Prescription database construction by data extraction from scanned files

1Noah Hallberg, 1Ryan Kwong, 2Ankur Malik, 1Saul Means, 1Udayan Pandey, 1Bharath Sadagopan, 2Kabir Snell, 1Varon Srinivasan, 1Margaret Wang
1Purdue University, 2University of California Santa Barbara

Introduction

- Our corporate mentors tasked the joint Purdue and UCSB research group this semester to help them evaluate different methods for analyzing bulk data.
- The goal was to help them choose the most optimal OCR solution for their needs.
- Optical Character Recognition (OCR) using computer models to analyze data that tends to be handwritten.

Methodology

Overview:
- Creating our own mock database
- Criteria for choosing a software
- Testing the software

Three main software considerations:
1. HIPAA compliance
2. Performance
3. Price

Handwriting

OCR software

Pros: Handwriting and circle/box filling
      Easy, Support from Base64.ai
      Circled Data
      Some issues with tables

Cons: Version 2023.0
      Inferior performance compared to Amazon Textract
      Requires a large amount of processing

Document AI - machine learning, OCR platform

Pros: Confident text extraction
      Identified symbols and room for ML
      Poor results with circle data
      Unclear data setting

Cons: Version 2023.0
      Low accuracy on OCR tests
      Identification of symbols and room for ML
      Requires a large amount of processing

Evaluation

- Tesseract OCR (Python):
  - Pros: Highly customizable
  - Cons: Lower accuracy on OCR tests

- FineReader (ABBYY):
  - Pros: High-speed scans
  - Cons: Sometimes interprets scribbled numbers

- Inogen:
  - Pros: Business oriented model
  - Cons: Lower accuracy on OCR tests

- Base64.ai:
  - Pros: Highly customizable
  - Cons: Inconsistent with flow setting and prescription

Future Goals

We are creating a tool that allows Inogen to generate a patient prescription database
This new database can help them answer questions like:

- What proportion of patients are adhering to their prescribed flow setting?
- What kind of patients are not adhering to their prescribed flow setting?
- In cases of inconsistency with flow setting and prescription, how is the flow setting being misused (higher or lower)?

Acknowledgements

Purdue Datamine Team
- Paula Salmentello M.Sc.
- Kush Malhotra
- David Glass M.Sc.
- Maggie Bitt
- Nicholas Lenfestey MPH

UCSB Team
- Trevor Stryer Ph.D.
- Tim Robinson

Inogen
- Norbert Leinfellner Ph.D.
- Shariab Glazer M.D.
- MRA
- Hurst Pickler M.Sc.
- Jothi Manian M.Sc.
- Niousar Rajabghasemi M.Sc.
- Eddy Choi

References

https://docs.aws.amazon.com/textract/index.html
https://guides.nyu.edu/tesseract/usage
https://docs.aws.amazon.com/textract/index.html
https://base64.ai/features/data
https://docs.aws.amazon.com/textract/index.html