### Objective

* Determine the frequency and context of keywords related to Section 508 digital accessibility compliance in public-facing HUD handbooks.

### Tools and Setup

* MacBook Air using GeckoDriver for Firefox.
* Scraping code pulled HREF data at specified intervals and appended it to a base url.
* Successfully ran headless in batches of 10 with no API issues.

### Methodology

#### **Scraping the “First Click”**

* Scraped the first page of 5 handbook types: Admin, CFO, OCHCO, CIO, CPO.
* Handled unique cases and empty links.
* Ensured quality by scraping each type one at a time.
* Compared expected vs. actual scraped policies.

#### **Scraping the “Second Click”**

*PDFs:*

* Subset direct links to PDFs, about ⅓ of the “first click” falls within this category.
* Checked URL validity and searchability. Column flags:
  + url\_test: a binary flag that indicates whether a URL was successfully accessed and its content (a PDF file, in this context) was retrieved.
  + searchable: a binary flag that indicates whether the text extracted from a PDF contains enough recognizable text to be considered "searchable". This typically means that the PDF contains selectable text, not just images of text, and that it includes a sufficient number of common words to indicate meaningful content.
* Keyword occurrences counted for searchable PDFs.

*Landing Pages:*

* Remaining ⅔ of links led to intermediary pages.
* Added columns: pdf\_name and pdf\_url, the latter of which refers to the document’s direct link rather than url, which is the intermediary page for instances where straight\_to\_pdf == False
  + **If pdf\_url is blank**, it means a link was not associated with that policy or information. For much of the statistical analysis, these rows are simply filtered out. There are not many of them.
* Handled broken links prefixed with "hudatwork,”of which there are only 3.

*Verification:*

* Manually verified the number of expected PDFs per landing page.
* Compared expected and actual counts for QA.

#### *Handling Exceptions*

* Addressed unique formatting and broken links.
* Rescraped problematic links individually.

#### **Stacking and Cleaning Data**

* Combined PDF and landing page data.
* Removed duplicates and handled NaN values.
* Final data frame: full\_stacked\_scrape. Contains additional columns such as:
  + The count of each keyword in every document scraped [508, disab, accessib, assistive tech, rehab, ICT]
  + straight\_to\_pdf: A binary flag that indicates whether or not the link went straight to a pdf at the first click.
  + destination\_document\_type: a field that indicates the type of document in pdf\_url

#### **Keyword Analysis**

* Subset of data with keyword pings, where at least one keyword had a number of pings > 0: raw\_keyword\_pings.
* Extracted sentences containing keywords and compared expected vs. actual keyword pings. These are now additional columns in the data frame below.
  + Mismatches here are likely due to variations in HTML formatting or instances where keywords occur within the same sentence.
* Final data frame: UPDATED\_keyword\_context\_df\_with\_missing.
  + In this data frame, the unique attribute for each row is going to be the context column, which is the sentence in which the keyword under column keyword was found.
  + expected\_pings and actual\_pings refer to the number of times the value in keyword was (expected to be) found in that particular pdf\_url, which again is the link for the document being scraped. Therefore, these values can repeat for each keyword and pdf\_url combination, as a keyword can be found in one pdf\_url more than once.