being referred to by that caption. Figure 1 illustrates how effective this concept can be; much of the time a human reader can locate regions containing figures within a page of a scholarly document even if all that is known is the locations of the captions and body text. This observation stems from our knowledge of how scholarly documents are typically formatted. Authors present information in a continuous flow across the document and so do not include extraneous elements or unneeded whitespace. Therefore regions in the document that do not contain body text must contain something else of importance, almost certainly a figure. A particular motivation for our approach is that figures in scholarly documents are the least structured elements in the document and thus the trickiest to parse effectively. Graphics can have large amounts of text, large amounts of white space, be composed of many separate elements, or otherwise contain content that is hard to anticipate. Tables can be formatted in grids, with only vertical lines, with no lines at all, or in many other variations. However most academic venues have strict guidelines on how body text and section titles should be formatted which tend to follow a narrow set of conventions, such as being left aligned and being in either a one or two column format. Such guidelines make positively identifying body text a much easier task. Once the body text is found, the regions of the document containing figures can be detected without making any assumptions as to the nature of those figures other than that they do not contain any elements that were identified as body text.

Our proposed algorithm has three phases:

1. Caption start identification. This step involves parsing the text of the document to find words like 'Figure 1:' or 'Table 1.' that indicate the start of a caption, while taking steps to avoid false positives. The scope of this phase is limited to identifying the first word of the caption, not the entire caption itself.
2. Region identification. This involves chunking the text in the PDF into blocks, then identifying which blocks of text are captions, body text, or part of a figure. This step also attempts to identify regions containing graphical components. The output is a number of bounding boxes labelled as body text, image text, caption text, or graphic region.
3. Caption assignment. This phase involves assigning, for each caption, the region of space within the document that it refers to, by making use of the regions found in the previous step.

Caption Start Identification

This phase of the algorithm identifies words that mark the beginning of captions within the document. We extract text from documents using Poppler (Poppler 2014). This step assumes that the PDFs being used as input have their body text and captions encoded as PDF text operators, not as part of embedded images. This assumption is almost always true for more recent scholarly PDFs, but exceptions exist for some older PDFs, such as those that were created by scanning paper documents[5].

The extracted text is scanned to find words of the form ⟨Figure| Fig|Table⟩ followed by either a number, or a period or colon and then a number. These phrases are collected as potential starts of captions. This first pass has high recall, but can also generate false positives. To remove false positives, we look for textual cues in combination with a simple consistency assumption: we assume that authors have labelled their figures in a consistent manner as is required by most academic venues. If we detect redundancy in the phrases found, for example if we find multiple phrases referring to 'Figure 1', we attempt to apply a number of filters that selectively remove phrases until we have a unique phrase for each figure mentioned. Filters are only applied if they would leave at least one mention left for each figure number found so far. We have been able to achieve high accuracy using only a handful of filters. These include: (I): Select only phrases that contain a period. (II): Select only phrases that contain a semicolon. (III): Select only phrases that have bold font. (IV): Select only phrases that have italic font. (V): Select only phrases that are of different font sizes than the words that follow them. (VI): Select only phrases that begin new lines, as judged by Poppler's text detection system.

Our filters can be noisy, for example selecting bold phrases can, in some papers, filter out the true captions starts while leaving incorrect ones behind. Detecting bold and italic font can itself be challenging because such fonts can be expressed within a PDF in a variety of ways. However we can usually detect when a filter is noisy by noting that a filter would remove all mentions of a particular figure, in which case the filter is not applied and a different filter can be tried to remove the false positives. In general we have found our caption identification system to be highly accurate, but occasionally our consistency assumption is broken which can lead to errors.

Region Identification

Having detected the caption starts, this phase identifies regions of the document that contain body text, caption text, figure text, or graphical elements. The first step in this phase is identifying blocks of continuous text. To do this we use the text grouping algorithm made available in Poppler (Poppler 2014) to find lines of text within each page. Individual lines are then grouped together by drawing the bounding boxes of each line on a bitmap, expanding these boxes by slight margins, and then running a connected component algorithm[6] to group nearby lines together.

Having identified text blocks we need to decide if those blocks are body text, captions, or part of a figure. We identify caption text by finding text blocks that contain one of the previously identified caption starts and labeling those blocks as captions. We have found it useful to post process these blocks by filtering out text that is above the caption word or not aligned well with the rest of the caption text. The remaining blocks are classified as body text or image text. To identify body text we have found an important cue is the page margins. Scholarly articles align body text down to fractions of an inch to the left page margin, while figure text is often allowed to float free from page margins. We locate margins by parsing the text lines throughout the entire document and detecting places where many lines share the same starting x coordinate. Text blocks that are not aligned with the mar-

[5]Using an OCR system might allow us to put such documents in the same pipeline, albeit with more noise, but is not currently implemented.

[6]http://www.leptonica.com/