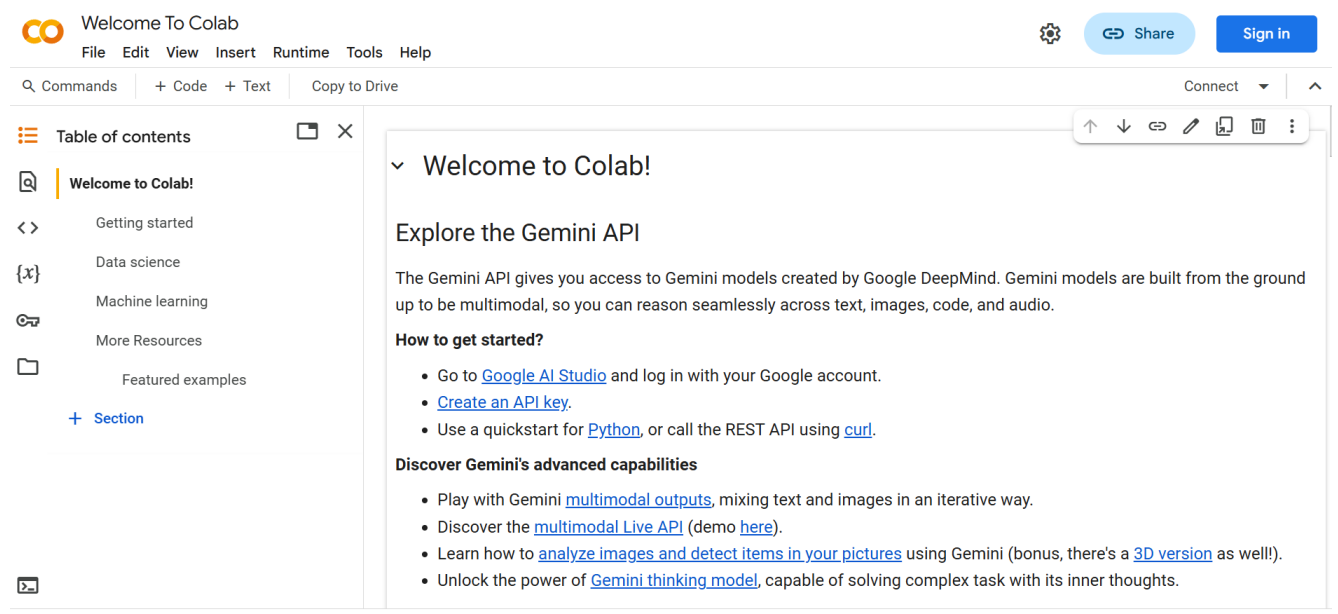


DATA PREP KIT(DPK)

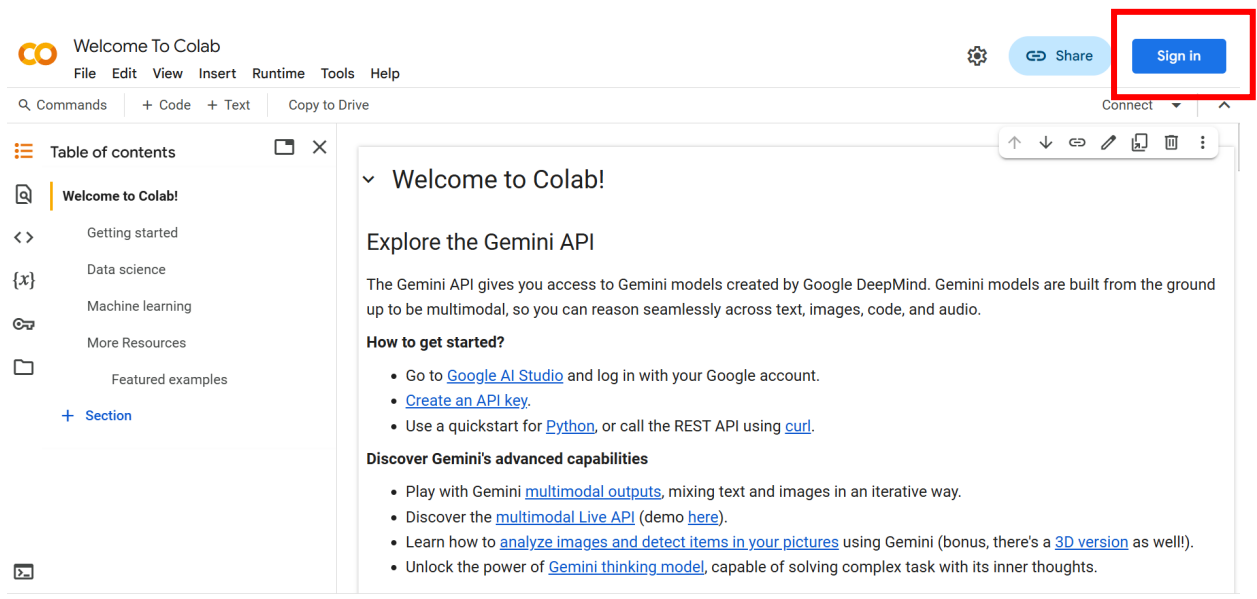
Data Prep Kit accelerates unstructured data preparation for LLM app developers. Developers can use Data Prep Kit to cleanse, transform, and enrich use case-specific unstructured data to pre-train LLMs, fine-tune LLMs, instruct-tune LLMs, or build retrieval augmented generation (RAG) applications for LLMs.

The following notebook example will allow you to test DPK, without cloning the repo. You can run it either on Google Colab or you can use your local environment (by downloading just the notebook). We use a temporary folder for input and output, but users are encouraged to use their own input folder.

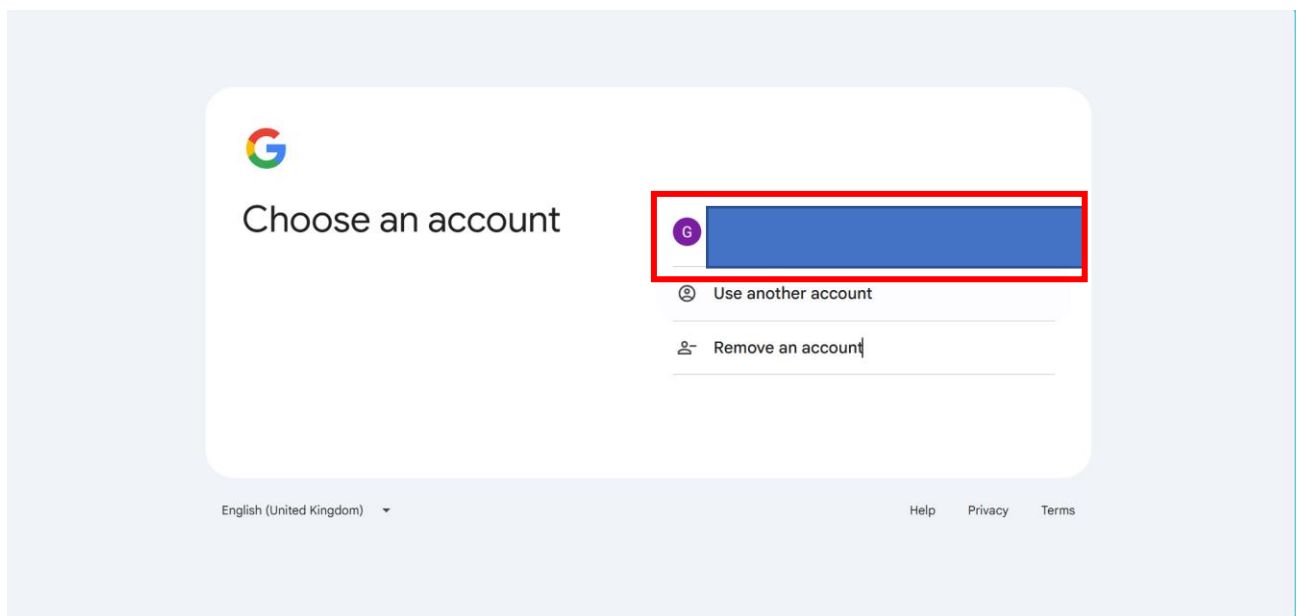
Step 1: Login to Google Colab (<https://colab.research.google.com/>)



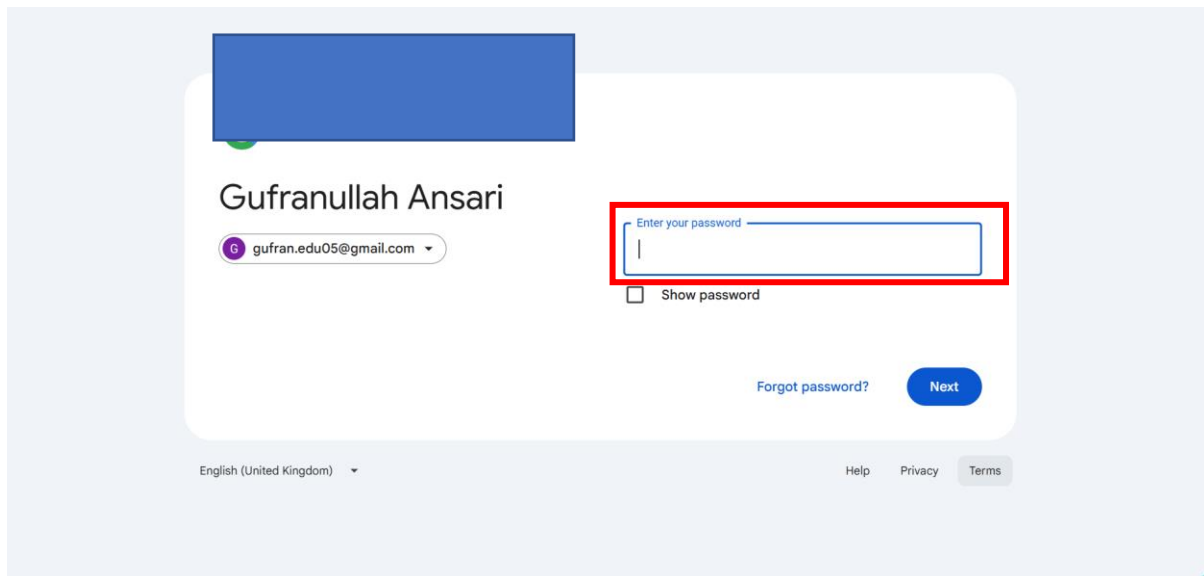
Step 2: Use your Gmail Account Id for Sign In



Step 3: Choose your Gmail account

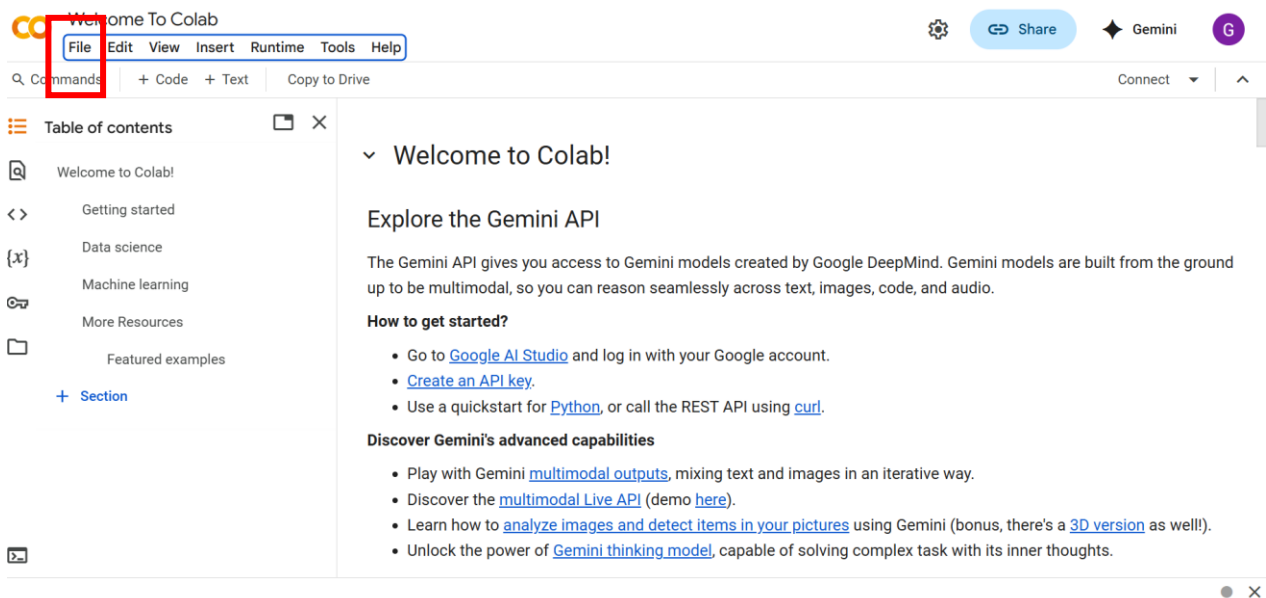


Step 4: Enter your password for the respective account



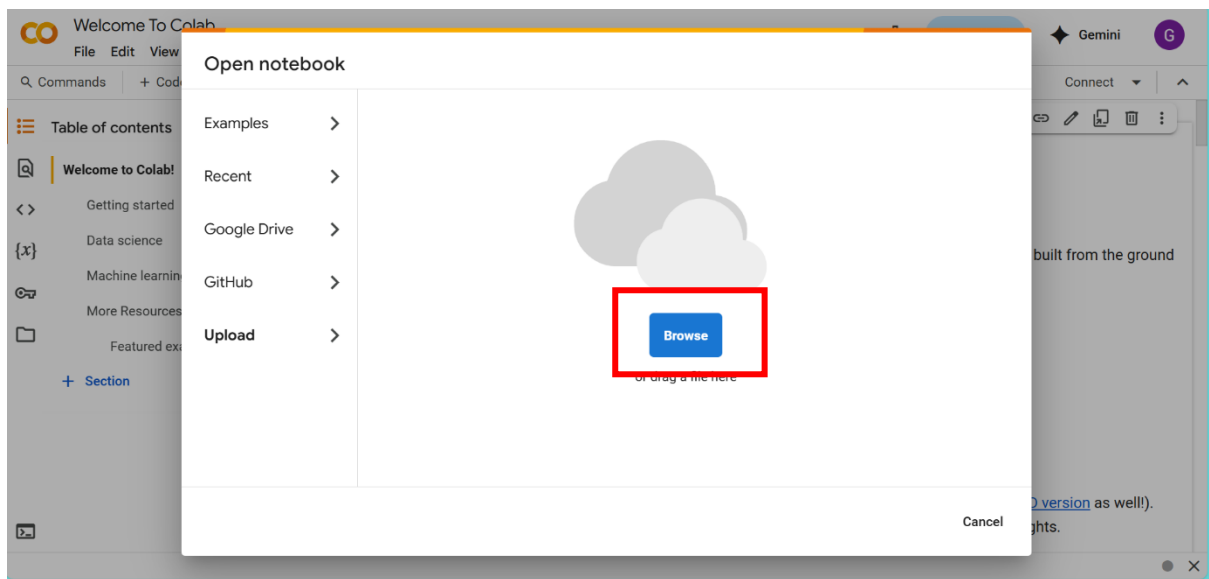
The image shows a Google account login interface. At the top, there is a blue rectangular placeholder for a profile picture. Below it, the name "Gufranullah Ansari" is displayed, followed by the email address "gufran.edu05@gmail.com" with a dropdown arrow. To the right of the email is a password input field with the placeholder text "Enter your password". This input field is highlighted with a red rectangular border. Below the password field is a checkbox labeled "Show password". At the bottom right, there is a "Next" button and a "Forgot password?" link. At the bottom left, there is a language selector showing "English (United Kingdom)". At the bottom right, there are links for "Help", "Privacy", and "Terms".

Step 5: Click on 'File'

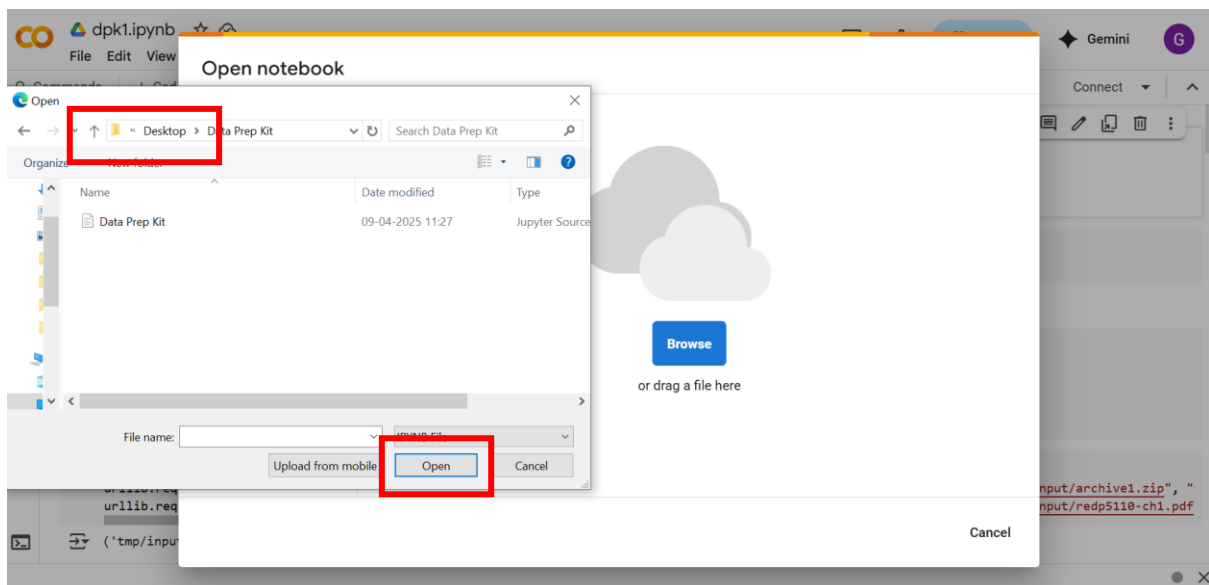


The image shows the Google Colab (Collaboratory) interface. At the top, there is a "Welcome To Colab" header. Below it is a menu bar with "File", "Edit", "View", "Insert", "Runtime", "Tools", and "Help". The "File" menu is highlighted with a red rectangular border. To the right of the menu bar are buttons for "Share", "Gemini", and a user profile icon. Below the menu bar is a search bar and a "Copy to Drive" button. On the left side, there is a "Table of contents" panel with a list of items: "Welcome to Colab!", "Getting started", "Data science", "Machine learning", "More Resources", and "Featured examples". The "Welcome to Colab!" item is selected. The main content area displays the "Welcome to Colab!" message, followed by "Explore the Gemini API" and a description of the Gemini API. Below this, there are two sections: "How to get started?" and "Discover Gemini's advanced capabilities", each with a list of bullet points. The "How to get started?" section includes links to "Google AI Studio", "Create an API key", and "Python". The "Discover Gemini's advanced capabilities" section includes links to "multimodal outputs", "multimodal Live API", "analyze images and detect items in your pictures", "3D version", and "Gemini thinking model".

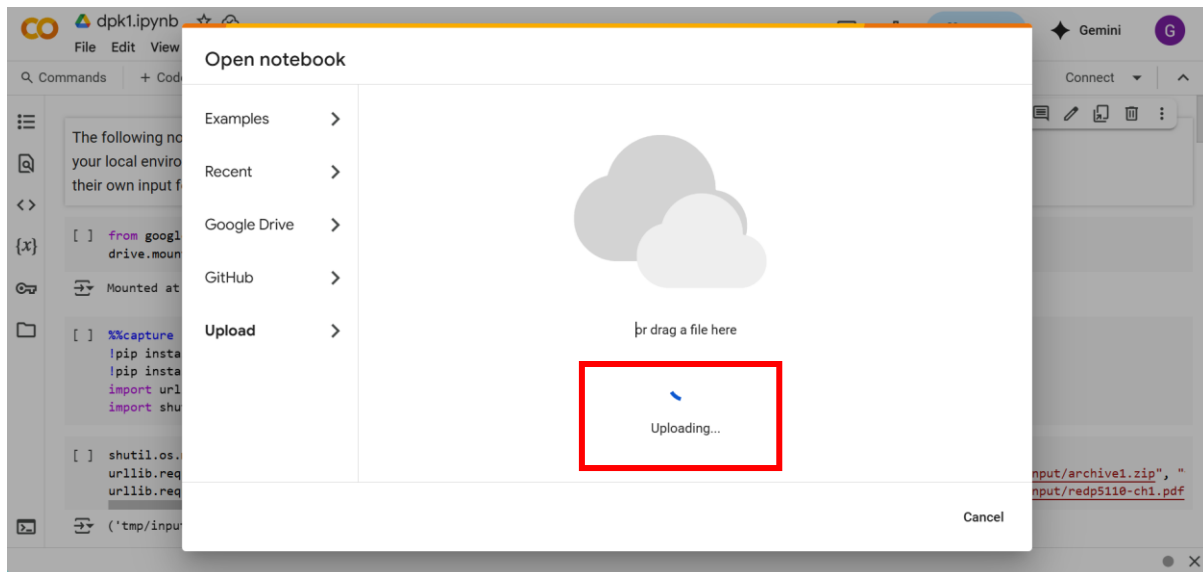
Step 6: Click on Browse



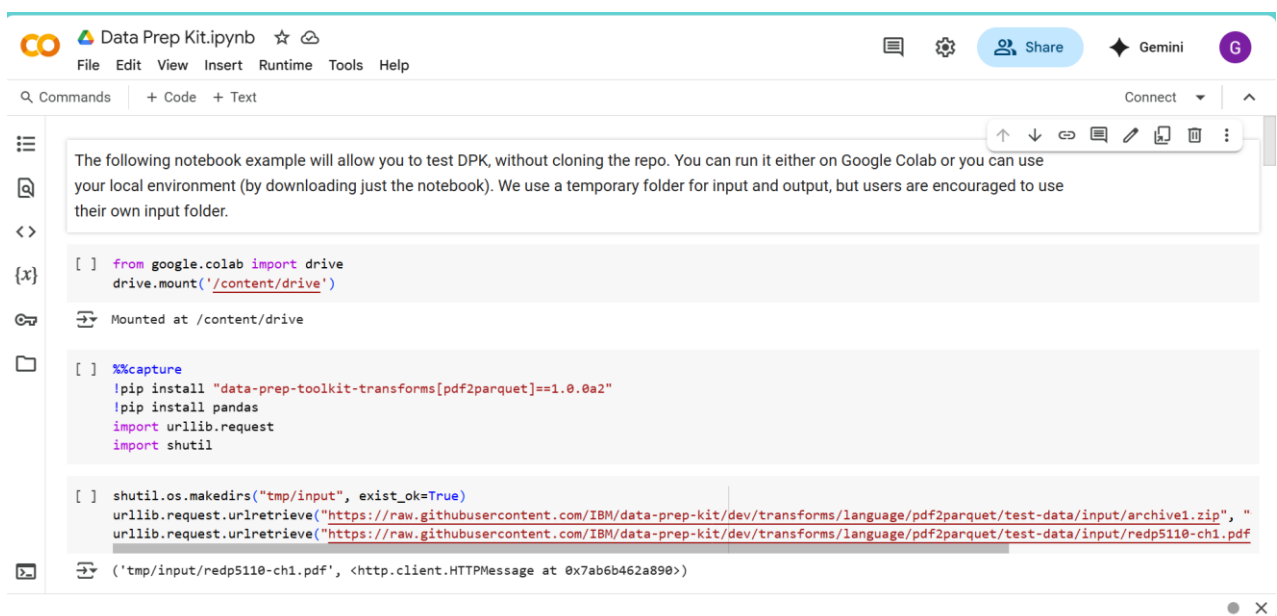
Step 7: Upload the Data Prep Kit *Jupyter Source File (.ipynb)* file from save location



Step 8: Uploading status wait here till file uploaded

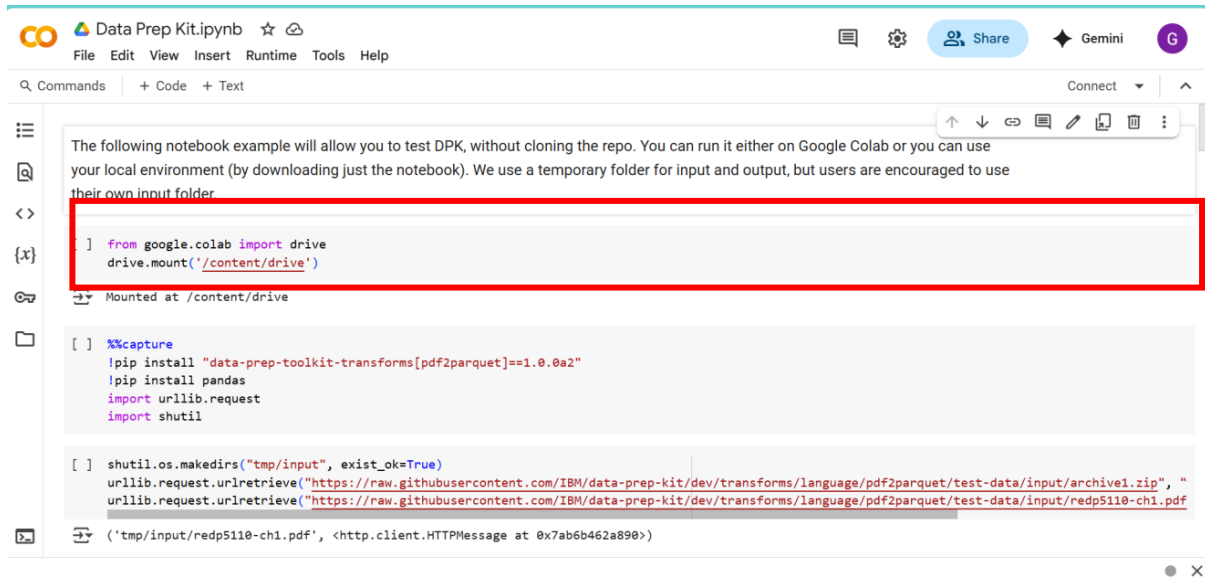


Step 9: After file uploaded below dashboard available



Step 10: Now Run each cell carefully

1.First cell Run and mount the your google drive



```
from google.colab import drive
drive.mount('/content/drive')
```

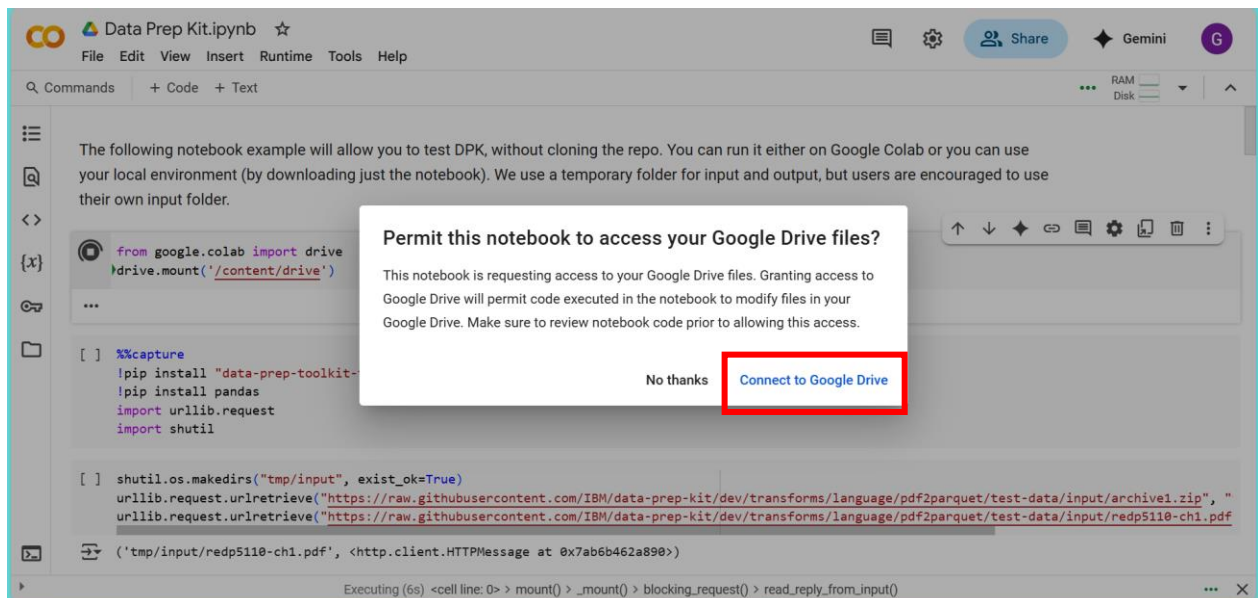
```
Mounted at /content/drive
```

```
!pip install "data-prep-toolkit-transforms[pdf2parquet]==1.0.0a2"
!pip install pandas
import urllib.request
import shutil

shutil.os.makedirs("tmp/input", exist_ok=True)
urllib.request.urlretrieve("https://raw.githubusercontent.com/IBM/data-prep-kit/dev/transforms/language/pdf2parquet/test-data/input/archive1.zip", "
urllib.request.urlretrieve("https://raw.githubusercontent.com/IBM/data-prep-kit/dev/transforms/language/pdf2parquet/test-data/input/redp5110-ch1.pdf"

('tmp/input/redp5110-ch1.pdf', <http.client.HTTPMessage at 0x7ab6b462a890>)
```

Step 10: Click on Connect to Google Drive



```
from google.colab import drive
drive.mount('/content/drive')
```

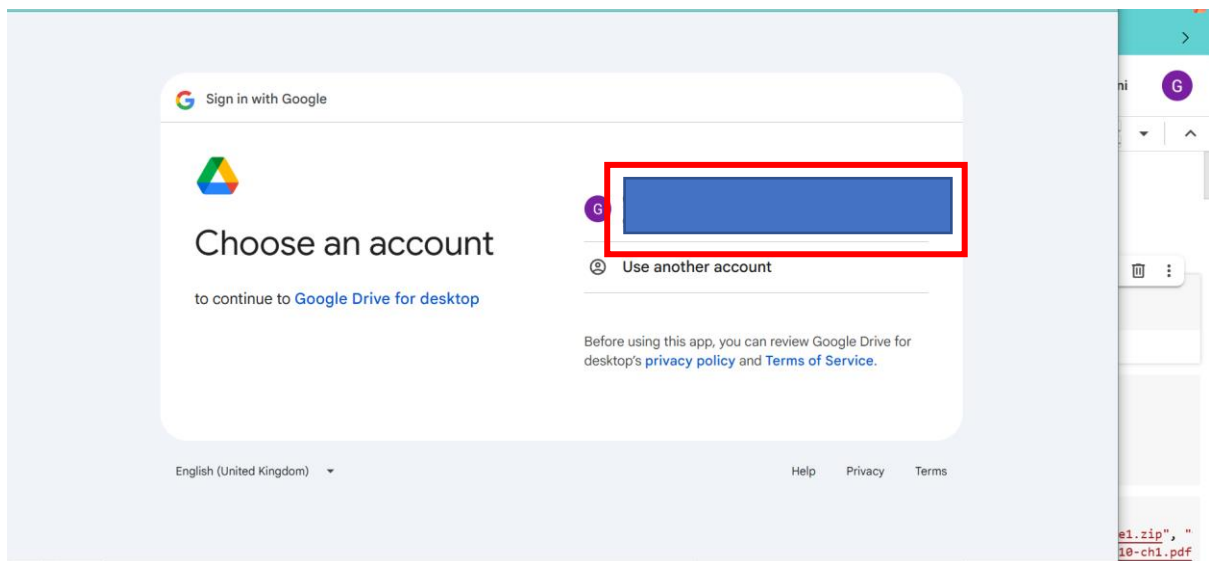
```
!pip install "data-prep-toolkit-transforms[pdf2parquet]==1.0.0a2"
!pip install pandas
import urllib.request
import shutil

shutil.os.makedirs("tmp/input", exist_ok=True)
urllib.request.urlretrieve("https://raw.githubusercontent.com/IBM/data-prep-kit/dev/transforms/language/pdf2parquet/test-data/input/archive1.zip", "
urllib.request.urlretrieve("https://raw.githubusercontent.com/IBM/data-prep-kit/dev/transforms/language/pdf2parquet/test-data/input/redp5110-ch1.pdf"

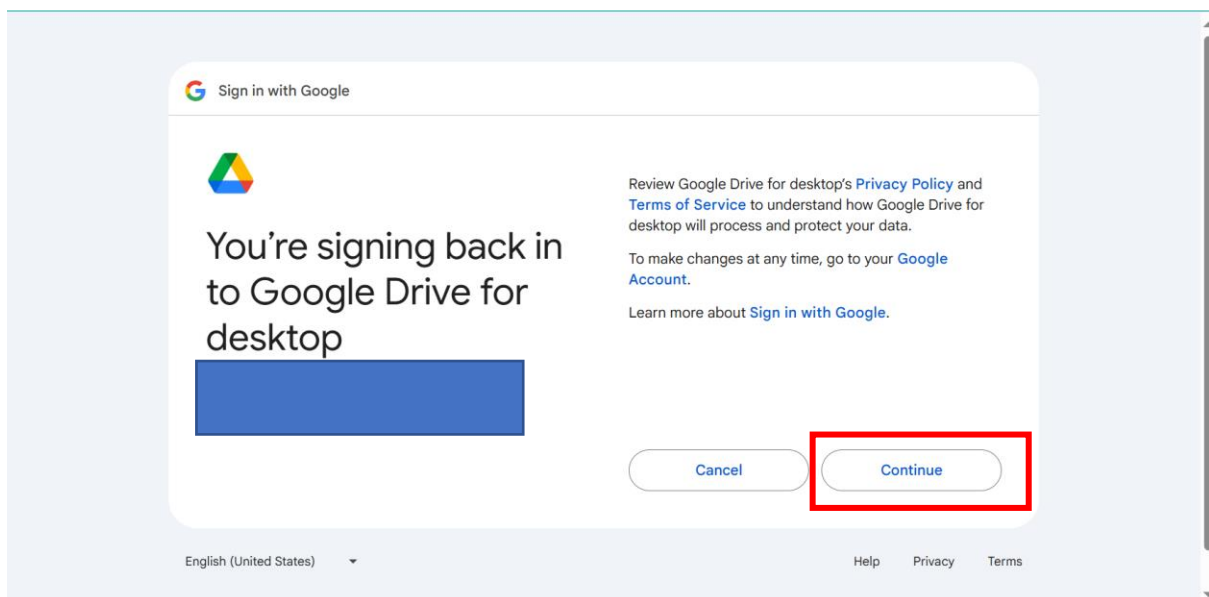
('tmp/input/redp5110-ch1.pdf', <http.client.HTTPMessage at 0x7ab6b462a890>)
```

Executing (6s) <cell line: 0> > mount() > _mount() > blocking_request() > read_reply_from_input()

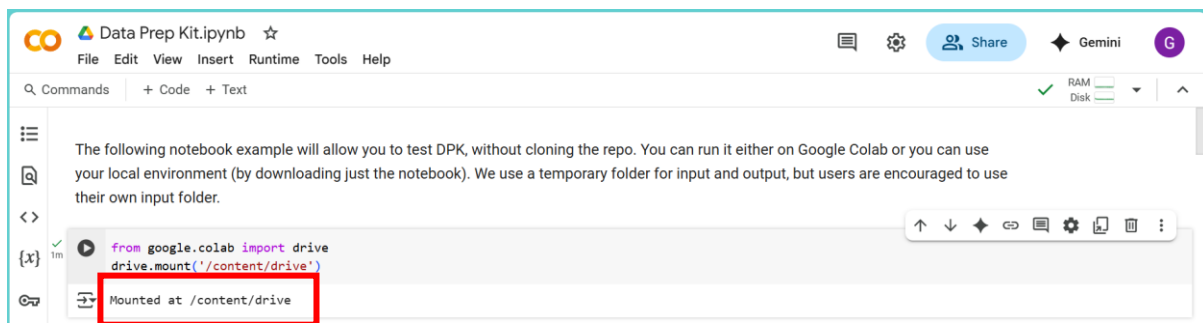
Step 11: Choose Gmail account



Step 12: Click on continue and follow the next step



Step 13: Now google drive mounted



The screenshot shows a Google Colab notebook titled "Data Prep Kit.ipynb". The code cell contains the following Python code:

```
from google.colab import drive
drive.mount('/content/drive')
```

The output of the code cell shows "Mounted at /content/drive".

Step 14: Run the Next cell one by one



The screenshot shows a Google Colab notebook with two code cells. The first cell contains the following code:

```
!pip install "data-prep-toolkit-transforms[pdf2parquet]==1.0.0a2"
!pip install pandas
import urllib.request
import shutil
```

The second cell contains the following code:

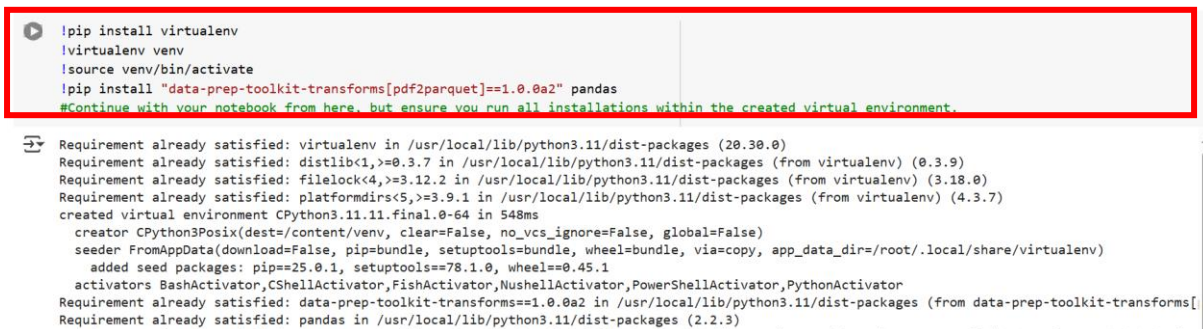
```
[ ] shutil.os.makedirs("tmp/input", exist_ok=True)
urllib.request.urlretrieve("https://raw.githubusercontent.com/IBM/data-prep-kit/dev/transforms/language/pdf2parquet/test-data/input/archive1.zip", "
urllib.request.urlretrieve("https://raw.githubusercontent.com/IBM/data-prep-kit/dev/transforms/language/pdf2parquet/test-data/input/redp5110-ch1.pdf", "
```



The screenshot shows a Google Colab notebook with a code cell containing the following code:

```
!pip install --upgrade numpy
!pip install --upgrade pandas
```

The output of the code cell shows the installation progress for numpy and pandas.



The screenshot shows a Google Colab notebook with a code cell containing the following code:

```
!pip install virtualenv
!virtualenv venv
!source venv/bin/activate
!pip install "data-prep-toolkit-transforms[pdf2parquet]==1.0.0a2" pandas
```

The output of the code cell shows the installation progress for virtualenv and the creation of a virtual environment.


```

▶ Pdf2Parquet(input_folder= "tmp/input",
               output_folder= "tmp/output",
               data_files_to_use=['.pdf', '.zip'],
               pdf2parquet_contents_type=pdf2parquet_contents_types.JSON).transform()

```

```
import pyarrow.parquet as pq
import pandas as pd
table = pq.read_table('tmp/output/archive1.parquet')
table.to_pandas()
```

	filename	contents	num_pages	num_tables	num_doc_elements	document_id	document_hash	ext
0	2305.03393v1-pg9.pdf	{ "schema_name": "DoclingDocument", "version": "1....	1	1	9	f5aa422e-6caa-4c44-90f1-b5e1a58cd364	3463920545297462180	pdf 467dcf63
1	2408.09869v1-pg1.pdf	{ "schema_name": "DoclingDocument", "version": "1....	1	0	12	4ba65f96-2a42-4963-a9d1-b94317728d1e	582377908831471240	pdf 8ed0cb6d

	filename	contents	num_pages	num_tables	num_doc_elements	document_id	document_hash	ext
0	redp5110-ch1.pdf	{'schema_name':'DoclingDocument','version':'1....	5	0	48	7c1422931f151-405d-b82-54436b17a7a9	74198560999363607	pdf 572c2937fa0e265

The screenshot displays the Data Prep Kit interface. On the left, a file explorer shows a directory structure with 'sample_data' containing a 'tmp' folder, which is highlighted with a red box. The main area on the right is a code editor showing a Python script that imports PyArrow and Pandas, reads a Parquet file from 'tmp/output/archive1.parquet', and converts it to a pandas DataFrame. Below the code, a table view displays the data with columns: filename, contents, num_pages, num_tables, num_doc_elements, and document. The table contains two rows of data. At the bottom, a status bar indicates the execution completed at 11:53 AM.

	filename	contents	num_pages	num_tables	num_doc_elements	document
0	2305.03393v1-pg9.pdf	{'schema_name': 'DoclingDocument', 'version': '1...	1	1	9	f5aa426caa-4c9cb5e1a58cd:
1	2408.09869v1-pg1.pdf	{'schema_name': 'DoclingDocument', 'version': '1...	1	0	12	4ba65f2a42-49a9b94317728:

The screenshot shows the Google Colab interface for a notebook titled "Data Prep Kit.ipynb". The left sidebar displays the file explorer with a directory structure: "drive" > "sample_data" > "tmp". Inside "tmp", there are "input" and "output" folders. The "input" folder contains "archive1.zip" and "redp5110-ch1.pdf". The "output" folder contains "archive1.parquet". A red box highlights the "tmp" directory and its contents.

The main code editor shows two code blocks. The first block reads a Parquet file from "tmp/output/archive1.parquet". The second block reads a Parquet file from "tmp/output/redp5110-ch1.parquet". The output of the second block is displayed as a table with columns "filename", "contents", "num_pages", and "num".

	filename	contents	num_pages	num
0	2305.03393v1-pg9.pdf	{ "schema_name": "DoclingDocument", "version": "1....	1	
1	2408.09869v1-pg1.pdf	{ "schema_name": "DoclingDocument", "version": "1....	1	

The bottom status bar indicates "0s completed at 12:05 PM".

Congratulations! You successfully Run Data Prep Kit using Google Colab.