



LLM Context Pipeline

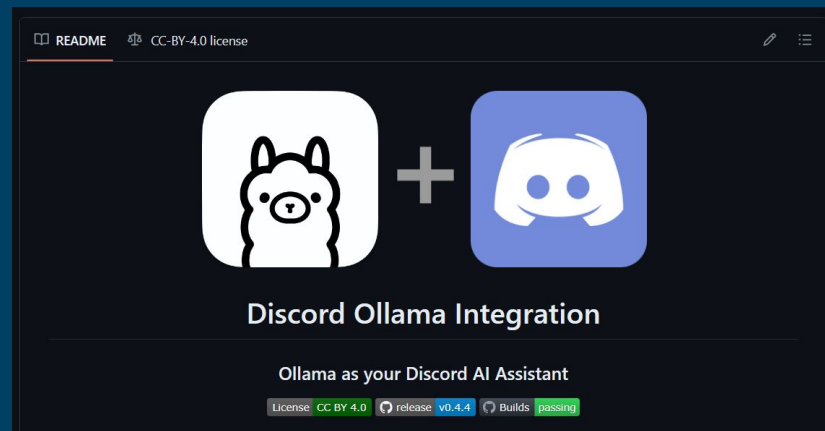


CPSC 324 - Big Data Analytics
Jonathan Smoley



Goal

- **Discord-Ollama**: open-source LLM chat bot
- **Provide Purpose**: take a cool project and give it a clear direction to move in
- **Automation**: simplify data collection for model file context



Examples

Basic Modelfile

An example of a `Modelfile` creating a mario blueprint:

```
FROM llama3
# sets the temperature to 1 [higher is more creative, lower is more coherent]
PARAMETER temperature 1
# sets the context window size to 4096, this controls how many tokens the LLM can use as context to generate the next token
PARAMETER num_ctx 4096

# sets a custom system message to specify the behavior of the chat assistant
SYSTEM You are Mario from super mario bros, acting as an assistant.
```

To use this:

1. Save it as a file (e.g. `Modelfile`)
2. `ollama create choose-a-model-name -f <location of the file e.g. ../Modelfile>`
3. `ollama run choose-a-model-name`
4. Start using the model!

Data Pipeline

Data Collection

- Web scrape www.gonzaga.edu
 - Alternative: University API
- Run on fixed interval
- Extract text and deeper links from accessible domains

Data Preparation

- Pull when current data is available
- Extract names, places, things, etc.
- Form contextual clues from information on Gonzaga University
- Standardize clues into a model file to be used in “training” an LLM

The Data Itself

During Web Scraping:

- Found in HTML documents available at www.gonzaga.edu
- HTML to Text, stored as separate instances found in documents

During Processing:

- Clue type, context found in, datetime found, etc.
- Fit for document storage (script of context clues)

Challenges:

- Not easy to recognize which text is valuable to an LLM, at least with code alone
 - An addition of NLP analysis would come in handy here

Services Involved

- Cloud Scheduler
- Pub/Sub
- Cloud Functions
- Cloud Storage

Cloud Scheduler										Jobs		CREATE JOB	REFRESH	FORCE RUN	EDIT	COPY	PAUSE	RESUME	DELETE	LEARN
SCHEDULER JOBS										APP ENGINE CRON JOBS										
Filter Filter jobs																				
<input checked="" type="checkbox"/>	Name	Status of last execution	Region	State	Description	Frequency	Target	Last run	Next run	Actions										
<input checked="" type="checkbox"/>	gonzaga-scraper	Success	us-west1	Enabled	trigger a web scraper for the gonzaga public website	0 */3 * * * (America/Los_Angeles)	Topic : projects/ollama-context-pipeline/topics/scrape-gonzaga-website	May 10, 2024, 9:37:46 PM	May 11, 2024, 12:00:00 AM	⋮										

Topics

CREATE TOPIC

DELETE

SHOW INFO PANEL

LIST

METRICS

Filter

Filter topics

<input type="checkbox"/>	Topic ID <div>↑</div>	Encryption key	Topic name	Retention	Ingestion source
<input type="checkbox"/>	scrape-gonzaga-website	Google-managed	projects/ollama-context-pipeline/topics/scrape-gonzaga-website <div></div>	—	— <div></div>

Cloud Functions										Functions		CREATE FUNCTION	REFRESH	LEARN	RELEASE NOTES
Filter Filter functions															
<input type="checkbox"/>	Environment	Name	Last deployed	Region	Recommendation	Trigger	Runtime	Memory allocated	Executed function	Actions					
<input type="checkbox"/>	2nd gen	data-parser	May 10, 2024, 9:30:46 PM	us-west1		Bucket: gonzaga-scraper-bucket	Python 3.12	512 MiB	handler	⋮					
<input type="checkbox"/>	2nd gen	web-scraper	May 10, 2024, 9:36:59 PM	us-west1		Topic: scrape-gonzaga-website	Python 3.12	512 MiB	handler	⋮					

Bucket details

GO TO PATH

REFRESH

LEARN

gonzaga-scraper-bucket

Location

Storage class

Public access

Protection

us-west1 (Oregon)

Standard

Not public

Soft Delete

OBJECTS

CONFIGURATION

PERMISSIONS

PROTECTION

LIFECYCLE

OBSERVABILITY

INVENTORY REPORTS

OPERATIONS

Folder browser

gonzaga-scraper-bucket

Buckets > gonzaga-scraper-bucket

UPLOAD FILES

UPLOAD FOLDER

CREATE FOLDER

TRANSFER DATA

MANAGE HOLDS

EDIT RETENTION

DOWNLOAD

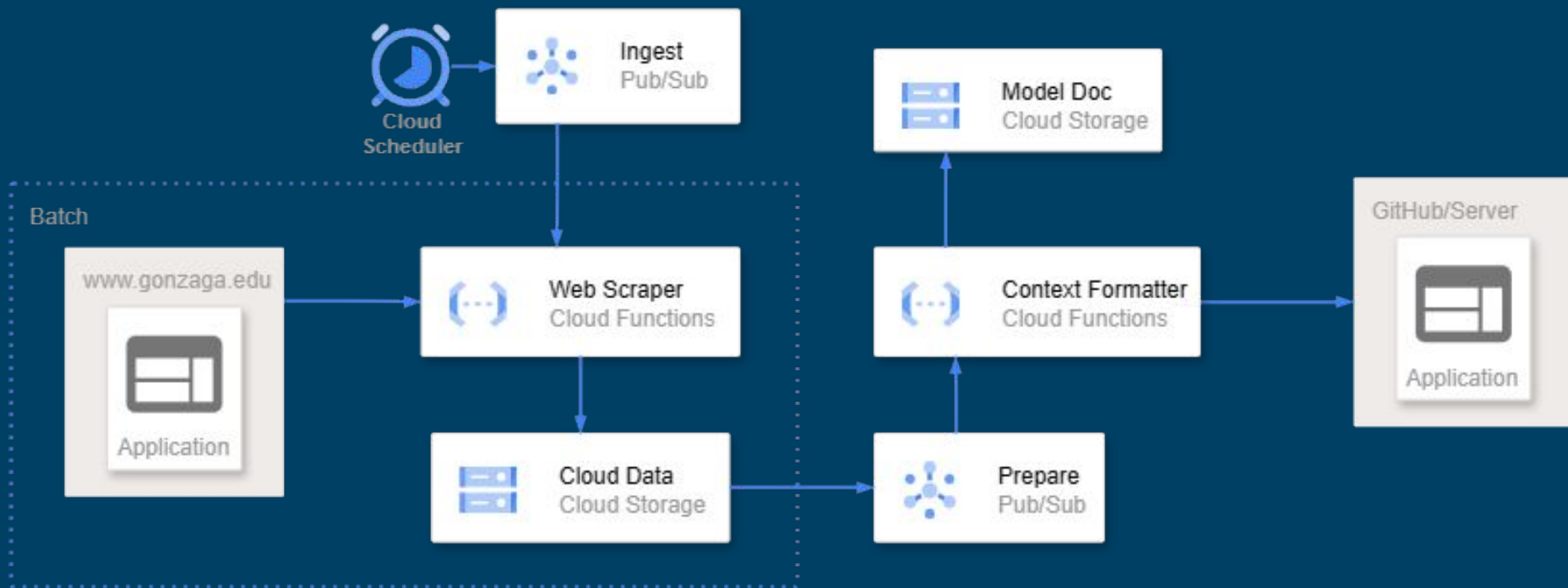
DELETE

Filter by name prefix only

Filter objects and folders

Show Live objects only

<input type="checkbox"/>	Name	Size	Type	Created	Storage class	Last modified	Public	
<input type="checkbox"/>	data.txt	24 B	text/plain	May 10, 2024, 9:37:50 PM	Standard	May 10, 2024, 9:37:50 PM	Not public	
<input type="checkbox"/>	parser.zip	1.3 KB	application/zip	May 10, 2024, 9:28:00 PM	Standard	May 10, 2024, 9:28:00 PM	Not public	
<input type="checkbox"/>	scraper.zip	2.6 KB	application/zip	May 10, 2024, 9:35:02 PM	Standard	May 10, 2024, 9:35:02 PM	Not public	



Demo

- Services working together
- Existing scraper work
- Automation

Future Work

Short-Term

- Finalize web scraper
 - Assess typical document structure on www.gonzaga.edu
 - Choose elements to scan for
- Decide a final destination for results
 - Cloud Storage -> GitHub
 - Model File -> GitHub

Long-Term

- Build out parser
 - ML/NLP solution to look for keywords in text collected
- Set up model file (context clues)
 - Design sentence structure

Resources

- Ollama: <https://ollama.com/>
- Discord-Ollama: <https://github.com/kevinthedang/discord-ollama>