Modules

Data connection

Document loaders

File Directory

File Directory

This covers how to load all documents in a directory.

Under the hood, by default this uses the UnstructuredLoader

```
from langchain.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)
```

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/microbiome-app/lib/python3.9/site-
```

```
packages (4.65.0)

Oit [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.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.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 big list of arbitrary files from a directory using the TextLoader class.

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

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

A. Default Behavior

```
loader.load()
```

```
color: #bfbf7f">/data/source/langchain/langchain/document_loaders/</span><sp</pre>
decoration-color: #808000; font-weight: bold">text.py</span>:<span style="co
#0000ff">29</span> in <span style="color: #00ff00; text-decoration-color: #0
<span style="color: #800000; text-decoration-color: #800000"> </span>
                                                                          <spa
decoration-color: #7f7f7f">26 </span><span style="color: #7f7f7f; text-decor
</span>text = <span style="color: #808000; text-decoration-color: #808000">"
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
                                                                          <spa
decoration-color: #7f7f7f">27 </span><span style="color: #7f7f7f; text-decor
<span style="color: #0000ff; text-decoration-color: #0000ff">with</span> <sp</pre>
decoration-color: #00ffff">open</span>(<span style="color: #00ffff; text-dec
#00ffff">self</span>.file_path, encoding=<span style="color: #00ffff; text-d
#00ffff">self</span>.encoding) <span style="color: #0000ff; text-decoration-
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
decoration-color: #7f7f7f">28 </span><span style="color: #7f7f7f; text-decor
</span><span style="color: #0000ff; text-decoration-color: #0000ff">try</spa</pre>
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span> <span</pre>
color: #800000">> </span>29 <span style="color: #7f7f7f; text-decoration-col
</span>text = f.read()
text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> |</span>
decoration-color: #7f7f7f">30 </span><span style="color: #7f7f7f; text-decor
</span><span style="color: #0000ff; text-decoration-color: #0000ff">except
text-decoration-color: #00ffff">UnicodeDecodeError</span> <span style="color
#0000ff">as</span> e:
                                                                        <span
color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
decoration-color: #7f7f7f">31 </span><span style="color: #7f7f7f; text-decor
</span><span style="color: #0000ff; text-decoration-color: #0000ff">if</span</pre>
decoration-color: #00ffff">self</span>.autodetect_encoding:
style="color: #800000; text-decoration-color: #800000"> | </span>
<span style="color: #800000; text-decoration-color: #800000"> |</span>
decoration-color: #7f7f7f">32 </span><span style="color: #7f7f7f; text-decor
</span>detected_encodings = <span style="color: #00ffff; text-decoration-col")</pre>
#00ffff">self</span>.detect_file_encodings()
                                                                     <span sty</pre>
color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span> <span</pre>
```

```
color: #bfbf7f">/home/spike/.pyenv/versions/3.9.11/lib/python3.9/</span><spa</pre>
decoration-color: #808000; font-weight: bold">codecs.py</span>:<span style="</pre>
#0000ff">322</span> in <span style="color: #00ff00; text-decoration-color: #
<span style="color: #800000; text-decoration-color: #800000"> </span>
                                                                         <spa
decoration-color: #7f7f7f"> 319 </span><span style="color: #7f7f7f; text-dec
<span style="color: #0000ff; text-decoration-color: #0000ff">def</span> <spa</pre>
decoration-color: #00ff00">decode</span>(<span style="color: #00ffff; text-d
<span style="color: #00ffff; text-decoration-color: #00ffff">input</span>, f
decoration-color: #0000ff">False</span>):
#800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
decoration-color: #7f7f7f"> 320 </span><span style="color: #7f7f7f; text-dec
<span style="color: #7f7f7f; text-decoration-color: #7f7f7f"># decode input
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
                                                                         <spa
decoration-color: #7f7f7f"> 321 </span><span style="color: #7f7f7f; text-dec
</span>data = <span style="color: #00ffff; text-decoration-color: #00ffff">s
#00ffff; text-decoration-color: #00ffff">input</span>
style="color: #800000; text-decoration-color: #800000"> | </span>
<span style="color: #800000; text-decoration-color: #800000">|</span> <span</pre>
color: #800000">>> </span> 322 <span style="color: #7f7f7f; text-decoration-d
consumed) = <span style="color: #00ffff; text-decoration-color: #00ffff">sel
style="color: #00ffff; text-decoration-color: #00ffff">self</span>.errors, f
style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
decoration-color: #7f7f7f"> 323 </span><span style="color: #7f7f7f; text-dec
<span style="color: #7f7f7f; text-decoration-color: #7f7f7f"># keep undecode
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
decoration-color: #7f7f7f"> 324 </span><span style="color: #7f7f7f; text-dec
<span style="color: #00ffff; text-decoration-color: #00ffff">self</span>.buf
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
decoration-color: #7f7f7f"> 325 </span><span style="color: #7f7f7f; text-dec
<span style="color: #0000ff; text-decoration-color: #0000ff">return</span> r
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color:</pre>
#800000"> L
<span style="color: #ff0000; text-decoration-color: #ff0000; font-weight: bo</pre>
style="color: #008000; text-decoration-color: #008000">'utf-8'</span> codec
#008080; text-decoration-color: #008080; font-weight: bold">0xca</span> in p
```

```
text-decoration-color: #008080; font-weight: bold">0</span>: invalid continu
<span style="font-style: italic">The above exception was the direct cause of
style="color: #800000; text-decoration-color: #800000; font-weight: bold">Tr
#bf7f7f; text-decoration-color: #bf7f7f; font-weight: bold">(most recent cal
#800000; text-decoration-color: #800000"> ---
<span style="color: #800000; text-decoration-color: #800000"> </span> in <sp</pre>
decoration-color: #00ff00"><module&gt;</span>:<span style="color: #0000ff
#0000ff">1</span>
style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span> <span</pre>
color: #800000">) </span>1 loader.load()
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
                                                                        <spa
decoration-color: #7f7f7f">2 </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> |</span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span> <span</pre>
color: #bfbf7f">/data/source/langchain/langchain/document_loaders/</span><sp</pre>
decoration-color: #808000; font-weight: bold">directory.py</span>:<span styl</pre>
color: #0000ff">84</span> in <span style="color: #00ff00; text-decoration-co
<span style="color: #800000; text-decoration-color: #800000"> </span>
                                                                        <spa
decoration-color: #7f7f7f">81 </span><span style="color: #7f7f7f; text-decor
    </span><span style="color: #0000ff; text-decoration-color: #0000ff">if
text-decoration-color: #00ffff">self</span>.silent_errors:
style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> |</span>
                                                                        <spa
decoration-color: #7f7f7f">82 </span><span style="color: #7f7f7f; text-decor
       </span>logger.warning(e)
decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> |</span>
decoration-color: #7f7f7f">83 </span><span style="color: #7f7f7f; text-decor
    </span><span style="color: #0000ff; text-decoration-color: #0000ff">else
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span> <span</pre>
color: #800000">) </span>84 <span style="color: #7f7f7f; text-decoration-col
```

```
</span><span style="color: #0000ff; text-decoration-color: #0000ff">raise</s</pre>
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
                                                                          <spa
decoration-color: #7f7f7f">85 </span><span style="color: #7f7f7f; text-decor
</span><span style="color: #0000ff; text-decoration-color: #0000ff">finally
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
                                                                          <spa
decoration-color: #7f7f7f">86 </span><span style="color: #7f7f7f; text-decor
    </span><span style="color: #0000ff; text-decoration-color: #0000ff">if
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
                                                                          <spa
decoration-color: #7f7f7f">87 </span><span style="color: #7f7f7f; text-decor
        </span>pbar.update(<span style="color: #0000ff; text-decoration-colo")</pre>
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> |</span>
<span style="color: #800000; text-decoration-color: #800000"> </span> <span</pre>
color: #bfbf7f">/data/source/langchain/langchain/document_loaders/</span><sp</pre>
decoration-color: #808000; font-weight: bold">directory.py</span>:<span styl
color: #0000ff">78</span> in <span style="color: #00ff00; text-decoration-co
<span style="color: #800000; text-decoration-color: #800000"> </span>
                                                                          <spa
decoration-color: #7f7f7f">75 </span><span style="color: #7f7f7f; text-decor
</span><span style="color: #0000ff; text-decoration-color: #0000ff">if</span</pre>
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
decoration-color: #7f7f7f">76 </span><span style="color: #7f7f7f; text-decor
</span><span style="color: #0000ff; text-decoration-color: #0000ff">if</span</pre>
style="color: #ff00ff; text-decoration-color: #ff00ff">or</span> <span style</pre>
color: #00ffff">self</span>.load hidden:
                                                                 <span style="</pre>
#800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> |</span>
decoration-color: #7f7f7f">77 </span><span style="color: #7f7f7f; text-decor
</span><span style="color: #0000ff; text-decoration-color: #0000ff">try</spa</pre>
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span> <span</pre>
color: #800000">> </span>78 <span style="color: #7f7f7f; text-decoration-col
</span>sub_docs = <span style="color: #00ffff; text-decoration-color: #00fff</pre>
style="color: #00ffff; text-decoration-color: #00ffff">str</span>(i), **<spa</pre>
decoration-color: #00ffff">self</span>.loader_kwargs).load()
                                                                   <span style</pre>
color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
                                                                          <spa
```

```
decoration-color: #7f7f7f">79 </span><span style="color: #7f7f7f; text-decor
      </span>docs.extend(sub docs)
decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
                                                                                                                         <spa
decoration-color: #7f7f7f">80 </span><span style="color: #7f7f7f; text-decor
</span><span style="color: #0000ff; text-decoration-color: #0000ff">except
text-decoration-color: #00ffff">Exception</span> <span style="color: #0000ff"
#0000ff">as</span> e:
                                                                                                                       <span
decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
                                                                                                                         <spa
decoration-color: #7f7f7f">81 </span><span style="color: #7f7f7f; text-decor
      </span><span style="color: #0000ff; text-decoration-color: #0000ff">if
text-decoration-color: #00ffff">self</span>.silent_errors:
style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000">|</span>
<span style="color: #800000; text-decoration-color: #800000"> |</span>
<span style="color: #800000; text-decoration-color: #800000"> </span> <span</pre>
color: #bfbf7f">/data/source/langchain/langchain/document_loaders/</span><sp</pre>
decoration-color: #808000; font-weight: bold">text.py</span>:<span style="color: #808000; font-weight: bold">text.py</span style="color: #808000; font-weight: bold style="color: #808000; font-weight: bold style="color: #808000; font-
#0000ff">44</span> in <span style="color: #00ff00; text-decoration-color: #0
<span style="color: #800000; text-decoration-color: #800000"> |</span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
                                                                                                                         <spa
decoration-color: #7f7f7f">41 </span><span style="color: #7f7f7f; text-decor
      </span><span style="color: #0000ff; text-decoration-color: #0000ff">exce
text-decoration-color: #00ffff">UnicodeDecodeError</span>:
style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> |</span>
                                                                                                                         <spa
decoration-color: #7f7f7f">42 </span><span style="color: #7f7f7f; text-decor
             </span><span style="color: #0000ff; text-decoration-color: #0000ff">
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
                                                                                                                         <spa
decoration-color: #7f7f7f">43 </span><span style="color: #7f7f7f; text-decor
</span><span style="color: #0000ff; text-decoration-color: #0000ff">else</sp
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> | </span> <span</pre>
color: #800000">> </span>44 <span style="color: #7f7f7f; text-decoration-col
</span><span style="color: #0000ff; text-decoration-color: #0000ff">raise</s
decoration-color: #00ffff">RuntimeError</span>(<span style="color: #808000;</pre>
loading {</span><span style="color: #00ffff; text-decoration-color: #00ffff"</pre>
style="color: #808000; text-decoration-color: #808000">}"</span>) <span styl</pre>
color: #0000ff">from</span> <span style="color: #00ffff; text-decoration-col"</pre>
underline">e</span>
                                                    <span style="color: #800000; text-decoration-</pre>
```

```
<span style="color: #800000; text-decoration-color: #800000"> |</span>
decoration-color: #7f7f7f">45 </span><span style="color: #7f7f7f; text-decor
</span><span style="color: #0000ff; text-decoration-color: #0000ff">except
text-decoration-color: #00ffff">Exception</span> <span style="color: #0000ff"
#0000ff">as</span> e:
decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
decoration-color: #7f7f7f">46 </span><span style="color: #7f7f7f; text-decor
</span><span style="color: #0000ff; text-decoration-color: #0000ff">raise</s</pre>
decoration-color: #00ffff">RuntimeError</span>(<span style="color: #808000;</pre>
loading {</span><span style="color: #00ffff; text-decoration-color: #00ffff"</pre>
style="color: #808000; text-decoration-color: #808000">}"</span>) <span styl</pre>
color: #0000ff">from</span> <span style="color: #00ffff; text-decoration-col
                                    <span style="color: #800000; text-decorat"</pre>
underline">e</span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
decoration-color: #7f7f7f">47 </span>
<span style="color: #800000; text-decoration-color: #800000"> </span>
<span style="color: #800000; text-decoration-color:</pre>
#800000"> ---
<span style="color: #ff0000; text-decoration-color: #ff0000; font-weight: bo</pre>
..<span style="color: #800080; text-decoration-color:</pre>
#800080">/../../tests/integration_tests/examples/</span><span style="c
#ff00ff">example-non-utf8.txt</span>
```

The file example-non-utf8.txt uses a different encoding 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/example-
non-utf8.txt

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

```
['../../../tests/integration_tests/examples/whatsapp_chat.txt',
   '../../../tests/integration_tests/examples/example-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/examples/example-non-
utf8.txt',
    '../../../tests/integration_tests/examples/whatsapp_chat.txt',
    '../../../tests/integration_tests/examples/example-utf8.txt']
```