Landis

Off-Site Coding Project 2018

Background

This project was created by our lead investor, Linus Liang, of Signia VC.

The purpose of the project is to allow a user to "scrape" DocSend (https://www.docsend.com/). DocSend is a simple platform that hosts documents (mostly PowerPoint decks). An example can be found here: https://docsend.com/view/2tgquda

As you can see from the link, there is no way to download the deck or even print it out. In addition, oftentimes these decks require an email or password before we can even view it. Examples of the same deck with those constraints can be found here:

- https://docsend.com/view/gai9mp3
- https://docsend.com/view/2ku54z9 (pw: testing123)

Linus often gets a lot of these links but his partners want a PDF that they can easily print it out. In order to generate a PDF, Linus wrote a scraper which can be found here:

https://github.com/linusliang/docsend_scraper

All the code is in application.py and uses Flask as the webserver with Bootstrap as the frontend. The code is meant for deployment on EC2 / Elastic Beanstalk but you shouldn't feel obligated to have to make it work on AWS.

With that context, here is the project with a few remaining features Linus would want built.

Goal: Create a better version of this DocSend Scraper.

There are 4 tasks left to do:

1. There are no error checks on the "Link" text field. Users can leave it empty or they can put in garbage text -- all of which crashes the script. The field should only accept validated Docsend links.
2. In addition, the user also needs to include https:// in the Link field. For example, the script doesn't work on "docsend.com/view/p8jxsqr", it only works for "https://docsend.com/view/p8jxsqr"
3. Currently while it's scraping, the "Scrape" button is disabled. However when it's done scraping, the button doesn't reset. One way to solve this problem is through threads + web sockets.
4. It doesn't currently give the user any feedback on the status of the scrape. The user does not know if the scrape has failed or if it is still processing. A simple status indicator or a UI that shows the user which slide the script is currently scraping needs to be added.

As for how to architect things, we're leaving that up to you. Feel free to use any server, front end package, 3rd party packages. etc. We are more interested in understanding how you eventually architect the system and the trade offs you considered.