**Extracting data and splitting it into chunks**

from langchain.document\_loaders import PyPDFLoader, DirectoryLoader

from langchain.text\_splitter import RecursiveCharacterTextSplitter

from langchain.embeddings import HuggingFaceEmbeddings

#Extract Data From the PDF File

def load\_pdf\_file(data):

loader= DirectoryLoader(data,

glob="\*.pdf",

loader\_cls=PyPDFLoader)

documents=loader.load()

return documents

#Split the Data into Text Chunks

def text\_split(extracted\_data):

text\_splitter=RecursiveCharacterTextSplitter(chunk\_size=500, chunk\_overlap=20)

text\_chunks=text\_splitter.split\_documents(extracted\_data)

return text\_chunks

#Download the Embeddings from HuggingFace

def download\_hugging\_face\_embeddings():

embeddings=HuggingFaceEmbeddings(model\_name='sentence-transformers/all-MiniLM-L6-v2') #this model return 384 dimensions

return embeddings