# Quantitative Analysis Pre-Processing

Before we can finish data processing to get the results for the quantitative analysis, we first need to “pre-process” the data into consistently formatted source files for each of the ~180 websites. The set of steps below walk through everything that needs to be done to generate/finish the 4 policy files for each company/website **X** - "**X**\_Terms\_of\_Service.txt", "**X**\_Terms\_of\_Service.docx", "**X**\_Privacy\_Policy.txt", and "**X**\_Privacy\_Policy.docx".

I have found that doing this was the most time-consuming part of the overall analysis process -- once we have the pre-formatted files available, the process of actually running them through the analysis tools to get results should be relatively fast and simple. The remaining work is extremely repetitive/tedious, but is thankfully super simple and straightforward enough for anyone to help divide and conquer it -- no technical background needed.

Required Tools

* **Notepad++:** <https://notepad-plus-plus.org/>
  + This is one of the best text editors to use for viewing/editing cross-platform code in many file formats
  + There are some alternative options, but this one provides the most reliable control to remove all possible text formatting that would potentially break custom code solutions *(we want plaintext files to be UTF-8)*
* **Firefox:** <https://www.mozilla.org/en-US/firefox/new/>
  + This is my preferred web browser for privacy/cybersecurity reasons, which has many convenient *(and trustworthy)* extension tools to help significantly speed up certain tasks *(so you don’t have to write your own custom code for everything)*
* **Firefox Browser Extension:** <https://addons.mozilla.org/en-US/firefox/addon/copy-selected-links/>
  + This is the best tool I found to automatically extract hyperlink URLs from web pages -- but only the ones we want from the policy content, not the hundreds of unrelated navigation/advertising links found on most website pages
* **Google Chrome Browser:** <https://www.google.com/chrome/>
  + This has certain convenient built-in features that Firefox doesn’t have, such as:
  + Feature #1: Copying/pasting from Chrome keeps the formatting compatible with word processing programs, but also keeps the pasted content visually similar to the source content, making the pre-processing a lot easier and more consistent
  + Feature #2: The automatic translation functionality provided is reliable for most content on international websites, so no third-party extensions or custom translation steps are needed

Template Instructions

1. Download/install all of the tools above, and change settings as needed
2. Open Notepad++ and new Incognito Mode windows for Chrome and Firefox *(mostly for organizational purposes)*
3. Go to the [main policy spreadsheet](https://docs.google.com/spreadsheets/d/1jaki362E3ET-Vj0K--75979hlzC-65QR0lGrIqqylu4/edit?usp=sharing) to select a “chunk” of websites from the list to work on *(like 3-10 at a time)*
4. Go to the [Pre-Formatted Policy Files](https://drive.google.com/drive/folders/1uhqHzCrRpIz-7XlgcMKlXFqviwGsTj-Z?usp=sharing) folder and download all the files corresponding to the current chunk of websites
   1. **Reminder:** there are supposed to be **4 total files** for each company/website **X** - "**X**\_Terms\_of\_Service.txt", "**X**\_Terms\_of\_Service.docx", "**X**\_Privacy\_Policy.txt", and "**X**\_Privacy\_Policy.docx"
5. For each company/website, start by updating the template files with the names/URLs *(might already be done)*
   1. Go back to the [main spreadsheet](https://docs.google.com/spreadsheets/d/1jaki362E3ET-Vj0K--75979hlzC-65QR0lGrIqqylu4/edit?usp=sharing) to lookup the URLs for the ToS and Privacy Policy *(in columns H and I respectively)*
   2. Open both plaintext files *("****X****\_Terms\_of\_Service.txt" and "****X****\_Privacy\_Policy.txt")* files with the same website/company name **X**, then copy/paste the ToS and PP URLs from the spreadsheet into the header line of both, after **"From:"**
   3. Also paste the URLs into the header of both Word-format template files *("****X****\_Terms\_of\_Service.docx" and "****X****\_Privacy\_Policy.docx")*
   4. Update the **"Date Accessed"** fields *(of both documents)* as needed to mark the policy version, just in case the online policy is updated in the near future
   5. Double-click the footer to add page numbers *(optional, I just forgot to do this before duplicating files)*
   6. Double-check the filenames, dates, and URLs, then save all 4 files for the company/website
6. Go to the webpage from the URL above *(in both Chrome and Firefox)*, open any collapsed sections, then click and drag to select the entire Policy text *(but don't include top/bottom toolbar navigation elements or icons)*, and right-click to copy it *(or just press Ctrl-C)*
7. First, right-click to paste the policy text into the corresponding plaintext file, starting on line 7 after the initially empty **"Embedded Hyperlinks Used:"** section
   1. If using the Firefox extension, click and drag to select the policy text as before, but this time right-click and paste it into line 4 of the text file
   2. Lastly, re-copy the policy text from Chrome to reset the clipboard for the next step
8. In the Word doc files, right-click the line after the … below and select "Keep Source Formatting" under "**Paste Options:**", to make sure the text content is similarly readable as the webpage *(Note: this step is important)*
9. If the text gets messed up when pasting, try going to the same URL in a different browser, and copy/pasting from there *(Chrome \*usually\* keeps the copy/pasted text's font formatting closest to the original, but Firefox is better at auto-formatting tables and other weird spacing found in some policies)*
10. Please try to clean up the pasted text's formatting as needed to make it easily readable *(i.e. add/remove empty spacing lines between sections or paragraphs and fix tables or numbering/bullet points)*
11. When you're finished making this document consistent/presentable, double-check the filenames, dates, and URLs again, delete these instructions, then save all 4 final file versions
12. Once finished with your current batch/chunk of files, re-upload them **\*with the same filenames\*** to the same Drive folder that you got them from *(by selecting the files in File Explorer, then click-and-drag into the Drive window for the folder)*
13. **Note:** don't forget to close the finished files for the company/website before moving on to the next one, to reduce the chances of accidentally overwriting the previous file data *(I've done that too many times already)*