```
In [1]: #!pip install langchain
         #!pip install openai
         #!pip install PyPDF2
         #!pip install faiss-cpu
         #!pip install tiktoken
 In [2]: from PyPDF2 import PdfReader
         from langchain.embeddings.openai import OpenAIEmbeddings
         from langchain.text splitter import CharacterTextSplitter
         from langchain.vectorstores import FAISS
In [3]:
         import getpass
         import os
         os.environ['OPEN API KEY'] =
In [4]: #provide the path of the pdf files
         pdfreader = PdfReader('Fun-Facts-About-Our-Universe.pdf')
 In [5]: from typing extensions import Concatenate
         #read text from pdf
         raw text =''
         for i,page in enumerate(pdfreader.pages):
             content = page.extract text() #extract text is the built-in function of concatenate
             if content:
                 raw text += content
 In [6]: #Glance at how the data looks
         raw text[:1000]
         'NASA Resource \nSlide Set\nThe Universe is vast and always changing. It includes \nplan
 Out[6]:
         ets, stars, galaxies, and all of space, matter, and \nenergy. Looking at the night sky,
         you can sense how \nsmall Earth is in the vastness of space. A million Earths \ncould fi
         t inside the Sun. The Sun is the only star in our \nSolar System but is one of over 100
         billion stars in a spiral \ngalaxy we call the Milky Way. Our Solar System is located \n
         about 2/3 of the way out from the galaxy's center. \nAstronomers think that most of the
         stars in our galaxy \nmay also have planets orbiting around them.\n2Milky Way GalaxyCons
         tellations\nIn contemporary astronomy, a constellation is one of 88 \nregions of the sky
         that cover the entire celestial sphere. \nThe ancient Sumerians, and later the Greeks (a
         s \nrecorded by Ptolemy), established most of the northern \nconstellations in internati
         onal use today.\nOrion is a prominent constellation located on the celestial \nequator a
         nd visible throughout the world. It was named afte'
 In [7]: | #We need to split the text using Character text split such that it should not increase t
         text splitter = CharacterTextSplitter(
             separator = ' n',
             chunk size = 800,
             chunk overlap = 200,
             length function = len,
         texts = text splitter.split text(raw text)
In [8]:
         len(texts)
Out[8]:
 In [9]: #Download embeddings from OpenAI to create vectors of these texts
         openai api key= 'sk-
         embeddings = OpenAIEmbeddings(openai api key= openai api key)
In [10]: #Creating embeddings for the text and storing in document search
         document search = FAISS.from texts(texts, embeddings)
```

```
In [11]: from langchain.chains.question_answering import load_qa_chain
    from langchain.llms import OpenAI

In [12]: openai_instance = OpenAI (openai_api_key=openai_api_key)

In [13]: chain = load_qa_chain(openai_instance, chain_type ='stuff')

In [14]: query = 'How mass is distributed in our solar system?'
    docs = document_search.similarity_search(query)
    chain.run(input_documents = docs, question = query)

Out[14]: "The vast majority of the Solar System's mass is in the Sun, with most of the remaining
    mass contained in Jupiter."

In [15]: query = 'Tell me something about Earth moon?'
    docs = document_search.similarity_search(query)
    chain.run(input_documents = docs, question = query)
```

Out[15]: "Earth's Moon is the 5th largest moon in the Solar System. Its surface is mostly dark a nd featureless lunar plains called maria, which were believed by ancient astronomers to be filled with water. The lighter-colored regions of the Moon are called highlands, sinc e they are higher than most maria. Several prominent mountain ranges on the near side ar e found along the periphery of the giant impact basins, many of which have been filled by mare basalt. It is the only celestial body humans have travelled to and landed on."