

**Μάθημα : Ανάκτηση Πληροφορίας**

**Εργασία : R.A.G**

**Γιώργος Χατζηλίγος AM 4835 , Στάθης Ανδρεοπουλος AM 4630**

[HTTPS://GITHUB.COM/STATHISAN4/ANAKTHSH](https://github.com/StathisAn4/ANAKTHSH)

## ΠΕΡΙΓΡΑΦΗ ΣΥΣΤΗΜΑΤΟΣ : ΒΑΣΙΚΟ PIPELINE RAG

Το pipeline χωριστικό σε τμήματα για την κατανόηση της αρχιτεκτονικής του RAG τα οποία θα αναλυθούν μετέπειτα λεπτομερώς τα κυρία κομμάτια σε βήματα είναι:

**Βήμα 1ο : download library & import dataset/corpus**

**Βήμα 2ο : chunking dataset/corpus**

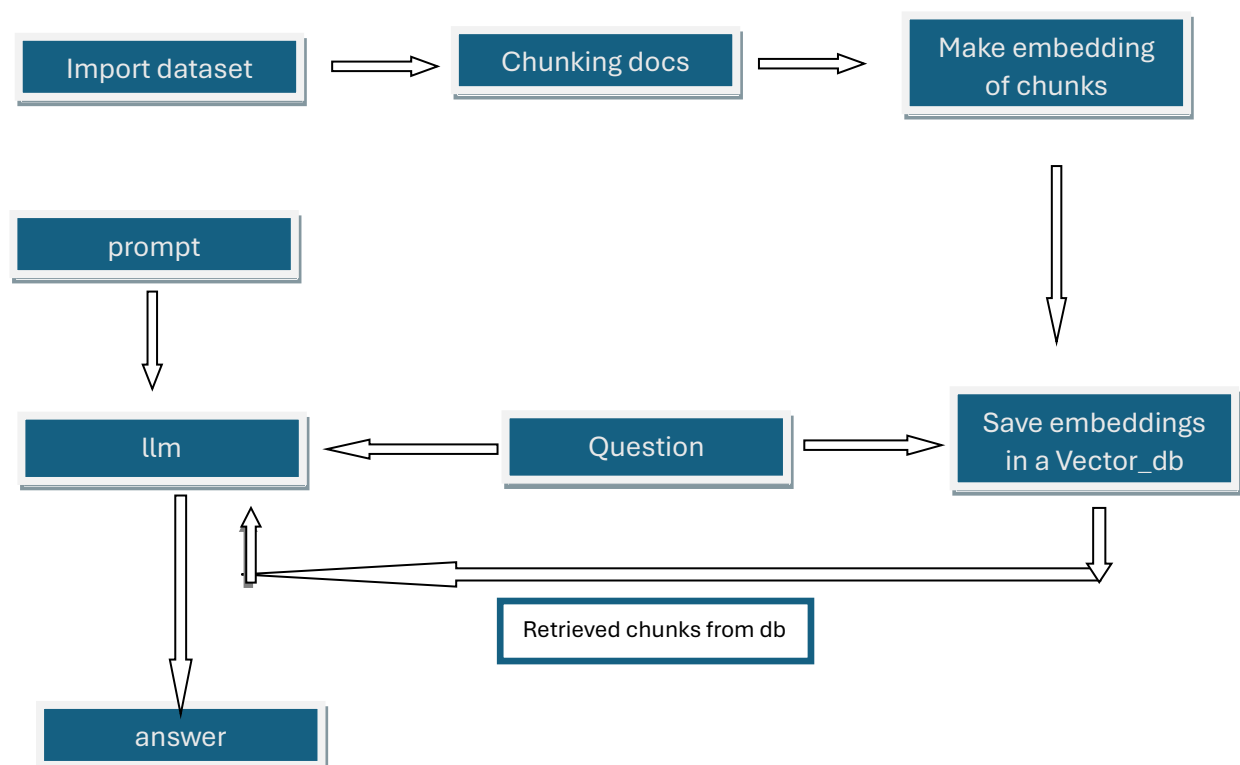
**Βήμα 3ο: embedding model και vector database**

**Βήμα 4ο: set up llm model**

**Βήμα 5ο: set up prompt for llm and question-answer chain**

**Βήμα 6ο: make a set of question and llm answer it**

**Βήμα 7ο: evaluate retrieved chunks with the question.**



## Μάθημα : Ανάκτηση Πληροφορίας

### Εργασία : R.A.G

#### Γιώργος Χατζηλίγος AM 4835 , Στάθης Ανδρεοπουλος AM 4630

Το σύστημά μας αποτελεί μια εξελιγμένη Retrieval-Augmented Generation (RAG) αρχιτεκτονική που έχει σχεδιαστεί για να επιτρέπει την υποβολή σύνθετων ερωτήσεων σε φυσική γλώσσα σε μια εκτεταμένη συλλογή ειδησεογραφικών άρθρων από το CNN. Η αρχιτεκτονική μας συνδυάζει τεχνολογίες ανάκτησης πληροφορίας με μεγάλα γλωσσικά μοντέλα για να παράγει ακριβείς, συνεκτικές και τεκμηριωμένες απαντήσεις βασισμένες σε πραγματικά δεδομένα.

Το βασικό πλεονέκτημα της RAG προσέγγισής μας είναι ότι αποφεύγει τα προβλήματα hallucination που συχνά παρουσιάζουν τα LLMs όταν λειτουργούν αυτόνομα, εξασφαλίζοντας ότι όλες οι απαντήσεις βασίζονται σε αληθινό περιεχόμενο από αξιόπιστες πηγές. Επιπλέον, το σύστημα παρέχει πλήρη διαφάνεια σχετικά με τις πηγές που χρησιμοποιεί, επιτρέποντας στους χρήστες να επαληθεύσουν την ορθότητα των πληροφοριών.

### Τεχνικές και παράμετροι αιτιολόγησής

| Παράμετρος | Τιμή | Αιτιολόγηση |

| Chunk Size | 1,000 chars | Optimal balance μεταξύ context completeness και processing efficiency |

| Chunk Overlap | 200 chars | Εξασφαλίζει continuity και αποφυγή information loss στα boundaries |

| Top-K Retrieval | 20 | Επαρκές context για comprehensive answers χωρίς noise introduction |

| Temperature | 0.0 ή 0.1 | Εξασφαλίζει consistent, reproducible responses για evaluation |

| Embedding Model | all-mpnet-base-v2 ή sentence-transformers/all-MiniLM-L6-v2 | Superior semantic understanding για news domain |

| Vector DB | FAISS Flat | Maximum accuracy για research purposes |

| Παράμετρος | Τιμή | Αιτιολόγηση |

| Chunk Size | 500 chars | Optimal balance μεταξύ context completeness και processing efficiency |

| Chunk Overlap | 800 chars | Εξασφαλίζει continuity και αποφυγή information loss στα boundaries |

| Top-K Retrieval | 10 | Επαρκές context για comprehensive answers χωρίς noise introduction |

| Temperature | 0.0 ή 0.1 | Εξασφαλίζει consistent, reproducible responses για evaluation |

| Embedding Model | all-mpnet-base-v2 ή sentence-transformers/all-MiniLM-L6-v2 | Superior semantic understanding για news domain |

## Μάθημα : Ανάκτηση Πληροφορίας

### Εργασία : R.A.G

### Γιώργος Χατζηλίγος AM 4835 , Στάθης Ανδρεοπουλος AM 4630

| Vector DB | FAISS Flat | Maximum accuracy για research purposes |

| Παράμετρος | Τιμή | Αιτιολόγηση |

| Chunk Size | 600 chars | Optimal balance μεταξύ context completeness και processing efficiency |

| Chunk Overlap | 120 chars | Εξασφαλίζει continuity και αποφυγή information loss στα boundaries |

| Top-K Retrieval | 10 | Επαρκές context για comprehensive answers χωρίς noise introduction |

| Temperature | 0.0 ή 0.1 | Εξασφαλίζει consistent, reproducible responses για evaluation |

| Embedding Model | all-mpnet-base-v2 ή sentence-transformers/all-MiniLM-L6-v2 | Superior semantic understanding για news domain |

| Vector DB | FAISS Flat | Maximum accuracy για research purposes |

## ΑΝΑΛΥΣΗ ΒΗΜΑΤΩΝ

### Βήμα 1ο : download library & import dataset/corpus

Χρησιμοποιήθηκαν τα συγκεκριμένα libraries τα οποία πρέπει να γίνουν install τοπικά σε κάθε υπολογιστή τα περισσότερα με την εντολή `pip install <lib_name>`. Για την εγκατάσταση του ollama ακολουθήσαμε τις οδηγίες του εργαστήριου(εγκαταστήσαμε και τεστάρουμε και τα 2) .

1. Για την εγκατάσταση ενός LLM θα χρειαστεί να τρέξετε την παρακάτω εντολή στο terminal: `ollama pull {model_name}`
2. Για να χρησιμοποιήσετε το LLM που εγκαταστήσατε θα πρέπει το OLLAMA να βρίσκεται σε λειτουργία. `ollama run {model_name}`
3. Τα προτεινόμενα μοντέλα είναι: llama3.2:1b (1.3G) ή llama3.2:3b (2G)

Απαραίτητες βιβλιοθήκες :

```
import os
```

```
import pandas as pd
```

```
import numpy as np
```

```
from tqdm import tqdm
```

## Μάθημα : Ανάκτηση Πληροφορίας

Εργασία : R.A.G

Γιώργος Χατζηλίγος AM 4835 , Στάθης Ανδρεοπουλος AM 4630

```
from langchain_text_splitters import RecursiveCharacterTextSplitter

from langchain_huggingface import HuggingFaceEmbeddings

from langchain_community.vectorstores import FAISS

from langchain_ollama import OllamaLLM

from langchain.chains import RetrievalQA

from langchain.prompts import PromptTemplate

from sklearn.metrics.pairwise import cosine_similarity
```

Εγκαταστήσουμε το dataset από το Kaggle διαβάζουμε το dataset σαν ένα csv το CNN\_Articles\_clean.csv

### Βήμα 2ο : chunking dataset/corpus

Το chunking είναι ο διαχωρισμός του κειμένου σε μικρότερα τμήματα . Πιο συγκεκριμένα εμείς χρησιμοποιήσαμε RecursiveCharacterTextSplitter . Αυτό μας προσφέρει την ευκολία κατά το pre-processing και την προετοιμασία του κάθε chunk διότι διαχωρίζει με βάση το μέγεθος και την επικάλυψη που θα έχει το προηγούμενο με το επόμενο και σημεία που μπορούν να σπάσουν δηλαδή να γίνουν split.

Τα features που χρησιμοποιήσαμε από το corpus/dataset ήταν Headline + Article text για το κάθε chunk .

Μετάπειτα για καθένα από αυτά χρησιμοποιήσαμε ένα metadata\_chunk ώστε να αποθηκεύουμε τα εξής features Headline , article\_index , chunk\_index .

Αρχικά, χρησιμοποιήσαμε `chunk_size=1000`, `chunk_overlap=200`,



```
Splitting articles into chunks: 100%|██████████| 4076/4076 [00:00<00:00, 5636.56it/s]
Total number of chunks: 32559
```

Figure 1: num\_of\_articles and num\_of\_chunks

Έπειτα, χρησιμοποιήσαμε `chunk_size=600`, `chunk_overlap=120`,



```
Splitting articles into chunks: 100%|██████████| 4076/4076 [00:00<00:00, 5775.61it/s]
Total number of chunks: 50993
```

### Βήμα 3ο: embedding model και vector database

#### 3.1 embedding model

## Μάθημα : Ανάκτηση Πληροφορίας

Εργασία : R.A.G

Γιώργος Χατζηλίγος AM 4835 , Στάθης Ανδρεοπουλος AM 4630

Σε αυτό το βήμα προσπαθούμε για την μετατροπή chunks σε embeddings .Στην τελευταία έκδοση του RAG δουλέψαμε με το εξής sentence-transformers/all-mpnet-base-v2 για καλύτερα αποτελέσματα με κόστος το χρόνο έναντι του sentence-transformers/all-MiniLM-L6-v2 που προτάθηκε και έβγαζε και αυτό ικανοποιητικά αποτελέσματα .Σημαντικό είναι να αναφέρουμε ότι τα διανύσματα έχουν κανονικοποιηθεί .

### 3.2 vector database

Χρησιμοποιήσαμε την διανυσματική βάση δεδομένων FAISS.

Χρησιμοποιεί το προηγούμενο embedding model = sentence-transformers/all-mpnet-base-v2 και δημιουργεί τα embeddings από τα chunked documents και αποθηκεύει τα embedding . Με παρόμοιο τρόπο δουλεύει και για τα metada που αναφέραμε .

### **Βήμα 4ο: set up llm model**

Σε αυτό το βήμα μπορούμε να χρησιμοποιήσουμε ελευθέρως κάποιο free llm model . Εμείς χρησιμοποιήσαμε και τα δυο(llama3.2:1b , llama3.2:3b) που προτάθηκαν και μείναμε στο llama3.2:3b με tune στο temperature είτε στο 0 είτε στο 0.1 .

Βλέποντας σε πειράματα που κάναμε ότι γενικότερα το temperature = 0 έβγαζε ελλιπής απαντήσεις στα ερωτήματα μας βάση του prompt που του δίναμε .

Βλέποντας σε πειράματα που κάναμε ότι γενικότερα το temperature = 0.1 έβγαζε μια απάντηση βασισμένη στο prompt .

Διαπιστώσαμε ότι το llama3.2:3b έχει καλύτερη κατανόηση context σε σχέση με το llama3.2:1b οπότε μπορεί να χρησιμοποιεί τα retrieved\_chunks χωρίς να πέφτουμε σε κάποιο λάθος .

### **Βήμα 5ο: set up prompt for llm and question-answer chain**

Κατασκευάσαμε ένα prompt το οποίο αρχικά δίνει μια ιδιότητα στο llm όπως αναφέρεται στις οδηγίες . Έπειτα ακολουθεί η ερώτηση που θέλουμε να απαντηθεί . Τέλος , το πως θέλουμε να απαντηθεί δηλαδή με βάση των retrieved\_chunks . Επίσης διαπιστώσαμε κατά την πορεία της άσκησης ότι ήταν αρκετά χρήσιμο να χρησιμοποιήσουμε επιπρόσθετα instructions .

Στη συνέχεια κατασκευάζουμε question-answer chain το οποίο λειτουργεί

1. βάζοντας το llm της επιλογής μας .

## Μάθημα : Ανάκτηση Πληροφορίας

### Εργασία : R.A.G

#### Γιώργος Χατζηλίγος AM 4835 , Στάθης Ανδρεοπουλος AM 4630

2. επιστρέφει όλα τα retrieve documents με την εντολή `chain_type="stuff"` (μερικές ενδεικτικές είναι "stuff": Όλα τα documents σε ένα prompt γρήγορο, απλό "map\_reduce": Επεξεργασία κάθε document ξεχωριστά, μετά σύνθεση, "refine": Βελτίωση της απάντησης σταδιακά με κάθε document "map\_rerank": Βαθμολόγηση απαντήσεων και επιλογή της καλύτερης) .

3. Επιστρέφει τα 10 πιο σχετικά documents από τη faiss με την εντολή `retriever=vectorstore.as_retriever(search_kwargs={"k": 10})` . Η τιμή 10 εναντι της τιμής 20 διαπιστώθηκε μετά από πειράματα ότι το llm μπορεί να τα διαχειριστεί καλύτερα των αριθμό των κείμενων

4. Τέλος χρησιμοποιεί το prompt που έχουμε δημιουργήσει και πάνω σε αυτό θα γίνει το retrieve και θα παραχθεί και η απάντηση του llm .

Η χρήση αυτής της αλυσίδας γίνεται με τη χρήση της συνάρτησης `invoke()`

```
def query_with_rag(question):  
    result = qa_chain.invoke({"query": question})
```

Figure 2 Χρήση της qa\_chain

```
def query_without_rag(question):  
    prompt = f"You are an artificial intelligence assistant who is an expert in analyzing from CNN article  
    answer = llm.invoke(prompt)
```

Figure 3 Λειτουργία με απλο περασμα prompt

#### Βήμα 6ο: make a set of question and llm answer it

Τα question αλλάζουν in code και έχουν γίνει κατά βάση με γεγονότα τα οποία ήταν επίκαιρα το έτος 2020- 2022 παγκοσμίως .Προσπαθήσαμε να καλύψουμε κάποια βασικά topic των άρθρων όπως football , economy , motosport , health .

Το σύστημα llm+rag χρησιμοποιεί την ερώτηση και τα retrieved chunks ώστε να παράξει μια απάντηση βάση των retrieved chunks

Προσθέσαμε και ένα llm το οποίο είναι σεταρισμένο με παρόμοιο prompt χωρίς το context μόνο με μια πανομοιότυπη ιδιότητα You are an artificial intelligence assistant who is an expert in analyzing from CNN articles 2020 – 2022 και να απαντάει στην ίδια ερώτηση . Χρησιμοποιήθηκε απλώς για μια γρήγορη σύγκριση .

#### Βήμα 7ο: evaluate retrieved chunks with the question.

Σημαντικό για το evaluation question με chunk είναι ότι το question θα πρέπει να περάσει πρώτα από την διαδικασία να γίνει embedding μέσα από το μοντέλο το οποίο έχουμε χρησιμοποιήσει και εμείς για να κάνουμε τα chunks embeddings .

Τα πειράματα μας είχαν την εξής δομή :

# Μάθημα : Ανάκτηση Πληροφορίας

## Εργασία : R.A.G

### Γιώργος Χατζηλίγος AM 4835 , Στάθης Ανδρεοπουλος AM 4630

1. Με βάση την ερώτηση πρώτα απαντάει llm+ RAG επιστρέφει τα πιο relevant chunk με πληροφορίες για το headline , chunk\_id , doc\_id από το dataset, Article text .Επίσης σημαντικό είναι το evaluation μέσω κάποιου similarity για αυτό το λόγο χρησιμοποιήσαμε cosine-similarity score μεταξύ ερώτησης(embedded) και του κάθε επιστρεφόμενου chunk .
2. μετέπειτα απαντάει το llm στο ερώτημα .
3. Κάποια χρήσιμα αποτελέσματα όπως ερώτηση , απάντηση llm+rag , απάντηση llm , min\_similarity , max\_similarity , mean\_similarity αποθηκεύονται σε ένα csv.

```
[WITH RAG]
Question: What are the main challenges in US policy?
Answer (with RAG): Based on the provided context, some of the main challenges in US policy include:

1. Diminished American influence in European defense and security: The unpredictability from the White House and the dismantling of US bureaucracy under Trump have led to a decrease in US influence in Europe.
2. Laborious and opaque hiring processes in the federal government: This has made it difficult for early-career individuals to enter the public sector, as they often prioritize factors like salary and benefits over career development.
3. Competing with private sector competitors for talent: The federal government's inability to compete with private sector employers in terms of hiring early-career individuals has hindered its ability to attract top talent.
4. Addressing the Great Resignation: The pandemic has caused workers across various sectors to re-evaluate their employment packages, leading to a surge in resignations and making it difficult for employers to fill positions.
5. Climate change negotiations: The US is facing significant pressure to address climate change, with world leaders gathering in Paris on November 30 to discuss the "carbon budget" and the need for a global agreement.

These challenges highlight the need for proactive policy-making and strategic planning to address these issues and ensure the long-term success of US policies.

Similarity scores between question and retrieved chunks:
-----
Chunk 1 (Article 876, Chunk 7)
Title: Trump has trashed America's most important alliance. The rift with Europe could take decades to repair - CNN
Similarity Score: 0.4450
-----
Chunk 2 (Article 592, Chunk 0)
Title: Analysis: The question now facing Democrats - CNN
Similarity Score: 0.4353
-----
Chunk 3 (Article 238, Chunk 5)
Title: Opinion: The government has a talent problem. This bill could help change that - CNN
Similarity Score: 0.4289
-----
Chunk 4 (Article 2976, Chunk 9)
Title: Ursula von der Leyen challenges US on climate and asserts Brussels' role in 'new international order' - CNN
Similarity Score: 0.4150
-----
```

Figure 4 ενδεικτικό αποτέλεσμα answer-rag-llm

```
[WITHOUT RAG]
Question: What are the main challenges in US policy?
Answer (without RAG): Based on my analysis of CNN articles from 2020 to 2022, I have identified some of the main challenges in US policy during this period:

1. **Partisan Gridlock**: The COVID-19 pandemic and subsequent economic crisis highlighted the deepening partisan divide in the US, making it increasingly difficult for lawmakers to pass legislation.
2. **Racial Tensions and Police Reform**: The Black Lives Matter movement and high-profile police brutality cases led to renewed calls for police reform and racial justice. However, the pandemic also exacerbated racial tensions and led to a surge in hate crimes.
3. **Economic Inequality and Job Creation**: The pandemic exacerbated existing economic inequalities, leading to concerns about job creation, income inequality, and access to healthcare.
4. **Climate Change and Energy Policy**: As climate change became an increasingly pressing issue, policymakers faced challenges in developing effective energy policies that balanced environmental concerns with economic growth.
5. **Immigration Reform**: The COVID-19 pandemic led to increased scrutiny of immigration policies, particularly those related to border security and asylum seekers. However, lawmakers also faced challenges in addressing the needs of immigrants and refugees.
6. **Gun Control and Public Safety**: Mass shootings and gun violence continued to be a major concern in the US, leading to renewed calls for stricter gun control laws. However, these efforts were often blocked by partisan gridlock.
7. **Voting Rights and Election Security**: The 2020 presidential election highlighted concerns about voting rights, election security, and the integrity of the electoral process. Policymakers faced challenges in addressing these concerns and ensuring the integrity of future elections.
8. **National Security and Foreign Policy**: The US continued to face complex national security challenges, including the COVID-19 pandemic's impact on global health, the rise of China, and the ongoing conflict in Ukraine.

These are some of the main challenges in US policy that I identified through my analysis of CNN articles from 2020 to 2022.
```

Figure 5 ενδεικτικό αποτέλεσμα answer-llm

**Μάθημα : Ανάκτηση Πληροφορίας**

**Εργασία : R.A.G**

**Γιώργος Χατζηλίγος AM 4835 , Στάθης Ανδρεοπουλος AM 4630**