Modern NLP

Based on Deep Learning and Language models.
Day 2 Afternoon

Alexandre Gazagnes

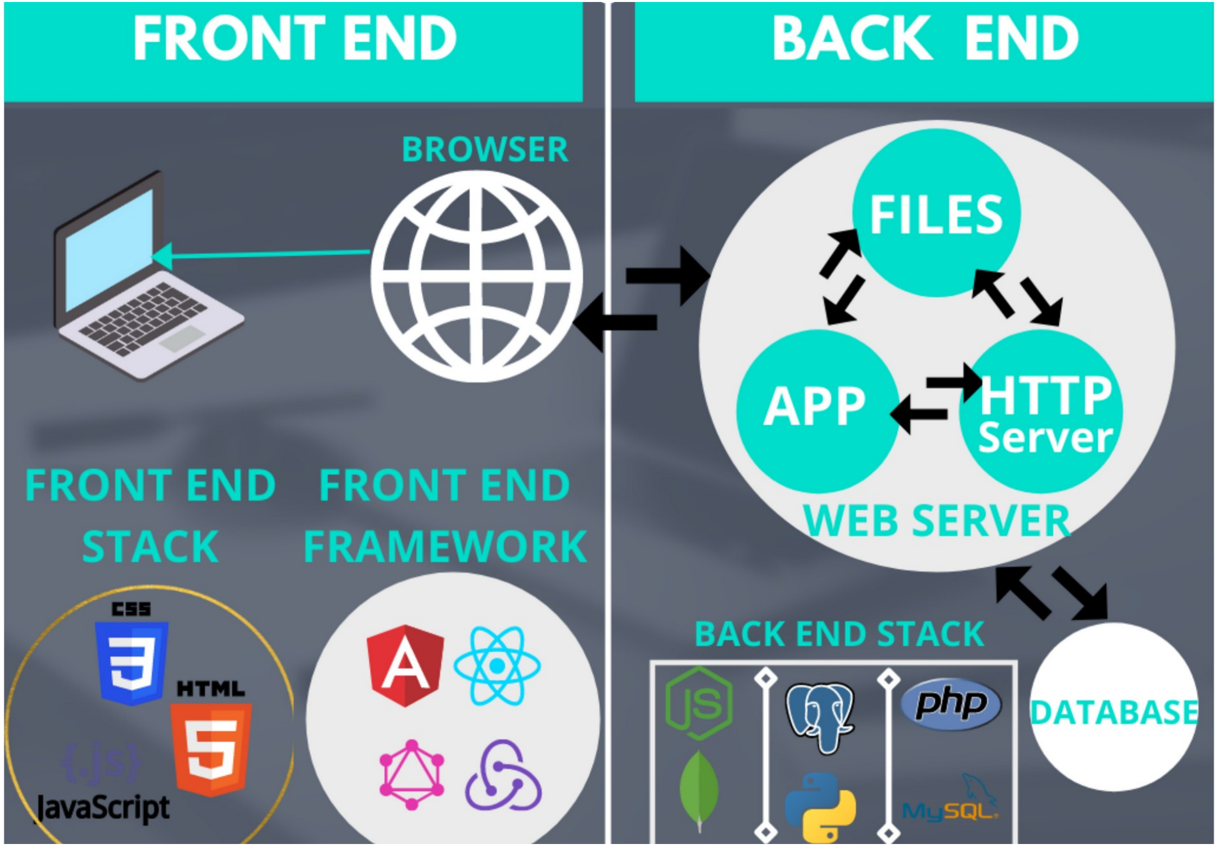
2nd Day

1. Morning (~ 2h)
   - Smallest remainder of Day 1
   - King – Man + Woman
   - Using advanced embedding techniques

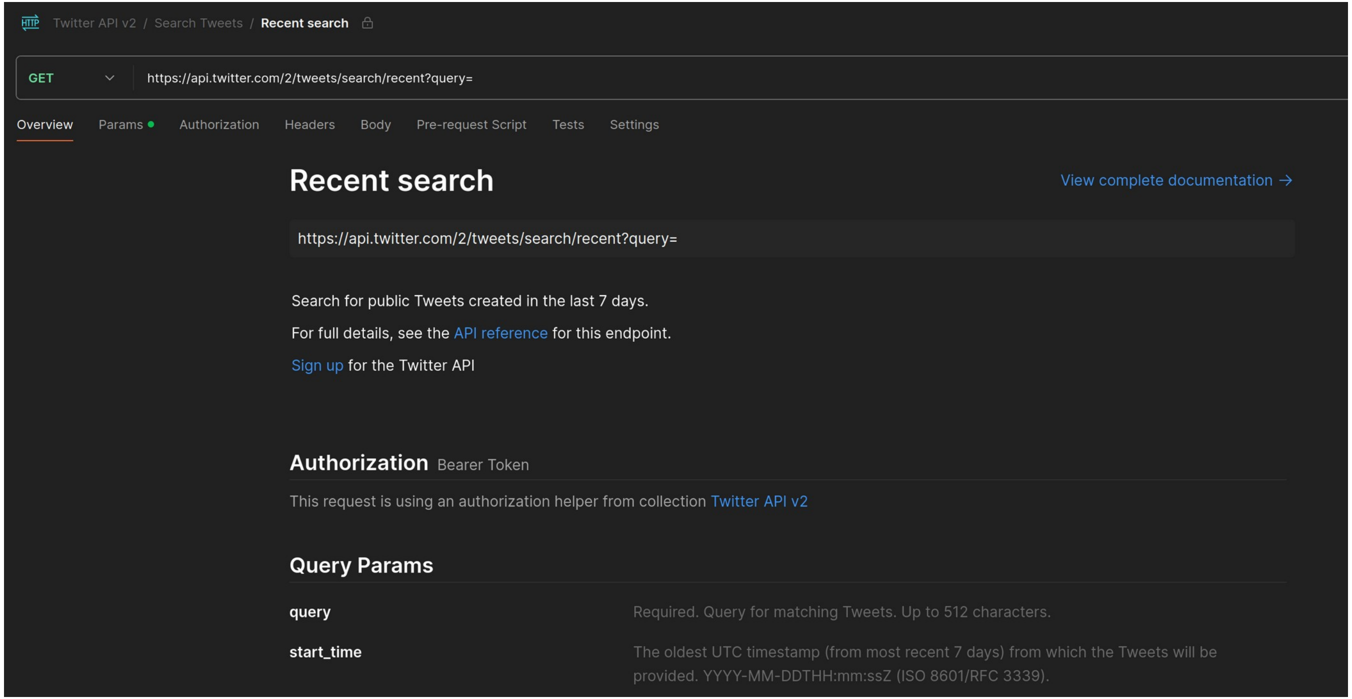
2. Afternoon (~ 4h)
   - Gentle introduction to delivery API + front End
   - Transfert learning
   - Fine Tuning

First … Let’s Talk !

Back vs Front ?



What is Back-End / API ?



Twitter API v2 / Search Tweets / Recent search
GET
Overview
Params
Authorization
Headers
Body
Pre-request Script
Tests
Settings
View complete documentation →

https://api.twitter.com/2/tweets/search/recent?query=

Recent search

https://api.twitter.com/2/tweets/search/recent?query=

Search for public Tweets created in the last 7 days.

For full details, see the API reference for this endpoint.

Sign up for the Twitter API

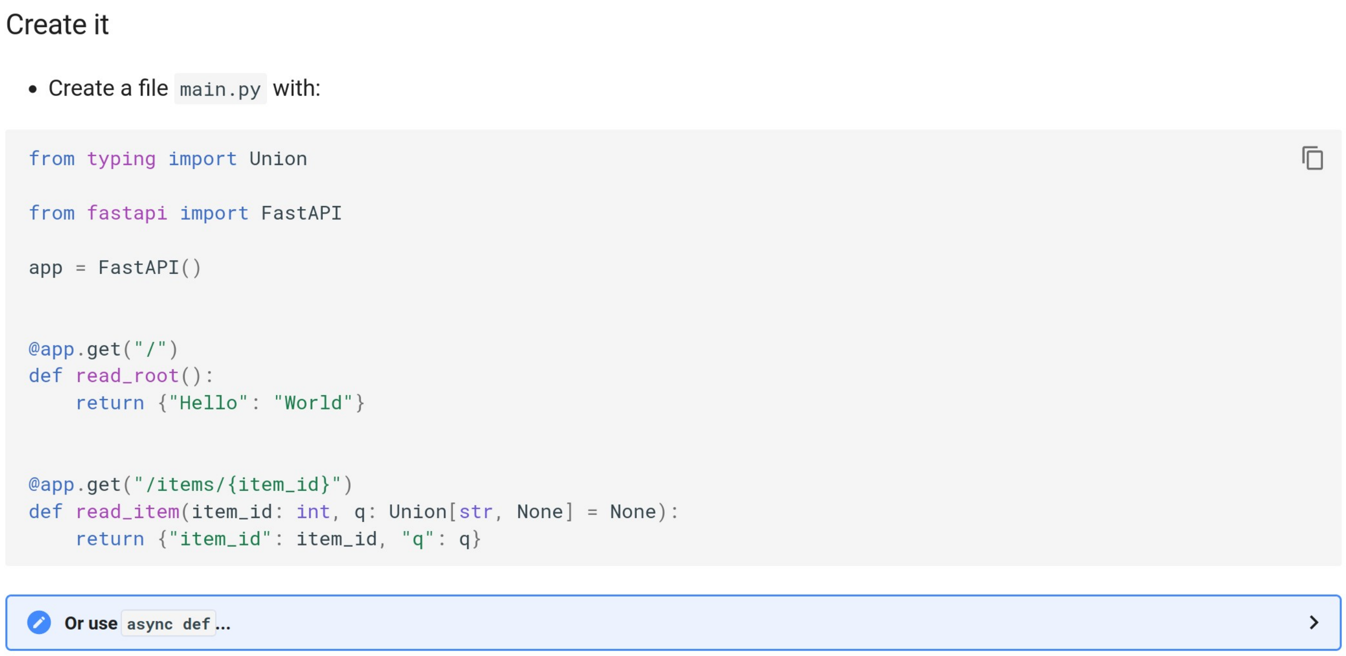
Authorization Bearer Token

This request is using an authorization helper from collection Twitter API v2

Query Params

query	Required. Query for matching Tweets. Up to 512 characters.
start_time	The oldest UTC timestamp (from most recent 7 days) from which the Tweets will be provided. YYYY-MM-DDTHH:mm:ssZ (ISO 8601/RFC 3339).

Fast API



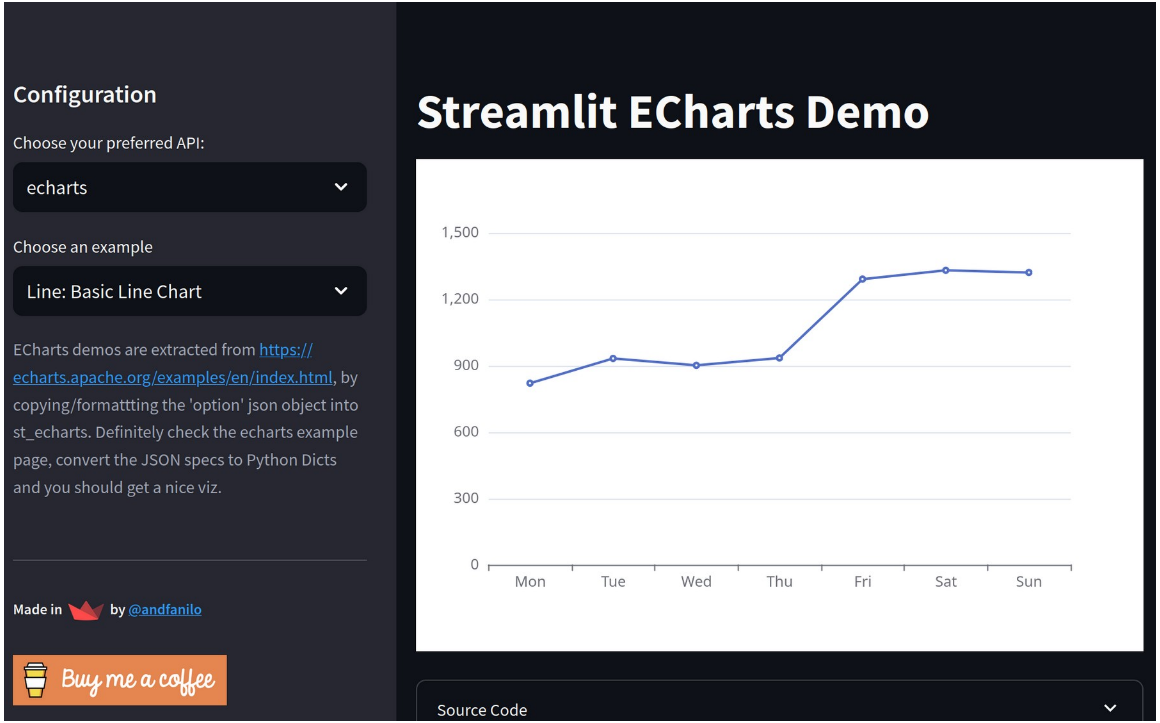
Create it

- Create a file `main.py` with:

```python
from typing import Union

from fastapi import FastAPI

app = FastAPI()


@app.get("/")
def read_root():
    return {"Hello": "World"}


@app.get("/items/{item_id}")
def read_item(item_id: int, q: Union[str, None] = None):
    return {"item_id": item_id, "q": q}
```

Or use `async def`...

What is Front End



Configuration

Choose your preferred API:

echarts

Choose an example

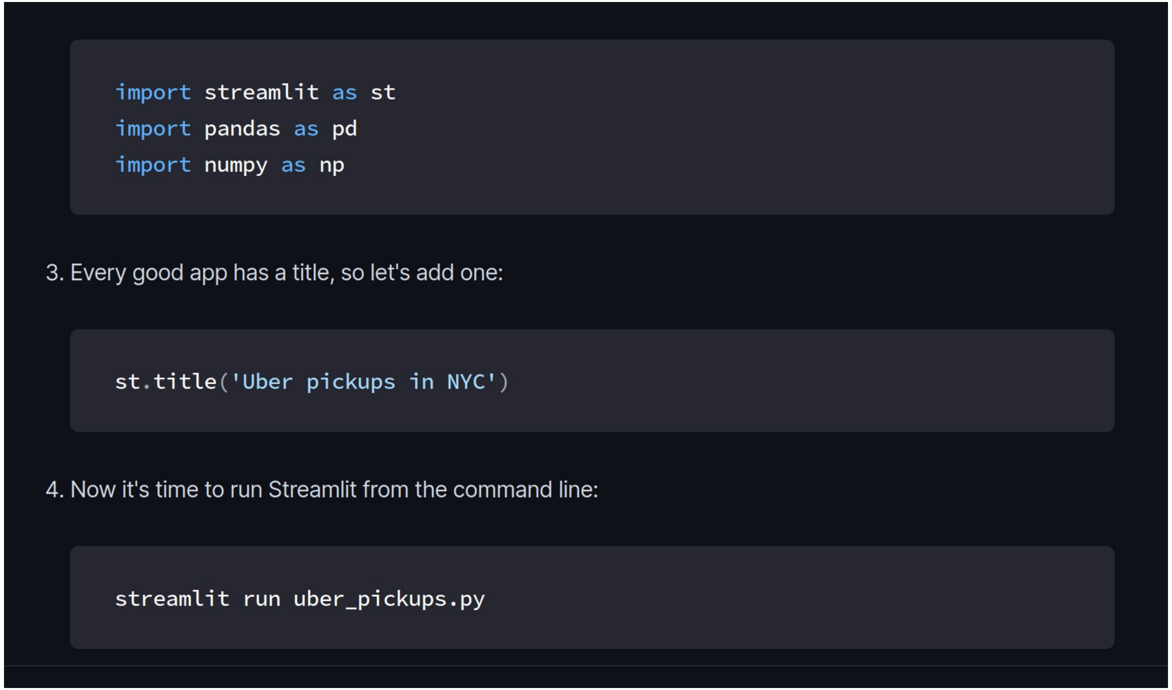
Line: Basic Line Chart

ECharts demos are extracted from https://echarts.apache.org/examples/en/index.html, by copying/formattting the 'option' json object into st_echarts. Definitely check the echarts example page, convert the JSON specs to Python Dicts and you should get a nice viz.

Made in by @andfanilo

Streamlit ECharts Demo

Streamlit



```
import streamlit as st
import pandas as pd
import numpy as np
```

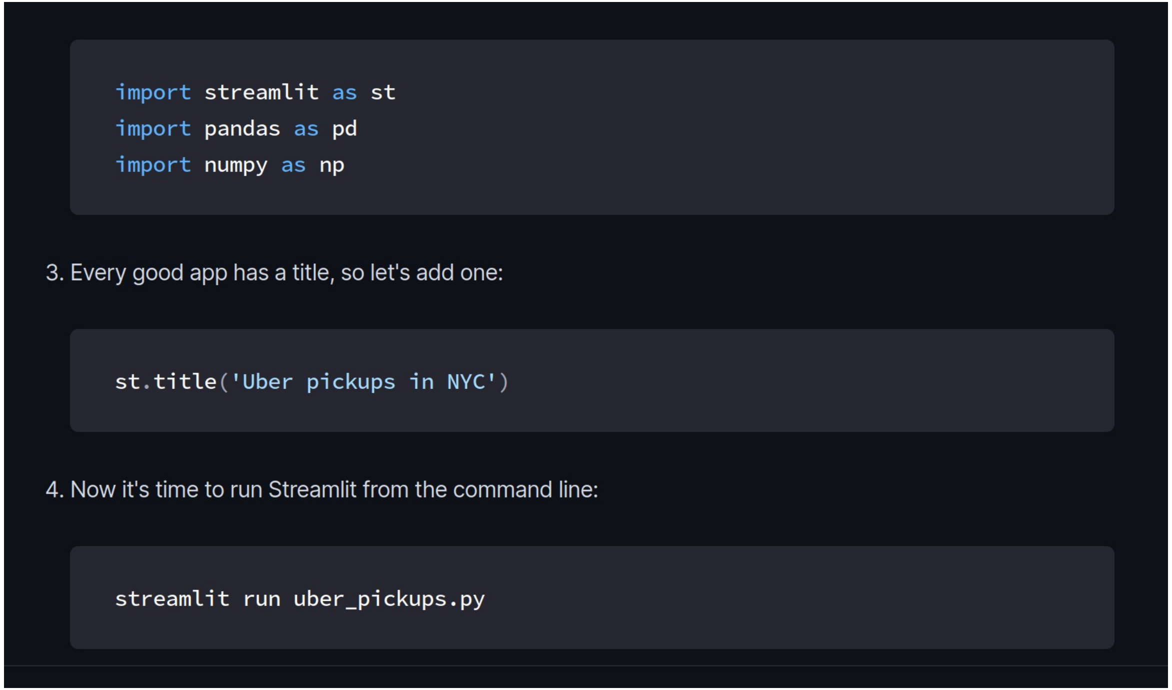
3. Every good app has a title, so let's add one:

```
st.title('Uber pickups in NYC')
```

4. Now it's time to run Streamlit from the command line:

```
streamlit run uber_pickups.py
```

Streamlit



```
import streamlit as st
import pandas as pd
import numpy as np
```

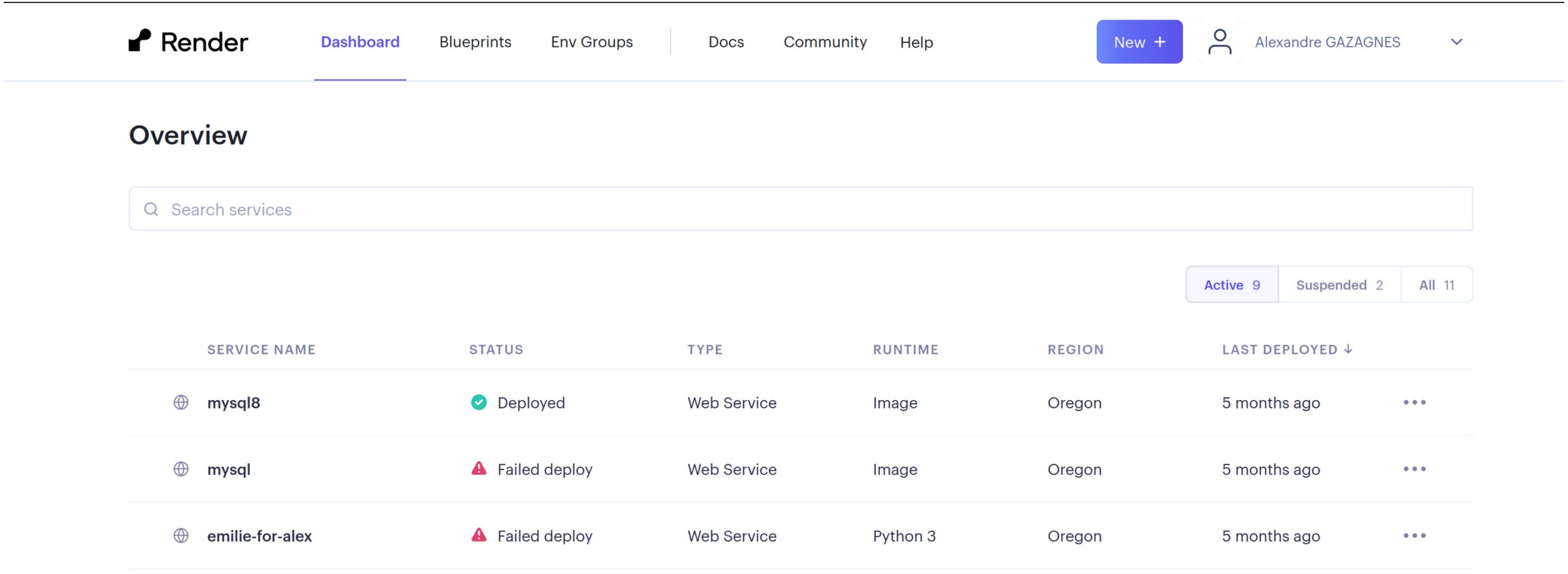
3. Every good app has a title, so let's add one:

```
st.title('Uber pickups in NYC')
```

4. Now it's time to run Streamlit from the command line:

```
streamlit run uber_pickups.py
```

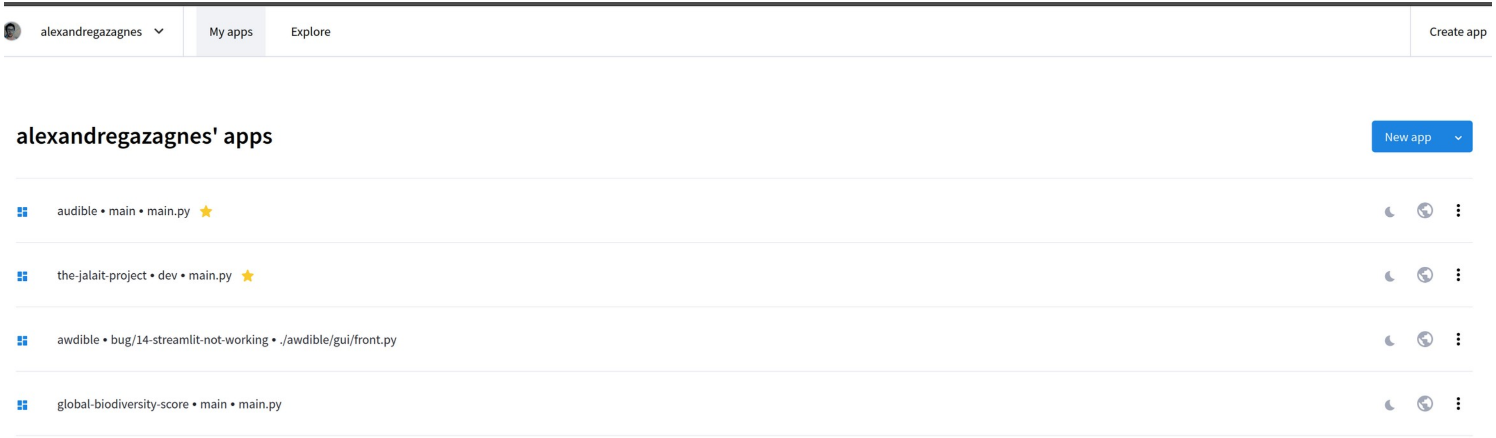
Production Back



Overview

SERVICE NAME	STATUS	TYPE	RUNTIME	REGION	LAST DEPLOYED ↓
mysql8	✓ Deployed	Web Service	Image	Oregon	5 months ago
mysql	⚠ Failed deploy	Web Service	Image	Oregon	5 months ago
emilie-for-alex	⚠ Failed deploy	Web Service	Python 3	Oregon	5 months ago

Production Front



alexandregazagnes' apps

audible • main • main.py ⭐

the-jalait-project • dev • main.py ⭐

awdible • bug/14-streamlit-not-working • ./awdible/gui/front.py

global-biodiversity-score • main • main.py

Practice !

(FreeStyle ;))

NLP Tasks with pretrained models

What is a pretrained model ?

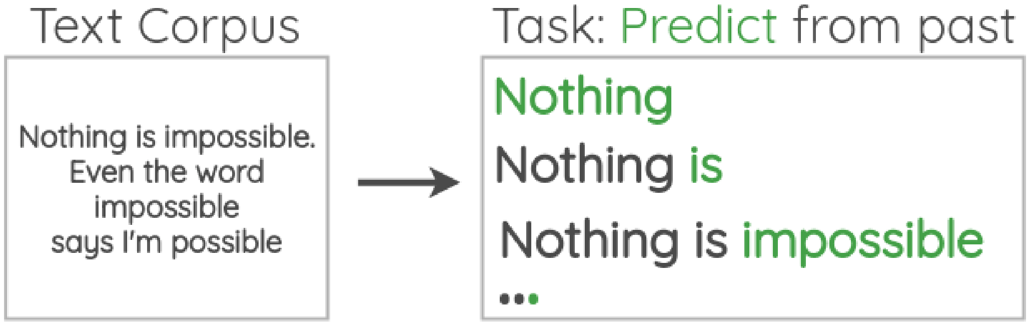
A **pretrained** model is a neural network that has been trained on a large dataset and can be used directly on your data or as a starting point for other tasks.

These models can s**ave a significant amount of time and computational resources,** as it is not necessary to train a model from scratch on a new task.

Finetuning is the process of further training a pretrained model on a specific task using task-specific data (usually small dataset). The degree of finetuning varies depending on the size and similarity of the new task to the pretraining task.

Large language models case

LLM model is presented with a large dataset of text and is then trained **to predict the next word** in a sequence of words (**Self-supervised learning)**.



Text Corpus

Nothing is impossible.
Even the word
impossible
says I'm possible

Task: Predict from past

Nothing
Nothing is
Nothing is impossible
...

Large language models have achieved **state-of-the-art performance** on a wide range of natural language processing tasks, including language translation, text summarization, and question answering.

Most used pretrained models in NLP

Bard: **Google's** Bard, initially powered by the LaMDA model with **137 billion** parameters, has been enhanced with the PaLM model, which has a staggering **540 billion parameters**. The upgrade with PaLM significantly boosts Bard's functionality, including improved reasoning, multi-step prompt handling

LLaMA: Meta's family of LLaMA models, ranging from **7 to 65 billion parameters**, provides a robust and efficient alternative for various NLP tasks, outperforming larger models in many benchmarks. Used in facebook for Content Moderation.

GPT: Developed by **OpenAI**, it is a transformer-based model with over **175 billion parameters for GPT3**, trained on a dataset of petabytes of data collected over 8 years of web crawling. GPT4 has almost one trillion parameters (the specific number of parameters for GPT-4 hasn't been publicly disclosed).

Classification & Sentiment Analysis

The process of d**etermining the category or a given text**. Sentiment Analysis is one of the most used case.

Examples:

- Determining whether a customer review of a product is positive or negative.
- Classifying a news article as belonging to a specific category (politics, sports, entertainment, etc.).

Zero-Shot Classification

Zero-shot classification is an NLP algorithm that can classify text data into a set of predefined categories or labels without the need for any labeled training data. Instead, the algorithm relies on natural language understanding and general knowledge to make predictions based on the input text and a set of given parameters.

Examples:

- Classifying news articles into predefined topics such as politics, sports, and entertainment, without being trained on any specific labeled data for those categories. The algorithm relies on its ability to understand the language and the context of the articles to make accurate predictions.
- Classifying customer feedback into different categories such as positive, negative, or neutral sentiments, without any labeled training data for sentiment analysis. The algorithm can use its understanding of language and common patterns in text to identify the sentiment expressed in the feedback.
- Classifying documents or texts into different languages, without any labeled data for language classification. The algorithm can analyze the input text's linguistic features and identify the most likely language based on its knowledge of language structure and patterns.
- Classifying products or services into predefined categories based on their features and characteristics, without any labeled data for product classification. The algorithm can analyze the text describing the product's features and characteristics and classify it into the most appropriate category based on its understanding of the product domain.

Information Extraction & Questing Answering

The process of extracting **specific information from a text** and answering questions based on that information.

Examples:

- Extracting information about a person from a Wikipedia article and answering questions about their education and career.
- Extracting information about a company from a financial report and answering questions about their revenue and profits.

Translation

The process of converting text from one language to another.

Example:

- Translating a Spanish news article into English.
- These models can also be adopted to generate voice : Text-to-Speech

Summarization

The process of **condensing a text** to its main points or **key information**.

Examples:

- Summarizing a long research paper into a brief summary for a conference presentation.
- Summarizing a news article about a complex legal case into a summary for a non-legal audience.

Topic Modeling

The goal of topic modeling is to uncover the underlying structure and patterns in the data, and to group together documents that are related to each other based on their content.

Topic modeling algorithms typically use techniques such as latent Dirichlet allocation (LDA) to identify the underlying topics present in a set of documents. These topics may be represented as a set of keywords or phrases, and each document may be assigned a probability distribution over the topics, indicating the degree to which it relates to each topic

Examples:

- A news organization may use topic modeling to identify the most common themes and issues discussed in their articles, and to group similar articles together for easier organization and analysis.
- A social media monitoring tool may use topic modeling to identify the most frequently discussed topics and hashtags on a particular platform, and to track changes in these topics over time.

Text Generation & Prompting

The process of creating new text based on a given input or set of parameters.

Prompting refers to the process of providing input, context and/or starting point for a text generation.

Examples:

- Generating personalized email responses based on the recipient's name and previous interactions with the company.
- Generating product descriptions for an e-commerce website based on the product's features and target audience.
- Generating creative headlines for news articles based on the content of the article.
- Generating social media posts for a brand based on current events and trending topics.

Evaluation

There are several common evaluation metrics used in natural language processing (NLP) tasks, depending on the specific task and dataset. Some of the most common metrics include:

- **Accuracy**: This is the most basic and commonly used metric, which simply calculates the percentage of correctly predicted labels or outputs.
- **Precision, Recall and F1-score**: These are used for classification tasks and take into account both true positives and false positives. Precision is the number of true positives divided by the number of true positives and false positives, recall is the number of true positives divided by the number of true positives and false negatives, and F1-score is the harmonic mean of precision and recall.
- **BLEU, ROUGE and METEOR:** These are used for text generation tasks such as machine translation, summarization, and text completion. **BLEU** compares the generated text to one or more reference texts, **ROUGE** compares the generated text to a reference text, and **METEOR** compares the generated text to a reference text, taking into account synonyms and stemming.

It is important to note that these metrics should be chosen based on the specific NLP task, and that a single metric may not provide a complete picture of the model's performance.

Practice !

Annexes

Annexes

To find dataset you can start by exploiting:

https://datasetsearch.research.google.com/

https://huggingface.co/datasets/viewer/

NLP in Practice: Project Work

Conduct a study about trends in healthy food in 2022. This project could involve the following steps:

- Collecting data on healthy food trends by scraping relevant information from websites, blogs, and social media platforms.
- Preprocessing the collected data by cleaning, normalizing, and tokenizing the text
- Using NLP techniques such as sentiment analysis, topic modeling, and named entity recognition to analyze the data and uncover trends in healthy food.
- Visualizing the results using charts and graphs to make the findings more clear and actionable.
- Creating a report or presentation summarizing the findings and discussing their implications for the healthy food industry.

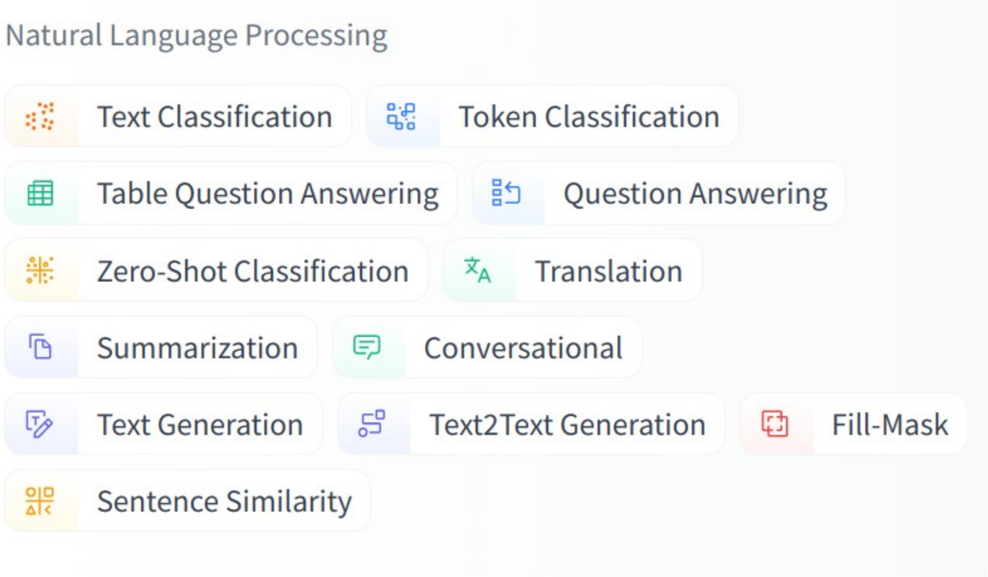
This project would allow students to apply their NLP skills to a real-world problem and gain a better understanding of trends in healthy food. Additionally, it could be possible to share the findings with relevant stakeholders in the food industry and potentially have an impact.

Collecting data on healthy food trends

One example of where you could collect data for this NLP project on healthy food trends in 2022 would be from social media platforms such as Twitter, Instagram, and Facebook. You could use the social media APIs to access data on posts and comments related to healthy food, such as those containing keywords like "organic," "plant-based," "gluten-free," etc. You could also use hashtags such as #healthyeating or #plantbaseddiet to find relevant content.

Another example would be scraping websites that specialize in healthy food recipes and nutrition information, such as Whole Foods, EatingWell, and Healthline. These websites contain a lot of useful information on healthy food trends, such as recipes, ingredients, and nutritional information.

HuggingFace NLP Tasks classification



Natural Language Processing

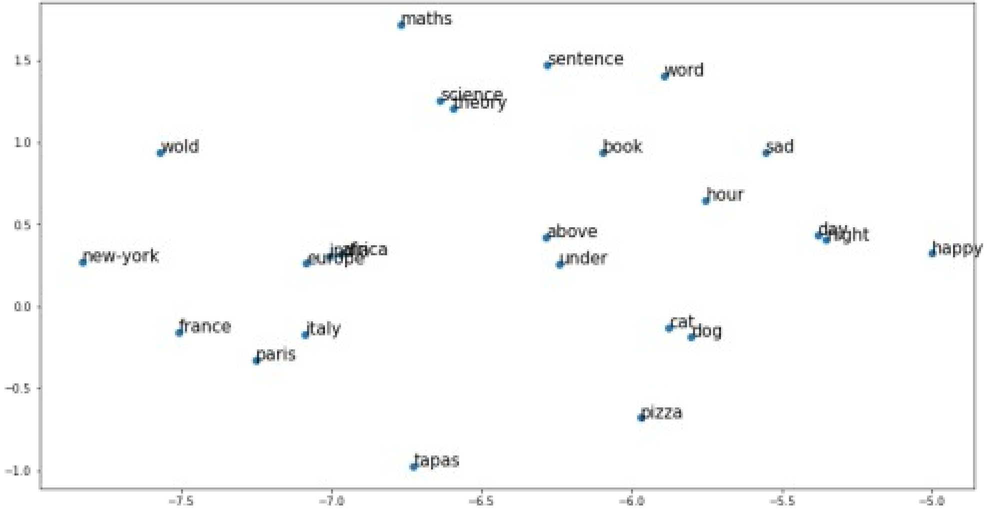
- Text Classification
- Token Classification
- Table Question Answering
- Question Answering
- Zero-Shot Classification
- Translation
- Summarization
- Conversational
- Text Generation
- Text2Text Generation
- Fill-Mask
- Sentence Similarity

https://github.com/wandb/examples

https://wandb.ai/ayush-thakur/huggingface/reports/How-To-Fine-Tune-HuggingFace-Transformers-on-a-Custom-Dataset--Vmlldzo0MzQ2MDc

Word2vec :

Word2Vec is a technique for creating dense, numerical representations of words, also called **"word embeddings."** These embeddings capture the meaning and context of a word in a continuous, multi-dimensional space. Word2Vec is trained using **neural networks** on large corpora of text.



Any Business ideas, techs projects to bring in ?

1. .
2. .
3. .
4. .

Usage in the food industry

Here are some examples:

- **Recipe generation:** Generate new recipes based on a given set of ingredients or dietary restrictions.
- **Nutritional analysis:** Extract nutritional information from food labels or recipes, making it easier for consumers to track their intake or for companies to comply with regulatory requirements.
- **Food safety monitoring**: Analyze social media posts or news articles for mentions of food safety incidents or outbreaks, allowing companies to identify and address potential issues more quickly.