File Directory

This covers how to load all documents in a directory.

Under the hood, by default this uses the UnstructuredLoader.

```
from langchain_community.document_loaders
import DirectoryLoader
```

We can use the `glob` parameter to control which files to load. Note that here it doesn't load the `.rst` file or the `.html` files.

```
loader = DirectoryLoader('../',
glob="**/*.md")
```

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

```
len(docs)
```

```
    1
```

Show a progress bar

By default a progress bar will not be shown. To show a progress bar, install the `tqdm` library (e.g. `pip install tqdm`), and set the `show_progress` parameter to `True`.

```
loader = DirectoryLoader('../', glob="**/*.md", show_progress=True)
docs = loader.load()
```

```
    Requirement already satisfied: tqdm in /Users/jon/.pyenv/versions/3.9.16/envs/microbiom
app/lib/python3.9/site-packages (4.65.0)


    0it [00:00, ?it/s]
```

Use multithreading

By default the loading happens in one thread. In order to utilize several threads set the `use_multithreading` flag to true.

```
loader = DirectoryLoader('../',
glob="**/*.md", use_multithreading=True)
docs = loader.load()
```

Change loader class

By default this uses the `UnstructuredLoader` class. However, you can change up the type of loader pretty easily.

```
from langchain_community.document_loaders
import TextLoader
```

```
loader = DirectoryLoader('../',
glob="**/*.md", loader_cls=TextLoader)
```

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

```
len(docs)
```

```
    1
```

If you need to load Python source code files, use the `PythonLoader`.

```
from langchain_community.document_loaders
import PythonLoader
```

```
loader = DirectoryLoader('../../../../../',
glob="**/*.py", loader_cls=PythonLoader)
```

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

```
len(docs)
```

```
    691
```

Auto-detect file encodings with TextLoader

In this example we will see some strategies that can be useful when loading a large list of arbitrary files from a directory using the `TextLoader` class.

First to illustrate the problem, let's try to load multiple texts with arbitrary encodings.

```
path =
'../../../../../tests/integration_tests/example
loader = DirectoryLoader(path, glob="**/*.txt",
loader_cls=TextLoader)
```

A. Default Behavior

```
loader.load()
```

```
<pre style="white-space:pre;overflow-x:auto;lin
Mono',consolas,'Courier New',monospace"><span s
#800000">╭──────────────────────────────── </spa
weight: bold">Traceback </span><span style="col
(most recent call last)</span><span style="colo
─────────────────────────────────╮</span>
<span style="color: #800000; text-decoration-co
color: #bfbf7f">/data/source/langchain/langchai
decoration-color: #808000; font-weight: bold">t
#0000ff">29</span> in <span style="color: #00ff
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f">26 </span><span styl
```

```
</span>text = <span style="color: #808000; text
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f">27 </span><span styl
<span style="color: #0000ff; text-decoration-co
decoration-color: #00ffff">open</span>(<span st
#00ffff">self</span>.file_path, encoding=<span 
#00ffff">self</span>.encoding) <span style="col
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f">28 </span><span styl
</span><span style="color: #0000ff; text-decora
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
color: #800000">❱ </span>29 <span style="color:
</span>text = f.read()
text-decoration-color: #800000">|</span>
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f">30 </span><span styl
</span><span style="color: #0000ff; text-decora
text-decoration-color: #00ffff">UnicodeDecodeEr
#0000ff">as</span> e:
color: #800000">|</span>
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f">31 </span><span styl
</span><span style="color: #0000ff; text-decora
decoration-color: #00ffff">self</span>.autodete
style="color: #800000; text-decoration-color: #
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f">32 </span><span styl
</span>detected_encodings = <span style="color:
```

```
#00ffff">self</span>.detect_file_encodings()
color: #800000">│</span>
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
color: #bfbf7f">/home/spike/.pyenv/versions/3.9
decoration-color: #808000; font-weight: bold">c
#0000ff">322</span> in <span style="color: #00f
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f"> 319 </span><span st
<span style="color: #0000ff; text-decoration-co
decoration-color: #00ff00">decode</span>(<span
<span style="color: #00ffff; text-decoration-co
decoration-color: #0000ff">False</span>):
#800000; text-decoration-color: #800000">│</spa
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f"> 320 </span><span st
<span style="color: #7f7f7f; text-decoration-co
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f"> 321 </span><span st
</span>data = <span style="color: #00ffff; text
#00ffff; text-decoration-color: #00ffff">input<
style="color: #800000; text-decoration-color: #
<span style="color: #800000; text-decoration-co
color: #800000">❱ </span> 322 <span style="colo
consumed) = <span style="color: #00ffff; text-d
style="color: #00ffff; text-decoration-color: #
```

```
style="color: #800000; text-decoration-color: #
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f"> 323 </span><span st
<span style="color: #7f7f7f; text-decoration-co
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f"> 324 </span><span st
<span style="color: #00ffff; text-decoration-co
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f"> 325 </span><span st
<span style="color: #0000ff; text-decoration-co
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
#800000">╰────────────────────────────────────
<span style="color: #ff0000; text-decoration-co
style="color: #008000; text-decoration-color: #
#008080; text-decoration-color: #008080; font-w
text-decoration-color: #008080; font-weight: bo

<span style="font-style: italic">The above exce

<span style="color: #800000; text-decoration-co
style="color: #800000; text-decoration-color: #
#bf7f7f; text-decoration-color: #bf7f7f; font-w
#800000; text-decoration-color: #800000"> ─────
<span style="color: #800000; text-decoration-co
decoration-color: #00ff00">&lt;module&gt;</span
#0000ff">1</span>
style="color: #800000; text-decoration-color: #
<span style="color: #800000; text-decoration-co
```

```
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
color: #800000">❱ </span>1 loader.load()
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f">2 </span>
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
color: #bfbf7f">/data/source/langchain/langchai
decoration-color: #808000; font-weight: bold">d
color: #0000ff">84</span> in <span style="color
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f">81 </span><span styl
│   </span><span style="color: #0000ff; text-de
text-decoration-color: #00ffff">self</span>.sil
style="color: #800000; text-decoration-color: #
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f">82 </span><span styl
│   │   </span>logger.warning(e)
decoration-color: #800000">│</span>
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f">83 </span><span styl
│   </span><span style="color: #0000ff; text-de
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
color: #800000">❱ </span>84 <span style="color:
```

```
</span><span style="color: #0000ff; text-decora
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f">85 </span><span styl
</span><span style="color: #0000ff; text-decora
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f">86 </span><span styl
│ </span><span style="color: #0000ff; text-de
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f">87 </span><span styl
│ │ </span>pbar.update(<span style="color:
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
color: #bfbf7f">/data/source/langchain/langchai
decoration-color: #808000; font-weight: bold">d
color: #0000ff">78</span> in <span style="color
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f">75 </span><span styl
</span><span style="color: #0000ff; text-decora
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f">76 </span><span styl
</span><span style="color: #0000ff; text-decora
style="color: #ff00ff; text-decoration-color: #
```

```
color: #00ffff">self</span>.load_hidden:
#800000">│</span>
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f">77 </span><span styl
</span><span style="color: #0000ff; text-decora
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
color: #800000">❱ </span>78 <span style="color:
</span>sub_docs = <span style="color: #00ffff;
style="color: #00ffff; text-decoration-color: #
decoration-color: #00ffff">self</span>.loader_k
color: #800000">│</span>
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f">79 </span><span styl
│   </span>docs.extend(sub_docs)
decoration-color: #800000">│</span>
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f">80 </span><span styl
</span><span style="color: #0000ff; text-decora
text-decoration-color: #00ffff">Exception</span
#0000ff">as</span> e:
decoration-color: #800000">│</span>
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f">81 </span><span styl
│   </span><span style="color: #0000ff; text-de
text-decoration-color: #00ffff">self</span>.sil
style="color: #800000; text-decoration-color: #
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
color: #bfbf7f">/data/source/langchain/langchai
```

```
decoration-color: #808000; font-weight: bold">t
#0000ff">44</span> in <span style="color: #00ff
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f">41 </span><span styl
│   </span><span style="color: #0000ff; text-de
text-decoration-color: #00ffff">UnicodeDecodeEr
style="color: #800000; text-decoration-color: #
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f">42 </span><span styl
│   │   </span><span style="color: #0000ff; tex
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f">43 </span><span styl
</span><span style="color: #0000ff; text-decora
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
color: #800000">❱ </span>44 <span style="color:
</span><span style="color: #0000ff; text-decora
decoration-color: #00ffff">RuntimeError</span>(
loading {</span><span style="color: #00ffff; te
style="color: #808000; text-decoration-color: #
color: #0000ff">from</span> <span style="color:
underline">e</span>            <span style="col
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f">45 </span><span styl
</span><span style="color: #0000ff; text-decora
text-decoration-color: #00ffff">Exception</span
#0000ff">as</span> e:
```

```
decoration-color: #800000">│</span>
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f">46 </span><span styl
</span><span style="color: #0000ff; text-decora
decoration-color: #00ffff">RuntimeError</span>(
loading {</span><span style="color: #00ffff; te
style="color: #808000; text-decoration-color: #
color: #0000ff">from</span> <span style="color:
underline">e</span>                 <span style=
<span style="color: #800000; text-decoration-co
decoration-color: #7f7f7f">47 </span>
<span style="color: #800000; text-decoration-co
<span style="color: #800000; text-decoration-co
#800000">╰──────────────────────────────────────
<span style="color: #ff0000; text-decoration-co
..<span style="color: #800080; text-decoration-
#800080">/../../../../tests/integration_tests/e
#ff00ff">example-non-utf8.txt</span>
</pre>
```

The file `example-non-utf8.txt` uses a different encoding, so the `load()` function fails with a helpful message indicating which file failed decoding.

With the default behavior of `TextLoader` any failure to load any of the documents will fail the whole loading process and no documents are loaded.

B. Silent fail

We can pass the parameter `silent_errors` to the `DirectoryLoader` to skip the files which could not be loaded and continue the load process.

```
loader = DirectoryLoader(path,
glob="**/*.txt", loader_cls=TextLoader,
silent_errors=True)
docs = loader.load()
```

```
    Error loading
../../../../../tests/integration_tests/examples
non-utf8.txt
```

```
doc_sources = [doc.metadata['source']  for
doc in docs]
doc_sources
```

```

['../../../../../tests/integration_tests/exampl
     '../../../../../tests/integration_tests/ex
utf8.txt']
```

C. Auto detect encodings

We can also ask `TextLoader` to auto detect the file encoding before failing, by passing the `autodetect_encoding` to the loader class.

```
text_loader_kwargs={'autodetect_encoding':
True}
loader = DirectoryLoader(path,
glob="**/*.txt", loader_cls=TextLoader,
loader_kwargs=text_loader_kwargs)
docs = loader.load()
```

```
doc_sources = [doc.metadata['source']  for
doc in docs]
doc_sources
```

```
    ['../../../../../tests/integration_tests/ex
utf8.txt',

'../../../../../tests/integration_tests/example
     '../../../../../tests/integration_tests/ex
utf8.txt']
```