OCR学習用データセット 文字種

OCR学習用データセットの文字種

基本的な文字種

- ひらがな
- カタカナ
- 数字
- JIS第一水準漢字・JIS第二水準漢字
- 下記の記号
  - 半角記号　,.- /()
  - 句読点　、。
  - 括弧　() 〔〕 『』 【】 「」
  - 丸付き文字　①②③④⑤⑥⑦⑧⑨⑩⑪⑫⑬⑭ ㊀㊁㊂㊃㊄㊅㊆㊇㊈㊉ ❶❷❸❹❺❻❼❽❾
  - 丸付き漢字　㊞
  - 括弧付き文字　㈱㈲㈾㈲
  - 繰り返し記号　＼ノゞゝ々〃ゝゞヽヾ
  - 図形　○●□◆▲△▽
  - その他の記号　― ・ ※↓ →↑←⇩⇨⇧⇦ ？～ =�please…

凸版印刷による事前調査で出現率3,000位までのJIS第一水準漢字・JIS第二水準漢字に包摂が可能なJIS第二水準外の漢字

欧文・ギリシア文字

- 本文中に出現する3文字程度までの欧文・ギリシア文字
  - 半角アルファベット52文字（U+0041-U+005A、U+0061-U+007A）
  - ギリシア文字48文字（U+0391-U+03A9、U+03B1-U+03C9）

留意事項

- 欧文・ギリシア文字は、十分な領域がありレイアウト認識ができるものについては、矩形情報のみのOCR学習用データを作成し、文字情報は入力対象外とする。
- 半角文字が存在する(ASCIIコード内)の英数字記号は、別紙2「性能評価対象とする資料の内訳及び判定基準」の仕様を踏まえ、すべて半角のコードポイントを割り当てる。
- 合略仮名や結合文字、特殊な丸付文字など、コードポイントがないため１文字で入力不可能な字形は本件OCR学習用データ作成の対象外とする。
- 訓点、図版内文字、表組内文字は文字種に関わらずOCR学習用データ作成の対象外とする。
- 変体仮名は対応する現代仮名遣いのコードポイントを割り当てる。
  漢字と変体仮名どちらとも取れる文字は漢字のコードポイントを割り当てる。
- 行頭字下げ、文中スペースは学習用データ作成の対象外とする。
- 包摂不可能なJIS第二水準外文字、および汚れやかすれなどで文字が判読不能な文字は「**二**」を割り当てる。