## File Extensions to Remove (Black List)

* png, jpg, svg, and other image formats
  + These files will have nothing to do with the creation of SBOMs, nor will they reasonably store SBOM related information
* css, html, zip, 7z
  + These files should have nothing to do with the creation of SBOMs, nor should they store any SBOM related information. It is most likely they were caught in our net with the “dependency” and “dependencies” keywords

## File Extensions to Keep (White List)

We can apply additional filters to these files, but these are the ones that should not be removed wholesale

* json, xml
  + The json and xml file extensions could and have been used for storing dependency related information, like SBOMs

## Reasons for False Positives

* BOM also stands for Byte Order Mark
  + Mitigation: also scan for keywords “unicode” or “encoding”
* Keywords like BOM and SBOM appear in context of other words
  + Mitigation: Apply more rigorous regular expressions that check for white-space or a delimiter character like “-,/,\,etc”
  + Issues: We would lose files that have terms like “sbomGenerator” because there is not a clear break between “sbom” and “generator”
    - Mitigation: Look for camel case in variable names and functions

## Additional Filtering Ideas

* Look for terms in conjunction. That is only classify a file as a true hit if term a appears with term b.

## File Types of Interest

* .gitignore
  + Is there useful information we can pull from these in the SBOM context? Probably not?

## Notes

* [According to SPDX](https://spdx.dev/faq/#using-10): **Is there a designated file extension for SPDX files?**
  + The SPDX specification does not specify which file extension a SPDX file should have. A common practice is to append the SPDX file with the extension .spdx.
* [According to CycloneDX](https://cyclonedx.org/specification/overview/): defined in JSON Schema, XML Schema, and Protocol Buffers
* No files have the term “cyclonedx”in the file name