі.



Використовуємо створений раніше токен для роботи з моделлю.



Відповідь Mistral:

*"I'm going as a robot. What do you think of these pants? They're tight and they make my butt look big. It's a problem. No, no it's not. No. No. I'm thinking about a robot. A robot that is all shiny and sparkly. That's just the costume. Not"*

**Висновок**:

В ході виконання лабораторної роботи було використано дві LLM для генерації тексту та виявлено сильні і слабкі сторони.

GPT-2 показав загальну здатність до текстової генерації, але його відповідь не відповідала конкретному запиту. Mistral, навпаки, намагався дати відповідь у контексті запиту, що демонструє перевагу новіших моделей у завданнях з високим рівнем контекстуальності. Але Mistral генерує незавершений текст, що вказує на необхідність корекцій під час використання.

**Контрольні питання.**

1. Що таке обробка природної мови і для чого вона використовується?

Це галузь штучного інтелекту, що займається взаємодією комп'ютерів із людською мовою. Мета NLP — навчити машини розуміти, аналізувати, інтерпретувати та генерувати текст так само, як це робить людина.  
Використовується в автоматичному перекладі, голосових асистентах, аналізі соцмереж, чатботах тощо.

1. Що таке Bag of Words та TF-IDF?

Bag of Words – це базова техніка перетворення тексту на числовий вектор. У цьому підході текст розбивається на окремі слова (або токени), а кожне слово фіксується в загальному списку. Далі створюється вектор, що відображає частоти появи слів у документі, при цьому порядок слів ігнорується.

TF-IDF (Term Frequency - Inverse Document Frequency) – це статистична міра, яка показує важливість слова в документі щодо всього корпусу текстів. Вона складається з двох компонентів: частотність терміну і інверсна частота документа (логарифм відношення загальної кількості документів до кількості документів, які містять це слово).

1. У чому відмінність між статистичними методами NLP і сучасними нейронними мережами?

Статистичні методи NLP базуються на ймовірнісному аналізі тексту. Вони розраховують частотність слів і послідовностей слів, але мають обмеження у роботі з довгими контекстами та складними структурами тексту. Такі методи активно використовувалися до 2010-х років.

Сучасні нейронні мережі (RNN, LSTM, трансформери) працюють з текстом на основі глибокого навчання. Вони навчаються розпізнавати складні залежності в даних та враховувати контекст у всьому тексті, що дозволяє їм генерувати більш точні результати. Трансформери (наприклад, GPT, BERT) значно перевершують статистичні моделі завдяки обробці великих обсягів даних і здатності працювати з довгим контекстом.

1. Що таке великі мовні моделі?

Великі мовні моделі – це нейронні моделі, натреновані на величезних обсягах текстових даних для розуміння та генерації природної мови. Вони використовують архітектуру трансформерів, що дозволяє їм враховувати контекст у великих текстових фрагментах та виконувати різноманітні завдання. Приклади LLM: GPT (Generative Pre-trained Transformer), BERT (Bidirectional Encoder Representations from Transformers), T5 (Text-to-Text Transfer Transformer)