HW#4 - Web Archiving, Part 2

Ethan Landers

Due: Sunday, October 27, 2024, by 11:59 PM

Q1

Q: What can you say about the relationship between the age of a URI-R and the number of its mementos?

This question was challenging to answer because the URI-Rs I collected TimeMaps for had either 0 or 3 mementos each. I excluded the URI-Rs with 0 mementos from my analysis, resulting in plots for those with only 3 mementos. A scatter plot was not effective in this case due to the lack of variability in the number of mementos. Instead, I opted to create a box plot, which effectively illustrates the age of each memento across all the URI-Rs for which I grabbed TimeMaps.

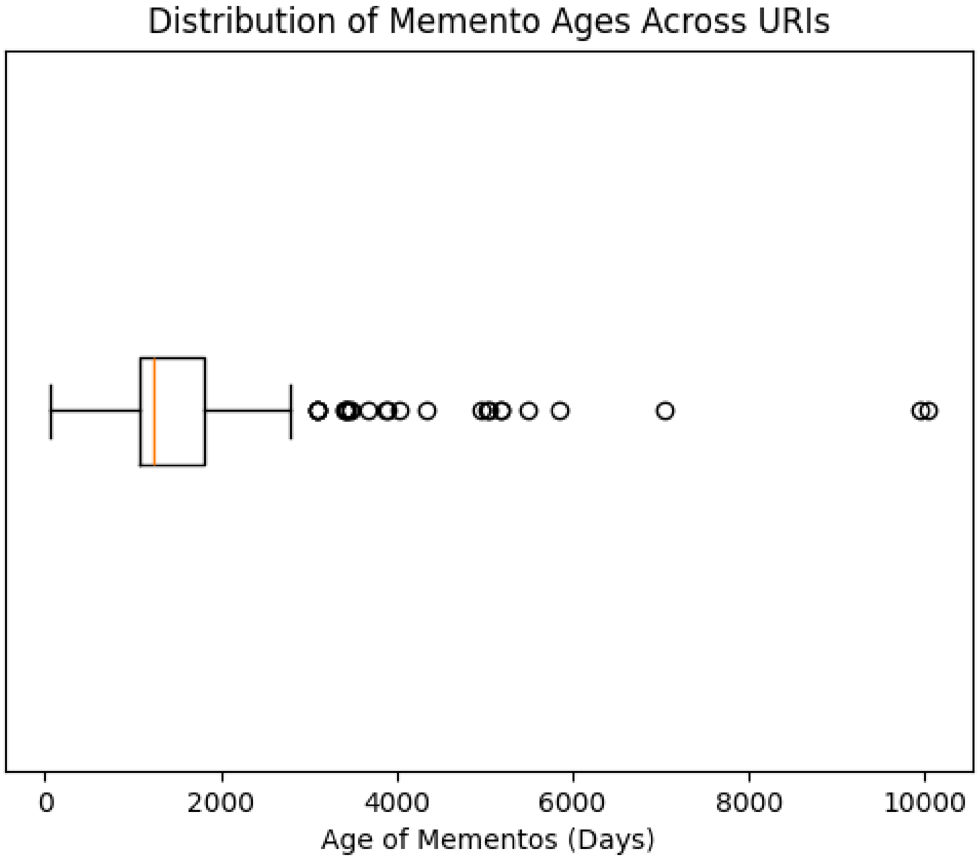


Figure 1: Distribution of Mementos Ages Across URIs

The results indicate that the majority of mementos for archived URI-Rs are between 1,000 and 2,000 days old, with a few outliers exceeding 2,000 days.

Q: What URI-R had the oldest memento? Did that surprise you?

The URI-R that has the oldest memento from the URI-Rs that were analyzed was https://www.unc.edu, with the earliest memento date being 1997-04-27 05:36:51. I'm not that surprised by the result as the University of North Carolina Chapel Hill is a prestigious state university that is known for conducting research.

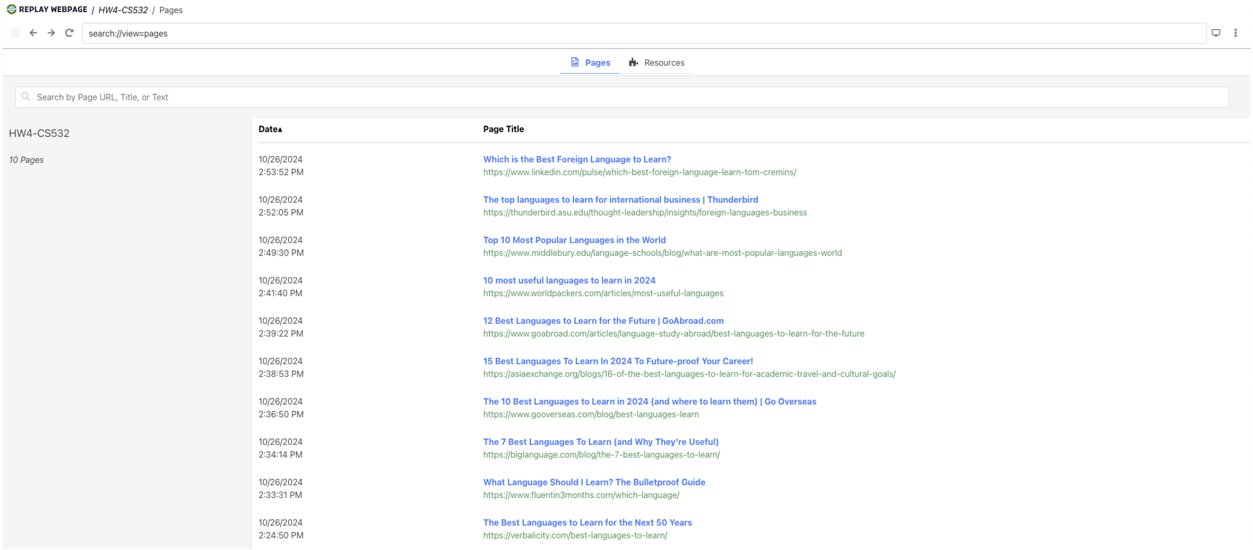
Q: How many URI-Rs had an age of < 1 week, meaning that their first memento was captured the same week you collected the data?

It was concluded that zero URI-Rs had an age of less than 1 week after analysis. All the URI-Rs that I looked at had mementos that were captured more than a week ago (as of 10/26/2024).

Q2

Click here to access my Conifer public collection for Q2 of this assignment.

Figure 2 shows the list of archived pages as well as the browser address bar after uploading the WARC file (created by archiving 10 webpages using Conifer) to ReplayWeb.page (`https://replayweb.page/`).


REPLAY WEBPAGE / HW4-CS532 / Pages
search://view=pages
Pages
Resources
Search by Page URL, Title, or Text
HW4-CS532
10 Pages

Date▴	Page Title
10/26/2024 2:53:52 PM	Which is the Best Foreign Language to Learn? https://www.linkedin.com/pulse/which-best-foreign-language-learn-tom-cremins/
10/26/2024 2:52:05 PM	The top languages to learn for international business | Thunderbird https://thunderbird.asu.edu/thought-leadership/insights/foreign-languages-business
10/26/2024 2:49:30 PM	Top 10 Most Popular Languages in the World https://www.middlebury.edu/language-schools/blog/what-are-most-popular-languages-world
10/26/2024 2:41:40 PM	10 most useful languages to learn in 2024 https://www.worldpackers.com/articles/most-useful-languages
10/26/2024 2:39:22 PM	12 Best Languages to Learn for the Future | GoAbroad.com https://www.goabroad.com/articles/language-study-abroad/best-languages-to-learn-for-the-future
10/26/2024 2:38:53 PM	15 Best Languages To Learn In 2024 To Future-proof Your Career! https://asiaexchange.org/blogs/16-of-the-best-languages-to-learn-for-academic-travel-and-cultural-goals/
10/26/2024 2:36:50 PM	The 10 Best Languages to Learn in 2024 (and where to learn them) | Go Overseas https://www.gooverseas.com/blog/best-languages-learn
10/26/2024 2:34:14 PM	The 7 Best Languages To Learn (and Why They're Useful) https://biglanguage.com/blog/the-7-best-languages-to-learn/
10/26/2024 2:33:31 PM	What Language Should I Learn? The Bulletproof Guide https://www.fluentin3months.com/which-language/
10/26/2024 2:24:50 PM	The Best Languages to Learn for the Next 50 Years https://verbalicity.com/best-languages-to-learn/

Figure 2: ReplayWeb.page WARC file ”Pages” Tab

Q: Why did you choose this particular topic?

I am a big language fan, and I speak Spanish self-taught. I want to learn a new language, but I have the hardest time choosing the next language to learn. Therefore, I chose to archive websites discussing the best foreign languages to learn.

Q: Did you have any issues in archiving the webpages?

For some websites, when I started the archiving process, a 404 error message would appear, and I couldn't archive the website that I wanted to. Otherwise, no issues. While capturing, I would scroll to the bottom of every webpage so the whole page could be captured.

Q: Do the archived webpages look like the original webpages?

For the most part, the archived webpages look quite similar to their original webpages. Sometimes there is a minor formatting difference, but nonetheless very similar.

Q: How many URLs were archived in the WARC file? How does this compare to the number of Pages?

A total of 1,140 URLs were archived in the WARC file, corresponding to 10 webpages. This indicates that each archived webpage contained a diverse array of resources, including images, scripts, and other types of content.

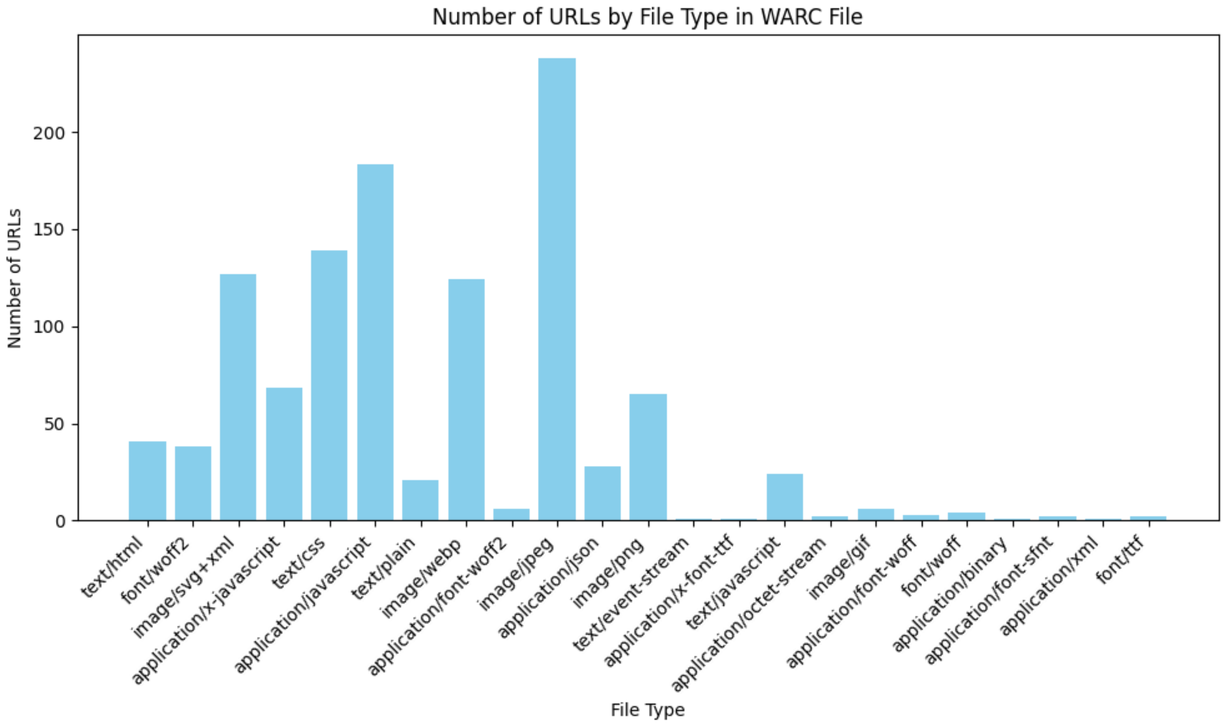


Figure 3: Number of URLs by File Type in WARC File

Q: Which file type had the most URLs? Were you surprised by this?

JPEG was the file type that had the most URLs. This doesn't surprise me as many web pages contain several JPEG image files.

References

- Counters in Python, `https://www.geeksforgeeks.org/counters-in-python-set-1/`
- Python — datetime.timedelta() function, `https://www.geeksforgeeks.org/python-datetime-timedelta-function/`
- Warcio, `https://github.com/webrecorder/warcio`