For starters I want to explain why I’m turning in three webcrawlers and three processing files. I had a pretty difficult time picking a topic that gave me enough websites, let alone ones that could be used to scrape from. None out of the three that I picked ended up generating 15 files due to anti-scraping bots, the greatest being 12 files generated for a single topic. I will be focusing on the webcrawler I made about Transgender legislation currently being proposed and passed in the United States as I am very passionate about this topic. I started this webcrawler at the google search for “trans legislation” to lessen the bias between news sources.

My Top Terms:



These were chosen as they’re necessary to acknowledge when discussing Transgender Healthcare (you can’t talk about this topic without knowledge on gender-affirming care, what transgender means, or the ‘anti-trans’ sentiment behind a lot of these bills). I also included ‘federal’ as there was recently a federal bill proposed (that will most likely pass on to the senate as our current house of representatives is Republican, but is likely to be thrown out by the senate as it is ruled by the Democrats), ‘malpractice’ as many anti-transgender bills frame gender affirming care as medical malpractice, and ‘proposed’ to get the bills currently being proposed nationwide.

My Knowledge Base:

Text

Description automatically generated

Text

Description automatically generated

For my knowledge base I pickled a dictionary of all the sentences that contain the terms I listed above. This should allow for a wide range of knowledge from all the sources. The all\_lines is a list of every single sentence from all of the scraped websites, so it checks every possible sentence that could contain the word and then returns a list of all sentences that contains the word. In the dictionary I used the term as the key.