```
!pip install sentence-transformers faiss-cpu gradio
```

```
Requirement already satisfied: tomlkit<0.14.0,>=0.12.0 in /usr/local/lib/python3.12/dist-packages (from gradio) (0.13.3)
Requirement already satisfied: typer<1.0,>=0.12 in /usr/local/lib/python3.12/dist-packages (from gradio) (0.20.0)
Requirement already satisfied: uvicorn>=0.14.0 in /usr/local/lib/python3.12/dist-packages (from gradio) (0.38.0)
Requirement already satisfied: fsspec in /usr/local/lib/python3.12/dist-packages (from gradio-client==1.13.3->gradio) (2025.3.
Requirement already satisfied: websockets<16.0,>=13.0 in /usr/local/lib/python3.12/dist-packages (from gradio-client==1.13.3->
Requirement already satisfied: idna>=2.8 in /usr/local/lib/python3.12/dist-packages (from anyio<5.0,>=3.0->gradio) (3.11)
Requirement already satisfied: sniffio>=1.1 in /usr/local/lib/python3.12/dist-packages (from anyio<5.0,>=3.0->gradio) (1.3.1)
Requirement already satisfied: annotated-doc>=0.0.2 in /usr/local/lib/python3.12/dist-packages (from fastapi<1.0,>=0.115.2->gra
Requirement already satisfied: certifi in /usr/local/lib/python3.12/dist-packages (from httpx<1.0,>=0.24.1->gradio) (2025.10.5
Requirement already satisfied: httpcore==1.* in /usr/local/lib/python3.12/dist-packages (from httpx<1.0,>=0.24.1->gradio) (1.0
Requirement already satisfied: h11>=0.16 in /usr/local/lib/python3.12/dist-packages (from httpcore==1.*->httpx<1.0,>=0.24.1->g
Requirement already satisfied: filelock in /usr/local/lib/python3.12/dist-packages (from huggingface-hub>=0.20.0->sentence-tra
Requirement already satisfied: requests in /usr/local/lib/python3.12/dist-packages (from huggingface-hub>=0.20.0->sentence-tra
Requirement already satisfied: hf-xet<2.0.0,>=1.1.3 in /usr/local/lib/python3.12/dist-packages (from huggingface-hub>=0.20.0->
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.12/dist-packages (from pandas<3.0,>=1.0->gradio
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.12/dist-packages (from pandas<3.0,>=1.0->gradio) (2025.2
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.12/dist-packages (from pandas<3.0,>=1.0->gradio) (2025
Requirement already satisfied: annotated-types>=0.6.0 in /usr/local/lib/python3.12/dist-packages (from pydantic<2.12,>=2.0->gra
Requirement already satisfied: pydantic-core==2.33.2 in /usr/local/lib/python3.12/dist-packages (from pydantic<2.12,>=2.0->grad
Requirement already satisfied: typing-inspection>=0.4.0 in /usr/local/lib/python3.12/dist-packages (from pydantic<2.12,>=2.0->
Requirement already satisfied: setuptools in /usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-transformer
Requirement already satisfied: sympy>=1.13.3 in /usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-transfor
Requirement already satisfied: networkx in /usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-transformers)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.6.77 in /usr/local/lib/python3.12/dist-packages (from torch>=1.11.0-
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.6.77 in /usr/local/lib/python3.12/dist-packages (from torch>=1.11.
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.6.80 in /usr/local/lib/python3.12/dist-packages (from torch>=1.11.0-
Requirement already satisfied: nvidia-cudnn-cu12==9.10.2.21 in /usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->se
Requirement already satisfied: nvidia-cublas-cu12==12.6.4.1 in /usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->se
Requirement already satisfied: nvidia-cufft-cu12==11.3.0.4 in /usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sen
Requirement already satisfied: nvidia-curand-cu12==10.3.7.77 in /usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->s
Requirement already satisfied: nvidia-cusolver-cu12==11.7.1.2 in /usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->
Requirement already satisfied: nvidia-cusparse-cu12==12.5.4.2 in /usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->
Requirement already satisfied: nvidia-cusparselt-cu12==0.7.1 in /usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->s
Requirement already satisfied: nvidia-nccl-cu12==2.27.3 in /usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->senten
Requirement already satisfied: nvidia-nvtx-cu12==12.6.77 in /usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sente
Requirement already satisfied: nvidia-nvjitlink-cu12==12.6.85 in /usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->
Requirement already satisfied: nvidia-cufile-cu12==1.11.1.6 in /usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->se
Requirement already satisfied: triton==3.4.0 in /usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-transfor
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.12/dist-packages (from transformers<5.0.0,>=4.41.0-
Requirement already satisfied: tokenizers<=0.23.0,>=0.22.0 in /usr/local/lib/python3.12/dist-packages (from transformers<5.0.0
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.12/dist-packages (from transformers<5.0.0,>=4.41.0
Requirement already satisfied: click>=8.0.0 in /usr/local/lib/python3.12/dist-packages (from typer<1.0,>=0.12->gradio) (8.3.0)
Requirement already satisfied: shellingham>=1.3.0 in /usr/local/lib/python3.12/dist-packages (from typer<1.0,>=0.12->gradio) (
Requirement already satisfied: rich>=10.11.0 in /usr/local/lib/python3.12/dist-packages (from typer<1.0,>=0.12->gradio) (13.9.
Requirement already satisfied: joblib>=1.2.0 in /usr/local/lib/python3.12/dist-packages (from scikit-learn->sentence-transform
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.12/dist-packages (from scikit-learn->sentence-tr
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.12/dist-packages (from python-dateutil>=2.8.2->pandas<3.0,>=
Requirement already satisfied: markdown-it-py>=2.2.0 in /usr/local/lib/python3.12/dist-packages (from rich>=10.11.0->typer<1.0
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in /usr/local/lib/python3.12/dist-packages (from rich>=10.11.0->typer<1
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.12/dist-packages (from sympy>=1.13.3->torch>=1.11.
Requirement already satisfied: charset_normalizer<4,>=2 in /usr/local/lib/python3.12/dist-packages (from requests->huggingface
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.12/dist-packages (from requests->huggingface-hub>=
Requirement already satisfied: mdurl~=0.1 in /usr/local/lib/python3.12/dist-packages (from markdown-it-py>=2.2.0->rich>=10.11.
Downloading faiss_cpu-1.13.0-cp39-abi3-manylinux_2_27_x86_64.manylinux_2_28_x86_64.whl (23.6 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 23.6/23.6 MB 38.2 MB/s eta 0:00:00
Installing collected packages: faiss-cpu
Successfully installed faiss-cpu-1.13.0
```

```
from sentence_transformers import SentenceTransformer
import faiss
import numpy as np
import gradio as gr
import textwrap

# === Real mini-corpus for ITAI-2277 (Capstone) ===
documents = [
    {
        "id": "doc1",
        "course": "ITAI 2277",
        "week": "Course Overview",
        "doc_type": "Syllabus",
        "title": "ITAI 2277 – Course Vision and Welcome",
        "text": """
Houston Community College’s vision is to deliver relevant, high-quality education that ensures
success for all students, the community, and the economy. ITAI 2277 – Artificial Intel Resource
(Capstone) is taught fully online (WW – Online Anytime).
```

```
The course is led by Professor Anna Devarakonda (Annapurna Rachapudi). The class focuses on
an AI Applications Capstone Project where students design and deploy real-world AI solutions in
domains such as healthcare, finance, sustainability, and more. Students work with tools like
TensorFlow, PyTorch, and cloud platforms, and practice AI application development, model deployment,
and professional collaboration.

The course does not use a traditional textbook. Instead, all instructional materials are provided
through Canvas, using curated, up-to-date articles, papers, videos, and resources. Students are
encouraged to check the course site and their HCC email at least once per day and to start from
the Modules section in Canvas.
"""
    },
    {
        "id": "doc2",
        "course": "ITAI 2277",
        "week": "Course Requirements",
        "doc_type": "Syllabus",
        "title": "ITAI 2277 – Assignments, Grading and Workload",
        "text": """
ITAI 2277 uses multiple graded components:

- Module Group Assignments (15%): change according to the lecture topic. Formats may include
  written Word/PDF documents, PowerPoint presentations, or multimedia submissions.

- Case Study Analysis (20%): tied to lecture topics and focused on applying concepts.

- Exams/Quizzes (20%): 4 separate assessments covering key concepts, using multiple-choice,
  true/false, and short-answer questions.

- Midterm (20%): group case study analysis that tests analytical skills related to ethical,
  philosophical, and practical applications of AI in a specific industry.

- Final Project (25%): students, working in groups, create a proposal for integrating AI into a
  new or existing process within a chosen sector. This is the main capstone-style deliverable.

There is also an optional Extra Credit Portfolio (5%) for uploading course work to GitHub to
continue building an AI portfolio.

The HCC grading system uses A (90–100), B (80–89), C (70–79), D (60–69), F or FX (failing),
W (withdrawn), and I (incomplete), according to standard HCC policies.
"""
    },
    {
        "id": "doc3",
        "course": "ITAI 2277",
        "week": "Policies",
        "doc_type": "Syllabus",
        "title": "ITAI 2277 – Incompletes, Attendance, Make-Up Work, Academic Integrity",
        "text": """
Incomplete grades (“I”) are only considered if the student has completed at least 85% of the work
in the course, and the instructor still has the discretion to decline the request.

Make-up exams and assignments are allowed only for documented emergencies, such as hospitalization
or auto accidents. They do not apply to reasons like forgetting the due date or being busy with work.
Documentation must be provided as soon as possible. All missed grades are recorded as zeros if
no approved make-up is arranged.

Online students must show satisfactory progress in the course. Students may be withdrawn if they
miss turning in assignments that total more than 12.5% of the course work before the final exam.
Students are responsible for contacting the instructor if they are having a problem.

Academic Integrity: Scholastic dishonesty results in referral to the Dean of Student Services.
Group work is allowed, but groups must not share the same files and then make minor changes to
submit as their own. Using copied work or unauthorized collaboration may result in a 0 on the
assignment and a disciplinary referral. Students must follow HCC academic integrity procedures.
"""
    },
    {
        "id": "doc4",
        "course": "ITAI 2277",
        "week": "Final Project",
        "doc_type": "Assignment",
        "title": "ITAI 2277 – Capstone Final Project and Presentation",
        "text": """
The Capstone Final Assignment for ITAI 2277 is titled “Capstone Project 2025.” Students must design,
develop, and submit a GitHub repository as their final class project, along with a PowerPoint or PDF.
```

```
This capstone course allows students to synthesize knowledge from the entire Associate degree in
Applied Technologies – AI and Robotics. Working in teams, students build a substantial project that
integrates multiple technologies from computer vision, natural language processing, robotics,
machine learning, deep learning, and related areas.

By the end of the course, students should be able to:
1. Plan and execute a comprehensive project that integrates multiple AI/ML technologies.
2. Apply knowledge from core courses to implement solutions to real-world problems.
3. Collaborate using industry-standard tools and professional practices.
4. Document and communicate technical work through professional reports and presentations.
5. Evaluate and refine solutions using technical metrics, stakeholder requirements, and ethical considerations.
6. Deliver a complete, portfolio-ready project that demonstrates readiness for professional AI/ML roles.

Final presentation requirements include:
– A public GitHub repo with a clear README and installation instructions.
– A formal project presentation (around 20 minutes per team).
– A live demonstration and Q&A session showcasing the system and its impact.
"""
    },
]
```

```
# load model
model = SentenceTransformer("sentence-transformers/all-MiniLM-L6-v2")

# Convert text embedding
corpus_texts = [doc["text"] for doc in documents]
embeddings = model.encode(corpus_texts, convert_to_numpy=True, show_progress_bar=True)

# Create faiss
dim = embeddings.shape[1]
index = faiss.IndexFlatL2(dim)
index.add(embeddings)

print("Index size:", index.ntotal)
```

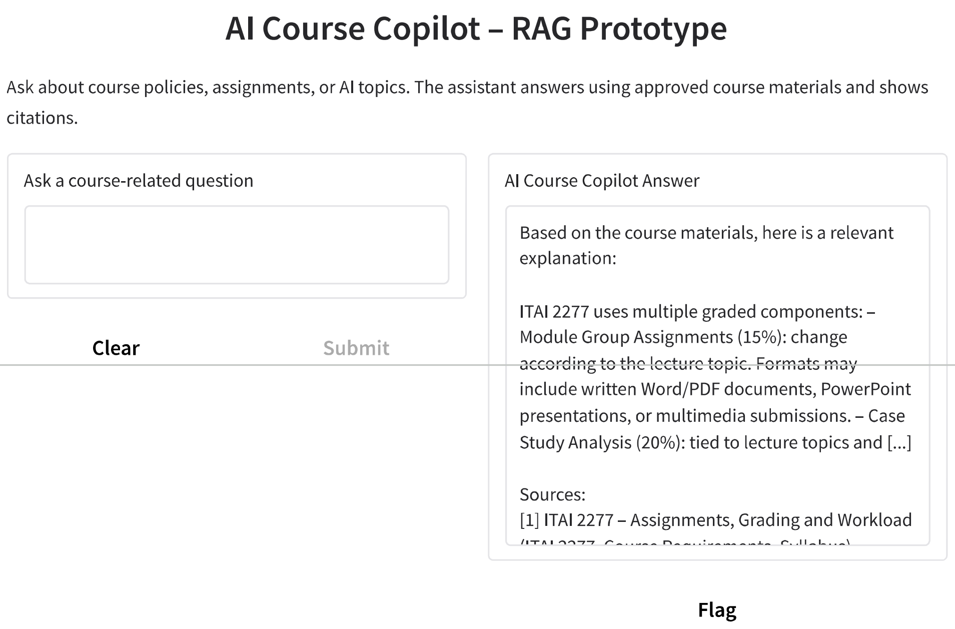
```
/usr/local/lib/python3.12/dist-packages/huggingface_hub/utils/_auth.py:94: UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab (https://huggingface.co/settings/tokens), set it
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public models or datasets.
  warnings.warn(
modules.json: 100%                                   349/349 [00:00<00:00, 27.0kB/s]
config_sentence_transformers.json: 100%              116/116 [00:00<00:00, 12.4kB/s]
README.md:      10.5k/? [00:00<00:00, 932kB/s]
sentence_bert_config.json: 100%                      53.0/53.0 [00:00<00:00, 5.47kB/s]
config.json: 100%                                    612/612 [00:00<00:00, 57.9kB/s]
model.safetensors: 100%                              90.9M/90.9M [00:00<00:00, 153MB/s]
tokenizer_config.json: 100%                          350/350 [00:00<00:00, 36.5kB/s]
vocab.txt:      232k/? [00:00<00:00, 10.2MB/s]
tokenizer.json:      466k/? [00:00<00:00, 25.1MB/s]
special_tokens_map.json: 100%                        112/112 [00:00<00:00, 12.0kB/s]
config.json: 100%                                    190/190 [00:00<00:00, 20.8kB/s]
Batches: 100%                                        1/1 [00:00<00:00,  1.62it/s]
Index size: 4
```

```
def retrieve(query, k=3):
    """Return top-k docs for a query with similarity scores."""
    query_emb = model.encode([query], convert_to_numpy=True)
    distances, indices = index.search(query_emb, k)
    results = []
    for rank, (idx, dist) in enumerate(zip(indices[0], distances[0]), start=1):
        doc = documents[idx]
        results.append({
            "rank": rank,
            "score": float(dist),
            "id": doc["id"],
            "title": doc["title"],
            "course": doc["course"],
            "week": doc["week"],
            "doc_type": doc["doc_type"],
            "snippet": textwrap.shorten(doc["text"].replace("\n", " "), width=280)
        })
    return results


def format_answer(query, k=3):
    """Generate a simple answer using retrieved snippets + citations."""
    results = retrieve(query, k=k)
    if not results:
        return "I couldn't find any relevant course materials for this question."

    # Parte tipo "respuesta" (extractive)
    top = results[0]
    answer_intro = (
        f"Based on the course materials, here is a relevant explanation:\n\n"
        f"{top['snippet']}\n\n"
    )

    # Citations
    citations_lines = []
    for r in results:
        citations_lines.append(
            f"[{r['rank']}] {r['title']} "
            f"({r['course']}, {r['week']}, {r['doc_type']})"
        )
    citations_text = "Sources:\n" + "\n".join(citations_lines)

    return answer_intro + citations_text


# Prueba rápida en la consola
print(format_answer("What is the late work policy in ITAI 1370?"))
```

```
Based on the course materials, here is a relevant explanation:

ITAI 2277 uses multiple graded components: – Module Group Assignments (15%): change according to the lecture topic. Formats may

Sources:
[1] ITAI 2277 – Assignments, Grading and Workload (ITAI 2277, Course Requirements, Syllabus)
[2] ITAI 2277 – Incompletes, Attendance, Make-Up Work, Academic Integrity (ITAI 2277, Policies, Syllabus)
[3] ITAI 2277 – Capstone Final Project and Presentation (ITAI 2277, Final Project, Assignment)
```

```
def rag_chat(query):
    return format_answer(query, k=3)

demo = gr.Interface(
    fn=rag_chat,
    inputs=gr.Textbox(lines=2, label="Ask a course-related question"),
    outputs=gr.Textbox(lines=12, label="AI Course Copilot Answer"),
    title="AI Course Copilot – RAG Prototype",
    description="Ask about course policies, assignments, or AI topics. The assistant answers using approved course materials and
)

demo.launch()
```

```
It looks like you are running Gradio on a hosted Jupyter notebook, which requires `share=True`. Automatically setting `share=Tru

Colab notebook detected. To show errors in colab notebook, set debug=True in launch()
* Running on public URL: https://3317d6bcebc2e48beb.gradio.live

This share link expires in 1 week. For free permanent hosting and GPU upgrades, run `gradio deploy` from the terminal in the wor
```



AI Course Copilot – RAG Prototype

Ask about course policies, assignments, or AI topics. The assistant answers using approved course materials and shows citations.

Ask a course-related question

AI Course Copilot Answer

Based on the course materials, here is a relevant explanation:

ITAI 2277 uses multiple graded components: – Module Group Assignments (15%): change according to the lecture topic. Formats may include written Word/PDF documents, PowerPoint presentations, or multimedia submissions. – Case Study Analysis (20%): tied to lecture topics and [...]

Sources:
[1] ITAI 2277 – Assignments, Grading and Workload
(ITAI 2277_Course Requirements_Syllabus)

Next steps: Deploy to Cloud Run