



AI and opportunities for archives

Henk Vanstappen, Datable

DLM Forum Members' Meeting
National Archives of Belgium
Brussels, May 28th 2024

Artificial intelligence?

AI is **technology** that enables computers and machines to **simulate human intelligence** and **problem-solving capabilities**.

(IBM)

./DATAFILE

Generative AI?

Generative AI refers to deep-learning **models that can generate** high-quality text, images, and other **content** based on the data they were trained on.

(IBM)

./DATAFILE



+ New Chat



Previous 30 Days

Match 'laars' to keyword 'sche

September

NLP Python Script Assistanc

Frie Leysen's cultural i

August

Regex Pattern: Structure Mat

Translate Text with Glossary

July

OCR Results CSV Export

Exclude "other decorative tec

Quasi-Mono RGB Detection

June

Legal Library Painting Gift.

XML Validation with XSD

Upgrade to Plus



Henk Vanstappen



aan het bevorderen van de kunsten werden zeer gewaardeerd in de Belgische en internationale culturele gemeenschap.

< 2 / 2 >



door wie werd ze in desingel opgevolgd?

< 2 / 2 >



Na haar vertrek als intendant van deSingel werd Frie Leysen opgevolgd door **Bart Vanvoorden**. Bart Vanvoorden was een Belgische theatermaker en cultuurmanager. Hij nam de leiding over deSingel op zich en zette het werk voort dat Frie Leysen had gestart, waarbij hij zich richtte op het promoten van innovatieve en diverse podiumkunsten en het bevorderen van culturele uitwisseling. Het culturele centrum bleef onder zijn leiding een belangrijk platform voor kunstenaars en publiek in Antwerpen en België. Het is echter mogelijk dat er na mijn laatste kennisupdate in september 2021 verdere ontwikkelingen hebben plaatsgevonden, dus het is raadzaam om de meest recente informatie te raadplegen voor de laatste updates over deSingel en zijn leiderschap.



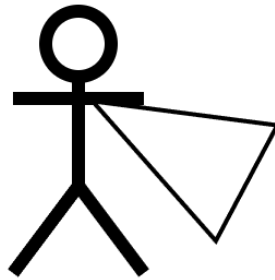
Regenerate

Send a message

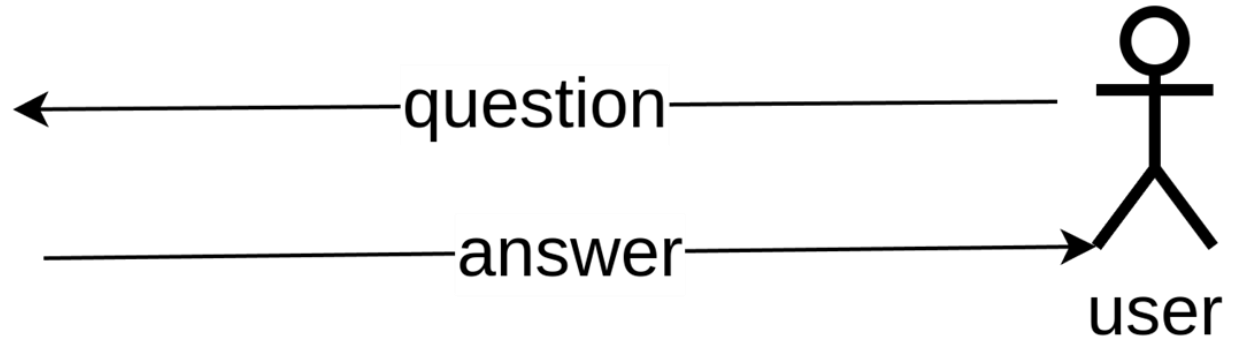


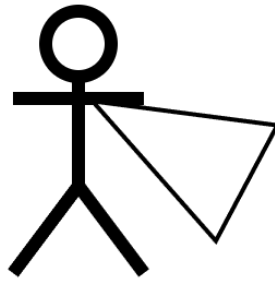
onderwerp

1. unstructured > structured

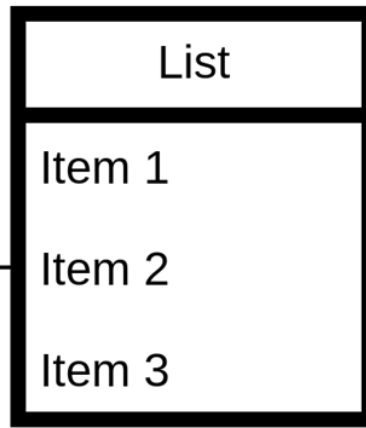


information
professional





information
professional



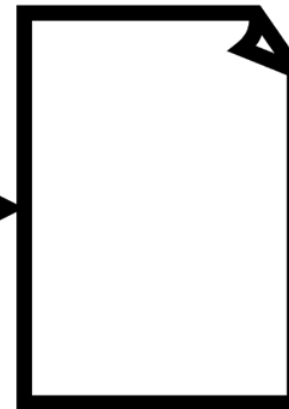
query

response



user

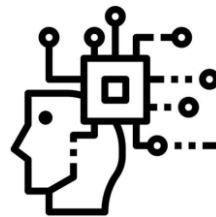
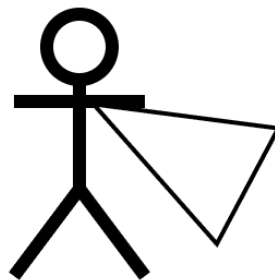
retrieve document



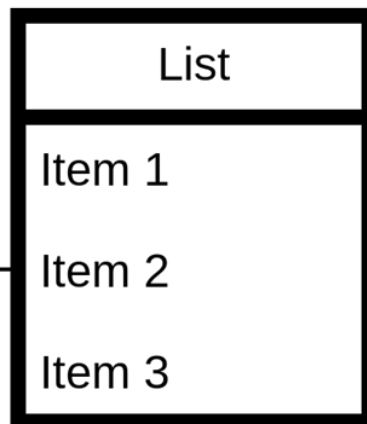
find answer

Information professional's toolbox

Vocabularies - Taxonomies - Ontologies -
Metadata - Controlled Vocabularies - Linked Open
Data - Semantic Web - Knowledge Graphs - UDC
- EAD - EAC - ISAD(G) - ISAAR(CPF) - ISBD -
AACR2 - BIBtex - ALTO - METS - BAG - MARC -
RDA - SKOS - PREMIS - LCSH - FRBR - JSON -
XML - IIIF - SPARQL - SQL - CDWA - CIDOC-
CRM - CSV - DACS - DDC - DOI - EXIF - FOAF -
GEDCOM - DACS - DAI - ISNI - ISILI - EDM -
ICONCLASS - ISDIAH - IEEE/LOM - 739.50



information
professional



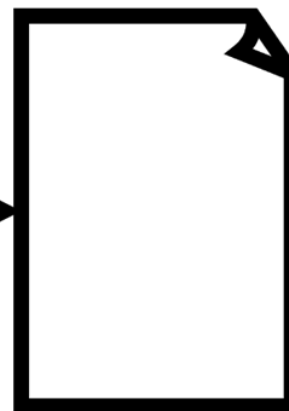
query

response



user

retrieve document



find answer

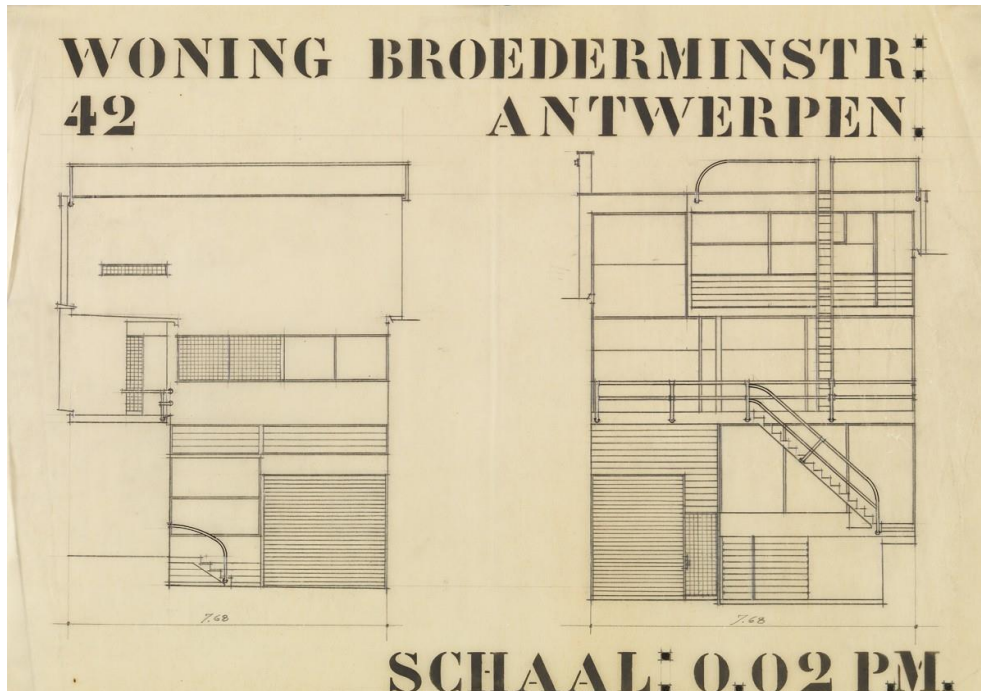
object detection



facade
building
city
bus
vehicle
window
landmark
plants
high building

Tagging 30K slides from the archives of the Vlaams Architecture Institute with Google Vision, © Christian Kieckens / VAI

image classification



- ☐ photo
- ☒ drawing
- ☐ building plan
- ☐ text

Classify 10K scans from the archives of the Vlaams Architecture Institute with ClarifAI, © VAI

document classification

Festus Pragnell
(FRANK WILLIAM PRAGNELL)

*Magazine Writer. Author of Novels of
Science Fiction and Adventure*

152, ST. MARY STREET
SOUTHAMPTON
ENGLAND

30th November, 1950.

Mr Forrest J Ackerman,
236 $\frac{1}{2}$ N New Hampshire,
Hollywood California.

Dear Forry,

Many ~~thaks~~ for your last letter, and for your efforts on my behalf. I enclose a letter and statement for Standard Publications, with a copy for you to retain. If you agree with this way of handling the matter perhaps you would forward the letter to them.

I shall be glad to see the new mag., Galaxy. Confidentially, the peculiar behaviour of Standard seem to me to be the desperate tricks of a firm that is getting into serious financial difficulties.

All the very best,

F W Pragnell

document classification

```
*classify_FB_bart_large.py - /home/henk/Python/text_classifiers/classify_FB_bart_large.py (3.10.12)*
File Edit Format Run Options Window Help
1 import json
2 from transformers import pipeline
3 classifier = pipeline("zero-shot-classification", model="facebook/bart-large-mnli")
4
5 sequence_to_classify = '''Seabied PRagnell
6
7 (FRANK WILLIAM PRAGNELL)
8
9 Magazine Writer. Author of Novels of
10 Science Fiction and Adventure
11
12 152, ST. MARY STREET
13
14 SOUTHAMPTON N
15 Mr Forrest J Ackerman, - pycranp 30th Nevember, 1950.
16
17 236% N New Hampshire,
18 Hollywood California.
19 Dear Forry,
20
21 Many thakks for your last letter, and for your efforts on my behalf. I enclose a letter
22 I shall be glad to see the new mag., Galaxy.
23 Confidentially, the peculiar behavidus of Standard seem to me to be the desperate tricks
24
25 All the very best, Q.
26 (19 (rag'''
27
28 candidate_labels = ['news', 'letter', 'manual', 'contract']
29 result = classifier(sequence_to_classify, candidate_labels, multi_label=True)
30 scores = result['scores']
31 labels = result['labels']
32 print(f'This is a {labels[0]} (score = {scores[0]})')
33
```

document classification

```
*classify_FB_bart_large.py - /home/henk/Python/text_classifiers/classify_FB_bart_large.py (3.10.12)*
File Edit Format Run Options Window Help
1 import json
2 from transformers import pipeline
3 classifier = pipeline("zero-shot-classification", model="facebook/bart-large-mnli")
4
5 sequence_to_classify = '''Seabied PRagnell
6
7 (FRANK WILLIAM PRAGNELL)'''
```

```
IDLE Shell 3.10.12 <2>
File Edit Shell Debug Options Window Help
Python 3.10.12 (main, Nov 20 2023, 15:14:05) [GCC 11.4.0] on l
Type "help", "copyright", "credits" or "license()" for more in
>>>
==== RESTART: /home/henk/Python/text_classifiers/classify_FB_
This is a letter (score = 0.9445213079452515)
>>>
```

```
27
28 candidate_labels = ['news', 'letter', 'manual', 'contract']
29 result = classifier(sequence_to_classify, candidate_labels, multi_label=True)
30 scores = result['scores']
31 labels = result['labels']
32 print(f'This is a {labels[0]} (score = {scores[0]})')
33
```


semantic search

```
*RAG_langchain_test_search_local.py - /home/henk/Python/THESAIV2_demo_en/RAG_langchain_test_search_local.py (3.10.12)*
File Edit Format Run Options Window Help
1 # SET model name and source dataset in parameters.py
2 # imports
3 from langchain_community.vectorstores import Chroma
4 from langchain_community.embeddings.sentence_transformer import (
5     SentenceTransformerEmbeddings,
6 )
7
8 #import parameters from parameters.py
9 from parameters import model_name, persist_directory
10 print(f'Model name: {model_name}')
11 print(f'embeddings: {persist_directory}')
12
13 embedding_function = SentenceTransformerEmbeddings(model_name=model_name)
14
15 query = "officer's clothes"
16
17 # load from disk
18 db = Chroma(persist_directory=persist_directory, embedding_function=embedding_
19 docs = db.similarity_search_with_score(query, k=5)
20 print("\n===Similarity results===")
21 print(f'Query was: {query}\n')
22 sources = set()
23 for doc in docs:
24     content = doc[0].page_content
25     source = doc[0].metadata.get('source')
26     key = f"{source.rsplit('/', 1)[-1]}"
27     print(f"content: {content}")
28     print(f"source: {key}\n")
29
```

Ln: 19 Col: 0

semantic search

```
*RAG_langchain_test_search_local.py - /home/henk/Python/THESAIv2_demo_en/RAG_langchain_test_search_local.py (3.10.12)*
File Edit Format Run Options Window Help
1 # SET model name and source dataset in parameters.py
2 # imports
3 from langchain community.vectorstores import Chroma
4 from
5
6 )
7
8 #imp
9 from
10 prin
11 prin
12
13 embe
14 source:
15 quer
16
17 # lo
18 db =
19 docs
20 prin
21 prin
22 sour
23 for
24
25
26
27
28
29
>>>
```

```
IDLE Shell 3.10.12 <2>
File Edit Shell Debug Options Window Help
===Similarity results===
Query was: officer's clothes

content:  men's clothing
source:   10435.json

content:  uniform (professional wear)
source:   10116.json

content:  boys' clothing
source:   11099.json

content:  women's clothing
source:   10434.json

content:  clothing by wearer
source:   10433.json
```

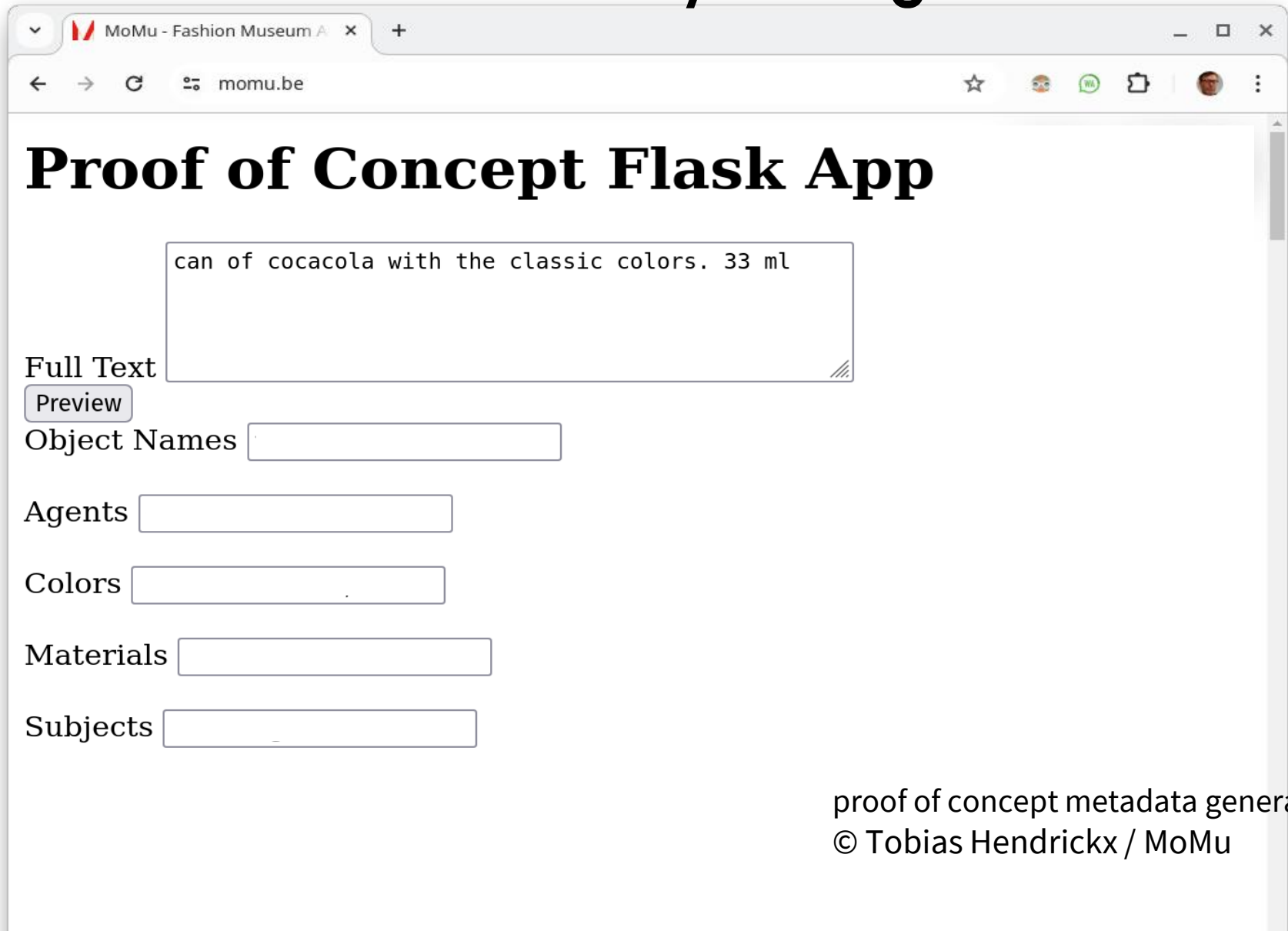

document Q&A

```
*text_QA.py - /home/henk/Python/text_QA/text_QA.py (3.10.12)*
File Edit Format Run Options Window Help
1 from transformers import pipeline
2
3 # Load the question answering pipeline
4 qa_model = pipeline("question-answering")
5
6 # Define the question and context
7 question1 = "Who wrote this letter?"
8 question2 = "To whom is this letter addressed?"
9 question3 = "When was the letter written?"
10
11 context =
12 Maandag, den 20ste februari 1920
13
14 Vriend Guido,
15 Lees ne keer dit briefke van Neut werk in gevolge van het geen gy daeruit peizen zult.
16 Ik heb vandage geschreven aen onzen vriend dat hy zou wachten het liedje te geven tot dat gy hem geantwoord hebt.
17 Indien het u mogelyk ware tegen morgen uchtend uwe verandering met den eersten post te zenden, het zou wonder wel zyn.
18 't eerste gesleep van den yzeren weg gaet ten 6,50 's morgens uit Brugge weg. Fraei, zet u aen 't werk.
19 De kalver pooten zyn iets dat in 't liedje tegen Mevrouw de Hemptine stond.
20 Neut leest pater voor sater 't is hem reeds gezeid.
21
22 Nu, in d' hope, groet ik u & vader ook, zeg het hem.
23
24 Tot ziens!
25 B Vanhove
26 Presbiter
27 '''
28
29 # Use the pipeline to find the answer
30 answer1 = qa_model(question=question1, context=context)
31 answer2 = qa_model(question=question2, context=context)
32 date = qa_model(question=question3, context=context)
33
34 # Print the answer
35 print(answer1)
36 print(answer2)|
```

document Q&A

```
*text_QA.py - /home/henk/Python/text_QA/text_QA.py (3.10.12)*
File Edit Format Run Options Window Help
1 from transformers import pipeline
2
3 # Load the question answering pipeline
4 qa_model = pipeline("question-answering")
5
6 # Define the question and context
7 question1 = "Who wrote this letter?"
8 question2 = "To whom is this letter addressed?"
9 question3 = "When was the letter written?"
10
11
12 IDLE Shell 3.10.12 <3>
13 File Edit Shell Debug Options Window Help
14 Python 3.10.12 (main, Nov 20 2023, 15:14:05) [GCC 11.4.0] on linux
15 Type "help", "copyright", "credits" or "license()" for more information.
16 >>>
17 ===== RESTART: /home/henk/Python/text_QA/text_QA.py =====
18 No model was supplied, defaulted to distilbert/distilbert-base-cased-distilled-squad and revision 626af31 (https://huggingface.co/distilbert/distilbert-base-cased-distilled-squad).
19 Using a pipeline without specifying a model name and revision in production is not recommended.
20 .
21 {'score': 0.3932872414588928, 'start': 667, 'end': 686, 'answer': 'B Vanhove\nPresbiter'}
22 {'score': 0.019157258793711662, 'start': 667, 'end': 686, 'answer': 'B Vanhove\nPresbiter'}
23 {'score': 0.11984480917453766, 'start': 29, 'end': 33, 'answer': '1920'}
24 >>>
25
26 Ln: 10 Col: 0
27
28
29 # Use the pipeline to find the answer
30 answer1 = qa_model(question=question1, context=context)
31 answer2 = qa_model(question=question2, context=context)
32 date = qa_model(question=question3, context=context)
33
34 # Print the answer
35 print(answer1)
36 print(answer2)
```

named entity recognition



MoMu - Fashion Museum A x momu.be

Proof of Concept Flask App

can of cocacola with the classic colors. 33 ml

Full Text

Preview

Object Names

Agents

Colors

Materials

Subjects

proof of concept metadata generator,
© Tobias Hendrickx / MoMu

named entity recognition

MoMu - Fashion Museum A x momu.be

Proof of Concept Flask App

can of cocacola with the classic colors. 33 ml

Full Text

Preview

Object Names 1. Can 2. Coca

Agents Coca-Cola

Colors Classic colors: Red, White

Materials 1. Can of Coca-Cola

Subjects 1. Beverages 2. Soft

proof of concept metadata generator,
© Tobias Hendrickx / MoMu

unstructured > structured

- image tagging
- image classification
- document classification
- semantic search
- document Q&A
- named entity recognition / topic detection
- face recognition
- optical character recognition
- summarization
- ...

onderwerp

2. structured > unstructured

Synthetic text



Fashion show France Andrevie

France, Paris, October 1978. Fashion show of France Andrevie, spring/summer 1979 women's ready-to-wear collection.

This item is provided and maintained by Paul van Riel

[Good to know](#)

[All metadata](#)

[Location](#)

Creator

 France Andrevie ; France Andrevie (Designer) ; Paul van Riel (Photographer) ; Paul van Riel

Type of item

Object Type: spring-summer ; Object Type: womenswear ; Object Type: ready-to-wear ; Object Type: fashion show ; [Fashion show](#) ; [Ready-to-wear](#) ; [Spring-summer](#) ; [Womenswear](#) ; spring-summer ; womenswear ; ready-to-wear ; fashion show

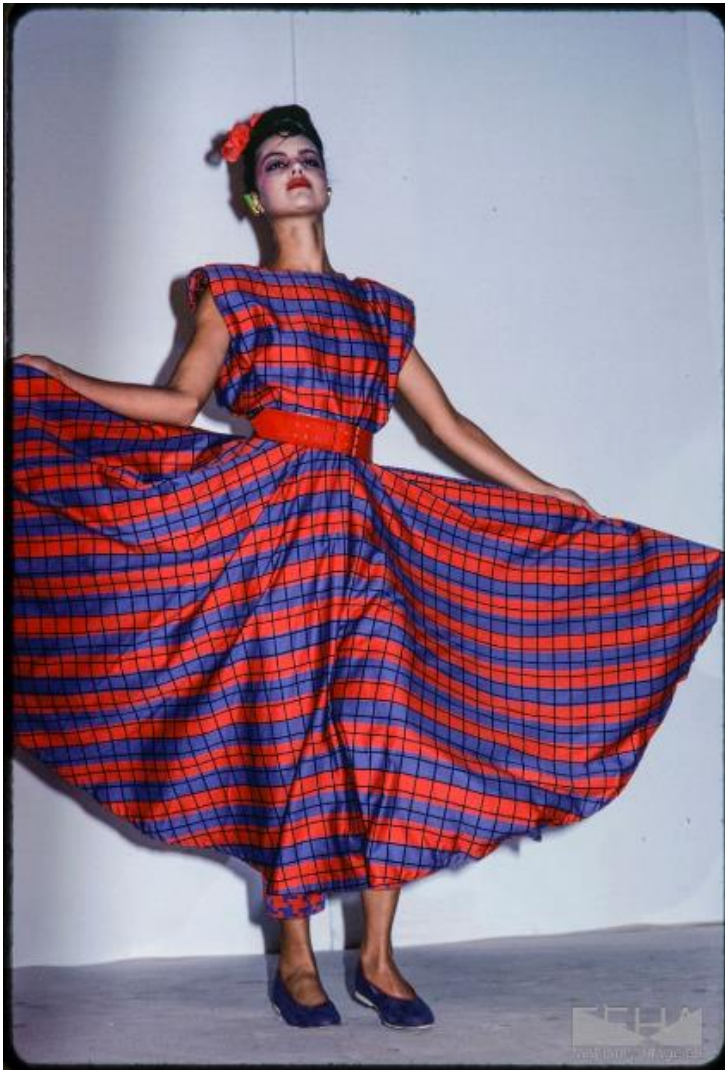
Date

15-10-1978 ; 1978-10-15

File Edit Format Run Options Window Help

```
6 from pathlib import Path
7 from gpt4all import GPT4All
8 import csv
9 import time
10 import os
11
12 model_name = 'llama-2-7b-chat.Q5_K_M.gguf'
13 model_path = Path.home() / '.local' / 'share' / 'nomic.ai' / 'GPT4All' #set mode
14 model = GPT4All(model_name, allow_download=False, device='cpu', verbose=True)
15 total_time = 0
16 temperature = 0.0
17
18 system_template = '''You are a helpfull assistant who writes factual description
19 You will be provided with some key value pairs, and transform them in a concise
20 Only use information that is available in the key-value pairs. Don't add estheti
21
22 prompt_template = '''DATA: "object type = handkerchief
23 designer = Delrue, Yves
24 creation date = between 1958 and 1958
25 creation location = Antwerp
26 DESCRIPTION: "Handkerchief designed by Yves Delrue in 1958 in Antwerp."
27
28 DATA: "object type = fashion collection, womenswear, spring-summer
29 creation date = 2013
30 designer = Emerson
31 creation location = Antwerp
32 photographer = Etienne Tordoir"
33 DESCRIPTION: "Look from the Emerson spring/summer 2013 womenswear collection, ph
34
35 DATA: "object type = fashion collection
36 designer = French fashion house and company Balenciaga
37 creation date = 2013
38 photographer = Etienne Tordoir."
39 DESCRIPTION: "Look from the 2013 collection from the French fashion house Balenc
40
```


Synthetic text



Fashion show France Andrevie

France, Paris, October 1978. Fashion show of France Andrevie, spring/summer 1979 women's ready-to-wear collection.

This item is provided and maintained by Paul van Riel

[Good to know](#)

[All metadata](#)

[Location](#)

Look from the 1979 spring/summer women's ready to wear collection, created by the French designer France Andrevie and photographed by dutch fashion photographer Paul van Riel.

onderwerp

3. unstructured > unstructured



+ New Chat



Previous 30 Days

Match 'laars' to keyword 'sche

September

NLP Python Script Assistanc

Frie Leysen's cultural i

August

Regex Pattern: Structure Mat

Translate Text with Glossary

July

OCR Results CSV Export

Exclude "other decorative tec

Quasi-Mono RGB Detection

June

Legal Library Painting Gift.

XML Validation with XSD

Upgrade to Plus



Henk Vanstappen



aan het bevorderen van de kunsten werden zeer gewaardeerd in de Belgische en internationale culturele gemeenschap.

< 2 / 2 >



door wie werd ze in desingel opgevolgd?

< 2 / 2 >



Na haar vertrek als intendant van deSingel werd Frie Leysen opgevolgd door **Bart Vanvoorden**. Bart Vanvoorden was een Belgische theatermaker en cultuurmanager. Hij nam de leiding over deSingel op zich en zette het werk voort dat Frie Leysen had gestart, waarbij hij zich richtte op het promoten van innovatieve en diverse podiumkunsten en het bevorderen van culturele uitwisseling. Het culturele centrum bleef onder zijn leiding een belangrijk platform voor kunstenaars en publiek in Antwerpen en België. Het is echter mogelijk dat er na mijn laatste kennisupdate in september 2021 verdere ontwikkelingen hebben plaatsgevonden, dus het is raadzaam om de meest recente informatie te raadplegen voor de laatste updates over deSingel en zijn leiderschap.



Regenerate

Send a message



RAG application

> Question:
Wie was algemeen directeur van deSingel na Frie Leysen?

> Answer (took 63.2 s.):

Based on Context Provided: The text does not provide information about who the general director of deSingel was after Frie Leysen. However, it mentions that Aerts was the programmer of classical music for the Antwerpse Elisabethzaal and also served as the cultural manager of deSingel during this time period. Therefore, we can infer that most likely Aerts took over as general director of deSingel after Frie Leysen's tenure ended in 1991.

> source_documents/FrieLeysen.txt:

Levensloop

Frie Leysen was een dochter van Bert Leysen (1920-1959) en de tweelingzus van acteur Johan Leysen. Ze studeerde middeleeuwse kunstgeschiedenis aan de Katholieke Universiteit Leuven. Tussen 1980 en 1991 was ze de eerste directeur van het cultureel centrum deSingel in Antwerpen dat onder haar

> source_documents/KoninklijkVlaamsConservatorium_1898-school-conservatorium-hogeschool-1998.pdf:

Leon Stynen heeft gewijd.

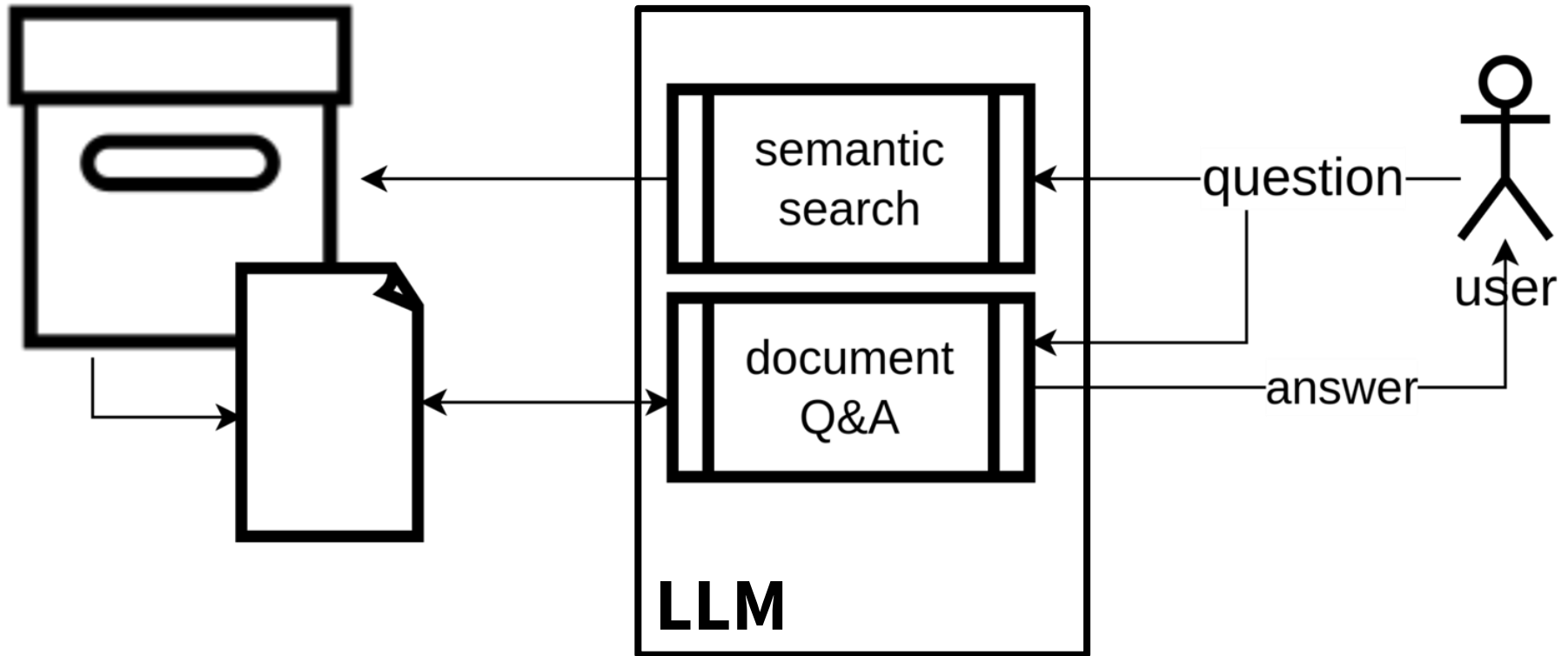
Korting Boudewijn in gesprek met Frie Leysen. algemeen directeur en artistiek beleid van deSingel.

> source_documents/Jerry_Aerts_(cultuurmanager).pdf:

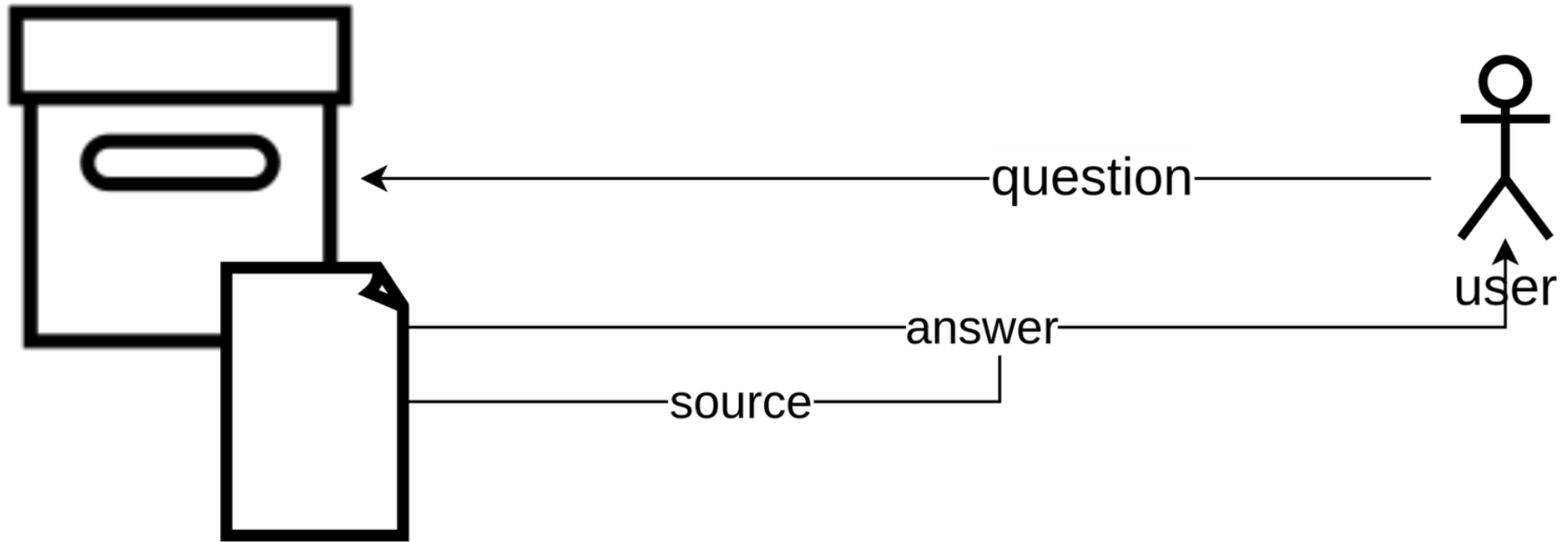
algemeen directeur over van Frie Leysen.[2] Tegelijkertijd

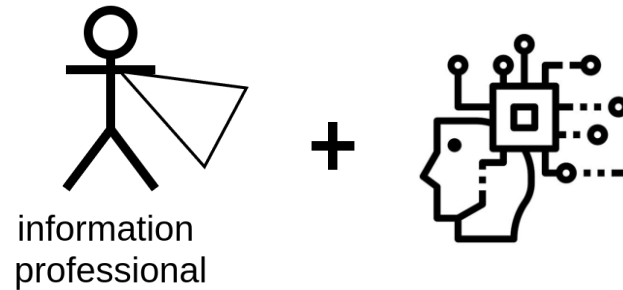
bleef hij programmator van klassieke muziek, wat hij ook deed voor de Antwerpse Elisabethzaal. Hij was dan, leraar, een sterke voeling met zijn publiek, waardoor hij

semantic search + document Q&A



semantic search + document Q&A





- try new approaches
- rethink processes
- it's not rocket science

MODAL: Metadata creation and Disclosure of Digital Archives using LLMs (2024-2026)

User stories and user needs

Technologies: Named entity recognition • Summarization • Speech-to-Text • Text classification • Text Q&A

Implementation scenarios

Knowledge network





AI Tools

Easily reusable and deployable

- Multilingual text recognition (OCR)
- Speech recognition / subtitles generation
- Machine translation



Open Datasets

Open labelled datasets for training and testing AI models

- Object detection / Image classification
- Color detection



Upskilling Material

Technical material (e.g. API documentation) and resources for professionals with lower IT skills (e.g. success stories)

- Semantic text annotation (NER)
- Platform for crowdsourcing for enrichment and validation

Q?

henk@datable.be