

Getting started

This tutorial will walk you through the process of installing python-documentcloud and making your first requests.

Installation

Provided that you have [pip](#) installed, you can install the library like so:

```
$ pip install documentcloud
```

Creating a client

Before you can interact with DocumentCloud, you first must import the library and initialize a client to talk with the site on your behalf.

```
>>> from documentcloud import DocumentCloud
>>> client = DocumentCloud()
```

Since we didn't provide any log-in credentials, the client above will only be able to access published documents. If have an account at DocumentCloud and want to use that, you can provide the credentials to the client.

```
>>> client = DocumentCloud(USERNAME, PASSWORD)
```

You can also specify a custom uri if you have installed your own version of DocumentCloud

```
>>> client = DocumentCloud(USERNAME, PASSWORD, base_uri="https://your.documentcloud.domain/api/",
auth_uri="https://your.account.server.domain/api/")
```

Searching for documents

You can now use client to interact with DocumentCloud. A search for documents about [journalist Ruben Salazar](#) would look like this:

```
>>> obj_list = client.documents.search("Ruben Salazar")
>>> # Let's grab the first one and look at it
>>> obj = obj_list[0]
>>> obj
<Document: Final OIR Report>
```

Interacting with a document

Once you have your hands on a document object, you can interact with the metadata stored at documentcloud.org. Here's a sample:

```
>>> print obj.title
Final OIR Report
>>> print obj.id
71072
>>> print obj.contributor_organization
Los Angeles Times
>>> print obj.canonical_url
http://www.documentcloud.org/documents/71072-oir-final-report.html
```

You can even download the PDF, page images and full text.

```
>>> obj.large_image_url
...
>>> obj.large_image
...
>>> obj.full_text
...
>>> obj.pdf
...
```

Uploading a document

You can upload a PDF document from your local machine to documentcloud.org. Here's how:

```
>>> from documentcloud import DocumentCloud
>>> client = DocumentCloud(DOCUMENTCLOUD_USERNAME, DOCUMENTCLOUD_PASSWORD)
>>> obj = client.documents.upload("/home/ben/pdfs/myfile.pdf")
```

And you don't have to provide a path, you can also upload a file object.

```
>>> from documentcloud import DocumentCloud
>>> client = DocumentCloud(DOCUMENTCLOUD_USERNAME, DOCUMENTCLOUD_PASSWORD)
>>> pdf = open("/home/ben/pdfs/myfile.pdf", "rb")
>>> obj = client.documents.upload(pdf)
```

You can also provide URLs that link to PDFs, if that's the kind of thing you're into.

```
>>> client.documents.upload("http://ord.legistar.com/Chicago/attachments/e3a0cbcb-044d-4ec3-9848-23c5692b1943.pdf")
```

Interacting with a newly uploaded public document

When you first upload a document, your local document object will not reflect some of the metadata and processing that happens in the first few seconds it is on the server. Documents set to public will be shown as private during that short interval. To interact with a document as soon as it is available, you can write a short loop to check whether it is ready.

First upload the document as normal.

```
>>> import time
>>> from documentcloud import DocumentCloud
>>> client = DocumentCloud(DOCUMENTCLOUD_USERNAME, DOCUMENTCLOUD_PASSWORD)
>>> obj = client.documents.upload("/home/ben/pdfs/myfile.pdf", access='public')
```

Then refresh your local document object from the server. If it is does not show up as public, then it is still processing, and you'll have to check again.

```
>>> obj = client.documents.get(obj.id)
>>> while obj.status != 'success':
>>>     time.sleep(5)
>>>     obj = client.documents.get(obj.id)
```

Uploading a directory of documents as a project

Here's how to upload a directory full of documents and add them all to a new project. Be warned, this will upload any documents in directories inside the path you specify.

```
>>> # Connect to documentCloud
>>> from documentcloud import DocumentCloud
>>> client = DocumentCloud(DOCUMENTCLOUD_USERNAME, DOCUMENTCLOUD_PASSWORD)
>>> # Create the project
>>> project, created = client.projects.get_or_create_by_title("Groucho Marx's FBI file")
>>> # Upload all the pdfs
>>> obj_list = client.documents.upload_directory('/home/ben/pdfs/groucho_marx/')
>>> # Add the newly created documents to the project
>>> project.document_list = obj_list
>>> # Save the changes to the project
>>> project.put()
```

Uploading a PDF from a URL

How to read a PDF document from a URL on the World Wide Web and upload it to DocumentCloud without saving it to your local hard drive.

```
>>> from documentcloud import DocumentCloud
>>> url = "http://myhost.org/interesting-doc.pdf"
>>> client = DocumentCloud(DOCUMENTCLOUD_USERNAME, DOCUMENTCLOUD_PASSWORD)
>>> # Upload the specified URL to the given client
>>> obj = client.documents.upload(url)
```