HTRやってみたver２

今後のフロー

目標

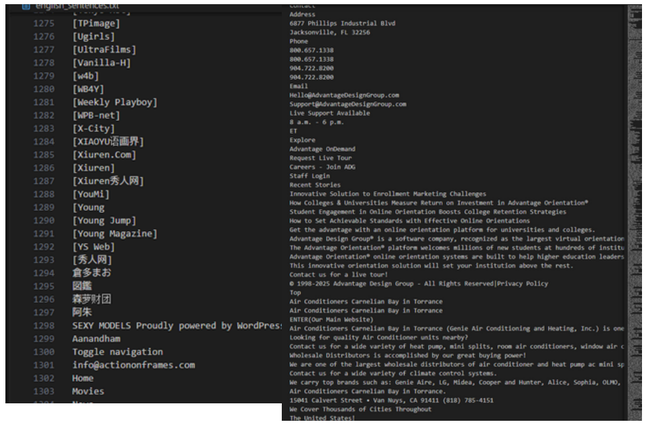
　7月中にHTR動かす

ステップ

- Common Crawlなどから英語のテキスト（1文単位）を抽出
- 抽出したテキストを手書き風画像に変換
- 画像と対応するテキストをペアにして学習データセットを作成
- 既存のHTRモデルをgithubから取得し学習させる

データセット探索

項目	Common Crawl	Project Gutenberg	Reddit
内容	インターネット全体のクロールデータ	主に英語の古典文学	ユーザー投稿のフォーラム
言語	多言語（英語多め）	英語中心	英語（ただし多言語あり）
構造	生HTMLベース、ノイズ多い	プレーンテキスト、構造化されている	JSON構造、メタ情報豊富
データサイズ	非常に大規模	比較的小さい	中規模
用途例	Web検索モデル訓練、トレンド抽出	文学的スタイルの分析、言語モデル微調整	会話の研究、感情分析、スラング辞書作成
クリーンさ	× ノイズが多い	○ 非常にクリーン	△ 投稿により異なる



今回の実験においてCommon Crawlはなんか微妙そう
日本語、中国語etc,あと英語もほんとに英語？？まず言語が多数
さらに英語がweb用の文章が多いので汎用的じゃなさそう

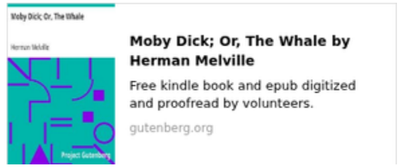
データセット

Project Gutenberg　意味不明な
英文もないぞ～！
よしこれにする！！

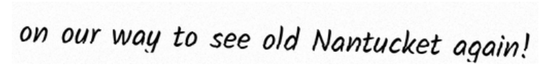
以下二冊の本の英文を取得した

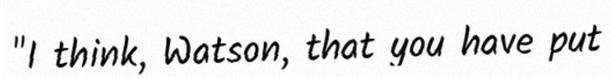
Gutenberg書籍ID：1661
タイトル:The Adventures of Sherlock Holmes by Arthur Conan Doyle
GutenbergのURL：

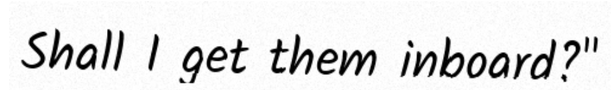
Gutenberg書籍ID：2701
実際の本：Moby-Dick; or, The Whale by Herman Melville
GutenbergのURL：



画像例







データ数(仮)

データ	train	val
枚数	9913	1102

HTRモデルどれにしようかな

プロジェクト	Star	備考
PyLaia	239 (github.com)	学術論文でよく比較対象になる BLSTM-CTC 系キット
TrOCR (rsommerfeld/trocr)	206 (github.com)	HuggingFace TrOCR ラッパー、Transformer 系を試すなら手軽
Loghi	123 (github.com)	Kraken と連携できるフルパイプライン／Gradio デモ付き
HTR-VT	83 (github.com)	Vision Transformer + CTC、最新論文の公式実装

transformer使ってるしなんかよさそお
trocr(transformer ocr)