# General course-related questions

## I just discovered the course. Can I still join?

Yes, but if you want to receive a certificate, you need to submit your project while we’re still accepting submissions.

## Course - I have registered for the [insert-zoomcamp-name]. When can I expect to receive the confirmation email?

You don't need it. You're accepted. You can also just start learning and submitting homework (while the form is **Open**) without registering. It is not checked against any registered list. Registration is just to gauge interest *before* the start date.

## What is the video/zoom link to the stream for the “Office Hours” or live/workshop sessions?

The zoom link is only published to instructors/presenters/TAs.

Students participate via Youtube Live and submit questions to Slido (link would be pinned in the chat when Alexey goes Live). The video URL should be posted in the [announcements channel on Telegram & Slack](https://t.me/dezoomcamp) before it begins. Also, you will see it live on the DataTalksClub [YouTube Channel](https://www.youtube.com/c/DataTalksClub).

Don’t post your questions in chat as it would be off-screen before the instructors/moderators have a chance to answer it if the room is very active.

## SaturnCloud - How do I get access?

Issue: I get the notice that due to traffic, I’m on a waitlist for new signups.

Answer: There was a form to submit our emails to, so Alexey can send it in bulk. If you missed that deadline, just sign up manually (or via request tech demo link) and use the chat to request for free hours for “*llm zoomcamp*”

Issue: I’m a pre-existing user from a different zoomcamp and I’m not awarded the free hours even though I’ve submitted my email in the form.

Answer: Just request it via their chat, after you’ve logged in using your pre-existing account, citing “*llm zoomcamp*” .

## SaturnCloud - How many free hours do we get?

We get 15 free hours per month, which might be limited to the free tier’s hardware configuration.

## SaturnCloud - Something went wrong. Max of 15 hours of resource usage per month

This message means you have used all allocated hours. Make sure to set Shutout After in settings. Also, do not leave your notebooks running. If your hours are out, try using Google Colab and Kaggle.

## Cloud alternatives with GPU

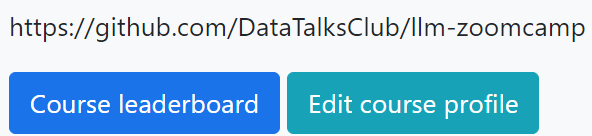
Check the quota and reset cycle carefully - is the free hours per month or per week? Usually if you change the configuration, the free hours quota might also be adjusted,or it might be billed separately.

1. Google Colab
2. Kaggle
3. Databricks (?), so many others.

Use GPTs to find out. Some might have restrictions on what you can and cannot install, so be sure to read what is included in a free vs paid tier.

## Leaderboard - I am not on the leaderboard / how do I know which one I am on the leaderboard?

When you set up your account you are automatically assigned a random name such as “Lucid Elbakyan” for example. Click on the *Jump to your record on the leaderboard* link to find your entry.

  
If you want to see what your Display name is, click on the Edit Course Profile button.

1. First field is your nickname/displayed-name, change it if you want to be known as your Slack username or Github username or whatever nickname of your choice, if you want to remain anonymous.
2. Unless you want “Lucid Elbakyan” on your certificate, it is **mandatory** that you change the second field to your **official name** as in your identification documents - passport, national ID card, driver’s license, etc. This is the name that is going to appear on your Certificate!

## Certificate - Can I follow the course in a self-paced mode and get a certificate?

No, you can only get a certificate if you finish the course with a “live” cohort.

We don't award certificates for the self-paced mode. The reason is you need to peer-review 3 capstone(s) after submitting your own project.

You can only peer-review projects at the time the course is running; after the form is closed and the peer-review list is compiled.

## I missed the first homework - can I still get a certificate?

Yes, you need to pass the Capstone project to get the certificate. Homework is not mandatory, though it is recommended for reinforcing concepts, and the points awarded count towards your rank on the leaderboard.

## I was working on next week’s homework/content - why does it keep changing?

This course is being offered for the first time, and things will keep changing until a given module is ready, at which point it shall be announced. Working on the material/homework in advance will be at your own risk, as the final version could be different.

## When will the course be offered next?

Summer 2025 (via Alexey).

## Are there any lectures/videos? Where are they?

Please check the bookmarks and pinned links, especially DataTalks.Club’s YouTube account.

## WSL2 - ResponseError: model requires more system memory (X.X GiB) than is available (Y.Y GiB). My system has more than X.X GiB.

Your WSL2 is set to use Y.Y GiB, not all your computer memory. Create .wslconfig file under your Windows user profile directory (C:\Users\YourUsername\.wslconfig) with the desired RAM allocation:

[wsl2]

memory=8GB

Restart WSL: wsl --shutdown

Run the free command to verify the changes. For more details, read [this article](https://www.aleksandrhovhannisyan.com/blog/limiting-memory-usage-in-wsl-2/).

# Module 1: Introduction

## OpenAI: Error when running OpenAI chat.completions.create command

You may receive the following error when running the OpenAI chat.completions.create command due to insufficient credits in your OpenAI account:

| NotFoundError: Error code: 404 - {'error': {'message': 'The model `gpt-4o` does not exist or you do not have access to it.', 'type': 'invalid\_request\_error', 'param': None, 'code': 'model\_not\_found'}} |
| --- |

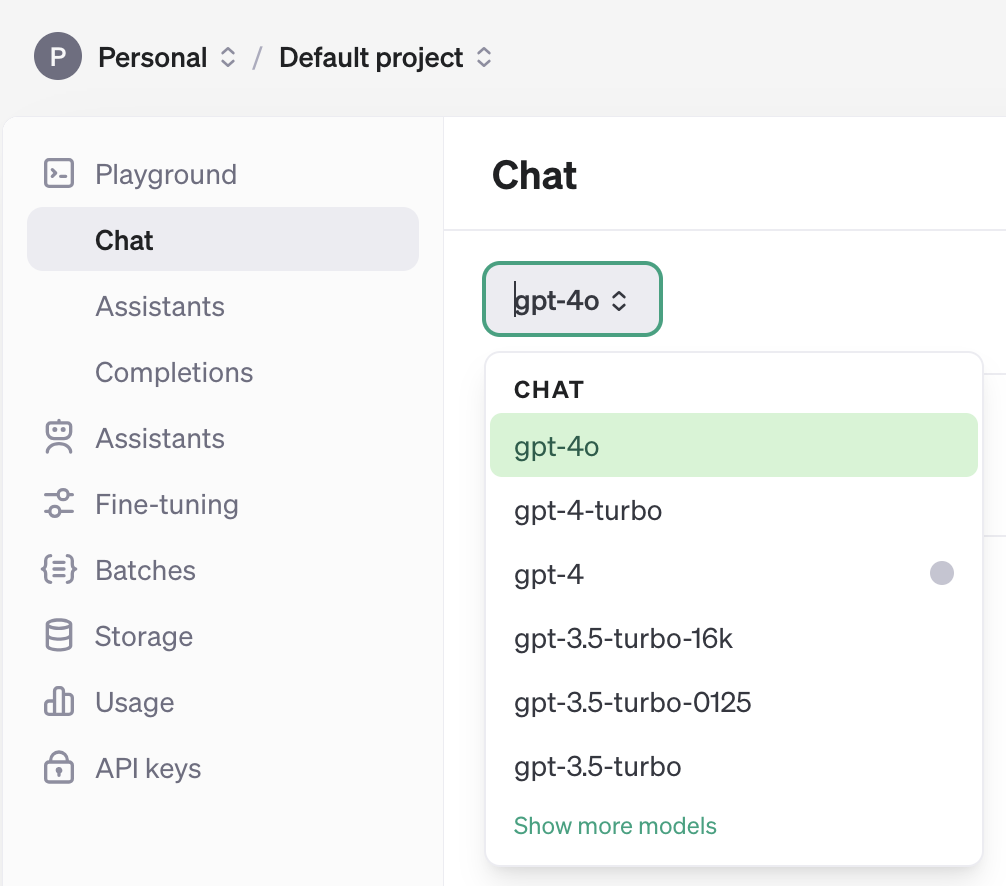
## OpenAI: Error: RateLimitError: Error code: 429 -

RateLimitError: Error code: 429 - {'error': {'message': 'You exceeded your current quota, please check your plan and billing details. For more information on this error, read the docs: <https://platform.openai.com/docs/guides/error-codes/api-errors.>', 'type': 'insufficient\_quota', 'param': None, 'code': 'insufficient\_quota'}

The above errors are related to your OpenAI API account’s quota.  
There is no free usage of OpenAI’s API so you will be required to add funds using a credit card (see pay as you go in the OpenAI settings at [platform.openai.com](http://platform.openai.com)). Once added, re-run your python command and you should receive a successful return code.

**Steps to resolve:**

1. Add credits to your account [here](https://platform.openai.com/settings/organization/billing/overview) (min $5)
2. In chat.completions.create(model='**gpt-4o**', …) specify one of the available for you models:



1. You might need to recreate an API key after adding credits to your account and update it locally.

## OpenAI: Error: 'Cannot import name OpenAI from openai'; How to fix?

Update openai version from 0.27.0 -> any 1.x version

## OpenAI: How much will I have to spend to use the Open AI API?

Using the Openai API does not cost much, you can recharge from 5 dollars. At least for what I spent on the first unit it was barely 5 cents.

## OpenAI: Do I have to subscribe and pay for Open AI API for this course?

No, you don't have to pay for this service in order to complete the course homeworks, you could use some of the alternatives free from this list posted into the course Github.

[llm-zoomcamp/01-intro/open-ai-alternatives.md at main · DataTalksClub/llm-zoomcamp (github.com)](https://github.com/DataTalksClub/llm-zoomcamp/blob/main/01-intro/open-ai-alternatives.md)

## ElasticSearch: ERROR: Elasticsearch exited unexpectedly

If you get this error, it’s likely that elasticsearch doesn’t get enough RAM

I specified the RAM size to the configuration (-m 4GB)

docker run -it \

--rm \

--name elasticsearch \

-m 4GB \

-p 9200:9200 \

-p 9300:9300 \

-e "discovery.type=single-node" \

-e "xpack.security.enabled=false" \

docker.elastic.co/elasticsearch/elasticsearch:8.4.3

Or give it \_less\_ RAM:

Tip for Github Codespace users

If you want to run elasticsearch server in a docker, then it may fail with the command in the documentation.

In that case, you can try inserting this line -e "ES\_JAVA\_OPTS=-Xms512m -Xmx512m".

This reduces the resource usage.

Full command:

docker run -it \

--rm \

--name elasticsearch \

-p 9200:9200 \

-p 9300:9300 \

-e "discovery.type=single-node" \

-e "xpack.security.enabled=false" \

-e "ES\_JAVA\_OPTS=-Xms512m -Xmx512m" \

docker.elastic.co/elasticsearch/elasticsearch:8.4.3

If it doesn't work, try this:

sudo sysctl -w vm.max\_map\_count=262144

And give the Java machine inside the container more RAM:

docker run -it \

--rm \

--name elasticsearch \

-p 9200:9200 \

-p 9300:9300 \

--ulimit nofile=65536:65536 \

--ulimit memlock=-1:-1 \

--memory=4g \

--cpus=2 \

-e "discovery.type=single-node" \

-e "xpack.security.enabled=false" \

-e "ES\_JAVA\_OPTS=-Xms2g -Xmx2g" \

[docker.elastic.co/elasticsearch/elasticsearch:8.4.3](http://docker.elastic.co/elasticsearch/elasticsearch:8.4.3)

Another possible solution may be to set the memory\_lock to false:

docker run -it \

--rm \

--name elasticsearch \

-p 9200:9200 \

-p 9300:9300 \

-e "discovery.type=single-node" \

-e "xpack.security.enabled=false" \

-e "ES\_JAVA\_OPTS=-Xms512m -Xmx512m" \

-e "bootstrap.memory\_lock=false" \

docker.elastic.co/elasticsearch/elasticsearch:8.4.3

## ElasticSearch: ERROR: Elasticsearch.index() got an unexpected keyword argument 'document'

Instead of **document** as used in the course video, use **doc**

## Docker: How do I store data persistently in Elasticsearch?

When you stop the container, the data you previously added to elastic will be gone. To avoid it, we can add volume mapping:

docker volume create elasticsearch\_data

docker run -it \

--rm \

--name elasticsearch \

-p 9200:9200 \

-p 9300:9300 \

-v elasticsearch\_data:/usr/share/elasticsearch/data \

-e "discovery.type=single-node" \

-e "xpack.security.enabled=false" \

docker.elastic.co/elasticsearch/elasticsearch:8.4.3

## Authentication: Safe and easy way to store and load API keys

You can store your different API keys in a yaml file that you will add in your .gitignore file. Be careful to never push or share this file.

* For example, you can create a new file named “api\_keys.yml” in your repository.
* Then, do not forget to add it in your .gitignore file:

*#api\_keys*

api\_keys.yml

* You can now fill your api\_keys.yml file:

OPENAI\_API\_KEY: “sk[...]”

GROQ\_API\_KEY: “gqk\_[...]”

* Save your file.
* You will need the pyyaml library to load your yaml file, so run this command in your terminal:

pip install pyyaml

* Now, open your jupyter notebook.
* You can load your yaml file and the associated keys with this code:

import yaml

# Open the file

with open('api\_keys.yml', 'r') as file:

# Load the data from the file

data = yaml.safe\_load(file)

# Get the API key (Groq example here)

groq\_api\_key = data['GROQ\_API\_KEY']

* Now, you can easily replace the “api\_key” value directly with the loaded values without loading your environment variables.

Added by Mélanie Fouesnard

## Authentication: Why is my OPENAI\_API\_KEY not found in the jupyter notebook?

### Option1: using direnv

created the .envrc file & added my API key, ran direnv allow in the terminal

was getting an error: "OpenAIError: The api\_key client option must be set either by passing api\_key to the client or by setting the OPENAI\_API\_KEY environment variable"

resolution: install dotenv & add the following to a cell in the notebook. You can install dotenv by running: pip install python-dotenv.

from dotenv import load\_dotenv

load\_dotenv('.envrc')

### Option 2: using Codespaces Secrets

* Log in to your GitHub account and navigate to Settings > Codespaces
* There is a section called secrets where you can create Secrets like OPENAI\_API\_KEY and select for which repositories the secret is supposed to be available.
* Once you set this up, the key will be available in your codespaces session

## OpenSource: How can I use Ollama open-source models locally on my pc without using any API?

Prior to using Ollama models in llm-zoomcamp tasks, you need to have ollama installed on your pc and the relevant LLM model downloaded with ollama from https://www.ollama.com

To download ollama for Ubuntu:

``` curl -fsSL https://ollama.com/install.sh | sh ```

To download ollama for Mac and Windows, follow the guide on this link:

<https://ollama.com/download/>

Ollama a number of open-source LLMs like:

* Llama3
* Phi3
* Mistral and Mixtral
* Gemma
* Qwen
* You can explore more models on [https://ollama.com/library/](https://ollama.com/download/)

To download a model in Ollama, simply open command prompt and type:

``` ollama run model\_name ```

e.g.

``` ollama run phi3 ```

It will automatically download the model and you can use it same way as above for later time.

To use Ollama models for inference and llm-zoomcamp tasks, use the following function:

import ollama

def llm(prompt):

response = ollama.chat(

model="llama3",

messages=[{"role": "user", "content": prompt}]

)

return response['message']['content']

For example, we can use it in the following way:

prompt = "When does the llm-zoomcamp course start?"

answer = llm(prompt)

print(answer)

## OpenSource: I am using Groq, and it doesn't provide a tokenizer library based on my research. How can we estimate the number of OpenAI tokens asked in homework question 6?

The question asks for the number of tokens in gpt-4o model. tiktoken is a python library that can be used to get the number of tokens. You don't need openai api key to to get the number of tokens. You can use the code provided in the question to get the number of tokens.

## OpenSource: Can I use Groq instead of OpenAI?

You can use any LLM platform for your experiments and your project. Also, the homework is designed in such a way that you don’t need to have access to any paid services and can do it locally. However, you would need to adjust the code for that platform. See their documentation pages.

## OpenSource: Can I use open-source alternatives to OpenAI API?

Yes. See module 2 and the [open-ai-alternatives.md](https://github.com/DataTalksClub/llm-zoomcamp/blob/main/01-intro/open-ai-alternatives.md) in module 1 folder.

## Returning Empty list after filtering my query (HW Q3)

This is likely to be an error when indexing the data. First you need to add the index settings before adding the data to the indices, then you will be good to go applying your filters and query.

## Question

Answer

# Module 2: Open-Source LLMs

## Saturn Cloud issues

Please see the General section or use CTRL+F to search this doc.

## SaturnCloud: How do you manage the changes from SaturnCloud to your Github repository?

Of course you should have first added your Github repository in SaturnCloud and the SSH Key in your Github account settings.

Once you are in jupyter notebook from SaturnCloud, open the terminal and write these lines:

1- Navigate to Your Project Directory:

cd /home/jovyan/my\_project

2- Configure GitHub Remote to Use SSH:

git remote set-url origin [git@github.com](mailto:git@github.com):username/repository.git

3- Stage, Commit and push your changes:

git add .

git commit -m "Your commit message"

git push

## SaturnCloud: How can I clean out the hugging face model cache on a saturn cloud notebook?

Clean out your cache using the following code:

from transformers import TRANSFORMERS\_CACHE

print(TRANSFORMERS\_CACHE)

import shutil

shutil.rmtree(TRANSFORMERS\_CACHE)

*Note:* Make sure to shutdown the notebook and restart the kernel

## ElasticSearch: Can I backup and restore my elasticsearch index from one to another docker container?

Yes, you can. Here the step to follow:

- Open a bash session in the elasticsearch container

```bash

docker exec -it elasticsearch bash

```

- Add path.repo configuration:

```bash

echo path.repo: ["/usr/share/elasticsearch/backup"] >> /usr/share/elasticsearch/config/elasticsearch.yml

```

- Restart container and verify it was created correctly:

```bash

docker restart elasticsearch

curl -X GET "localhost:9200/\_snapshot/my\_backup?pretty"

```

- Create the snapshot (this is the backup ;) )

```bash

curl -X PUT "localhost:9200/\_snapshot/my\_backup/snapshot\_1?wait\_for\_completion=true" -H 'Content-Type: application/json' -d'

{

"indices": "your\_index\_name",

"ignore\_unavailable": true,

"include\_global\_state": false

}

'

```

- Copy the backup to my machine:

```bash

docker cp elasticsearch:/usr/share/elasticsearch/backup /path/to/local

```

- Now create the new container or use docker-compose just in case you are following the module 2:

```bash

docker compose up -d

```

- Add de path.repo configuration in the new one, same as before:

```bash

docker exec -it new\_elasticsearch bash

echo path.repo: ["/usr/share/elasticsearch/backup"] >> /usr/share/elasticsearch/config/elasticsearch.yml

```

- Restart the docker container and copy the snapshot in it:

```bash

docker restart new\_elasticsearch

docker cp /path/to/local/backup new\_elasticsearch:/usr/share/elasticsearch

```

- Register the Snapshot Repository in the New Container:

```bash

curl -X PUT "localhost:9200/\_snapshot/my\_backup" -H 'Content-Type: application/json' -d'

{

"type": "fs",

"settings": {

"location": "/usr/share/elasticsearch/backup"

}

}

'

```

- Verify if it exists:

```bash

curl -X GET "localhost:9200/\_snapshot/my\_backup/snapshot\_1?pretty"

```

- Restore the snapshot:

```bash

curl -X POST "localhost:9200/\_snapshot/my\_backup/snapshot\_1/\_restore" -H 'Content-Type: application/json' -d'

{

"indices": "your\_index\_name",

"ignore\_unavailable": true,

"include\_global\_state": false

}

'

```

- Show your indexes:

```bash

curl -X GET "localhost:9200/\_cat/indices?v"

```

- Extra point: If you want to change the original index name by other when you restore the snapshot:  
```bash

curl -X POST "localhost:9200/\_snapshot/my\_backup/snapshot\_1/\_restore?pretty" -H 'Content-Type: application/json' -d'

{

"indices": "old\_index",

"ignore\_unavailable": true,

"include\_global\_state": false,

"rename\_pattern": "old\_index",

"rename\_replacement": "new\_index"

}

'

```

## ElasticSearch: How can I limit the memory used by the ElasticSearch container?

You can limit the amount of memory used in the ElasticSearch container by adding the next line to the environment section of your docker-compose. Choose the amount of your preference, e.g.:

- "ES\_JAVA\_OPTS=-Xms1g -Xmx1g" # Set Java heap size to 1GB

- You can limit CPU usage for an Elasticsearch service within a docker-compose.yaml file, you can utilize the resource configuration options available in Docker Compose. This includes cpus to limit the number of CPUs that the container can utilize. You can configure your Elasticsearch section in the docker-compose.yaml to restrict CPU usage:

services:

elasticsearch:

image: docker.elastic.co/elasticsearch/elasticsearch:8.4.3

container\_name: elasticsearch

environment:

- discovery.type=single-node

- xpack.security.enabled=false

ports:

- "9200:9200"

- "9300:9300"

deploy:

resources:

limits:

cpus: '1.0' # Limits to 1 CPU

reservations:

cpus: '0.5' # Reserves 0.5 CPUs

## 

## Docker: How to inspect the content of a file inside a Docker container ?

You have several ways to inspect the content of a file when you are inside a Docker container.

* First, make sure you ran the docker container interactively using bash:

docker exec -it <container> bash

* Then, you are able to use bash commands. For this case, I propose two solutions:
  + Use “cat” and the file you want to see the content: cat your\_file . This will directly print the content in your terminal.
  + Install vim or nano using apt get and open the file using vim or nano (this can be more suitable for larger files):
    - apt-get install vim
    - vim your\_file
    - Then, you can exit your file in vim by pressing ESC then typing “:q” and finally press ENTER

Added by Mélanie Fouesnard

## Docker: Error: Docker mounted volume adds ;C to end of windows path

Use the following line instead in mounting the current volume to docker for Q4:  
`-v "/${PWD}/ollama\_files:/root/.ollama"`

## Docker: Why does inferring using Phi 3 locally take so long on Macbook Air M1?

In Docker Desktop, try to increase the resource.

Go to the Dashboard > Settings > Resources. Raise the memory limit to 15GB and swap to 4GB - be generous. Applied and restarted the changes

Added by Dandy Arif Rahman

## Docker: How can to clean docker cache?

docker system prune -a

## Ollama: “Error: pull model manifest: 503: no healthy upstream” when pulling a model with Ollama

A network connection failure usually causes this error and if you try to repeat the operation immediately it’ll still fail. It’s a temporary error, you should wait for 2 or 3 minutes before attempting to pull the model again. Then some minutes later, the operation will success.

Added by Eduardo Muñoz

## Ollama: Error: NotFoundError: Error code: 404 - {'error': {'message': "model XXX not found, try pulling it first" …

To solve this you need to pull one of these models first: <https://ollama.com/library> . Also check the proper name of the module.

Added by Taras Goriachko

Ollama: Running Ollama locally on Colab gives error after the llm() line

APIConnectionError: Connection error.

It seems to be running at [localhost:11434](http://localhost:11434/) however [localhost:11434](http://localhost:11434/)/v1/ gives 404

Found a solution in the Medium article and this link:

<https://medium.com/@mauryaanoop3/running-ollama-on-google-colab-free-tier-a-step-by-step-guide-9ef74b1f8f7a>

<https://github.com/ollama/ollama/issues/703>

Added by Hanaa

## Ollama: How can remove Ollama model?

ollama list  
ollama rm [model\_name]

## 

## Ollama: Error code 500 InternalServerError

InternalServerError: Error code: 500 - {'error': {'message': 'model requires more system memory (5.6 GiB) than is available (1.5 GiB)', 'type': 'api\_error', 'param': None, 'code': None}}.

Running elastic search with the docker-compose is the cause of the RAM memory issue. To fix this you need to change the docker-compose.yaml file to limit the RAM usage of elastic search

version: '3.8'

services:

elasticsearch:

image: docker.elastic.co/elasticsearch/elasticsearch:8.4.3

container\_name: elasticsearch

environment:

- discovery.type=single-node

- xpack.security.enabled=false

- ES\_JAVA\_OPTS=-Xms1g -Xmx1g *# change 1*

ports:

- "9200:9200"

- "9300:9300"

deploy:

resources:

limits:

memory: 2G *# change 2*

ollama:

image: ollama/ollama

container\_name: ollama

volumes:

- ollama:/root/.ollama

ports:

- "11434:11434"

volumes:

ollama:

Added by Zoe Zelkha

## Mistral AI: Unable to get Mistral-7B-v0.1 access despite accepting terms on HF

Manually set the token as below:

access\_token = <your\_token>

model = AutoModelForCausalLM.from\_pretrained("mistralai/Mistral-7B-v0.1", token=access\_token)

tokenizer = AutoTokenizer.from\_pretrained("mistralai/Mistral-7B-v0.1", token=access\_token)

## Python: Error: ModuleNotFoundError: No module named 'transformers.cache\_utils'

To solve just install transformers directly from github

!pip install git+<https://github.com/huggingface/transformers>

## Python: Exception: data did not match any variant of untagged enum PyPreTokenizerTypeWrapper at line 40 column 3

To solve just install transformers directly from github

!pip install git+<https://github.com/huggingface/transformers>

## Python: from google.protobuf.pyext import \_message / TypeError: bases must be types

pip install protobuf==3.20.1

Added by Ibai Irastorza

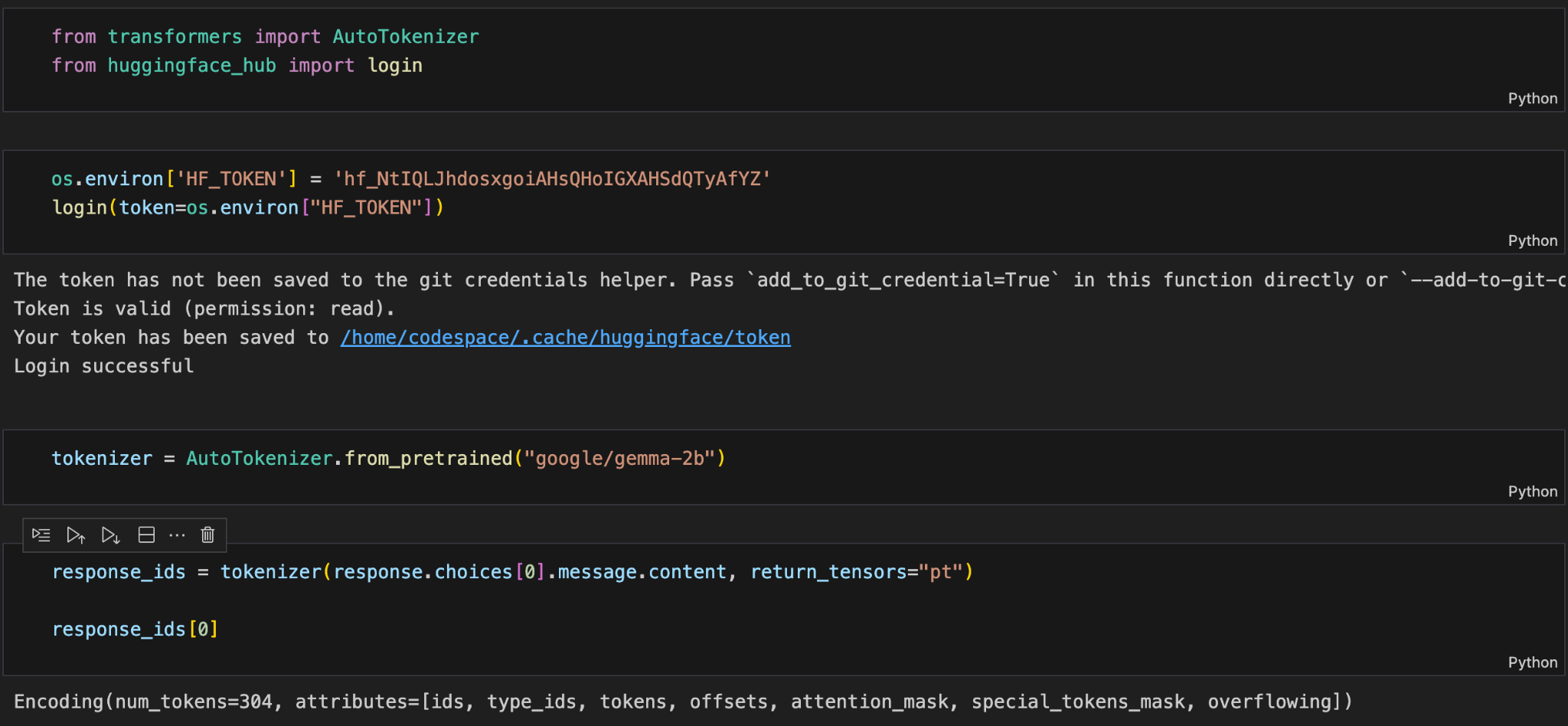
## HuggingFace: How to get the number of tokens in a certain string related to a certain model on hugging face?

1. search with the model name on hugging face.

2. get the transformer used on the model.

3. using the transformer, encode the string you want.

4. calculate the length of the outputted tensor.



The previous code snippet uses the tokenizer of google/gemma-2b LLM.   
Don’t forget to make your token secret.

Added by kamal

## How to run a model using CUDA for GPU usage?

The last version I checked for CUDA was 12.5 using a cloud environment like Saturn Cloud. Then the torch package for python should be on supported for that version of CUDA, is followed by cu121 which means that version of torch supports cuda 12.1. Check this page to find the package and version available for CUDA (remember to search the keyword “cu”

In my case I focused on using a torch==2.3.1 and the last cuda version supported was 12.1 (it works on Saturn Cloud)

To install all the needed packages use this command:

!pip install transformers accelerate torch==2.3.1+cu121 torchvision==0.18.1+cu121 torchaudio==2.3.1+cu121 --trusted-host download.pytorch.org --index-url https://download.pytorch.org/whl/cu121

And after that just executed this command:

!pip install --upgrade transformers

# Module 3: X

## ElasticSearch: Error: Elasticsearch.index() got an unexpected keyword argument 'document'

Upgrade elasticsearch 7.13.3 to 8.14.0 or any 7.x installation to 8.x. The earlier modules used a docker image of elasticsearch 8.4.3 so the python installation of elasticsearch must also be at least 8.x.

Or use the keyword *‘body’* instead of *‘document’*

For conda users, if you’re trying to update to elasticsearch 8.x using conda install elasticsearch==8.4.3 but getting a “PackagesNotFoundError", try this:  
  
$ conda config --add channels conda-forge

$ conda config --set channel\_priority strict

$ conda install -c conda-forge elasticsearch==8.4.3

## ElasticSearch: TypeError: Elasticsearch.search() got an unexpected keyword argument 'knn'

This worked for me:

## ElasticSearch: ConnectionError: Connection error caused by: ConnectionError(Connection error caused by: NewConnectionError(<urllib3.connection.HTTPConnection object at 0x7c455bb94ac0>: Failed to establish a new connection: [Errno 111] Connection refused)) in elastic search

Try to running docker container based on first course module like this :

docker run -it \

--rm \

--name elasticsearch \

-p 9200:9200 \

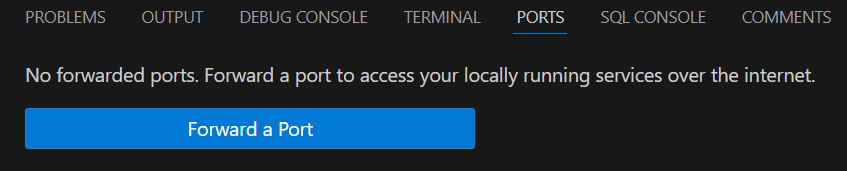
-p 9300:9300 \

-e "discovery.type=single-node" \

-e "xpack.security.enabled=false" \

docker.elastic.co/elasticsearch/elasticsearch:8.4.3

And don’t forget to forwarding your port 9200 if you’re using github codespace or run locally in vscode



## 

## Why do I get scores greater than 1 on my hits after querying my ElasticSearch database ?

As seen in this video: <https://www.youtube.com/watch?v=ptByfB_YcEg&t=102s>, we can get scores on obtained hits that are greater than 1 despite having a “cosine” similarity measure in our index settings. We would thus expect scores between -1 and 1. However, in the case of the final query, we have several scores additionned together to provide the final score:

* The KNN related score, which is between -1 and 1 (cosine similarity)
* The text relevance score: BM25 algorithm scores which can be any positive number, including above 1. This is a “ranking function which calculates score to represent a document's relevance with respect to query” (source: <https://stackoverflow.com/questions/43794749/what-is-bm25-and-why-elasticsearch-chose-this-algorithm-for-scoring-in-version-5>).

Since we have a “match” filter in our query, this triggers the usage of the BM25 ranking algorithm and the final score contains this information.

To get more details about the final scores, you can modify the search query and add an “explain” parameter:

response = es\_client.search(

index=index\_name,

query={

"match": {"section": "General course-related questions"},

},

knn=knn\_query,

size=5,

explain=True

)

Added by Mélanie Fouesnard

## 

## Not module named “sentence\_transformers”

For this module homework make sure you install the package sentence-transformers it can be installed as simply as:

**pip install sentence-transformers**

## 

## Can not create the index: Connection timeout.

I was getting this error at this step: es\_client.indices.create(index=index\_name, body=index\_settings)  
I checked the log of the elasticsearch server and running this command, the status was red: curl -X GET "<http://localhost:9200/_cluster/health?pretty>"

My problem was that I did not have enough disk space in my computer for docker images. I ended up removing unused ones, manually and pruning:

docker image prune

docker volume prune

docker container prune

Added by Ibai Irastorza

## TypeError: unsupported operand type(s) for \*: 'float' and 'dict' when running the vector search function within the evaluate function

Make sure your search function receives a query vector, not a dictionary. To resolve this, ensure that the q passed to the search\_function within evaluate is correctly transformed into an embedding vector. The following code can help:

v\_query = embedding\_model.encode(query\_text)

results = search\_function(v\_query)

## Find maximum of an numpy array (of any dimension):

max\_value = numpy\_array.max()

## What is the cosine similarity?

Cosine similarity is a measure used to calculate the similarity between two non-zero vectors, often used in text analysis to determine how similar two documents are based on their content. This metric computes the cosine of the angle between two vectors, which are typically word counts or TF-IDF values of the documents. The cosine similarity value ranges from -1 to 1, where 1 indicates that the vectors are identical, 0 indicates that the vectors are orthogonal (no similarity), and -1 represents completely opposite vectors.

## What are documents in ElasticSearch?

A “document” is a collection of fields, which are the key-value pairs that contain your data, that have been serialized as a JSON object.

## runing docker docker: Error response from daemon: Conflict. The container name "/elasticsearch" is already in use by container "20467e6723d78ff2e4e9e0c9a8b9580c07f070e4c852d12c585b1d71aefd6665". You have to remove (or rename) that container to be able to reuse that name. See 'docker run --help'.

docker stop elasticsearch

docker rm elasticsearch

How to scale Elastic search scores from [0, 1] to [-1, 1] to compare its results with your own ones, example calculating ranks using dot\_product metric ?

score = (es\_score - 0.5) \* 2

# Module 4: Monitoring

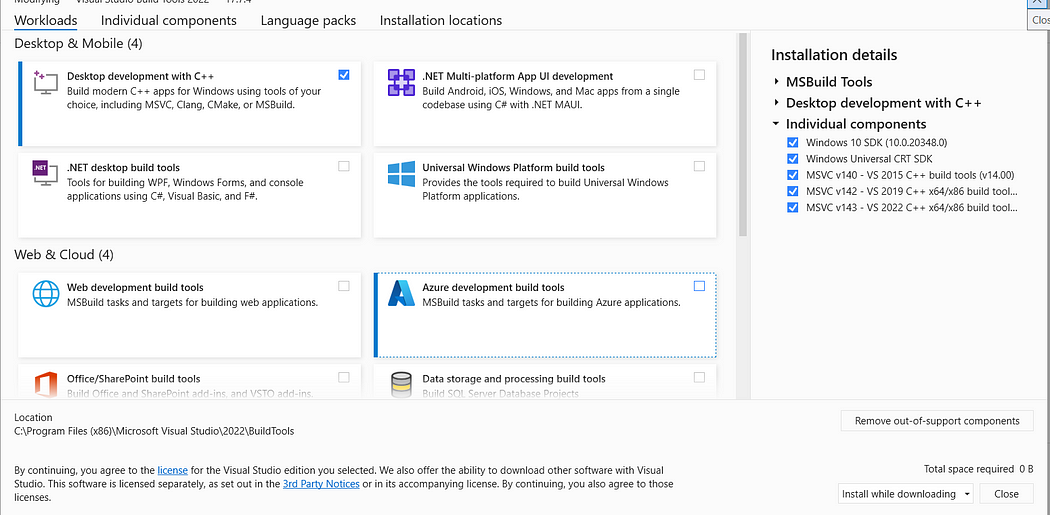
## Warning: 'model "multi-qa-mpnet-base-dot-v1" was made on sentence transformers v3.0.0 bet' how to suppress?

Upgrade `sentence-transformers` to v3.0.0>= e.x pip install sentence-transformers>=3.0.0 to avoid the warnings

## In Windows OS : OSError: [WinError 126] The specified module could not be found. Error loading "C:\Users\USER\AppData\Local\Programs\Python\Python310\lib\site-packages\torch\lib\fbgemm.dll" or one of its dependencies.

Solution 1 : Install Visual C++ Redistributable

Solution 2 : Install Visual Studio, not Visual Studio Code. Like in this depicted below and restart your system. For more details, please follow this link : https://discuss.pytorch.org/t/failed-to-import-pytorch-fbgemm-dll-or-one-of-its-dependencies-is-missing/201969

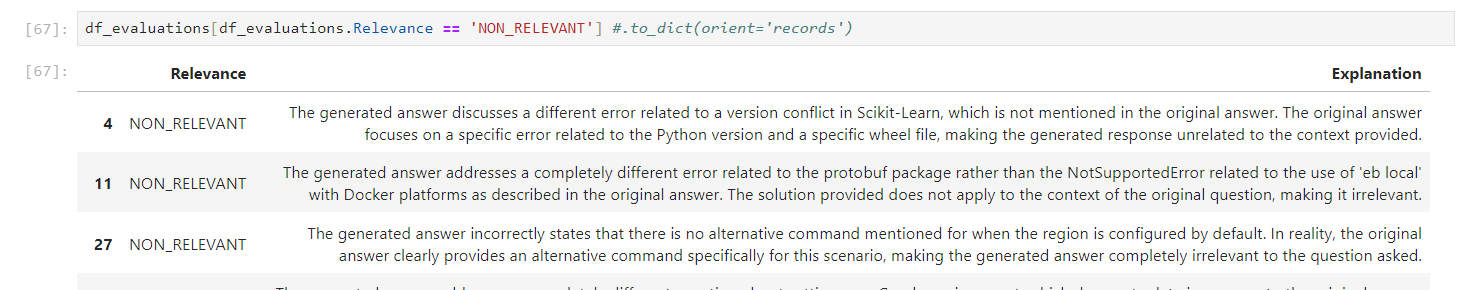


## OperationalError when running python prep.pypsycopg2. OperationalError: could not translate host name "postgres" to address: No such host is known. How do I fix this issue?

Inside .env file change POSTGRES\_HOST=localhost

## How set Pandas to show entire text content in a column. Useful to view the entire Explanation column content in the LLM-as-judge section of the offline-rag-evaluation notebook

By default, in the dataframe visualization, Pandas truncate the text content in a column to 50 characters. In order to view the entire explanation given by the judge llm for a NON RELEVANT answer, as in figure:



The instruction to show the results must be preceded by:

pd.set\_option('display.max\_colwidth', None)

Here are the specs for the display\_max\_colwidth option, as describide in the [official docs](https://pandas.pydata.org/docs/user_guide/options.html):

*display.max\_colwidth : int or None*

*The maximum width in characters of a column in the repr of*

*a pandas data structure. When the column overflows, a "..."*

*placeholder is embedded in the output. A 'None' value means unlimited.*

*[default: 50] [currently: 50]*

## How to normalize vectors in a Pandas DataFrame column (or Pandas Series)?

import numpy as np

normalize\_vec = lambda v: v / np.linalg.norm(v)

df["new\_col"] = df["org\_col"].apply(norm\_vec)

## How to compute the quantile or percentile of Pandas DataFrame column (or Pandas Series)?

To compute the 75% percentile or 0.75 quantile:

quantile: int = df["col"].quantile(q=0.75)

## How can I remove all Docker containers, images, and volumes, and builds from the terminal?

1. Delete all containers (including running ones):

```

docker rm -f

```

2. Remove all images:

```

docker rmi -f

```

3. Delete all volumes:

```

docker volume rm

```

# Module 5: X

## I have reached the orchestration pipeline's export and I’m facing a connection error at the stage of exporting to the vector database. Can someone help with the connection string?

Use the service name and port provided in the docker-compose.yaml file for the elasticsearch, e.g <http://><docker-compose-service-name>:<port> <http://elasticsearch:9200>

## Question

Answer

# Module 6: X

## Question

Answer

## Question

Answer

# Capstone Project

## Is it a group project?

No, the capstone is a solo project.

## Do we submit 2 projects, what does attempt 1 and 2 mean?

* You only need to submit 1 project.   
  If the submission at the first attempt fails, you can improve it and re-submit during **attempt#2** submission window.
* If you want to submit 2 projects for the experience and exposure, you must use different datasets and problem statements.
* If you can’t make it to the **attempt#1** submission window, you still have time to catch up to meet the **attempt#2** submission window
* Remember that the submission does not count towards the certification if you do not participate in the peer-review of 3 peers in your cohort

## Does the competition count as the capstone?

No, it does not (answered in office hours Jul 1st, 2024). You can participate in the [math-kaggle-llm-competition](https://datatalks-club.slack.com/archives/C0791HB4A58) as a group if you want to form teams; but capstone is an individual attempt.

## How is my capstone project going to be evaluated?

* Each submitted project will be evaluated by 3 (three) randomly assigned students who have also submitted the project.
* You will also be responsible for grading the projects from 3 fellow students yourself. Please be aware that: not complying to this rule also implies you failing to achieve the Certificate at the end of the course.
* The final grade you get will be the median score of the grades you get from the peer reviewers.
* And of course, the peer review criteria for evaluating or being evaluated must follow the guidelines defined here (TBA for link).

## Do I have to use ElasticSearch or X library?

Answer: No, you don’t have to use ElasticSearch. You can use any library you want. Just make sure it is documented so your peer-reviewers can reproduce your project.

# Certificates

## [IMPORTANT!] See [names on certificates](#_naukgwju1z8z)

## Question

Answer

# Workshops: dlthub

## Can I use the workshop materials for my own projects or share them with others?

Since dlt is open-source, we can use the content of this workshop for a capstone project. Since the main goal of dlt is to load and store data easily, we can even use it for other zoomcamps (mlops zoomcamp project for example). Do not hesitate to ask questions or use it directly in your projects.

Added by Mélanie Fouesnard

## How much free time does Google Colab gives for T4 GPU resource type?

## Google colab offers only 1 hour every 24h for using the T4 GPU resource type. But you can still use the CPU which is a bit slower than T4, especially while running the RAG.

## 

## There is an error when opening the table using dbtable = db.open\_table("notion\_pages\_\_\_homework"): FileNotFoundError: Table notion\_pages\_\_\_homework does not exist.Please first call db.create\_table(notion\_pages\_\_\_homework, data)

The error indicates that you have not changed all instances of “employee\_handbook” to “homework” in your pipeline settings

## There is an error when running main(): FileNotFoundError: Table notion\_pages\_\_\_homework does not exist.Please first call db.create\_table(notion\_pages\_\_\_homework, data)

Make sure you open the correct table in line 3: dbtable = db.open\_table("notion\_pages\_\_\_homework")

## How do I know which tables are in the db

You can use the db.table\_names() to list all the tables in the db

## Does DLT have connectors to ClickHouse or StarRocks?

Currently, DLT does not have connectors for ClickHouse or StarRocks but are open to contributions from the community to add these connectors.

## Notebook does not have secret access or 401 Client Error: Unauthorized for url: https://api.notion.com/v1/search

If you get this error

## 

Or 401 Client Error , then you either need to grant access to the key or the key is wrong.

## Error: How to fix requests library only installs v2.28 instead of v2.32 required for lancedb?

Install directly from source E.g `pip install "requests @ https://github.com/psf/requests/archive/refs/tags/v2.32.3.zip"`

# Workshops: X

## Connection refused error on prompting the ollam RAG?

If you get this error while doing the homework , simply restart the ollama server using nohup y running this line of the notebook !nohup ollama serve > nohup.out 2>&1 &

If you do stop and restart the cell, you will need to rerun the cell containing ollama serve first.

Added by Abiodun Gbadamosi

## Question

Answer