**Scraping without bulk data**

Luke can handle scrapers when bulk data is available. Others should only select states where bulk data is not freely available.

The scraper for each SoS website performs 4 steps for each respective record from the state in step\_3\_work\output\full\_retailer\_list.csv:

1. Crawling: Enter the name of the DBA into SoS search engine, and save all records that are returned by the SoS query as an excel file.
2. Cleaning: Clean up the crawling output so it can be run through the matching function
3. Matching: Determine which, if any, returned records are in fact matched to the DBA. My built-in matching function generates scores based on name.
4. Document Retrieval: For the SoS records matched to FDA records, save a screenshot of the website record, and download any available documents associated with the record.

**California Example**

As an example, walk through the California scraper’s files, read the comments, and follow along with what it does:

[\\afsv03\Research\FDA\2732 - FDA Tribal Tobacco Retailers\Technical\Tribal\_Master\state\_scrapers\CA](file:///\\afsv03\Research\FDA\2732%20-%20FDA%20Tribal%20Tobacco%20Retailers\Technical\Tribal_Master\state_scrapers\CA)