

Web Scrapping with OpenAI

Eli the Computer Guy - August 8, 2024

Is This Legal???

- ???
- SQUIRREL!!!
- Tech Professionals answer the questions their clients do not know to ask.
- Tech Professionals are responsible for laws, regulations, and licensing...
- I'm just a guy in a basement with a web cam....

OpenAI

- OpenAI is not ChatGPT!!!
- Different Account!!!
- API Key Should be an Environment Variable
 - <<— Needs to document Environment Variables
 - You can use script defined variable
- Example = basic-request.py

Requests Module and Get Function

- `Python3 -m pip install requests`
- Requests Module allows Python to Access Web Pages
- Example = `get-request.py`

Beautiful Soup 4

- `python3 -m pip install beautifulsoup4`
- Beautiful Soup Parses HTML
- `soup = BeautifulSoup(result, 'html.parser')`
 - Provide Text and Parser
 - There are different parsers
- <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>
- Example = `soup.py`

Beautiful Soup to OpenAI

- Requests -> Beautiful Soup -> Open AI
- Example = bs-openai.py

Feedparser

- `python3 -m pip install feedparser`
- <https://feedparser.readthedocs.io/en/latest/>
- <https://arstechnica.com/rss-feeds/>
- <https://feeds.arstechnica.com/arstechnica/index>
- Note: May need to use `requests.get()` to access feed
- Example = `feedparser-basic.py`

Feedparser and OpenAI

- Example = `feedparser-openai.py`

YouTube Transcription

- `python3 -m pip install youtube-transcript-api`
- Don't reinvent wheels...
- YouTube automatically creates text transcripts of videos
- Example = `yt-transcript.py`

Transcript and OpenAI

- Example = yt-openai.py

Labs

- lab-auto-post.py
 - We write a single blog post and create a featured image
- lab-auto-blog.py
 - Write numerous posts by scrapping an RSS feed
- lab-news.py
 - Ask questions about todays news from RSS feed