Google Drive

> Google Drive is a file storage and synchronization service developed by Google.

This notebook covers how to load documents from `Google Drive`. Currently, only `Google Docs` are supported.

Prerequisites

1. Create a Google Cloud project or use an existing project
2. Enable the Google Drive API
3. Authorize credentials for desktop app
4. `pip install --upgrade google-api-python-client google-auth-httplib2 google-auth-oauthlib`

🧑 Instructions for ingesting your Google Docs data

By default, the `GoogleDriveLoader` expects the `credentials.json` file to be `~/.credentials/credentials.json`, but this is configurable using the `credentials_path` keyword argument. Same thing with `token.json` - `token_path`. Note that `token.json` will be created automatically the first time you use the loader.

`GoogleDriveLoader` can load from a list of Google Docs document ids or a folder id. You can obtain your folder and document id from the URL:

- Folder: https://drive.google.com/drive/u/0/folders/1yucgL9WGgWZdM1TOuKkeghlPizuzMYb5 -> folder id is `"1yucgL9WGgWZdM1TOuKkeghlPizuzMYb5"`
- Document: https://docs.google.com/document/d/1bfaMQ18_i56204VaQDVeAFpqEijJTgvurupdEDiaUQw/edit -> document id is `"1bfaMQ18_i56204VaQDVeAFpqEijJTgvurupdEDiaUQw"`

```
pip install --upgrade google-api-python-client google-auth-httplib2 google-auth-oauthlib
```

```
from langchain.document_loaders import GoogleDriveLoader
```

```
loader = GoogleDriveLoader(
    folder_id="1yucgL9WGgWZdM1TOuKkeghlPizuzMYb5",
    # Optional: configure whether to recursively fetch files from subfolders. Defaults to False.
    recursive=False,
)
```

```
docs = loader.load()
```

When you pass a `folder_id` by default all files of type document, sheet and pdf are loaded. You can modify this behaviour by passing a `file_types` argument

```
loader = GoogleDriveLoader(
    folder_id="1yucgL9WGgWZdM1TOuKkeghlPizuzMYb5",
    file_types=["document", "sheet"]
    recursive=False
)
```

Passing in Optional File Loaders

When processing files other than Google Docs and Google Sheets, it can be helpful to pass an optional file loader to `GoogleDriveLoader`. If you pass in a file loader, that file loader will be used on documents that do not have a Google Docs or Google Sheets MIME type. Here is an example of how to load an Excel document from Google Drive using a file loader.

```
from langchain.document_loaders import GoogleDriveLoader
from langchain.document_loaders import UnstructuredFileIOLoader
```

```
file_id="1x9WBtFPWMEAdjcJzPScRsjpjQvpSo_kz"
loader = GoogleDriveLoader(
    file_ids=[file_id],
    file_loader_cls=UnstructuredFileIOLoader,
    file_loader_kwargs={"mode": "elements"}
)
```

```
docs = loader.load()
```

```
docs[0]
```

```
    Document(page_content='\n  \n    \n      Team\n      Location\n      Stanley Cups\n    \n    \n
Blues\n      STL\n      1\n    \n    \n      Flyers\n      PHI\n      2\n    \n    \n      Maple Leafs\n
TOR\n      13\n    \n  \n', metadata={'filetype': 'application/vnd.openxmlformats-
officedocument.spreadsheetml.sheet', 'page_number': 1, 'page_name': 'Stanley Cups', 'text_as_html': '<table
border="1" class="dataframe">\n  <tbody>\n    <tr>\n      <td>Team</td>\n      <td>Location</td>\n
<td>Stanley Cups</td>\n    </tr>\n    <tr>\n      <td>Blues</td>\n      <td>STL</td>\n      <td>1</td>\n
</tr>\n    <tr>\n      <td>Flyers</td>\n      <td>PHI</td>\n      <td>2</td>\n    </tr>\n    <tr>\n
<td>Maple Leafs</td>\n      <td>TOR</td>\n      <td>13</td>\n    </tr>\n  </tbody>\n</table>', 'category':
'Table', 'source': 'https://drive.google.com/file/d/1aA6L2AR3g0CR-PW03HEZZo4NaVlKpaP7/view'})
```

You can also process a folder with a mix of files and Google Docs/Sheets using the following pattern:

```
folder_id="1asMOHY1BqBS84JcRbOag5LOJac74gpmD"
loader = GoogleDriveLoader(
    folder_id=folder_id,
    file_loader_cls=UnstructuredFileIOLoader,
    file_loader_kwargs={"mode": "elements"}
)
```

```
docs = loader.load()
```

```
docs[0]
```

```
    Document(page_content='\n  \n    \n      Team\n      Location\n      Stanley Cups\n    \n    \n
Blues\n      STL\n      1\n    \n    \n      Flyers\n      PHI\n      2\n    \n    \n      Maple Leafs\n
TOR\n      13\n    \n  \n', metadata={'filetype': 'application/vnd.openxmlformats-
officedocument.spreadsheetml.sheet', 'page_number': 1, 'page_name': 'Stanley Cups', 'text_as_html': '<table
border="1" class="dataframe">\n  <tbody>\n    <tr>\n      <td>Team</td>\n      <td>Location</td>\n
<td>Stanley Cups</td>\n    </tr>\n    <tr>\n      <td>Blues</td>\n      <td>STL</td>\n      <td>1</td>\n
</tr>\n    <tr>\n      <td>Flyers</td>\n      <td>PHI</td>\n      <td>2</td>\n    </tr>\n    <tr>\n
<td>Maple Leafs</td>\n      <td>TOR</td>\n      <td>13</td>\n    </tr>\n  </tbody>\n</table>', 'category':
'Table', 'source': 'https://drive.google.com/file/d/1aA6L2AR3g0CR-PW03HEZZo4NaVlKpaP7/view'})
```