Acrobat SDK Folder-Level JavaScript Sample

This JavaScript adds a menu item under Edit->Acrobat SDK JavaScript: Copy Annotated Words.

This script returns to the console the words covered by highlight annotations. The script can be adapted to output the text elsewhere, such as to a file. It can also be edited to include other text-related annotations, such as underline or cross-out.

Annotations are processed one page at a time. Each page's words are stored in the pageWords array of Word objects.

This is a brute force method; other storage solutions might be considered. This script is limited to single-column, left-to-right, top-to-bottom, horizontal, non-overlapping text. Vertical text, text bound to a shape, right-to-left text, and so on will require special handling or may not work at all.

Note that text is recovered annotation-by-annotation in the order that the annotations were applied to the document, not in reading order or any other word order on the page.

Text returned may not always match exactly the text covered by the highlight. This is mainly dependent on two things: 1) whether a word is adjacent to punctuation or not, and 2) whether whole words or partial words are highlighted or not.

When a word is followed by a space, getPageNthWordQuads() returns the quads of only the word's characters, but if the word is adjacent to punctuation, the quads returned include the punctuation. Thus, an annotation that includes a word and a space will never have quads that exactly match those returned by getPageNthWordQuads. This script manages this by providing a vertMargin variable to set a range within which a match might be obtained.

Additionally, depending on line spacing, the vertical boundaries of an annotation may not exactly match those of the word. The horizMargin variable sets the range within which a match is considered.