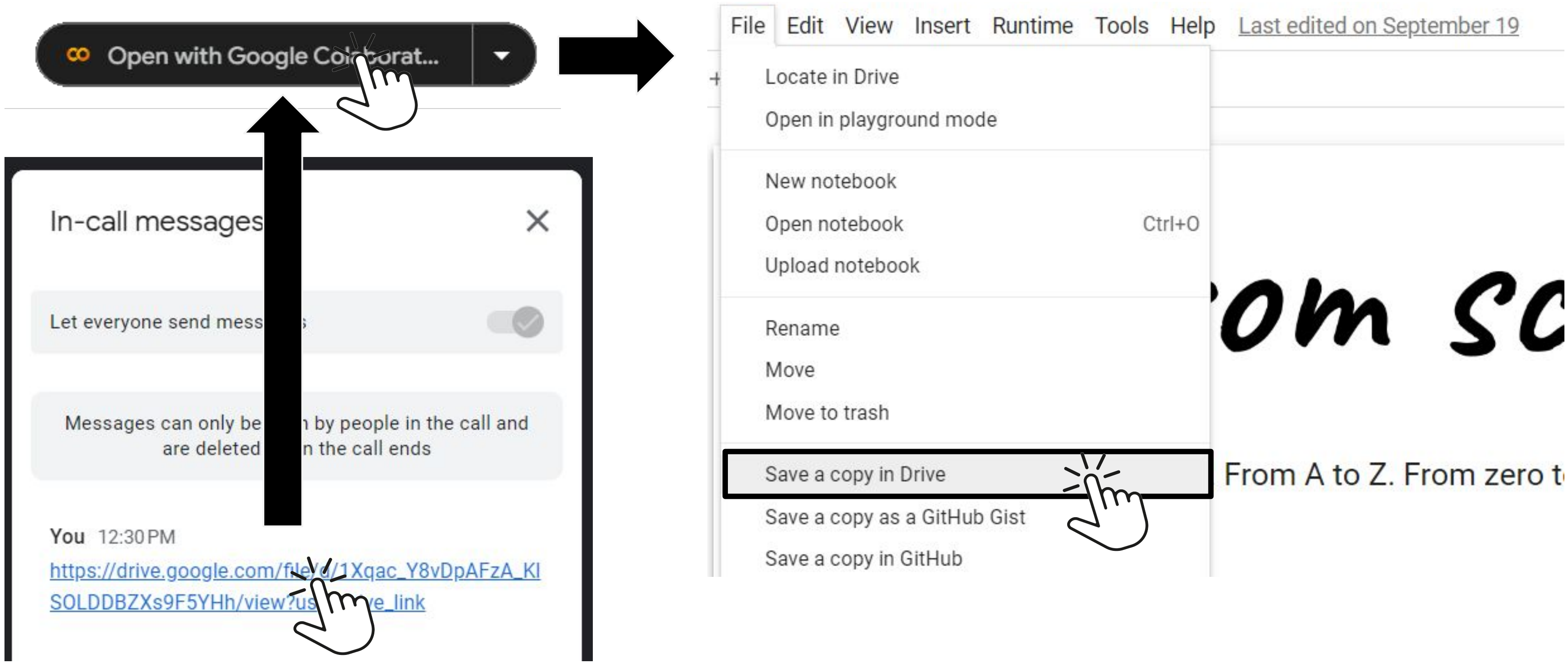
Before we get started:

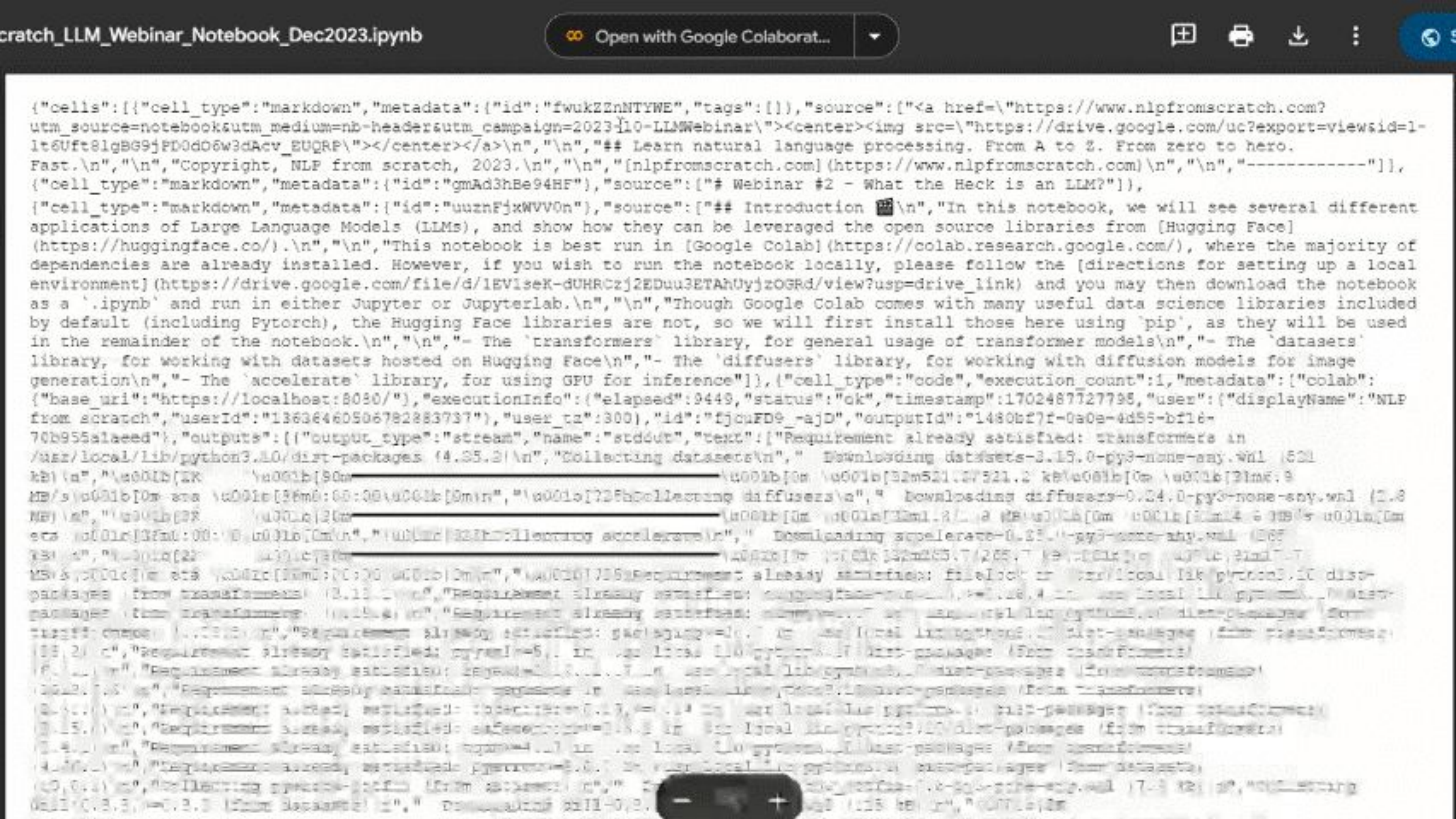


In-call messages

Let everyone send mess s

Messages can only be n by people in the call and are deleted n the call ends

You 12:30 PM
https://drive.google.com/file/d/1Xqac_Y8vDpAFzA_KISOLDDBZXs9F5YHh/view?us ve_link


cratch_LLM_Webinar_Notebook_Dec2023.ipynb
Open with Google Colaborat...

{"cells":[{"cell_type":"markdown","metadata":{"id":"fwukZZnNTYWE","tags":[]},"source":["<a href=\"https://www.nlpfromscratch.com?
utm_source=notebook&utm_medium=nb-header&utm_campaign=2023-10-LLMWebinar\"><center><img src=\"https://drive.google.com/uc?export=view&id=1-
1t6Uft8lgBG9jPD0dO6w3dAcv_EUQRP\"></center></a>\n","\n","## Learn natural language processing. From A to Z. From zero to hero.
Fast.\n","\n","Copyright, NLP from scratch, 2023.\n","\n","[nlpfromscratch.com](https://www.nlpfromscratch.com)\n","\n","------------"]},
{"cell_type":"markdown","metadata":{"id":"gmAd3hBe94HF"},"source":["# Webinar #2 - What the Heck is an LLM?"]},
{"cell_type":"markdown","metadata":{"id":"uuznFjxWVV0n"},"source":["## Introduction 🎬\n","In this notebook, we will see several different
applications of Large Language Models (LLMs), and show how they can be leveraged the open source libraries from [Hugging Face]
(https://huggingface.co/).\n","\n","This notebook is best run in [Google Colab](https://colab.research.google.com/), where the majority of
dependencies are already installed. However, if you wish to run the notebook locally, please follow the [directions for setting up a local
environment](https://drive.google.com/file/d/1EV1seK-dUHRCzj2EDuu3ETAhUyjzOGRd/view?usp=drive_link) and you may then download the notebook
as a `.ipynb` and run in either Jupyter or Jupyterlab.\n","\n","Though Google Colab comes with many useful data science libraries included
by default (including Pytorch), the Hugging Face libraries are not, so we will first install those here using `pip`, as they will be used
in the remainder of the notebook.\n","\n","- The `transformers` library, for general usage of transformer models\n","- The `datasets`
library, for working with datasets hosted on Hugging Face\n","- The `diffusers` library, for working with diffusion models for image
generation\n","- The `accelerate` library, for using GPU for inference"]},{"cell_type":"code","execution_count":1,"metadata":{"colab":
{"base_uri":"https://localhost:8080/"},"executionInfo":{"elapsed":9449,"status":"ok","timestamp":1702487727795,"user":{"displayName":"NLP
from scratch","userId":"13636460506782883737"},"user_tz":300},"id":"fjcuFD9_-ajD","outputId":"1480bf7f-0a0e-4d55-bf16-
70b955a1aeed"},"outputs":[{"output_type":"stream","name":"stdout","text":["Requirement already satisfied: transformers in
/usr/local/lib/python3.10/dist-packages (4.35.2)\n","Collecting datasets\n"," Downloading datasets-2.15.0-py3-none-any.whl (521
kB)\n","\u001b[2K   \u001b[90m━━\u001b[0m \u001b[32m521.2/521.2 kB\u001b[0m \u001b[31m6.9
MB/s\u001b[0m eta \u001b[36m0:00:00\u001b[0m\n","\u001b[?25hCollecting diffusers\n"," Downloading diffusers-0.24.0-py3-none-any.whl (1.8
MB)\n","\u001b[2K   \u001b[90m━━\u001b[0m \u001b[32m1.8/1.8 MB\u001b[0m \u001b[31m24.6 MB/s\u001b[0m
eta \u001b[36m0:00:00\u001b[0m\n","\u001b[?25hCollecting accelerate\n"," Downloading accelerate-0.25.0-py3-none-any.whl (265
kB)\n","\u001b[2K   \u001b[90m━━\u001b[0m \u001b[32m265.7/265.7 kB\u001b[0m [illegible]
MB/s\u001b[0m eta \u001b[36m0:00:00\u001b[0m\n","\u001b[?25hRequirement already satisfied: filelock in /usr/local/lib/python3.10/dist-
packages (from transformers) (3.13.1)\n","Requirement already satisfied: [illegible] in /usr/local/lib/python3.10/dist-
packages (from transformers) [illegible]\n","Requirement already satisfied: [illegible] in /usr/local/lib/python3.10/dist-packages (from
transformers) [illegible]\n","Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from transformers)
(23.2)\n","Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-packages (from transformers)
[illegible]\n","Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.10/dist-packages (from transformers)
[illegible]\n","Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (from transformers)
[illegible]\n","Requirement already satisfied: tokenizers<0.19,>=0.14 in /usr/local/lib/python3.10/dist-packages (from transformers)
(0.15.0)\n","Requirement already satisfied: safetensors>=0.3.1 in /usr/local/lib/python3.10/dist-packages (from transformers)
[illegible]\n","Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.10/dist-packages (from transformers)
[illegible]\n","Requirement already satisfied: pyarrow>=8.0.0 in /usr/local/lib/python3.10/dist-packages (from datasets)
[illegible]\n","Collecting pyarrow-hotfix (from datasets)\n"," [illegible] pyarrow_hotfix-0.6-py3-none-any.whl (7.9 kB)\n","Collecting
dill<0.3.8,>=0.3.0 (from datasets)\n"," Downloading dill-0.3. [illegible] .whl (115 kB)\n","[illegible]

NLP from scratch

From Zero to NLP in 60

Monthly Webinar

www.nlpfromscratch.com

Housekeeping

- Camera on if comfortable doing so
- This meeting will not be recorded
- Stay muted unless speaking
- Be professional

Who am I?

- Data Scientist
- Career consultant (SapientNitro, PwC, Accenture)
- Trainer
- Human

ChatGPT

[illegible]

[illegible]

[illegible]

[illegible]

Capabilities

Remembers what user said earlier in the conversation

Allows user to provide follow-up corrections

Trained to decline inappropriate requests

Limitations

May occasionally generate incorrect information

May occasionally produce harmful instructions or biased content

What is Natural Language Processing?

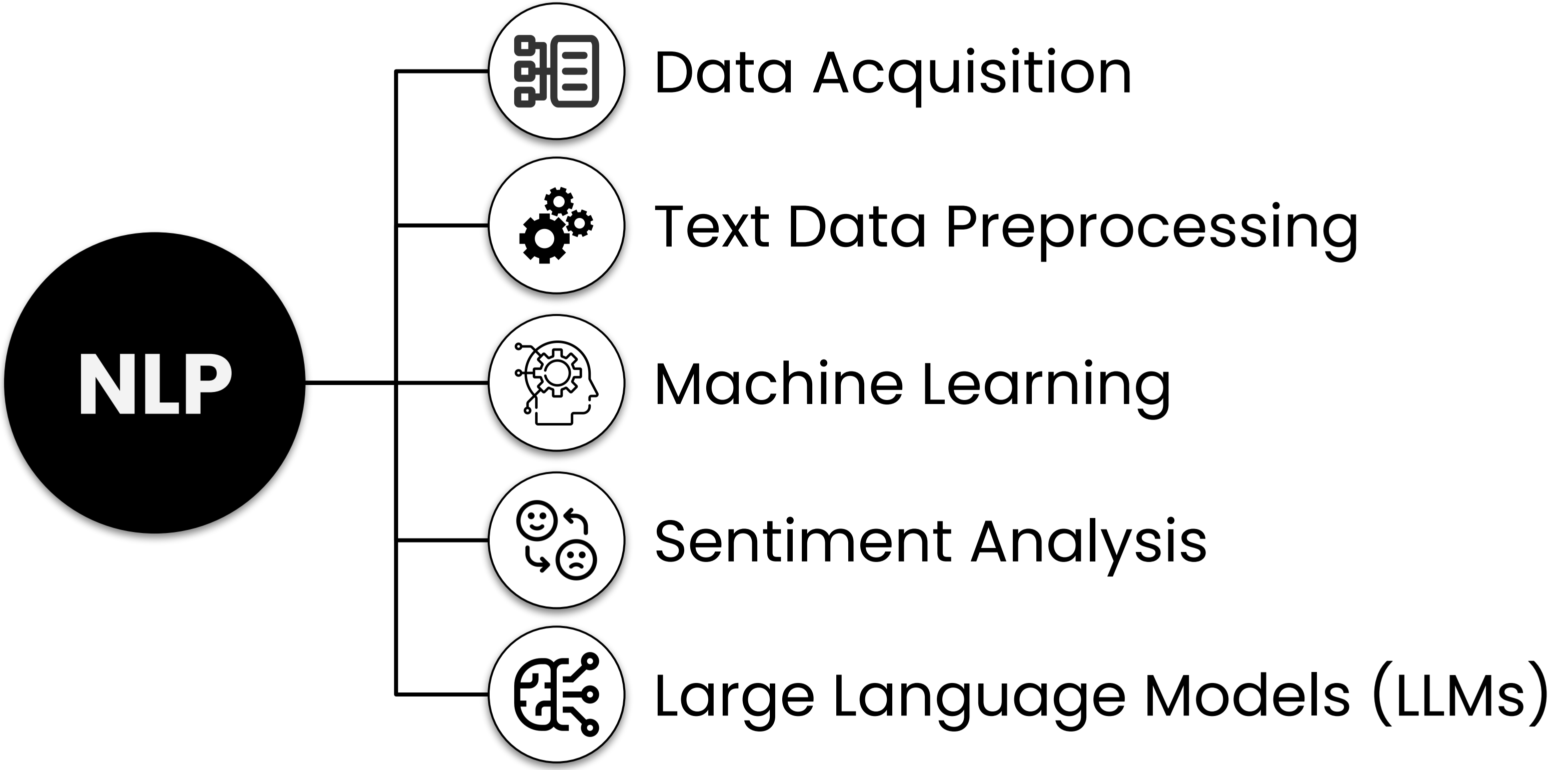
Natural language processing (NLP) is situated at the intersection of the fields of computational linguistics, computer science, and artificial intelligence.

Within the context of data science and AI, NLP aims to enable computers to work with - and potentially even understand - human language for various tasks.

Recently NLP has very much come to the forefront of AI within the popular consciousness, giving the popularity of generative text applications such as ChatGPT.

Today we will explore some of what is possible in natural language processing with building from a simple example of data acquisition and processing to fitting a machine learning model.

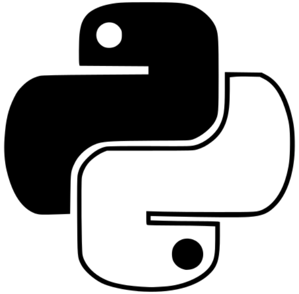
Today’s Coverage



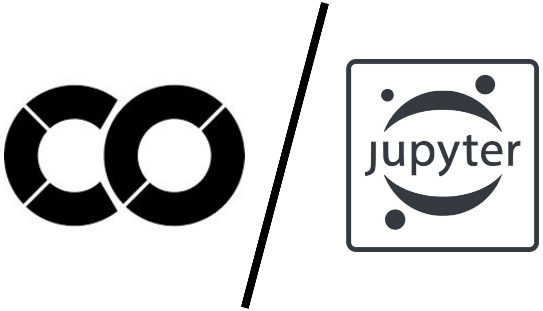
NLP

- Data Acquisition
- Text Data Preprocessing
- Machine Learning
- Sentiment Analysis
- Large Language Models (LLMs)

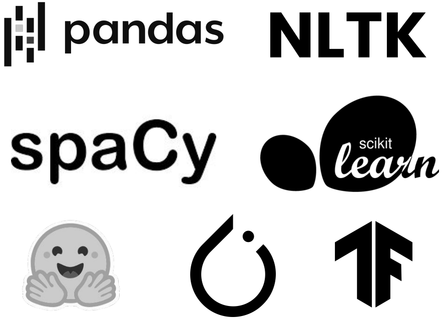
Tools of the Trade



Python 3



Google Colab
/ Jupyter



NLP and DS Libraries

Python Fundamentals

- Python is a powerful programming language and has become the de facto standard for doing data science work (and a majority of NLP)
- It is easy to learn even for the non-technical or those without prior programming experience
- Working with text at a fundamental level is built into base python and modules in the **pandas library**
- For natural language processing, there is a wide array of free, open source libraries that cover a wide variety of use cases and natural language processing tasks

Pandas Library

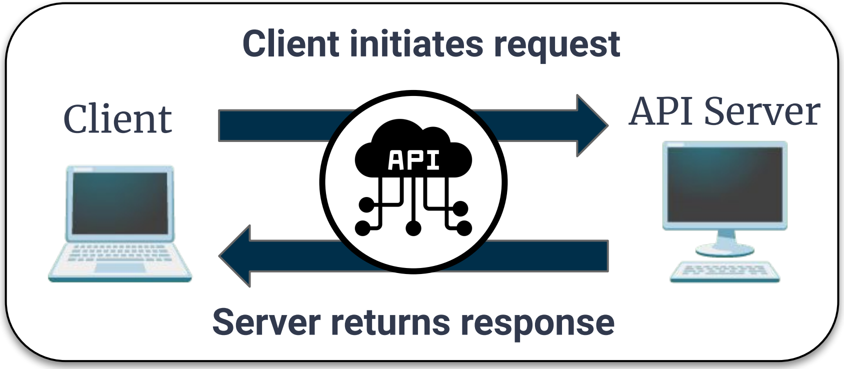
- Pandas is the core library for data manipulation in python and part of the "data science stack"
- Stores data in abstractions of columns (*Series*) and tables of data (*DataFrames*)
- Fast and optimized for working with datasets row-wise with array operations
- For NLP, built-in string accessors for doing text data manipulation on columns easily

Data Acquisition

- We can acquire data either from **requesting it directly from an API**, or extracting that locked in websites by doing **web scraping.**
- This usually requires writing code or using software or a third-party service designed for this task
- A **web service** is an application running on a computer which can provide data or perform transactions when you interact with over the wire
- The machine hosting the service is referred to as a **server** and a machine interacting with it a **client**

Data Acquisition - API vs. Web Scraping

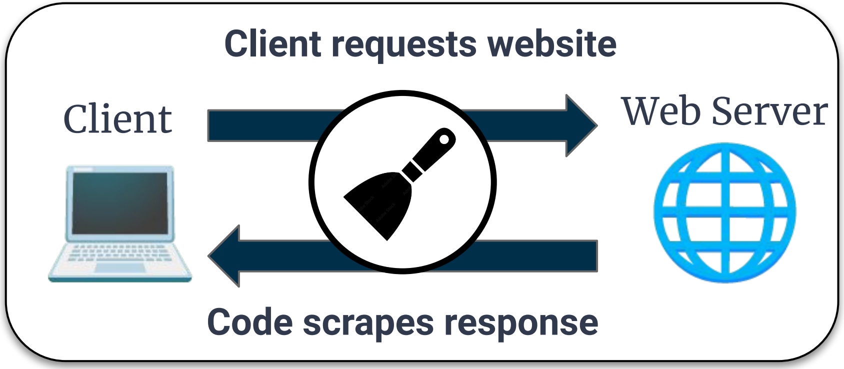
From an API



Client initiates request

Server returns response

Web scraping



Client requests website

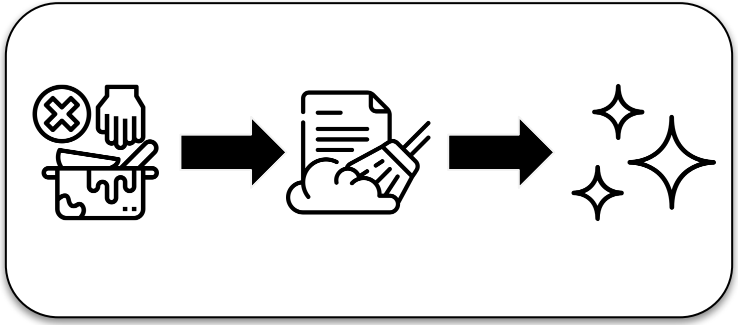
Code scrapes response

Text Preprocessing

- Refers to cleaning and transforming the original text data into **structured data** to make it suitable for machine learning (or other uses).

- While some preprocessing steps may need to be considered carefully as they may be specific to your use case, there are developed **standard approaches** that work well

- This kind of preprocessing is for traditional NLP approaches; **cutting-edge machine learning methods have their own preprocessing methods** which are more advanced (*e.g.* deep learning, LLMs)



Text Preprocessing Steps

Tokenization

Break free-form text documents down into tokens: constituent units of language (usually words)

Normalization

Apply techniques to reduce the noise and variance in the language data and standardize

Vectorization

Convert text data to numeric features: structured data suitable for machine learning or analytics

NLTK and scikit-learn

- NLTK is the *natural language toolkit*, a free open source python library for working with NLP
- Originally developed at University of Pennsylvania for teaching purposes but now standard in NLP workflows
- Scikit-learn (sklearn for short) is the standard open source library for machine learning in Python
- Covers the complete gamut of ML and also has modules and classes for text-specific tasks (datasets, vectorization, metrics, etc.)

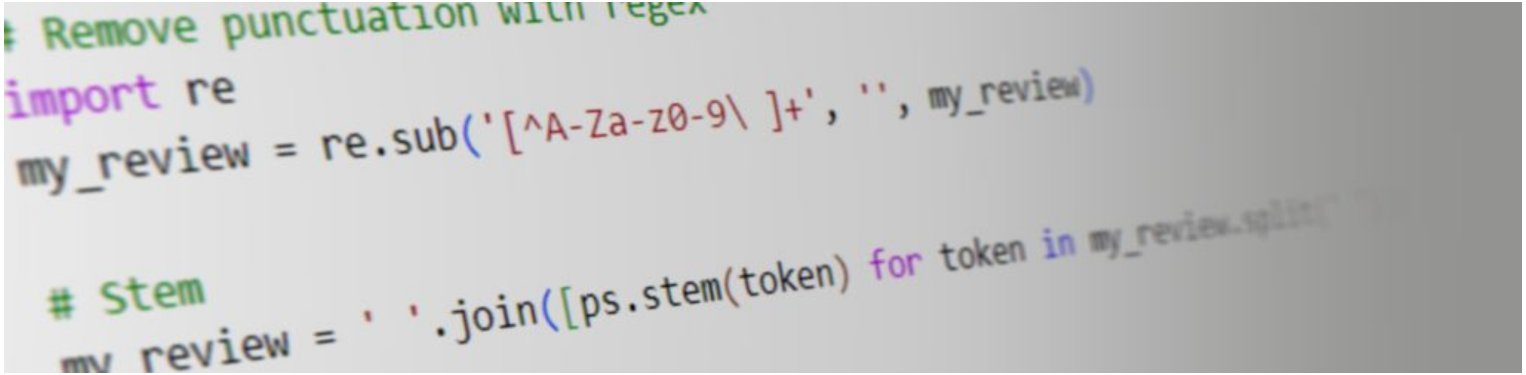
Machine Learning for Sentiment Analysis

- In **supervised learning**, we use our set of input features, X, and an associated set of data labels, y, to train a model to make predictions about unseen data.
- For natural language, a common use case is for that of **sentiment analysis** - predicting the emotional quality and strength of strings of text based on existing labelled data
- Though they are not explicitly trained for this task, given their generalizability, large language models (LLMs) **can perform sentiment analysis tasks** as a result of "zero-shot" learning.
- Here we will be testing Meta's LLaMA 2 chat model which was released in July 2023 and interacting with the through a Hugging Face space

NLP4Free 🔤⚡🤖🧠😃

https://github.com/nlpfromscratch/nlp4free

A Free Natural Language Processing (NLP) microcourse, from basics to deep learning



```
# Remove punctuation with regex
import re
my_review = re.sub('[^A-Za-z0-9\ ]+', '', my_review)

 # Stem
 my_review = ' '.join([ps.stem(token) for token in my_review.split(
```

LLM and Generative AI Workshops

Generative Text Models & Fine-tuning LLMs

- Intro to Hugging Face
- LLMs for generative text
- Fine-tuning models
- PEFT (LoRA) and quantization

Building GenAI Apps with OpenAI and GPT

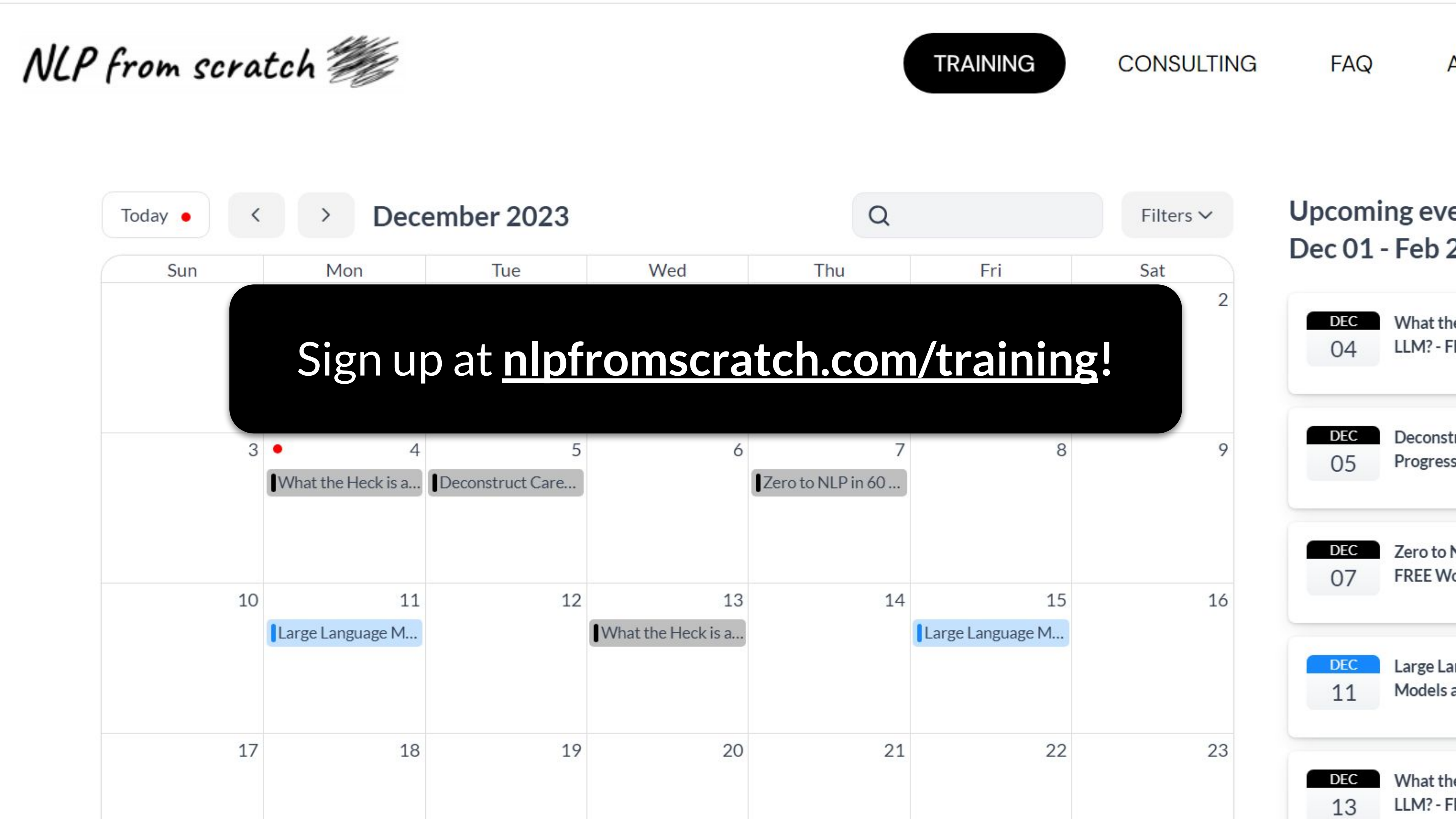
- Intro to APIs and OpenAI
- Working with the OpenAI API
- Setting up a dev environment
- Build a streaming chat app
- Open source alternatives to GPT

Intro to Python & Natural Language Processing

- Intro to NLP
- Intro to Python
- Working with text in Python
- Manipulating text in Pandas

GenAI for Work (1 hr)

- Introduction to GenAI
- Generative AI landscape
- Everyday use cases for GenAI tools
- Prompting and prompt engineering
- Productivity with GenAI



Sign up at nlpfromscratch.com/training!

December 2023

Sun	Mon	Tue	Wed	Thu	Fri	Sat
						2
3	● 4 What the Heck is a...	5 Deconstruct Care...	6	7 Zero to NLP in 60 ...	8	9
10	11 Large Language M...	12	13 What the Heck is a...	14	15 Large Language M...	16
17	18	19	20	21	22	23

Upcoming eve
Dec 01 - Feb 2

DEC 04 What the
LLM? - F

DEC 05 Deconst
Progress

DEC 07 Zero to N
FREE Wo

DEC 11 Large La
Models a

DEC 13 What the
LLM? - F

Manifesto

Knowledge is only valuable if it is useful.

The best way to learn is by doing.

Learning is a non-linear process.

Learning is exploration, not a journey.

Teaching and learning are complementary.

I would value your feedback.

NLP from scratch

Learn natural language processing. From A to Z. From zero to hero. Fast.

www.nlpfromscratch.com