

# test

|  |   |
|--|---|
|  | <!DOCTYPE html>   |
|  | <html lang="en">  |
|  | <head>  |
|  |   |
|  | <title>LangChain &lt;&gt;<br>Unstructured</title>   |
|  | <meta charset="utf-8" />  |
|  | <meta http-equiv="X-UA-Compatible"<br>content="IE=edge" />                                      |
|  | <meta name="HandheldFriendly"<br>content="True" />  |
|  | <meta name="viewport"<br>content="width=device-width, initial-<br>scale=1.0" />                 |
|  |   |
|  | <link rel="stylesheet" type="text/css"<br>href=" /assets/built/screen.css?<br>v=2df407f2e9 " /> |
|  |   |

|  |   |
|--|---|
|  | <pre>&lt;link rel="icon" href="https://blog.langchain.dev/content /images/size/w256h256/2023/01/parro ticon-1.png" type="image/png" /&gt;</pre>   |
|  | <pre>&lt;link rel="canonical" href="https://blog.langchain.dev/langch ain-unstructured/" /&gt;</pre>  |
|  | <pre>&lt;meta name="referrer" content="no- referrer-when-downgrade" /&gt;</pre>   |
|  |   |
|  | <pre>&lt;meta property="og:site_name" content="LangChain" /&gt;</pre>   |
|  | <pre>&lt;meta property="og:type" content="article" /&gt;</pre>  |
|  | <pre>&lt;meta property="og:title" content="LangChain &amp;lt;&amp;gt; Unstructured" /&gt;</pre>   |
|  | <pre>&lt;meta property="og:description" content="One of the core value props of LangChain is the ability to combine Large Language Models with your own text data. There are multiple (four!) different methods of doing so, and many different applications this can power."</pre> |
|  |   |
|  | <pre>A step that sits upstream of using text data is the ability to" /&gt;</pre>  |

|  |  |
|--|--|
|  | <pre>&lt;meta property="og:url" content="https://blog.langchain.dev/langchain-unstructured/" /&gt;</pre>   |
|  | <pre>&lt;meta property="og:image" content="https://blog.langchain.dev/content/images/2023/02/Screen-Shot-2023-02-05-at-11.32.11-PM.png" /&gt;</pre>  |
|  | <pre>&lt;meta property="article:published_time" content="2023-02-06T07:32:46.000Z" /&gt;</pre>   |
|  | <pre>&lt;meta property="article:modified_time" content="2023-06-24T22:48:36.000Z" /&gt;</pre>  |
|  | <pre>&lt;meta name="twitter:card" content="summary_large_image" /&gt;</pre>  |
|  | <pre>&lt;meta name="twitter:title" content="LangChain &amp;lt;&amp;gt; Unstructured" /&gt;</pre>   |
|  | <pre>&lt;meta name="twitter:description" content="One of the core value props of LangChain is the ability to combine Large Language Models with your own text data. There are multiple (four!) different methods of doing so, and many different applications this can power."</pre> |
|  |  |

|  |  |
|--|--|
|  | A step that sits upstream of using text data is the ability to" />   |
|  | <meta name="twitter:url"<br>content=" <a href="https://blog.langchain.dev/langchain-unstructured/">https://blog.langchain.dev/langchain-unstructured/</a> " />   |
|  | <meta name="twitter:image"<br>content=" <a href="https://blog.langchain.dev/content/images/2023/02/Screen-Shot-2023-02-05-at-11.32.11-PM.png">https://blog.langchain.dev/content/images/2023/02/Screen-Shot-2023-02-05-at-11.32.11-PM.png</a> " /> |
|  | <meta name="twitter:label1"<br>content="Written by" />   |
|  | <meta name="twitter:data1"<br>content="LangChain" />   |
|  | <meta name="twitter:site"<br>content="@LangChainAI" />   |
|  | <meta name="twitter:creator"<br>content="@LangChainAI" />  |
|  | <meta property="og:image:width"<br>content="1335" />   |
|  | <meta property="og:image:height"<br>content="710" />   |
|  |  |
|  | <script type="application/ld+json">  |
|  | {  |
|  | "@context": " <a href="https://schema.org">https://schema.org</a> ",   |

|  |  |
|--|--|
|  | "@type": "Article",  |
|  | "publisher": {   |
|  | "@type": "Organization",   |
|  | "name": "LangChain",   |
|  | "url": " <a href="https://blog.langchain.dev/">https://blog.langchain.dev/</a> ",  |
|  | "logo": {  |
|  | "@type": "ImageObject",  |
|  | "url":<br>" <a href="https://blog.langchain.dev/content/images/size/w256h256/2023/01/parroticon-1.png">https://blog.langchain.dev/content/images/size/w256h256/2023/01/parroticon-1.png</a> ", |
|  | "width": 60,   |
|  | "height": 60   |
|  | }  |
|  | },   |
|  | "author": {  |
|  | "@type": "Person",   |
|  | "name": "LangChain",   |
|  | "image": {   |
|  | "@type": "ImageObject",  |
|  | "url":<br>" <a href="https://blog.langchain.dev/content/images/2023/01/parroticon.png">https://blog.langchain.dev/content/images/2023/01/parroticon.png</a> ",                                 |

|  |  |
|--|--|
|  | "width": 448,  |
|  | "height": 448  |
|  | },   |
|  | "url":<br>" <a href="https://blog.langchain.dev/author/langchain/">https://blog.langchain.dev/author/langchain/</a> ",             |
|  | "sameAs": [  |
|  | " <a href="https://twitter.com/LangChainAI">https://twitter.com/LangChainAI</a> "  |
|  | ]  |
|  | },   |
|  | "headline": "LangChain &lt;&gt;<br>Unstructured",  |
|  | "url":<br>" <a href="https://blog.langchain.dev/langchain-unstructured/">https://blog.langchain.dev/langchain-unstructured/</a> ", |
|  | "datePublished": "2023-02-<br>06T07:32:46.000Z",   |
|  | "dateModified": "2023-06-<br>24T22:48:36.000Z",  |
|  | "image": {   |
|  | "@type": "ImageObject",  |

|  |  |
|--|--|
|  | "url":<br>" <a href="https://blog.langchain.dev/content/images/2023/02/Screen-Shot-2023-02-05-at-11.32.11-PM.png">https://blog.langchain.dev/content/images/2023/02/Screen-Shot-2023-02-05-at-11.32.11-PM.png</a> ",   |
|  | "width": 1335,   |
|  | "height": 710  |
|  | },   |
|  | "description": "One of the core value props of LangChain is the ability to combine Large Language Models with your own text data. There are multiple (four!) different methods of doing so, and many different applications this can power.\n\nA step that sits upstream of using text data is the ability to get your data into a text form. This can be rather tricky due to the multitude of different formats that exist out there.\n\nEnter... unstructured.io.\n\nUnstructured is a company with a mission of transforming natural l", |
|  | "mainEntityOfPage": {  |
|  | "@type": "WebPage",  |
|  | "@id": " <a href="https://blog.langchain.dev/">https://blog.langchain.dev/</a> "   |
|  | }  |
|  | }  |
|  | </script>  |

|  |   |
|--|---|
|  |   |
|  | <pre>&lt;meta name="generator" content="Ghost 5.30" /&gt;</pre>   |
|  | <pre>&lt;link rel="alternate" type="application/rss+xml" title="LangChain" href="https://blog.langchain.dev/rss/" /&gt;</pre>   |
|  |   |
|  | <pre>&lt;script defer src="https://cdn.jsdelivr.net/ghost/sodo- search@~1.1/umd/sodo-search.min.js" data- key="eaa48df0e56a1183c31a3a39e8" data- styles="https://cdn.jsdelivr.net/ghost/so do-search@~1.1/umd/main.css" data- sodo- search="https://blog.langchain.dev/" crossorigin="anonymous"&gt;&lt;/script&gt;</pre> |
|  | <pre>&lt;link href="https://blog.langchain.dev/webme ntion" rel="webmention" /&gt;</pre>  |
|  | <pre>&lt;script defer src="/public/cards.min.js? v=2df407f2e9"&gt;&lt;/script&gt;&lt;style&gt;:root { --ghost-accent-color: #12133a;}&lt;/style&gt;</pre>   |
|  | <pre>&lt;link rel="stylesheet" type="text/css" href="/public/cards.min.css? v=2df407f2e9"&gt;</pre>   |



|  |   |
|--|---|
|  |   |
|  | </head>   |
|  | <body class="post-template is-head-left-logo has-cover">  |
|  | <div class="viewport">  |
|  |   |
|  | <header id="gh-head" class="gh-head outer">   |
|  | <div class="gh-head-inner inner">   |
|  | <div class="gh-head-brand">   |
|  | <a class="gh-head-logo no-image" href="https://blog.langchain.dev">   |
|  | LangChain   |
|  | </a>  |
|  | <button class="gh-search gh-icon-btn" data-ghost-search><svg xmlns="http://www.w3.org/2000/svg" fill="none" viewBox="0 0 24 24" stroke="currentColor" stroke-width="2" width="20" height="20"><path stroke-linecap="round" stroke-linejoin="round" d="M21 21l-6-6m2-5a7 7 0 11-14 0 7 7 0 0114 0z"></path></svg></button> |
|  | <button class="gh-burger"></button>   |
|  | </div>  |

|  |  |
|--|--|
|  |  |
|  | <nav class="gh-head-menu">   |
|  | <ul class="nav">   |
|  | <li class="nav-home"><a href=" <a href="https://blog.langchain.dev/">https://blog.langchain.dev/</a> ">Home</a></li>                                 |
|  | <li class="nav-about"><a href=" <a href="https://blog.langchain.dev/about/">https://blog.langchain.dev/about/</a> ">About</a></li>                   |
|  | <li class="nav-github"><a href=" <a href="https://github.com/hwchase17/langchain">https://github.com/hwchase17/langchain</a> ">GitHub</a></li>       |
|  | <li class="nav-docs"><a href=" <a href="https://langchain.readthedocs.io/en/latest/">https://langchain.readthedocs.io/en/latest/</a> ">Docs</a></li> |
|  | </ul>  |
|  |  |
|  | </nav>   |
|  |  |
|  | <div class="gh-head-actions">  |

|  |   |
|--|---|
|  | <pre> &lt;button class="gh-search gh-icon-btn" data-ghost-search&gt;&lt;svg xmlns=" http://www.w3.org/2000/svg " fill="none" viewBox="0 0 24 24" stroke="currentColor" stroke-width="2" width="20" height="20"&gt;&lt;path stroke- linecap="round" stroke-linejoin="round" d="M21 21l-6-6m2-5a7 7 0 11-14 0 7 7 0 0114 0z"&gt;&lt;/path&gt;&lt;/svg&gt;&lt;/button&gt; </pre> |
|  | </div>  |
|  | </div>  |
|  | </header>   |
|  |   |
|  | <div class="site-content">  |
|  |   |
|  |   |
|  |   |
|  |   |
|  | <main id="site-main" class="site-main">   |
|  | <article class="article post ">   |
|  |   |
|  | <header class="article-header gh- canvas">  |
|  |   |

|  |   |
|--|---|
|  | <div class="article-tag post-card-tags">  |
|  | </div>  |
|  |   |
|  | <h1 class="article-title">LangChain<br>&lt;&gt; Unstructured</h1>   |
|  |   |
|  |   |
|  | <div class="article-byline">  |
|  | <section class="article-byline-content">  |
|  |   |
|  | <ul class="author-list">  |
|  | <li class="author-list-item">   |
|  | <a href=" /author/langchain/ " class="author-avatar">   |
|  |  |
|  | </a>  |
|  | </li>   |
|  | </ul>   |
|  |   |
|  | <div class="article-byline-meta">   |

|  |  |
|--|--|
|  | <h4 class="author-name"><a href=" /author/langchain/">LangChain</a></h4>                   |
|  | <div class="byline-meta-content">  |
|  | <time class="byline-meta-date" datetime="2023-02-05">Feb 5, 2023</time>                    |
|  | <span class="byline-reading-time"><br><span class="bull">&bull;</span> 1 min read</span>   |
|  | </div>   |
|  | </div>   |
|  |  |
|  | </section>   |
|  | </div>   |
|  |  |
|  | <figure class="article-image">   |
|  | /content/images/size/w2000/2023/02/Screen-Shot-2023-02-05-at-11.32.11-PM.png</a> "   |
|  | alt="LangChain &lt;&gt; Unstructured"   |
|  | />  |
|  | </figure>   |
|  |   |
|  | </header>   |
|  |   |
|  | <section class="gh-content gh-canvas">  |
|  | <p>&lt;p&gt;One of the core value props of LangChain is the ability to combine Large Language Models with your own text data. There are multiple (<a href="https://python.langchain.com/docs/modules/chains/document/">https://python.langchain.com/docs/modules/chains/document/</a>&gt;four! &lt;/a&gt;) different methods of doing so, and &lt;a</p> |

href=" [https://langchain.readthedocs.io/en/latest/use\\_cases/question\\_answering.html](https://langchain.readthedocs.io/en/latest/use_cases/question_answering.html) ">many</a> <a href=" [https://python.langchain.com/docs/use\\_cases/question\\_answering/](https://python.langchain.com/docs/use_cases/question_answering/) ">different</a> applications this can power.

</p><p>A step that sits upstream of using text data is the ability to get your data into a text form. This can be rather tricky due to the multitude of different formats that exist out there. </p>

<p>Enter... <a href=" <https://www.unstructured.io/> ">unstructured.io</a>.</p><p>Unstructured is a company with a mission of transforming natural language data from raw to machine ready. One of the main ways they do this is with an <a href=" <https://github.com/Unstructured-IO/unstructured> ">open source Python package</a>. This package as support for <a href=" <https://github.com/Unstructured-IO/unstructured#document-parsing> ">MANY</a> different types of file extensions: <code>.txt</code>,<br><code>.docx</code>,<br><code>.pptx</code>,<br><code>.jpg</code>,<br><code>.png</code>,<br><code>.eml</code>,<br><code>.html</code>, and<br><code>.pdf</code> documents.</p>

<p>After playing around with

Unstructured, we realized that by integrating with it we could easily start to build out first class support for loading documents of all types into a format that LangChains could work with. So we created the [https://python.langchain.com/docs/modules/data\\_connection/document\\_loaders/](https://python.langchain.com/docs/modules/data_connection/document_loaders/) Document Loaders module, a large part of which is powered by Unstructured.

There are currently two loaders that are powered by Unstructured. Both seem rather simple, but are quite powerful.

The first is the [https://python.langchain.com/docs/modules/data\\_connection/document\\_loaders/integrations/unstructured\\_file](https://python.langchain.com/docs/modules/data_connection/document_loaders/integrations/unstructured_file) UnstructuredFileLoader. This has a simple interface (you just pass it a file path) but under the hood Unstructured is doing a lot of smart logic to infer which data type it is (PDF, PowerPoint, image, etc) and extract text.

The second is the [https://python.langchain.com/docs/modules/data\\_connection/document\\_loaders/how\\_to/file\\_directory](https://python.langchain.com/docs/modules/data_connection/document_loaders/how_to/file_directory) Directory Loader. Again, this has a pretty simple interface: it takes only a path to a directory and an optional regex to glob for files against. But under the hood it is looping over all files and using the above UnstructuredFileLoader to



|  |   |
|--|---|
|  | <p>load them. This makes it possible to load files of all types in a single call.&lt;/p&gt;&lt;p&gt;We're incredibly excited to have made this integration with Unstructured. With their focus on transforming raw data into clean text, it makes it incredibly easy to combine language models with your data, no matter what form it is in.&lt;/p&gt;&lt;/section&gt;</p> |
|  |   |
|  |   |
|  | </article>  |
|  | </main>   |
|  |   |
|  |   |
|  |   |
|  |   |
|  | <aside class="read-more-wrap outer">  |
|  | <div class="read-more inner">   |
|  |   |
|  | <article class="post-card post">  |
|  |   |

|  |   |
|--|---|
|  | <a class="post-card-image-link" href=" /prem-challenge-with-langchain/ "> |
|  |   |
|  |   |
|  |   |
|  |   |
|  | </a>  |
|  |   |

|  |  |
|--|--|
|  | <div class="post-card-content">  |
|  |  |
|  | <a class="post-card-content-link"<br>href="/premise-challenge-with-langchain/">  |
|  | <header class="post-card-header">  |
|  | <div class="post-card-tags">   |
|  | </div>   |
|  | <h2 class="post-card-title">   |
|  | 🎉 Prem Challenge 🎉   |
|  | </h2>  |
|  | </header>  |
|  | <div class="post-card-excerpt">We're<br>excited to announce a challenge hosted<br>by Prem in collaboration with LangChain.   |
|  |  |
|  | At LangChain we try to make it easy as<br>possible to experiment with as many<br>different models as possible. That<br>includes the incredible number of<br>emerging open source models. We've<br>made efforts to make our framework<br>as</div> |
|  | </a>   |
|  |  |

|  |   |
|--|---|
|  | <footer class="post-card-meta">   |
|  | <time class="post-card-meta-date"<br>datetime="2023-06-26">Jun 26,<br>2023</time> |
|  | <span class="post-card-meta-<br>length">6 min read</span>                         |
|  | </footer>   |
|  |   |
|  | </div>  |
|  |   |
|  | </article>  |
|  |   |
|  | <article class="post-card post">  |
|  |   |
|  | <a class="post-card-image-link"<br>href=" <a href="#">/langchain/</a> ">          |
|  |   |
|  | <img class="post-card-image"  |
|  | srcset=" <a href="#">/content/images/size/w300/2023/06/mongo.webp 300w</a> ,      |
|  | <a href="#">/content/images/size/w600/2023/06/mongo.webp 600w</a> ,               |
|  | <a href="#">/content/images/size/w1000/2023/06/mongo.webp 1000w</a> ,             |

|  |   |
|--|---|
|  | /content/images/size/w2000/2023/06/mongo.webp 2000w"  |
|  | sizes="(max-width: 1000px) 400px, 800px"  |
|  | src=" <a href="/content/images/size/w600/2023/06/mongo.webp">/content/images/size/w600/2023/06/mongo.webp</a> " |
|  | alt="LangChain &lt;&gt; MongoDB Atlas"  |
|  | loading="lazy"  |
|  | />  |
|  |   |
|  |   |
|  | </a>  |
|  |   |
|  | <div class="post-card-content">   |
|  |   |
|  | <a class="post-card-content-link" href=" <a href="/langchain/">/langchain/</a> ">                               |
|  | <header class="post-card-header">   |
|  | <div class="post-card-tags">  |
|  | </div>  |
|  | <h2 class="post-card-title">  |
|  | LangChain &lt;&gt; MongoDB Atlas  |

|  |  |
|--|--|
|  | </h2>  |
|  | </header>  |
|  | <div class="post-card-excerpt">Today we're announcing LangChain's integration with MongoDB Atlas, adding support for one of the most popular developer data platforms in the world. This is an integration so anticipated that a few developers added the integration before we were ready to announce it :) |
|  |  |
|  |  |
|  | Overview   |
|  |  |
|  | One of the key components</div>  |
|  | </a>   |
|  |  |
|  | <footer class="post-card-meta">  |
|  | <time class="post-card-meta-date" datetime="2023-06-22">Jun 22, 2023</time>  |
|  | <span class="post-card-meta-length">3 min read</span>  |
|  | </footer>  |
|  |  |

|  |   |
|--|---|
|  | </div>  |
|  |   |
|  | </article>  |
|  |   |
|  | <article class="post-card post">  |
|  |   |
|  | <a class="post-card-image-link"<br>href=" <a href="#">/data-driven-characters/</a> "> |
|  |   |
|  | <img class="post-card-image"  |
|  | srcset=" <a href="#">/content/images/size/w300/2023/06/teaser_chatbot.jpg_300w</a> ,  |
|  | <a href="#">/content/images/size/w600/2023/06/teaser_chatbot.jpg_600w</a> ,           |
|  | <a href="#">/content/images/size/w1000/2023/06/teaser_chatbot.jpg_1000w</a> ,         |
|  | <a href="#">/content/images/size/w2000/2023/06/teaser_chatbot.jpg_2000w</a> "         |
|  | sizes="(max-width: 1000px) 400px,<br>800px"   |
|  | src=" <a href="#">/content/images/size/w600/2023/06/teaser_chatbot.jpg</a> "          |
|  | alt="Data-Driven Characters"  |

|  |  |
|--|--|
|  | loading="lazy"   |
|  | />   |
|  |  |
|  |  |
|  | </a>   |
|  |  |
|  | <div class="post-card-content">  |
|  |  |
|  | <a class="post-card-content-link"<br>href=" /data-driven-characters/ ">  |
|  | <header class="post-card-header">  |
|  | <div class="post-card-tags">   |
|  | </div>   |
|  | <h2 class="post-card-title">   |
|  | Data-Driven Characters   |
|  | </h2>  |
|  | </header>  |
|  | <div class="post-card-excerpt">Data-<br>driven-characters is a repo for creating,<br>debugging, and interacting your own<br>chatbots conditioned on your own story<br>corpora.</div> |
|  | </a>   |



|  |   |
|--|---|
|  |   |
|  | <footer class="post-card-meta">   |
|  | <time class="post-card-meta-date"<br>datetime="2023-06-19">Jun 19,<br>2023</time> |
|  | <span class="post-card-meta-<br>length">11 min read</span>                        |
|  | </footer>   |
|  |   |
|  | </div>  |
|  |   |
|  | </article>  |
|  | </div>  |
|  | </aside>  |
|  |   |
|  |   |
|  |   |
|  | </div>  |
|  |   |
|  | <footer class="site-footer outer">  |
|  | <div class="inner">   |

|  |   |
|--|---|
|  | <code>&lt;section class="copyright"&gt;&lt;a href="https://blog.langchain.dev"&gt;Lang Chain&lt;/a&gt; &amp;copy; 2023&lt;/section&gt;</code> |
|  | <code>&lt;nav class="site-footer-nav"&gt;</code>  |
|  | <code>&lt;ul class="nav"&gt;</code>   |
|  | <code>&lt;li class="nav-sign-up"&gt;&lt;a href="#/portal/"&gt;Sign up&lt;/a&gt;&lt;/li&gt;</code>   |
|  | <code>&lt;/ul&gt;</code>  |
|  |   |
|  | <code>&lt;/nav&gt;</code>   |
|  | <code>&lt;div&gt;&lt;a href="https://ghost.org/" target="_blank" rel="noopener"&gt;Powered by Ghost&lt;/a&gt;&lt;/div&gt;</code>              |
|  | <code>&lt;/div&gt;</code>   |
|  | <code>&lt;/footer&gt;</code>  |
|  |   |
|  | <code>&lt;/div&gt;</code>   |
|  |   |
|  | <code>&lt;script</code>   |
|  | <code>src="https://code.jquery.com/jquery-3.5.1.min.js"</code>  |

|  |   |
|--|---|
|  | integrity="sha256-9/aliU8dGd2tb6OSsuzixeV4y/faTqgFtohetphbbj0=" |
|  | crossorigin="anonymous">  |
|  | </script>   |
|  | <script src="/assets/built/casper.js?v=2df407f2e9"></script>    |
|  | <script>  |
|  | \$(document).ready(function () {                                |
|  | // Mobile Menu Trigger  |
|  | \$('.gh-burger').click(function () {                            |
|  | \$('body').toggleClass('gh-head-open');                         |
|  | });   |
|  | // FitVids – Makes video embeds responsive                      |
|  | \$(".gh-content").fitVids();                                    |
|  | });   |
|  | </script>   |
|  |   |
|  |   |
|  |   |
|  | </body>   |

|  |         |
|--|---------|
|  | </html> |
|  |         |