



**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 Brennan.")
db.insert("Jobs named his masterpiece computer after Lisa name.")
db.insert("Lisa name 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

Test

- DemoLLM
- Hello
- Demo

Vision

- Form

Loader

msciab

08:47

BOT

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

08:47

YOU

@

BOT

Collections: linkedin stevejobs default test
Current: default

Enter your message...

Send

Side View

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

The screenshot shows a terminal window with two tabs open: 'Connections.txt' and 'lcon2txt.py'.

Connections.txt:

```
1 ; the role System Analyst works at the company InfoCamere his linkedin page is https://www.linkedin.com/in/...
2 ;o has the role Direttore marketing works at the company Netalia - Il cloud italiano
3 ;o has the role Co-Founder and CEO works at the company peoplerank his linkedin page is https://www.linkedin.com/in/...
4 ;ias the role Insegnante di Elettrotecnica ed Elettronica works at the company Istituto Superiore di Catania has the role Junior Consultant (Audit) works at the company COREAS Srl
5 ;has the role IT Manager works at the company InfoCamere his linkedin page is https://www.linkedin.com/in/...
6 ;ias the role CIO Group MI.GA.L works at the company MI.GA.L GROUP his linkedin page is https://www.linkedin.com/in/...
7 ;elli has the role Capo Settore Informatica works at the company Ministero dell'Industria
8 ;ni has the role AI Research Scientist works at the company UniCredit his linkedin page is https://www.linkedin.com/in/...
9 ;has the role Co-Founder works at the company Anoki srl his linkedin page is https://www.linkedin.com/in/...
10 ;ia has the role CEO/CIO works at the company RATPACK his linkedin page is https://www.linkedin.com/in/...
11 ;. has the role Area Manager | Gruppi di Acquisto e Associazioni di Categoria works at the company Brand Genesi
12 ;ino has the role Creative Director & Founder works at the company Brand Genesi
13 ;ini has the role Offensive Security Specialist works at the company Yarix his linkedin page is https://www.linkedin.com/in/...
14 ;itta has the role Data Strategist works at the company PagoPA S.p.A. his linkedin page is https://www.linkedin.com/in/...
15 ;io has the role Principal works at the company Prometeia his linkedin page is https://www.linkedin.com/in/...
16 ;te has the role Dirigente Sviluppo Infrastrutturale Tecnologiche works at the company
17 ; role Funzionario informatico works at the company Provincia di Monza e della Brianza
18 ;ni has the role Publisher Manager works at the company Getfluence his linkedin page is https://www.linkedin.com/in/...
19 ;i has the role Segretario comunale presso il Comune di San Paolo Bel Sito works at the company
20 ;is the role Proprietario works at the company MatRos Consulting his email address is
21 ;as the role Founder & CEO works at the company Netabolics his linkedin page is https://www.linkedin.com/in/...
22 ;i has the role Private Banker works at the company Fideuram - Intesa Sanpaolo
23 ;; the role Enterprise Cloud Marketing works at the company Akamai Technologies his linkedin page is https://www.linkedin.com/in/...
24 ;is the role CEO & Co-founder works at the company Worldy his linkedin page is https://www.linkedin.com/in/...
25 ;is the role Mentee at LeadTheFuture works at the company LeadTheFuture Mentorship
26 ;itta has the role Research Fellow works at the company University of Birmingham
```

lcon2txt.py:

```
11 fileout = filein.rsplit(".",1)[0] + ".txt"
12
13 #file = open(filein)
14 with open(filein) as file:
15
16     #lines = Path(filein).read_text().split("\n")
17     reader = csv.reader(file)
18     # skip empty lines
19     next(reader) ; next(reader) ; next(reader)
20     header = next(reader)
21     res = ""
22     count = 0
23     for line in reader:
24         try:
25             [name, surname, url, email, company] = line
26             if name + surname == "":
27                 continue
28         except:
29             print("\nskip", line)
30             continue
31
32         sent = f"- The contact {name} {surname}"
33         sent += f" has the role {job}" if job else ""
34         sent += f" works at the company {company}" if company else ""
35         sent += f" his email address is {email}" if email else ""
36         sent += f" his linkedin page is {url}" if url else ""
```

Terminal Output:

```
28300.
28400.
28500.
*** saved Connections.txt
[gpu:~/mastrogt2/training:main]
$ 
$ 
[gpu:~/mastrogt2/training:main]
$ 
$ 
[gpu:~/mastrogt2/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]
```

Bottom Status Bar:

```
Ln 1, Col 1 Spaces: 4 UTF-8 LF Plain Text
```

Page Number: 13

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.

The screenshot shows a web browser window titled "Pinocchio - Nuvolaris" with the URL "msciab.openserverless.dev". The interface is a conversational AI tool. On the left, there's a sidebar with a "Pinocchio" logo and sections for Milvus, S3, Ollama, RAG, Test, DemoLLM, Hello, Demo, Vision, Form, and Store. The "RAG" section is currently selected. The main area has a teal header with the user handle "msciab" and the title "RAG". A "Side View" button is in the top right. The conversation log shows a message from "YOU" at 10:32: "@l list all the contacts in a CEO role". Below it, the "BOT" responds at 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 Venturepreneurial
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."

The BOT continues: "However, based on LinkedIn profiles and job titles, the following contacts could potentially be considered as CEOs:"

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 description
- Modify the RAG to be able to find images by their description