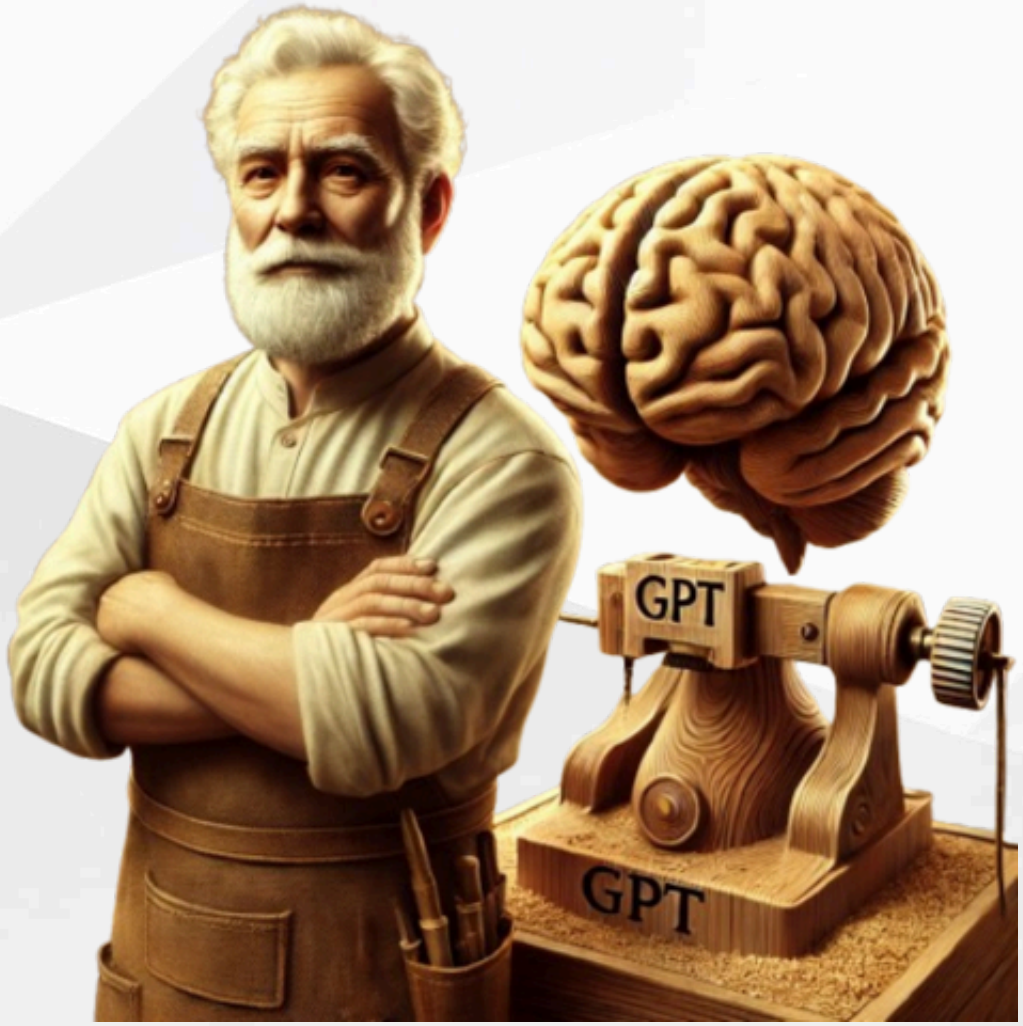
Developing Open LLM applications with

Apache OpenServerless

Lesson 7

A RAG system



A RAG system

- Implementing a RAG
- VectorDB loader for PDF documents
- A multi document, multi LLM RAG action

Implementing a RAG

RAG Concepts

- RAG is an acronym:
  - Retrieval-Augmented Generation
  - You augment your questions with context informations automatically provided
- You use your question to find relevant content
  - Embedding creates "semantic" vectors
  - Search the relevant content with Vector Search

Initializing VectorDB

```
import sys ; sys.path.append("packages/rag/rag")
import rag, vdb
```

Insert some content

```
args = {}
db = vdb.VectorDB(args, "jobs")
db.remove_by_substring("")
db.insert("Lisa is the first daughter of Steve Jobs.")
db.insert("Lisa full name is Lisa Brennon.")
db.insert("Jobs named his masterpiece computer after Lisa name.")
db.insert("Lisa name is is inspired by Mona Lisa.")
```

Creating a Context

Vector Search:

```
res = db.vector_search("Lisa")
res
```

Creating a context:

```
context = "Consider the following text:\n"
for i in res: context += f"{i[1]}\n"
context += "Answer to this prompt:\n"
print(context)
```

Query the LLM with Context

```
import rag
MODEL="llama3.1:8b"
question = "Who is Lisa?"
```

Without Context:

```
print(rag.llm(args, MODEL, question))
```

With context:

```
print(rag.llm(args, MODEL, context+question))
```

Best Practices for RAG

- Create significant chunks:
  - Typically 500-800 words, around 4000 characters
  - Better if they are semantically related
- Consider the context size:
  - `llama3.1:8b` has 128k context
  - so you can put around 30 chunks
- It is **essential** to structure the content properly for searches

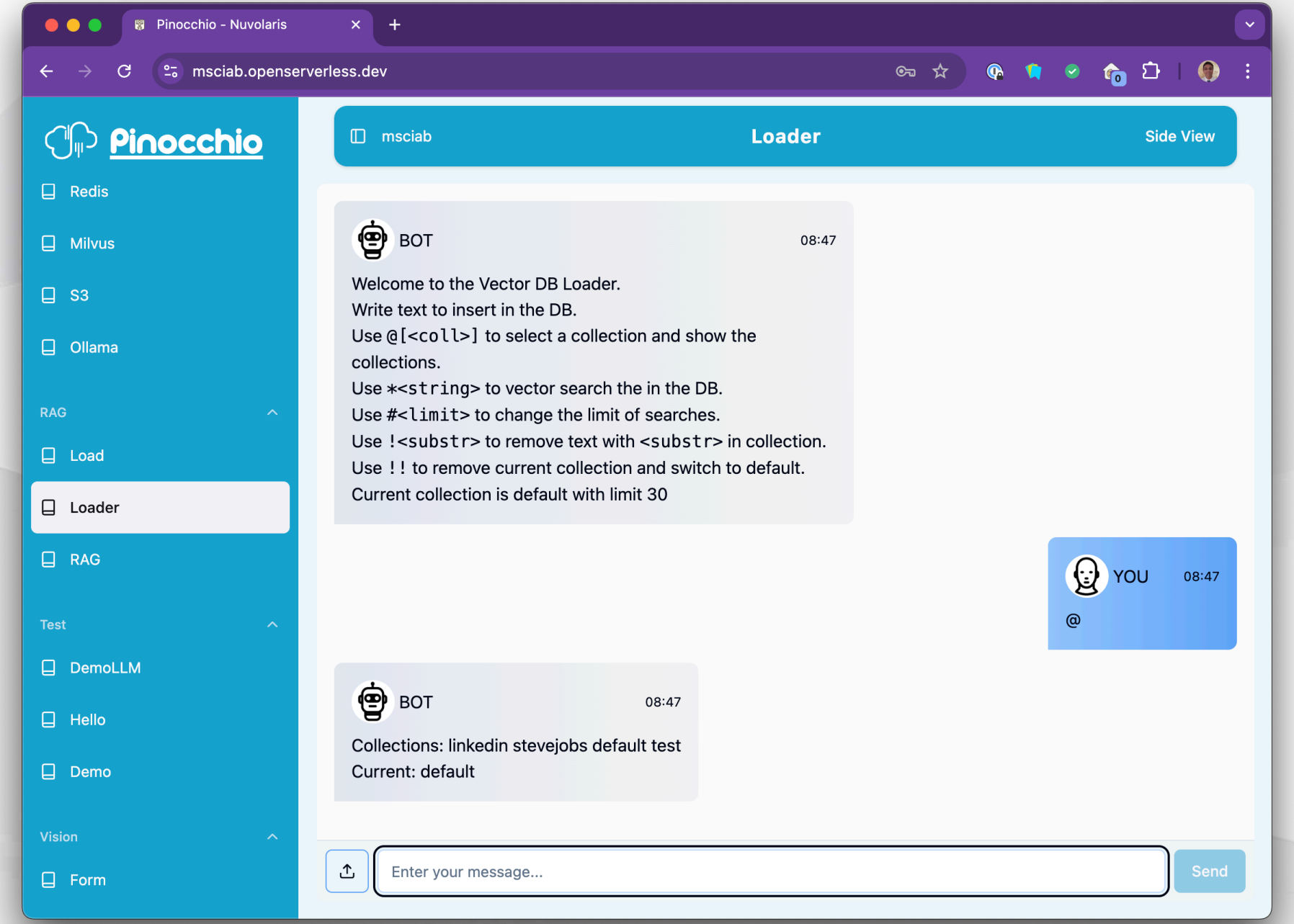
RAG Loader

RAG loader

A action to load the Vector DB:

```
Welcome to the Vector DB Loader.
Write text to insert in the DB.
Use `@[<coll>]` to select a collection and show the collections.
Use `#<limit>`  to change the limit of searches.
Use `*<string>` to vector search the <string>  in the DB.
Use `!<substr>` to remove text with `<substr>` in collection.
Use `!![<collection>]` to remove `<collection>` and switch to default.
```

- support multiple collections
- allows direct input of content
- also test searches and remove content and collections


Pinocchio - Nuvolaris
msciab.openserverless.dev
Pinocchio
Redis
Milvus
S3
Ollama
RAG
Load
Loader
RAG
Test
DemoLLM
Hello
Demo
Vision
Form
msciab
Loader
Side View
BOT
08:47
Welcome to the Vector DB Loader.
Write text to insert in the DB.
Use @[<coll>] to select a collection and show the collections.
Use *<string> to vector search the in the DB.
Use #<limit> to change the limit of searches.
Use !<substr> to remove text with <substr> in collection.
Use !! to remove current collection and switch to default.
Current collection is default with limit 30
YOU
08:47
@
BOT
08:47
Collections: linkedin stevejobs default test
Current: default
Enter your message...
Send

Rag Loader CLI

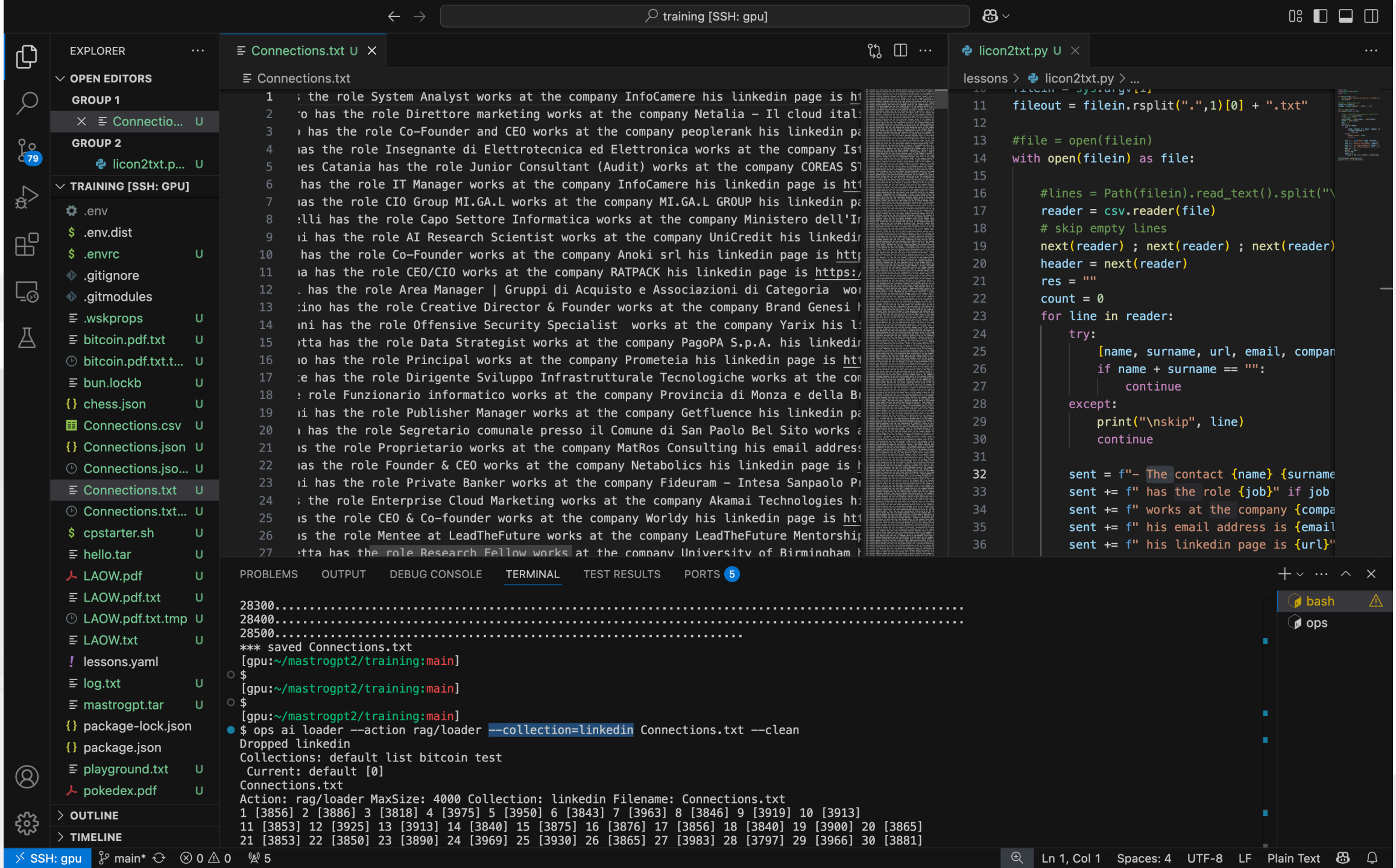
- Parses PDF in Text
- Import text in chunks

```
$ ops ai | grep loader
ai loader [--action=<action>] [--chunksize=<size>]
          [--collection=<name>] [--clean]
          <file>...
```

Use `--action=rag/loader` to select the loader action

Use `--action=<collection>` to select collection

Chunksize defaults to 4000


training [SSH: gpu]
EXPLORER
OPEN EDITORS
GROUP 1
Connectio... U
GROUP 2
licon2txt.p... U
TRAINING [SSH: GPU]
.env
.env.dist
.envrc U
.gitignore
.gitmodules
.wskprops U
bitcoin.pdf.txt U
bitcoin.pdf.txt.t... U
bun.lockb U
chess.json U
Connections.csv U
Connections.json U
Connections.jso... U
Connections.txt U
Connections.txt... U
cpstarter.sh U
hello.tar U
LAOW.pdf U
LAOW.pdf.txt U
LAOW.pdf.txt.tmp U
LAOW.txt U
lessons.yaml
log.txt U
mastrogpt.tar U
package-lock.json
package.json
playground.txt U
pokedex.pdf U
OUTLINE
TIMELINE
Connections.txt U
Connections.txt
licon2txt.py U
lessons > licon2txt.py > ...
fileout = filein.rsplit(".",1)[0] + ".txt"
#file = open(filein)
with open(filein) as file:
#lines = Path(filein).read_text().split("\
reader = csv.reader(file)
# skip empty lines
next(reader) ; next(reader) ; next(reader)
header = next(reader)
res = ""
count = 0
for line in reader:
try:
[name, surname, url, email, compan
if name + surname == "":
continue
except:
print("\nskip", line)
continue
sent = f"- The contact {name} {surname
sent += f" has the role {job}" if job
sent += f" works at the company {compa
sent += f" his email address is {email
sent += f" his linkedin page is {url}"
PROBLEMS
OUTPUT
DEBUG CONSOLE
TERMINAL
TEST RESULTS
PORTS 5
bash
ops
28300.....
28400.....
28500.....
*** saved Connections.txt
[gpu:~/mastrogpt2/training:main]
$
[gpu:~/mastrogpt2/training:main]
$
[gpu:~/mastrogpt2/training:main]
$ ops ai loader --action rag/loader --collection=linkedin Connections.txt --clean
Dropped linkedin
Collections: default list bitcoin test
Current: default [0]
Connections.txt
Action: rag/loader MaxSize: 4000 Collection: linkedin Filename: Connections.txt
1 [3856] 2 [3886] 3 [3818] 4 [3975] 5 [3950] 6 [3843] 7 [3963] 8 [3846] 9 [3919] 10 [3913]
11 [3853] 12 [3925] 13 [3913] 14 [3840] 15 [3875] 16 [3876] 17 [3856] 18 [3840] 19 [3900] 20 [3865]
21 [3853] 22 [3850] 23 [3890] 24 [3969] 25 [3930] 26 [3865] 27 [3983] 28 [3797] 29 [3966] 30 [3881]
SSH: gpu
main*
Ln 1, Col 1
Spaces: 4
UTF-8
LF
Plain Text

RAG Loader code

Loader Action

```
!code packages/rag/loader/loader.py
```

VectorDB Class

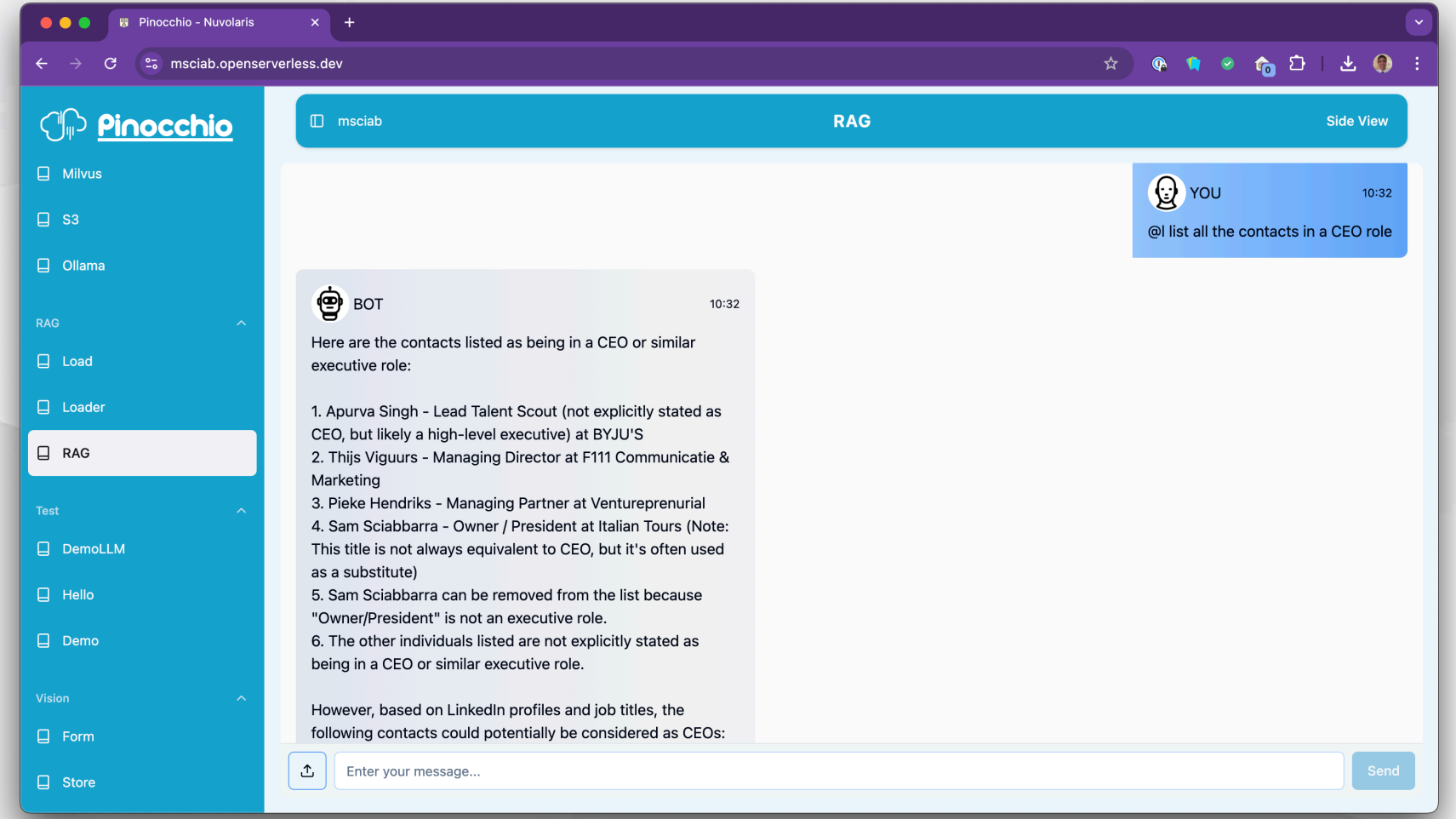
```
!code packages/rag/loader/vdb.py
```

RAG Query

RAG Query

- Extract text from DB and ask an LLM
- Supports multiple LLM, batch sizes and collections

```
Start with `@[LPM][<size>][<collection>]`
 to select the model then add `<size>` sentences
 from the `<collection>` to the context.
Models: L=llama P=phi4 D=deepseek M=mistral.
You can shorten collection names,
it will use the first one starting with the name.
Your query is then passed to the LLM
with the sentences for an answer.
```


Pinocchio - Nuvolaris
msciab.openserverless.dev
Pinocchio
Milvus
S3
Ollama
RAG
Load
Loader
RAG
Test
DemoLLM
Hello
Demo
Vision
Form
Store
msciab
RAG
Side View
YOU
10:32
@l list all the contacts in a CEO role
BOT
10:32
Here are the contacts listed as being in a CEO or similar executive role:
1. Apurva Singh - Lead Talent Scout (not explicitly stated as CEO, but likely a high-level executive) at BYJU'S
2. Thijs Viguurs - Managing Director at F111 Communicatie & Marketing
3. Pieke Hendriks - Managing Partner at Ventureprenurial
4. Sam Sciabbarra - Owner / President at Italian Tours (Note: This title is not always equivalent to CEO, but it's often used as a substitute)
5. Sam Sciabbarra can be removed from the list because "Owner/President" is not an executive role.
6. The other individuals listed are not explicitly stated as being in a CEO or similar executive role.
However, based on LinkedIn profiles and job titles, the following contacts could potentially be considered as CEOs:
Enter your message...
Send

RAG Query code

Parser

```
code packages/rag/rag/rag.py -g 20
```

Action

```
code packages/rag/rag/rag.py -g 110
```

Final Exercise

Modify the RAG loader to import images

- Modify the Database to accept also an image
- When an image is uploaded, process the image with image recognition and store its descritption
- Modify the RAG to be able to find images by their description