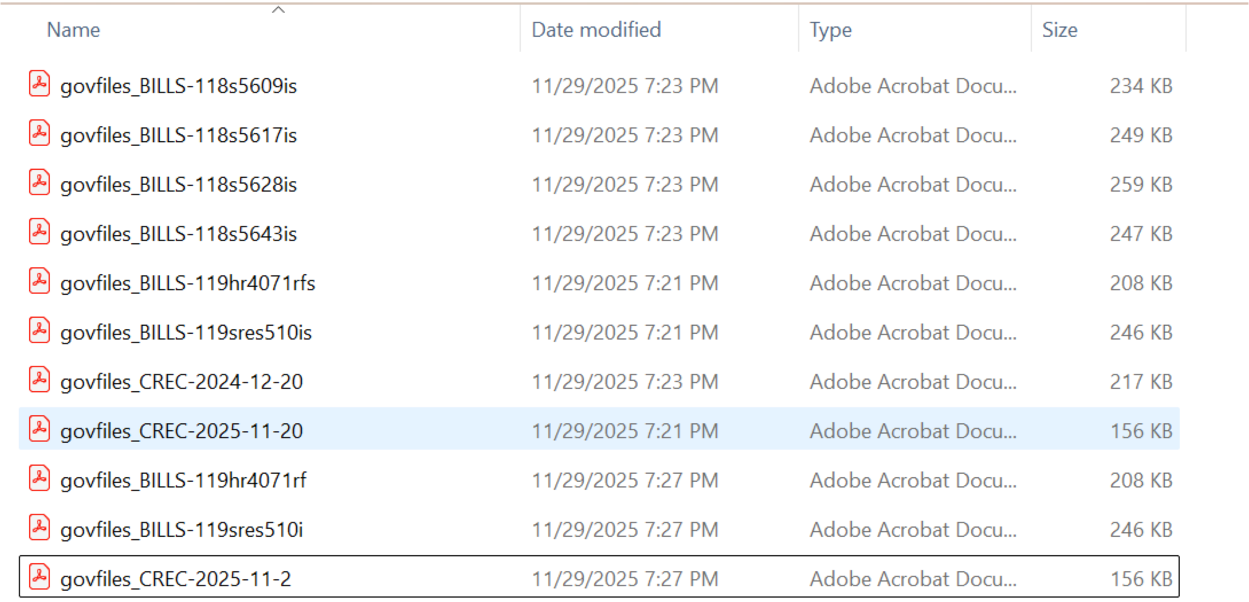
Write simple report on difficulties encountered in the scraping process.

- How useable is the scraped data?
- How to improve?

Answer

During the scraping process, I encountered several challenges that affected both the efficiency of the workflow and the number of usable documents I was able to obtain. I began by downloading the JSON search results for the 2025 records from the Committee on Foreign Relations. Although the file contained hundreds of entries, only three records had working PDF links that were successfully downloaded. Many entries included a pdfLink field, but the link was either empty, broken, or produced a zero-byte file. To increase the number of usable documents, I downloaded an additional JSON file for 2024, repeated the same filtering and downloading steps, and was able to obtain eight more working PDFs. Even though each JSON file contained hundreds of results, only a small portion actually linked to valid PDF documents, so multiple years had to be used to reach a reasonable number of downloads.



Name	Date modified	Type	Size
govfiles_BILLS-118s5609is	11/29/2025 7:23 PM	Adobe Acrobat Docu...	234 KB
govfiles_BILLS-118s5617is	11/29/2025 7:23 PM	Adobe Acrobat Docu...	249 KB
govfiles_BILLS-118s5628is	11/29/2025 7:23 PM	Adobe Acrobat Docu...	259 KB
govfiles_BILLS-118s5643is	11/29/2025 7:23 PM	Adobe Acrobat Docu...	247 KB
govfiles_BILLS-119hr4071rfs	11/29/2025 7:21 PM	Adobe Acrobat Docu...	208 KB
govfiles_BILLS-119sres510is	11/29/2025 7:21 PM	Adobe Acrobat Docu...	246 KB
govfiles_CREC-2024-12-20	11/29/2025 7:23 PM	Adobe Acrobat Docu...	217 KB
govfiles_CREC-2025-11-20	11/29/2025 7:21 PM	Adobe Acrobat Docu...	156 KB
govfiles_BILLS-119hr4071rf	11/29/2025 7:27 PM	Adobe Acrobat Docu...	208 KB
govfiles_BILLS-119sres510i	11/29/2025 7:27 PM	Adobe Acrobat Docu...	246 KB
govfiles_CREC-2025-11-2	11/29/2025 7:27 PM	Adobe Acrobat Docu...	156 KB

Another difficulty was inconsistency in metadata fields, especially with publication dates, since not all records used the same date variable. This made sorting and selecting the most recent documents more complicated. Overall, the scraped data is usable, but only after substantial cleaning, removing empty or broken links, checking file sizes, and manually combining multiple years of data. The process could be improved by adding automated checks for valid PDF links, removing zero-byte files, standardizing date fields, and enabling the script to automatically process several years of data until the target number of PDFs is reached.