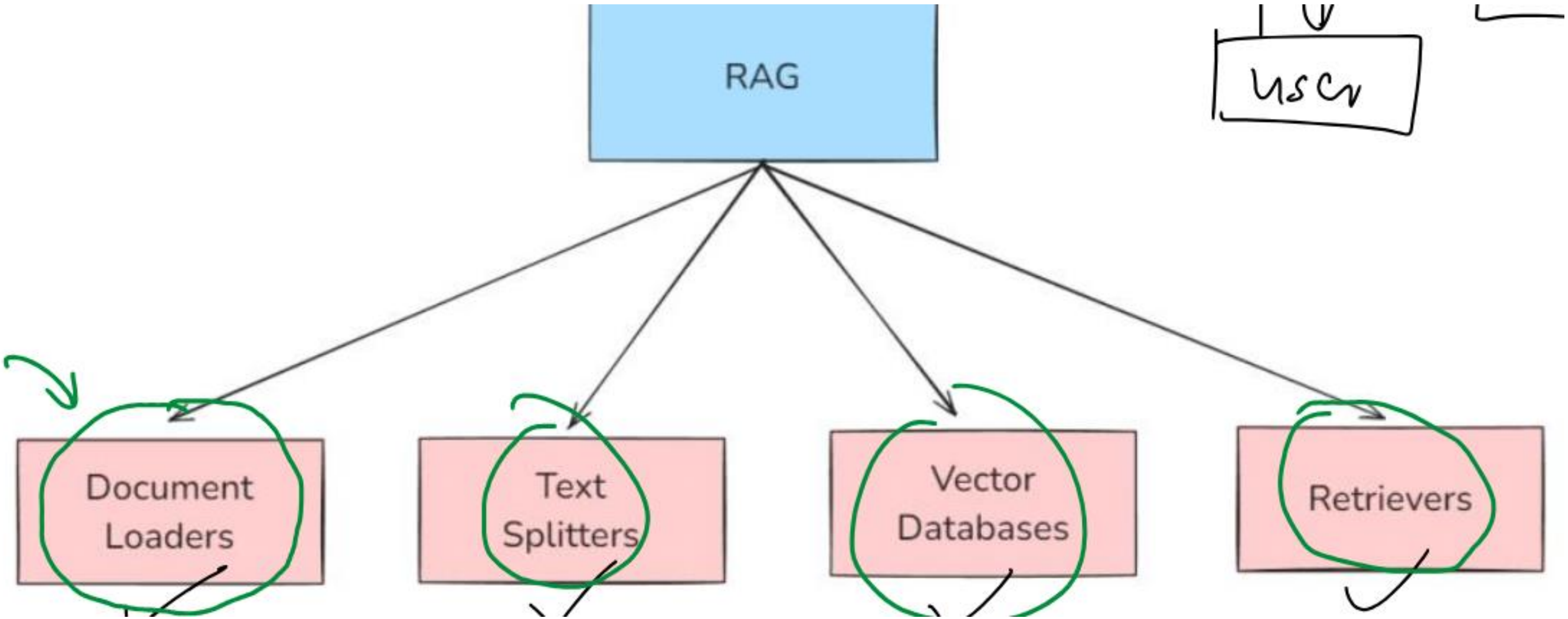
RAG

- RAG is a technique that combines information retrieval with language generation, where a model retrieves relevant documents from a knowledge base and then uses them as context to generate accurate and grounded responses.
- **Benefits of using RAG**
  1. Use of up-to-date information.
  2. Better privacy.
  3. No limit of document size



RAG

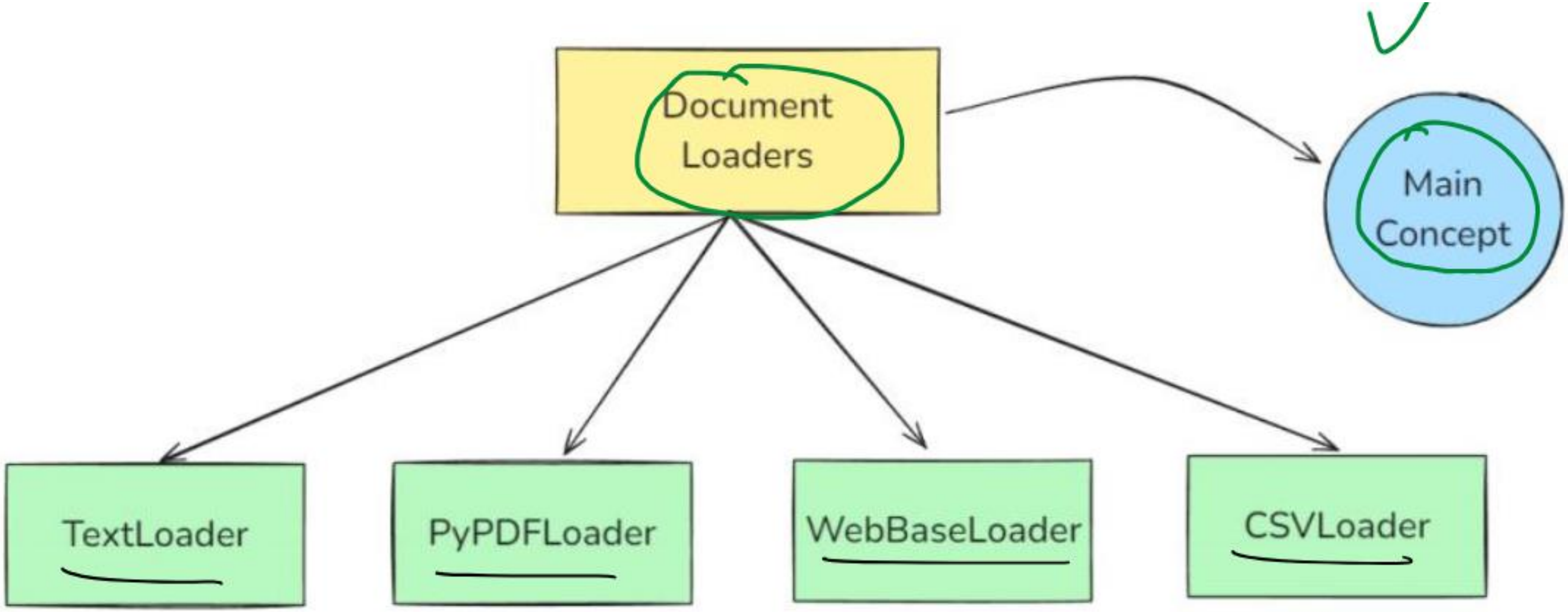
Document Loaders

Text Splitters

Vector Databases

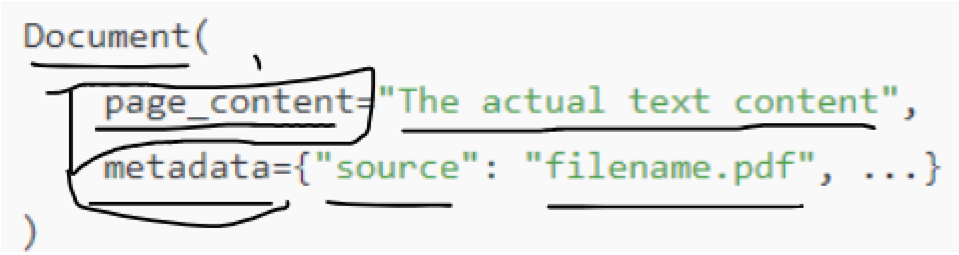
Retrievers

Document Loaders



Document Loaders in LangChain

- Document loaders are components in LangChain used to load data from various sources into a standardized format (usually as Document objects), which can then be used for chunking,embedding, retrieval, and generation.
  - Pdf
  - Txt
  - DB
  - S3



```
Document(
    page_content="The actual text content",
    metadata={"source": "filename.pdf", ...}
)
```

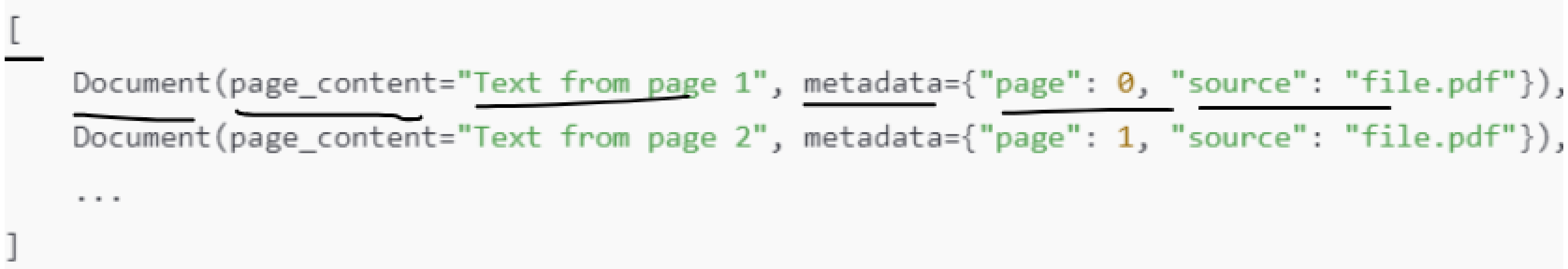
Text Loader

- TextLoader is a simple and commonly used document loader in LangChain that reads plain text(.txt) files and converts them into LangChain Document objects.
- Use Case
  - Ideal for loading chat logs, scraped text, transcripts, code snippets, or any plain text data into a LangChain pipeline.
- Limitation
  - Works only with .txt files

PyPDFLoader

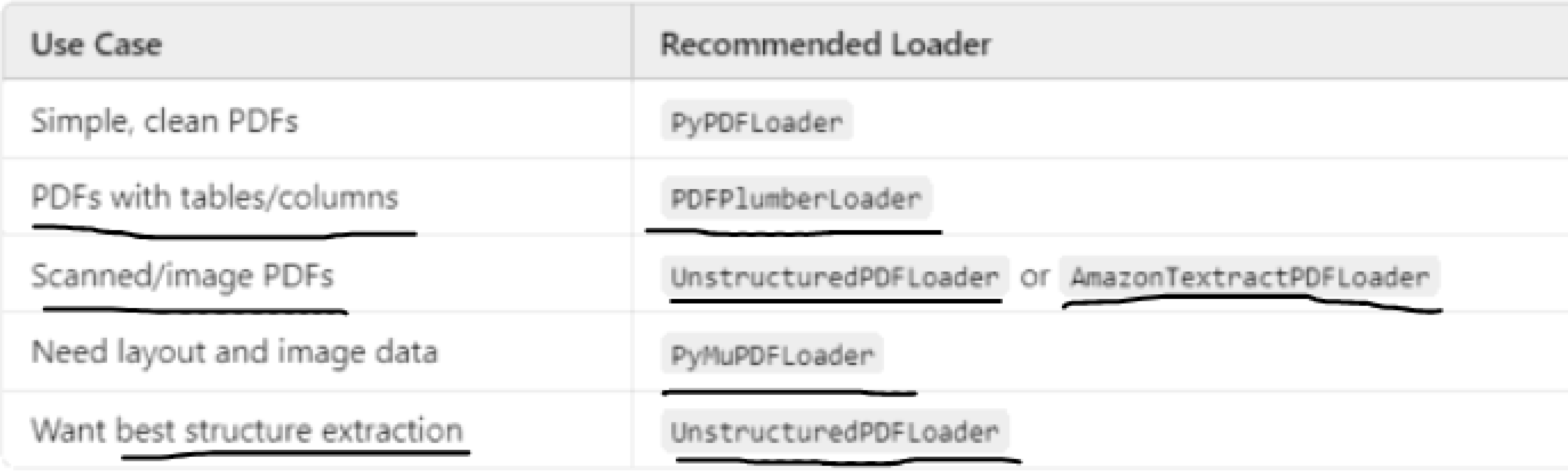
- PyPDFLoader is a document loader in LangChain used to load content from PDF files and convert each page into a Document object.

- Limitations:
- It uses the PyPDF library under the hood — not great with scanned PDFs or complex layouts.



```
[
    Document(page_content="Text from page 1", metadata={"page": 0, "source": "file.pdf"}),
    Document(page_content="Text from page 2", metadata={"page": 1, "source": "file.pdf"}),
    ...
]
```

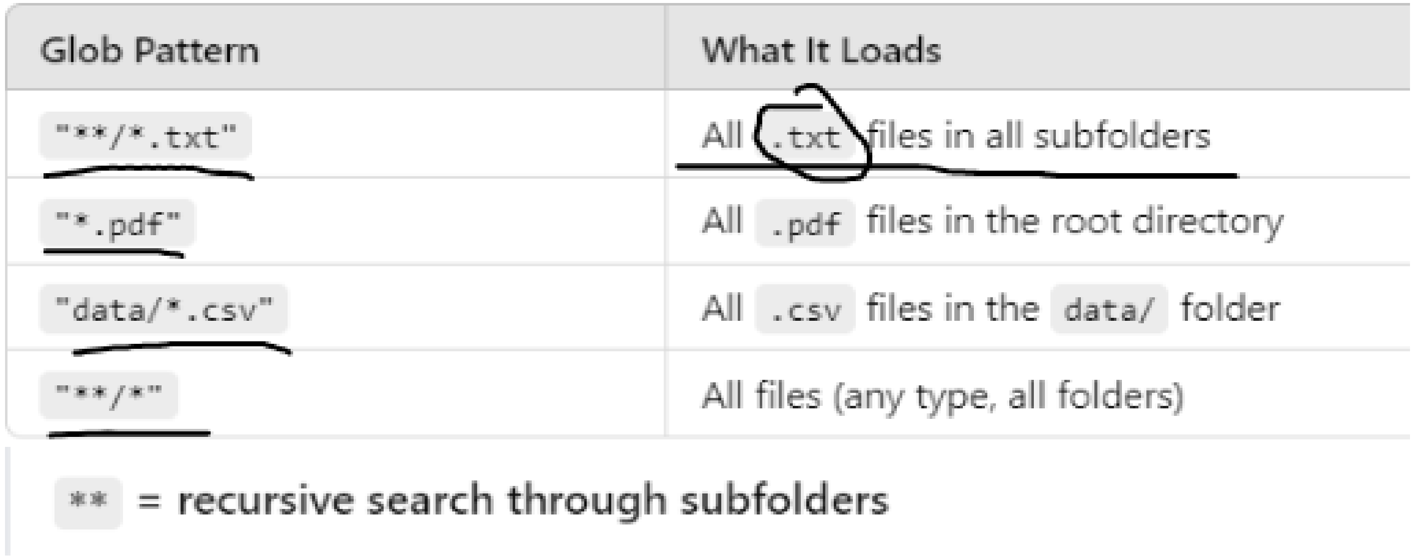
usecase



Use Case	Recommended Loader
Simple, clean PDFs	`PyPDFLoader`
PDFs with tables/columns	`PDFPlumberLoader`
Scanned/image PDFs	`UnstructuredPDFLoader` or `AmazonTextractPDFLoader`
Need layout and image data	`PyMuPDFLoader`
Want best structure extraction	`UnstructuredPDFLoader`

DirectoryLoader

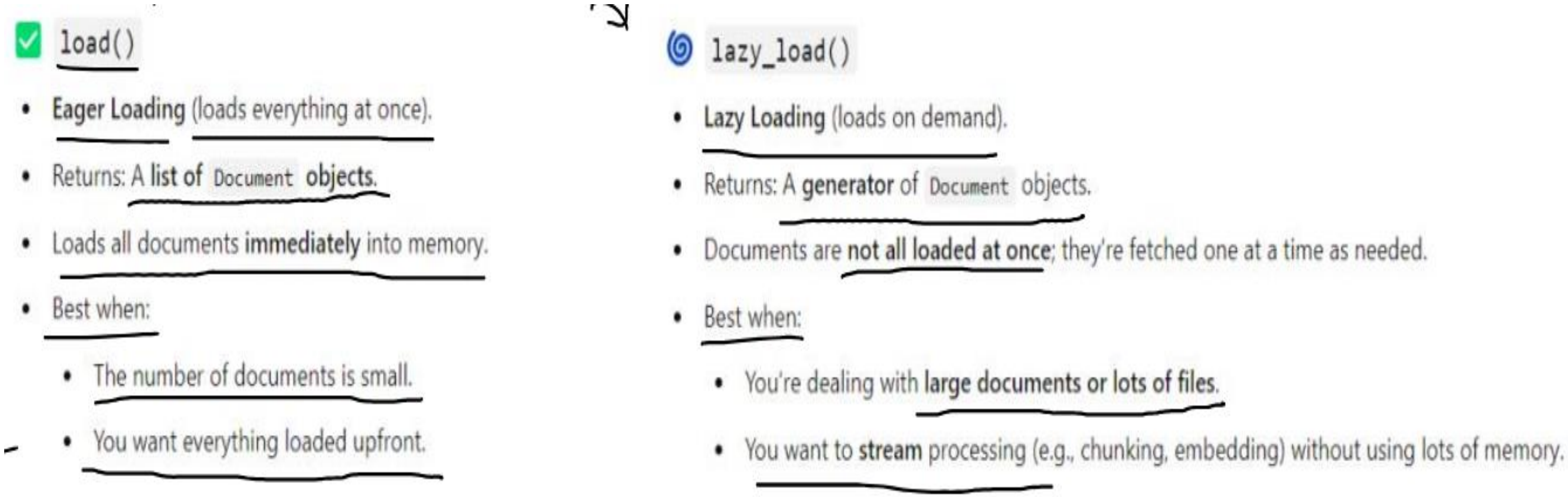
- DirectoryLoader is a document loader that lets you load multiple documents from a directory (folder) of files.



Glob Pattern	What It Loads
`"**/*.txt"`	All `.txt` files in all subfolders
`"*.pdf"`	All `.pdf` files in the root directory
`"data/*.csv"`	All `.csv` files in the `data/` folder
`"**/*"`	All files (any type, all folders)

`**` = recursive search through subfolders

Load vs Lazy load



✅ `load()`

- **Eager Loading** (loads everything at once).
- Returns: A **list of** `Document` **objects.**
- Loads all documents **immediately** into memory.
- Best when:
  - The number of documents is small.
  - You want everything loaded upfront.

🌀 `lazy_load()`

- **Lazy Loading** (loads on demand).
- Returns: A **generator** of `Document` objects.
- Documents are **not all loaded at once**; they're fetched one at a time as needed.
- Best when:
  - You're dealing with **large documents or lots of files.**
  - You want to **stream** processing (e.g., chunking, embedding) without using lots of memory.

WebBaseLoader

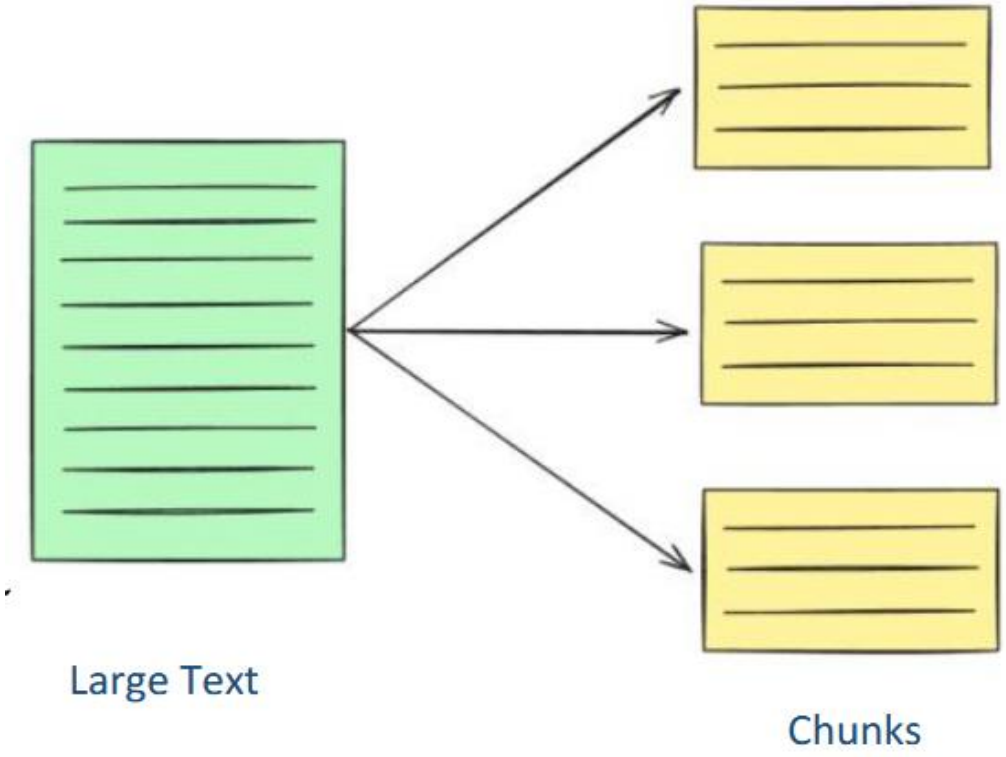
- **WebBaseLoader** is a document loader in LangChain used to load and extract text content from web pages (URLs).
- It uses BeautifulSoup under the hood to parse HTML and extract visible text.
- **When to Use:**
- For blogs, news articles, or public websites where the content is primarily text-based and static.
- **Limitations:**
- Doesn’t handle JavaScript-heavy pages well (use SeleniumURLLoader for that).
- Loads only static content (what's in the HTML, not what loads after the page renders)

CSVLoader

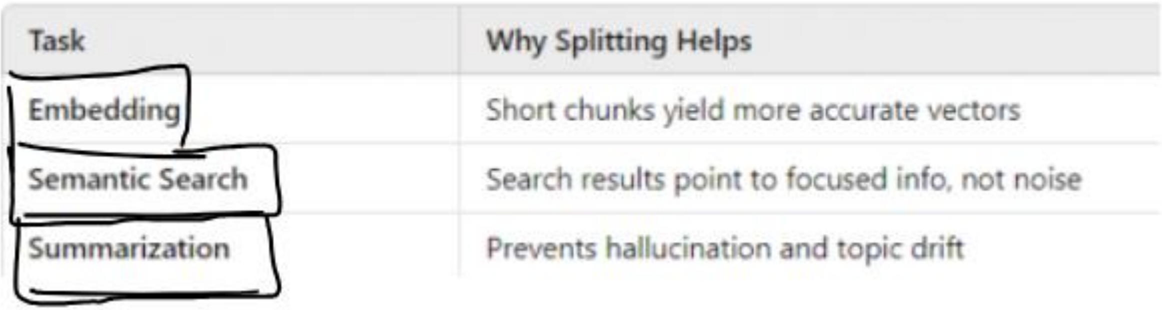
- **CSVLoader** is a document loader used to load CSV files into LangChain Document objects — one per row, by default.
- **Other Document Loaders**

Text Splitting

- **Text Splitting** is the process of breaking large chunks of text (like articles, PDFs, HTML pages, or books) into smaller, manageable pieces (chunks) that an LLM can handle effectively

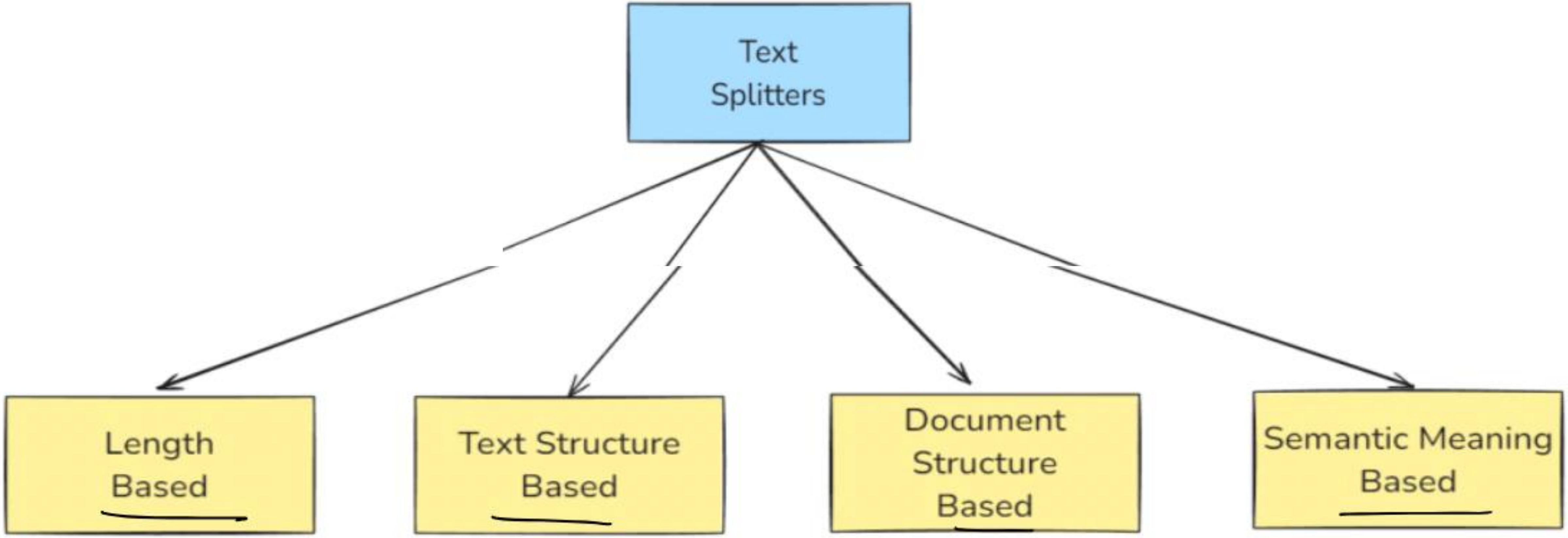


- **Overcoming model limitations:** Many embedding models and language models have maximum input size constraints. Splitting allows us to process documents that would otherwise exceed these limits.
- Downstream tasks - Text Splitting improves nearly every LLM powered task



Task	Why Splitting Helps
Embedding	Short chunks yield more accurate vectors
Semantic Search	Search results point to focused info, not noise
Summarization	Prevents hallucination and topic drift

- Optimizing computational resources: Working with smaller chunks of text can be more memory-efficient and allow for better parallelization of processing tasks.



Text Splitters

Length Based

Text Structure Based

Document Structure Based

Semantic Meaning Based

Length Based Text Splitting

Space exploration has led to incredible scientific discoveries. From landing on the Moon to exploring Mars, humanity continues to push the boundaries of what's possible beyond our planet.

These missions have not only expanded our knowledge of the universe but have also contributed to advancements in technology here on Earth. Satellite communications, GPS, and even certain medical imaging techniques trace their roots back to innovations driven by space program

Space exploration has led to incredible scientific discoveries. From landing on the Moon to explorin → c1

g Mars, humanity continues to push the boundaries of what's possible beyond our planet. These missi c2

ons have not only expanded our knowledge of the universe but have also contributed to advancements in c3

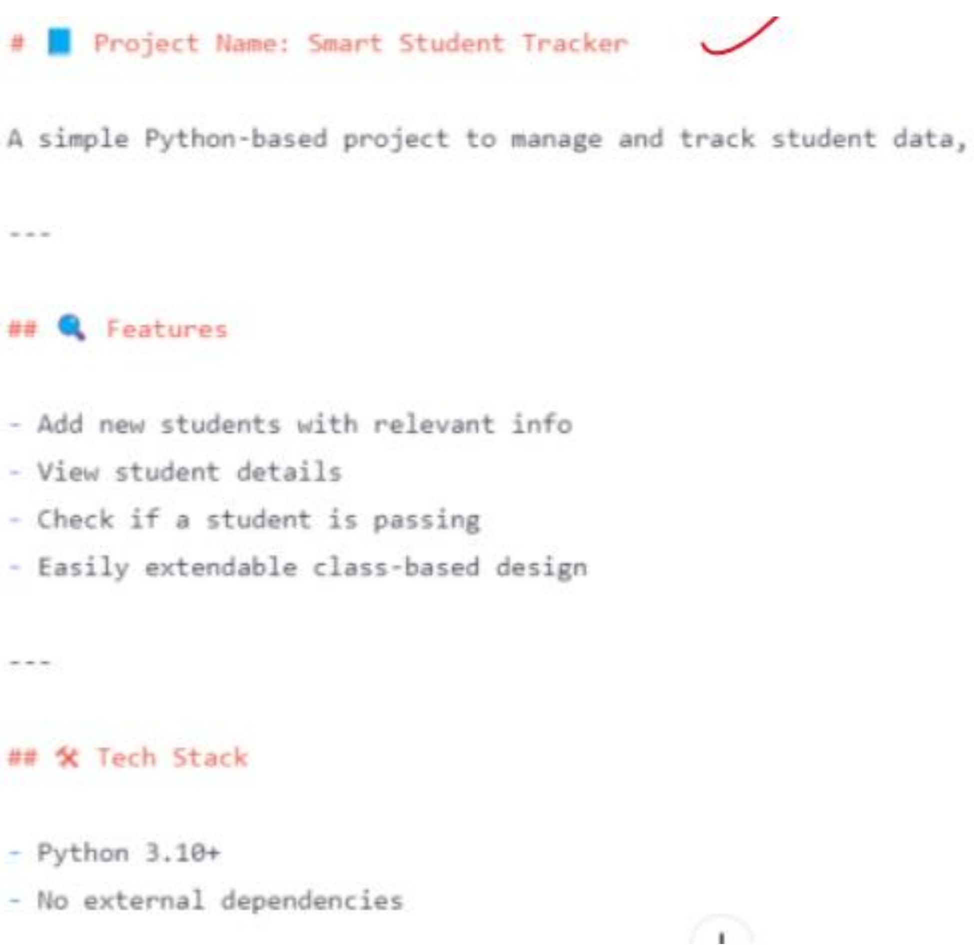
n technology here on Earth. Satellite communications, GPS, and even certain medical imaging techniqu c4

es trace their roots back to innovations driven by space programs. c5

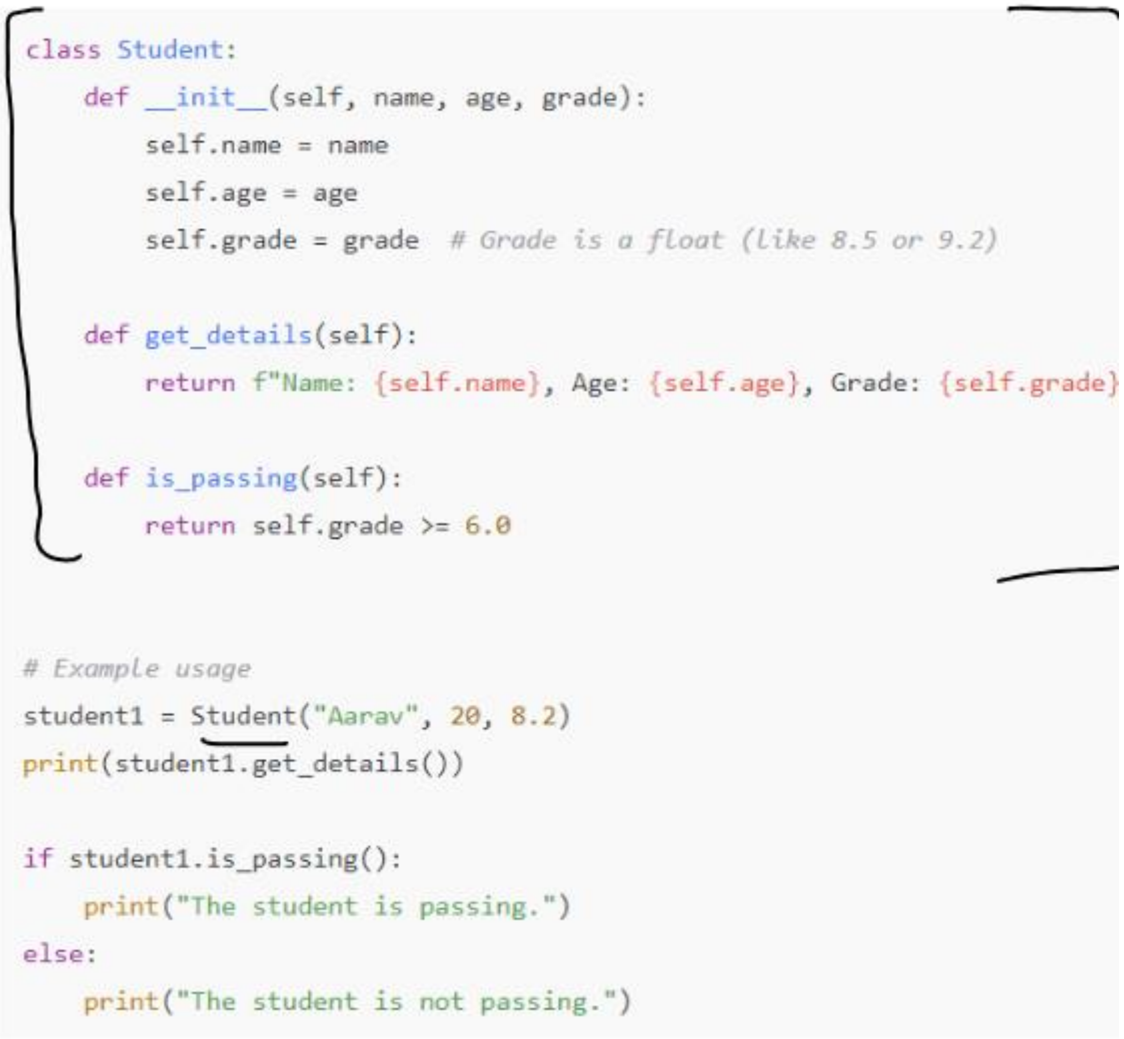
2. Text-Structured Based

- My name is Mehmed
- I am 35 years old
- I live in Lahore
- How are you

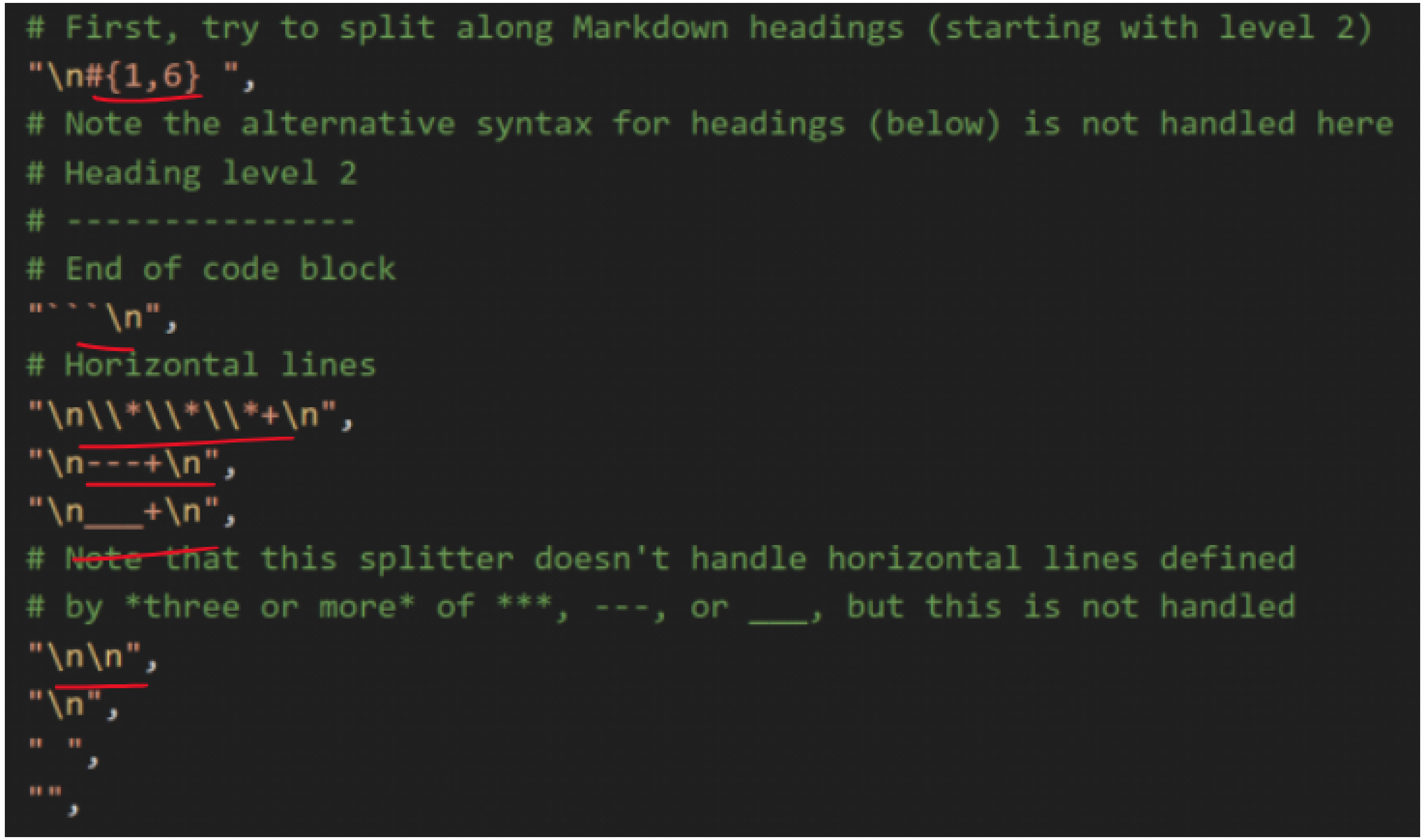
3. Document-Structured Based



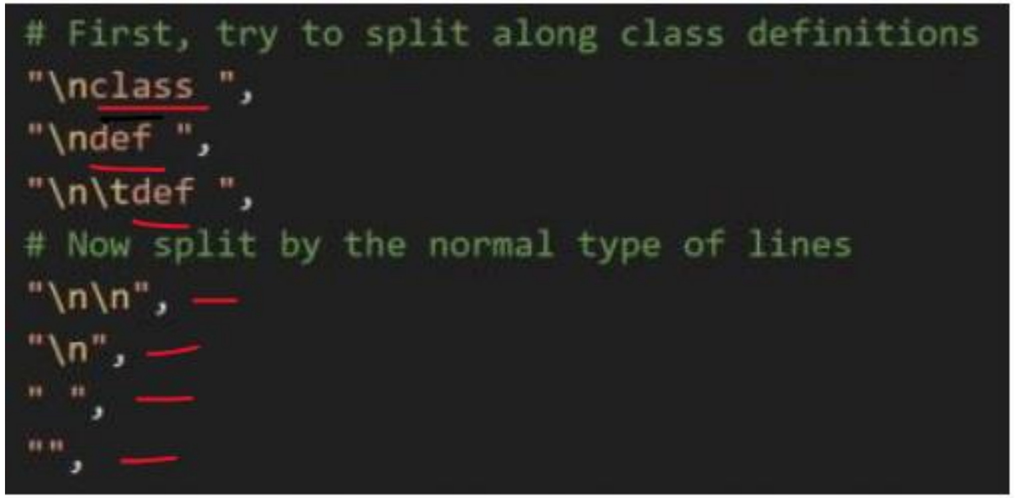
```
# 📘 Project Name: Smart Student Tracker

A simple Python-based project to manage and track student data,

---

## 🔍 Features

- Add new students with relevant info
- View student details
- Check if a student is passing
- Easily extendable class-based design

---

## 🛠 Tech Stack

- Python 3.10+
- No external dependencies
```



```python
class Student:
    def __init__(self, name, age, grade):
        self.name = name
        self.age = age
        self.grade = grade  # Grade is a float (like 8.5 or 9.2)

    def get_details(self):
        return f"Name: {self.name}, Age: {self.age}, Grade: {self.grade}

    def is_passing(self):
        return self.grade >= 6.0


# Example usage
student1 = Student("Aarav", 20, 8.2)
print(student1.get_details())

if student1.is_passing():
    print("The student is passing.")
else:
    print("The student is not passing.")
```



```
# First, try to split along Markdown headings (starting with level 2)
"\n#{1,6} ",
# Note the alternative syntax for headings (below) is not handled here
# Heading level 2
# ---------------
# End of code block
"```\n",
# Horizontal lines
"\n\\*\\*\\*+\n",
"\n---+\n",
"\n___+\n",
# Note that this splitter doesn't handle horizontal lines defined
# by *three or more* of ***, ---, or ___, but this is not handled
"\n\n",
"\n",
" ",
"",
```



```
# First, try to split along class definitions
"\nclass ",
"\ndef ",
"\n\tdef ",
# Now split by the normal type of lines
"\n\n",
"\n",
" ",
"",
```

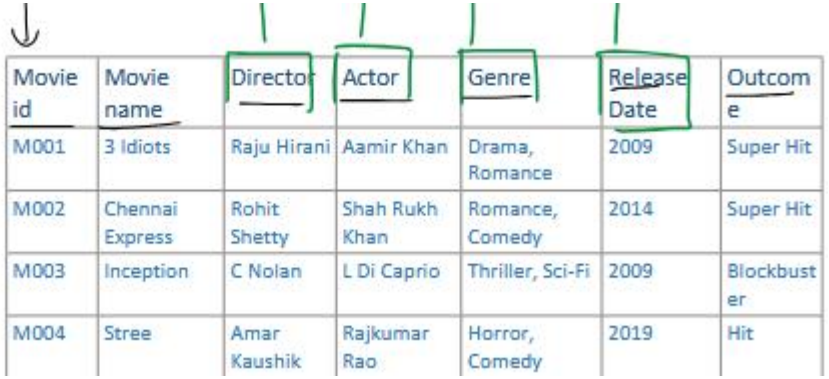
Semantic Meaning Based

- Farmers were working hard in the fields, preparing the soil and planting seeds for the next season. The sun was bright, and the air smelled of earth and fresh grass.The Pakistan Premier League (PSL) is the biggest cricket league in the world. People all over the world watch the matches and cheer for their favourite teams.

- Terrorism is a big danger to peace and safety. It causes harm to people and creates fear in cities and villages. When such attacks happen, they leave behind pain and sadness. To fight terrorism, we need strong laws, alert security forces, and support from people who care about peace and safety.

Vector DataBases

Why Vector Stores?



Movie id	Movie name	Director	Actor	Genre	Release Date	Outcome
M001	3 Idiots	Raju Hirani	Aamir Khan	Drama, Romance	2009	Super Hit
M002	Chennai Express	Rohit Shetty	Shah Rukh Khan	Romance, Comedy	2014	Super Hit
M003	Inception	C Nolan	L Di Caprio	Thriller, Sci-Fi	2009	Blockbuster
M004	Stree	Amar Kaushik	Rajkumar Rao	Horror, Comedy	2019	Hit