sort layout objects in page (text, lines, etc.) by vertical position

find = text // to start

find\_q = 0 // i.e. q0 is rubric

find\_sect = a || 1

sect\_numeric = True || False

traverse sorted lt\_objs

if find == text and lt\_obj is text

if find\_q != 0

if !sect\_numeric and text\_match <section\_header\_letter>

print/store ‘new section <letter> before <q>’

find\_sect++

sect\_numeric = False

elif sect\_numeric and text\_match <section\_header\_num>

print/store ‘new section <num> before <q>’

find\_sect++

sect\_numeric = True

elif find\_q == 0 and text\_match <time> OR find\_q != 0 and text\_match <question\_i>

crop\_top = lt\_obj.bbox[3]

crop\_left = bbox[0]

crop\_right = bbox[2] // to be updated

find = divider

elif find == divider

if lt\_obj is text

// perhaps need to be checking in here to get ‘rough’ area of question parts?

// also what if question finishes on an image/table/not text?

if (bbox[2] > crop\_right)

update crop\_right

if not whitespace

pos = bbox[1]

if not pos <= footer

crop\_bottom = pos

elif lt\_obj is divider

crop(page\_num, look\_for\_q, crop\_top, crop\_left, crop\_bottom, crop\_right)

find = text; find\_q++

if find == divider (at end of page) // assume EOP as divider

crop(page\_num, look\_for\_q, crop\_top, crop\_left, crop\_bottom, crop\_right)

find = text; find\_q++

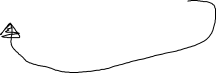
if find\_q == 0 (after all pages processed)

throw fatal error: rubric not found

elif find\_q == 1

throw fatal error: no questions found

order of finding things is: <rubr> <div>|<EOP> (<sect>) <q\_i>



an idea: rather than concretely extract ‘parts’ of questions (which is likely to be messy/not work properly due to layout objects being slightly out of order), use the ‘rough’ parts areas to blur out rest of question e.g. on hover over a part, or use to overlay the number of marks/time for a part