CS 410 Final Project

Pathology Report Search

Jared Coberly

coberly2@illinois.edu

Fall 2024

1. Background

Hospitals often employ third-party pathology services to make medical diagnoses on tissue removed during surgery. These pathology laboratories use a diversity of software to generate reports and most frequently the hospitals receive these reports as bundled digital faxes of scanned paper reports (PDFs containing images of paper reports). A hospital may receive multiple faxes per day of completed pathology reports. Finding specific patient or diagnostic information in this collection is often a manual process of opening a PDF (representing a single fax and containing numerous reports) and browsing it for the desired information. In this project, I utilize existing tool kits and pipelines to develop a system that will allow a user to search this collection.

The system was built with, and for use on, actual scanned pathology reports from a single commercial laboratory. As patient identifiable information is protected from disclosure under federal law, I cannot share these actual data; however, the source code can be ran on any scanned pathology report of a similar format.

1.1 The Report

In the example report below, identifiable information is redacted; however, one can appreciate the general structure of a surgical pathology report. The information itself is largely prescribed by federal law (Title 42 CFR 493.1291). The report includes the name and address of the performing laboratory, patient demographic data, information about the ordering physician, accession details for the specimen (unique accession number, dates associated with the surgery and the report), the diagnoses, the name of the physician making the diagnosis (i.e., the pathologist), any clinical information provided, a microscopic description (if warranted), and a gross description. Disclaimers regarding the use of FDA and non-FDA approved testing are present as an endnote.

Of particular interest are the diagnoses. This text is technical and concise and written as pharases instead of complete sentences -- devoid of many “stop words” found in general text. Other interesting features of the report is the mode of transmission. They are received as digital faxes. This pipeline converts the original report to an image for transmission and the resultant received facsimile is an Adobe Portable Document Format (PDF) containing images, not text. Use of gray- and black-scale backgrounds for section headers, which will make the use of optical character recognition challenging. Additionally, the image quality is quite poor making OCR even more challenging.

[redacted]

PATIENT INFORMATION

[redacted]

PHYSICIAN / CLIENT INFORMATION

[redacted]

SPECIMEN INFORMATION

ACCESSION # [redacted]

Collected: 8/21/2024 Received: 8/22/2024
Accessioned: 8/22/2024 Reported: 8/23/2024

SURGICAL PATHOLOGY REPORT

DIAGNOSIS

A. Stomach, not specified, biopsy ([redacted]):
- **Oxyntic and non-oxyntic mucosa with mild chronic gastritis**
- **No evidence of active H. pylori infection**

B. Esophagus, not specified, biopsy:
- **Specialized intestinal metaplasia without dysplasia**

C. Colon, sigmoid, biopsy:
- **Invasive moderately differentiated adenocarcinoma**
- **MMR IHC pending**

TC07 : All Other Cancers. SNOMED CODE P-11300 New malignancy second pathologist review.

Primary Pathologist: [redacted] **MD, Electronic Signature, 8/23/2024**

CLINICAL INFORMATION

Barrett's esophagus, duodenitis and sigmoid colon mass. ICD: Not provided.

SPECIMEN DATA

MICROSCOPIC DESCRIPTION:
Microscopic examination performed. [redacted]

SPECIMEN:
A: Stomach, gastric, biopsy [redacted]
B: Esophagus, biopsy
C: Colon, sigmoid, biopsy

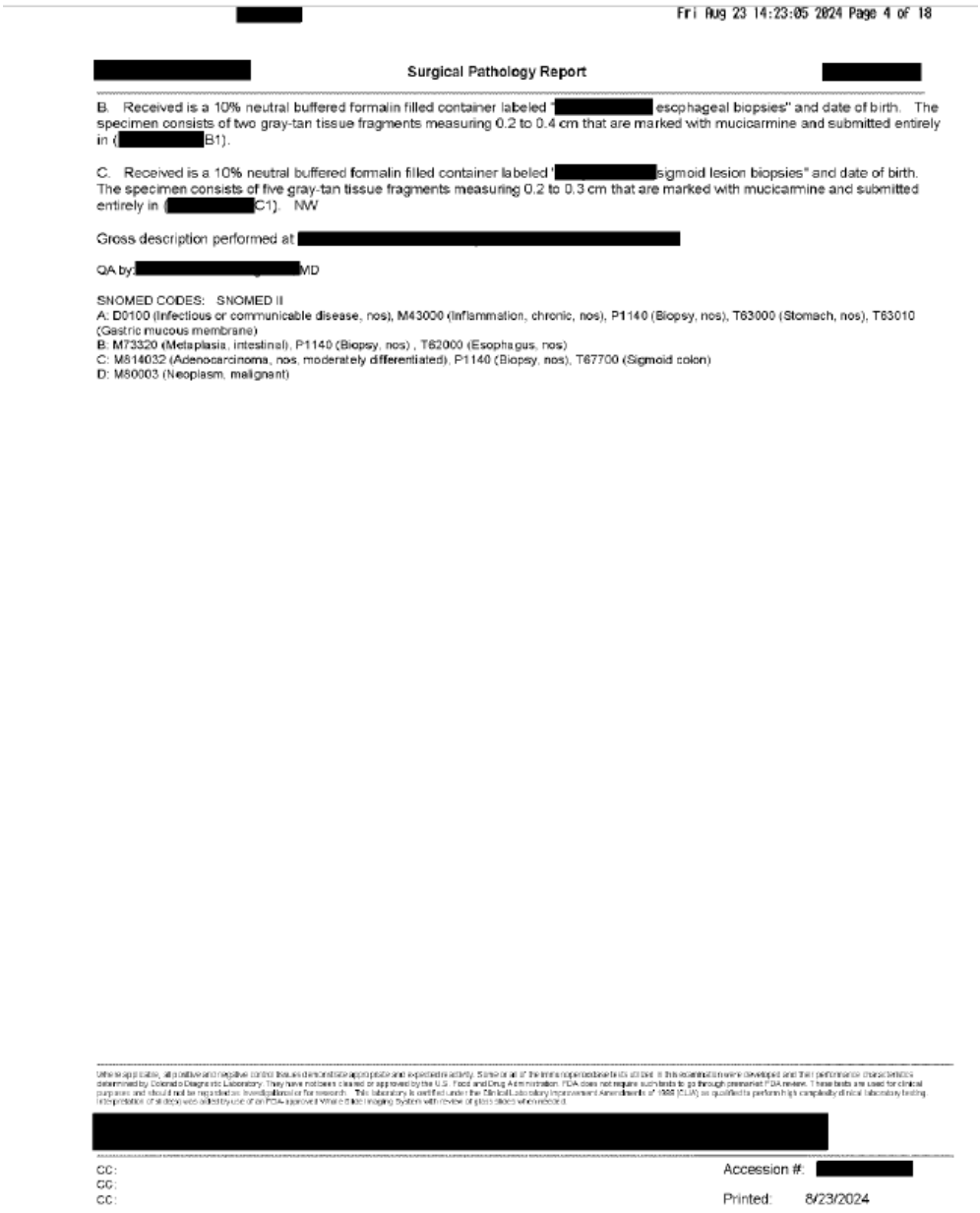
GROSS DESCRIPTION:
A. Received is a 10% neutral buffered formalin filled container labeled "[redacted] gastric biopsies" and date of birth. The specimen consists of seven gray-tan tissue fragments measuring 0.1 to 0.3 cm that are marked with mucicarmine and submitted entirely in [redacted] A1). Please note: Smallest specimens may not survive processing.

[redacted]

CC:
CC:
CC:

Accession#: [redacted]

Printed: 8/23/2024



[redacted] Fri Aug 23 14:23:05 2024 Page 4 of 18

[redacted] **Surgical Pathology Report** [redacted]

B. Received is a 10% neutral buffered formalin filled container labeled "[redacted] esophageal biopsies" and date of birth. The specimen consists of two gray-tan tissue fragments measuring 0.2 to 0.4 cm that are marked with mucicarmine and submitted entirely in ([redacted]B1).

C. Received is a 10% neutral buffered formalin filled container labeled "[redacted] sigmoid lesion biopsies" and date of birth. The specimen consists of five gray-tan tissue fragments measuring 0.2 to 0.3 cm that are marked with mucicarmine and submitted entirely in ([redacted]C1). NW

Gross description performed at [redacted]

QA by: [redacted] MD

SNOMED CODES: SNOMED II
A: D0100 (Infectious or communicable disease, nos), M43000 (Inflammation, chronic, nos), P1140 (Biopsy, nos), T63000 (Stomach, nos), T63010 (Gastric mucous membrane)
B: M73320 (Metaplasia, intestinal), P1140 (Biopsy, nos) , T62000 (Esophagus, nos)
C: M814032 (Adenocarcinoma, nos, moderately differentiated), P1140 (Biopsy, nos), T67700 (Sigmoid colon)
D: M80003 (Neoplasm, malignant)

Where applicable, all positive and negative control tissues demonstrate appropriate and expected reactivity. Since not all of the immunoreagents listed in CD Dist. II this examination were developed and their performance characteristics determined by Colorado Diagnostic Laboratory. They have not been cleared or approved by the U.S. Food and Drug Administration. FDA does not require such tests to go through premarket FDA review. These tests are used for clinical purposes and should not be reported as investigational or for research. This laboratory is certified under the Clinical Laboratory Improvement Amendments of 1988 (CLIA) as qualified to perform high complexity clinical laboratory testing. Interpretation of all slides was aided by use of an FDA-approved Whole Slide Imaging System with review of glass slides when needed.

[redacted]

CC: [redacted] Accession #: [redacted]
CC:
CC: Printed: 8/23/2024

2. Implementation

The system is implemented in python on Windows. The system works in three major phases:

1) Crawling the PDF documents, leveraging optical character recognition (OCR) to convert the report images to text

2) Generation of indexes using a standard pipeline

3) Creation of a search interface

2.1 Optical character recognition

As the reports are received as PDFs of images, the first step is to convert those images into machine-readable text. This requires two libraries: one to interact with the PDF document format and another to convert those images to text. Poppler is a utility that converts PDFs into images. Python's "pdf2image" is a python library that wraps poppler. Tesseract is a well-established OCR engine, first developed in the 1980's and released as open source in 2005. As this project is written in python, I used the python wrapper, pytesseract, to "read" text from the embedded images.

2.2 Indexing

Pathology diagnoses are often very concise, often a dozen words or less. The diagnosis is divided into two section: the specimen and the diagnoses. The specimen is a comma delimited string of anatomic location, specific site, and operative method. For example, "stomach, fundus, biopsy:" The diagnosis is a phrase describing the pathologic features or findings: "mild chronic gastritis."

These structure of these reports lend themselves well to bag-of-words representations. For this project, I utilized the Whoosh library for indexing and searching and indexed the diagnoses. Whoosh was chosen for it's native python implementation and it's flexibility. I used Whoosh's implementation of the Okapi BM25 ranking algorithm, which will be a strong performer on this text set and our defined use cases.

2.3 Search interface

The purpose of this project is to allow a user to search the library of diagnoses, so a user interface is required. As this is a relatively simple utility, a simple interface was designed in python using the tkinter library.

3. Use

3.1 Set-Up

This project was written for Windows. There are a few dependencies.

- **Poppler**. This is required for converting PDF to images. Windows distribution can be found here: https://github.com/oschwartz10612/poppler-windows/releases/

After installation, update the source code in cs410.py to include the path to the poppler engine:

```
poppler_engine_path =
r"C:/Users/VHACMOCoberJ/AppData/Local/Poppler/poppler-
24.07.0/Library/bin"
```

- **Tesseract.** This is required for converting images to text. Windows distribution can be found here: https://sourceforge.net/projects/tesseract-ocr.mirror/

  After installation, update the source code in cs410.py to include the path to the tesseract engine:

  ```
  pytesseract.pytesseract.tesseract_cmd =
  (r"C:/Users/VHACMOCoberJ/AppData/Local/Programs/Tesseract-OCR/
  tesseract.exe")
  ```

- **Report Directory.** Path to the report directory must point to the location of the scanned PDF reports. Update the source code in cs410.py to include the path to the tesseract engine:

  ```
  ameripath_dir = r"C:/Users/VHACMOCoberJ/Desktop/Reports"
  ```

Python dependencies are as follows:

- image library (PIL)
- tesseract wrapper (pytesseract)
- PDF to image converter (pdf2image)
- spell checker (pyspellchecker)
- indexing and search engine (whoosh)
- graphical user interface (tkinter)

Ensure libraries are installed by using the python installer program.

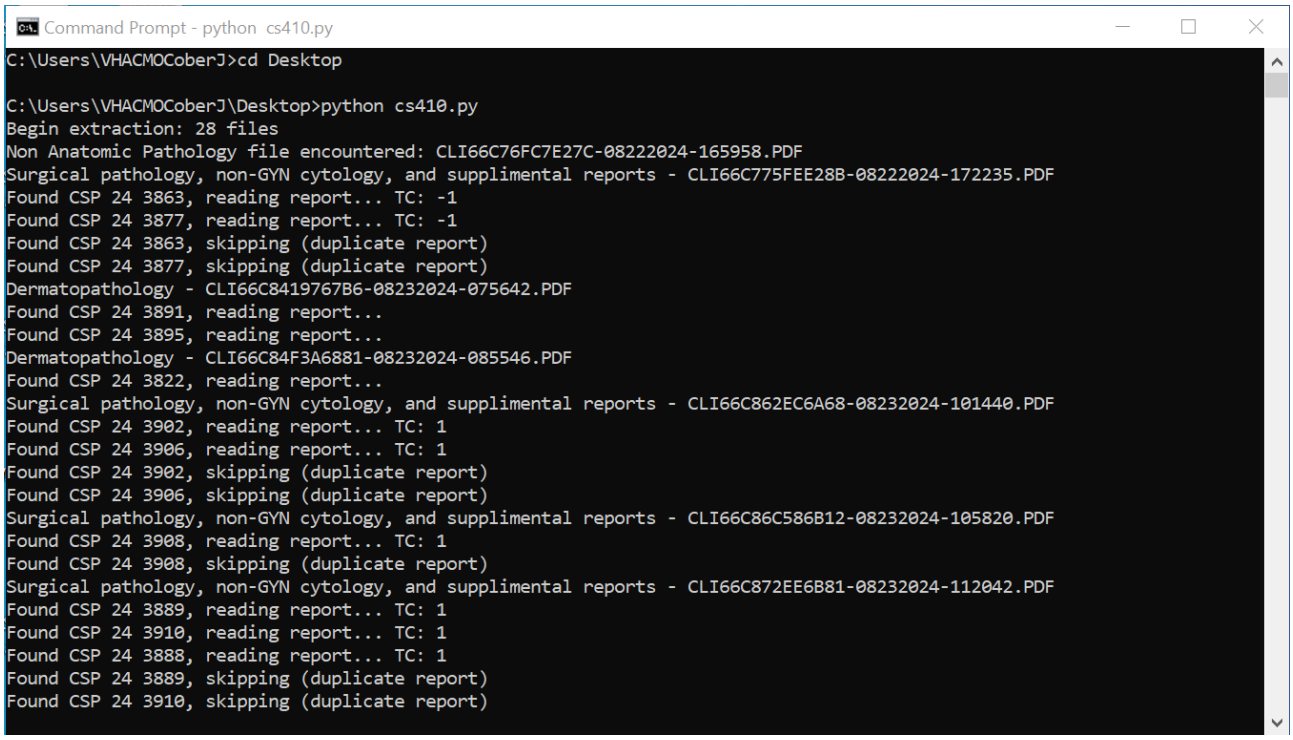
```
pip install PIL
```

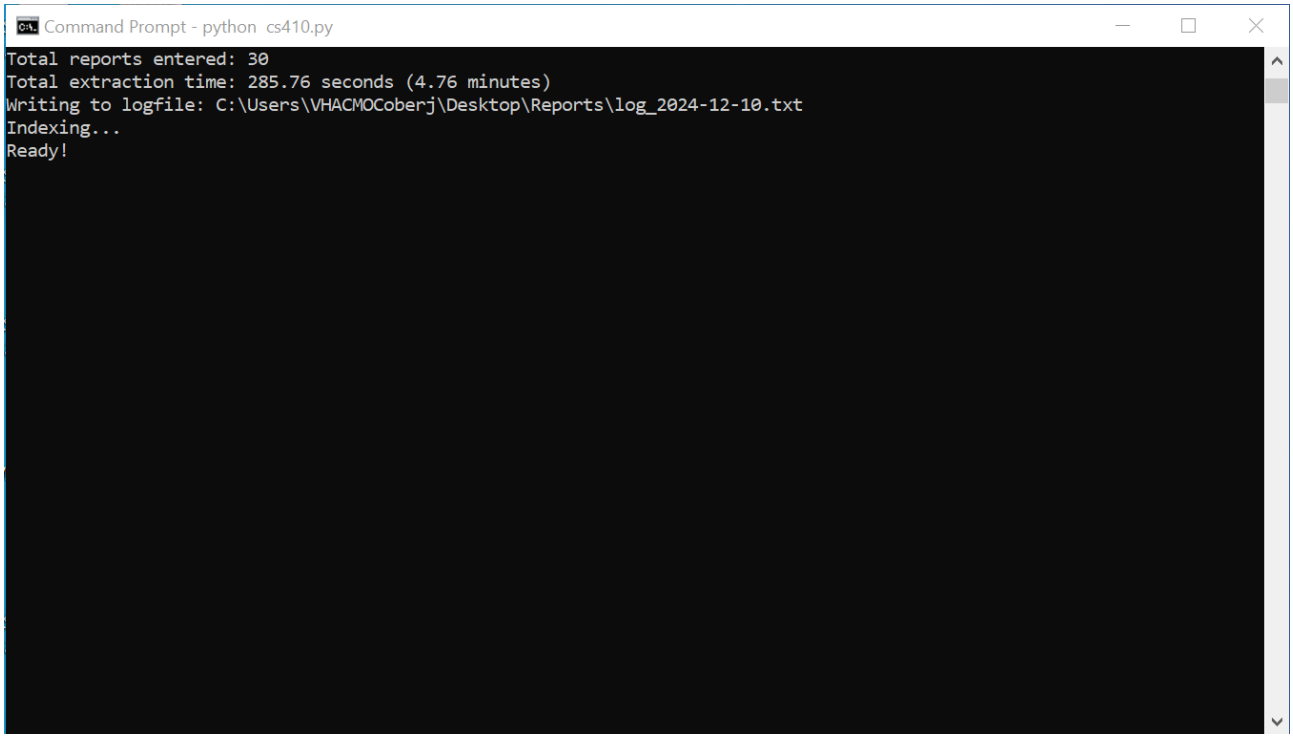
```
pip install pytesseract
pip install pdf2image
pip install pyspellchecker
pip install whoosh
pip install tkinter
```

3.2 Execution

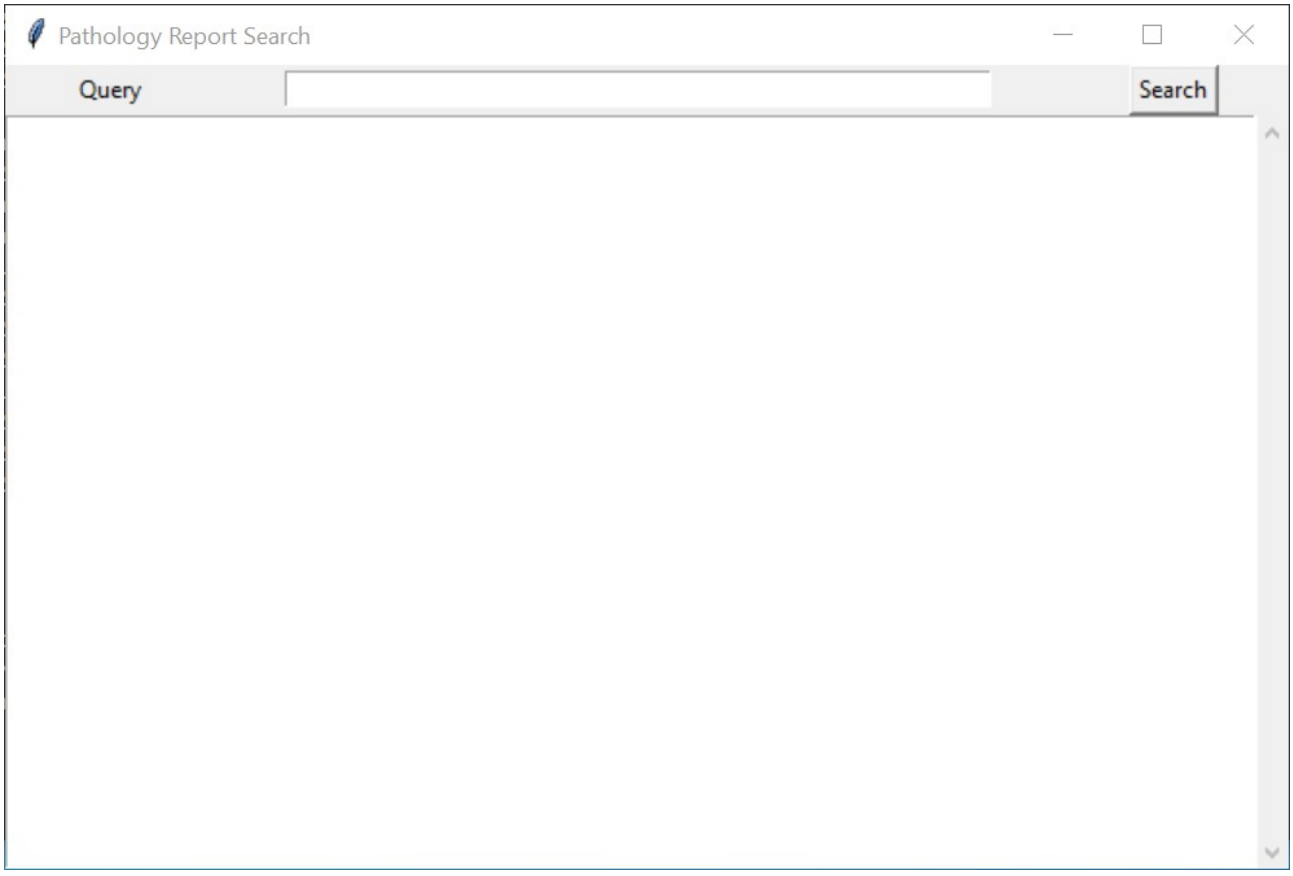
Run `python cs410.py` from the command line

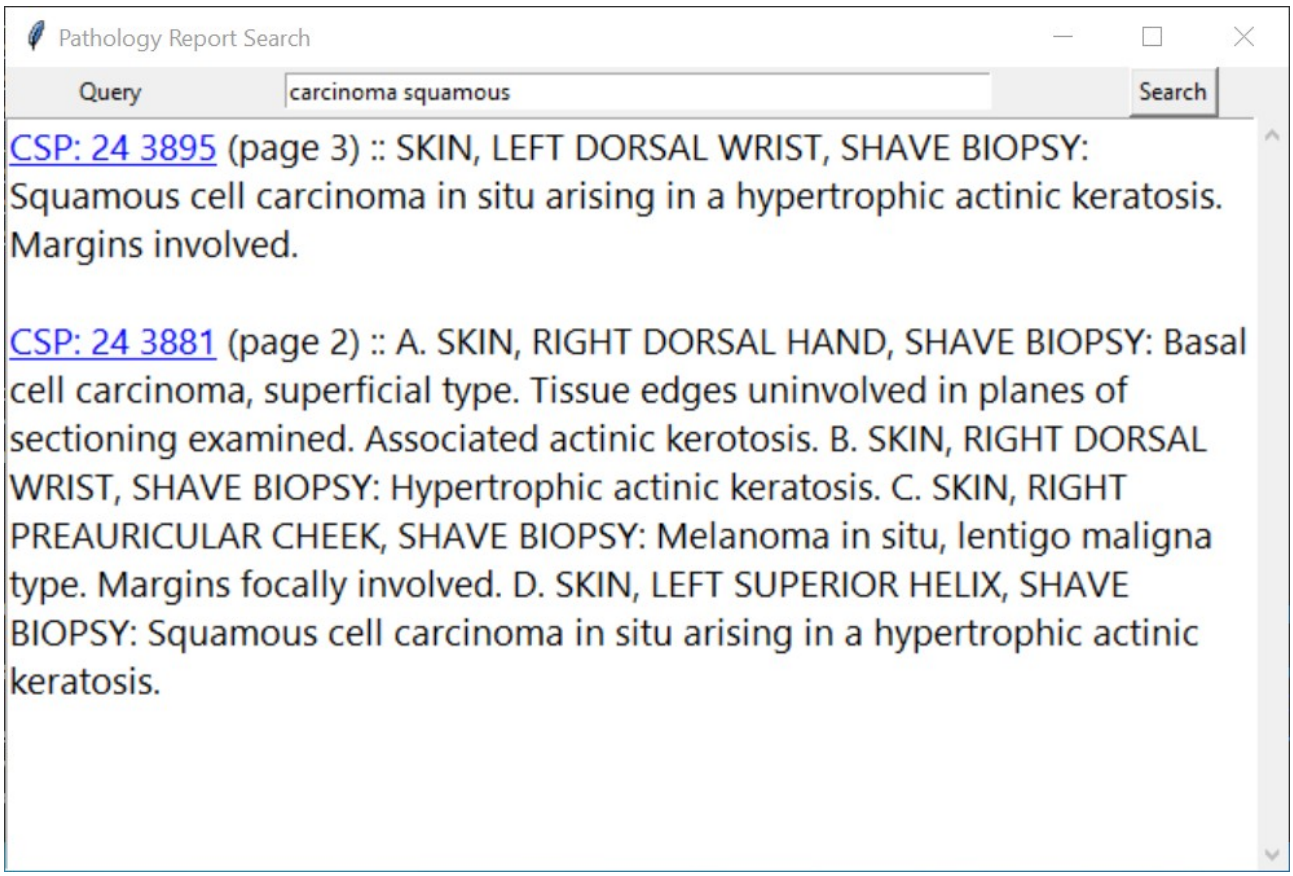
This will read the report directory, open each PDF file, convert the images to text, index the resultant documents, and then launch a GUI for searching the collection.





Once the GUI is visible, the user may enter search terms in the query field. Pressing the Enter key or the Search button, will submit the query and return a list of results. A new search will clear the results field and display the new search results. The results annotate the page number the diagnosis is found on in the original report and a hyperlink to directly open the PDF report.





4. Performance

4.1 PDF to Image to Text

Accuracy of this step is moderate. The report images contained in the PDF are of poor quality, no doubt the low resolution is an effort to keep transmission sizes small. The structure of the reports themselves, with its tabular format and shaded headers, is not the most straight forward for OCR rendering. Particularly the gray-scale shading on the section headers proved challenging for the OCR software. An equally challenging issue is the accuracy of the OCR itself. It sometimes struggles with spaces between words and confuses characters and numerals (zero and the letter O, for example). Additional complications were that the faxes often contain duplicate reports. In fact, almost always did I find two copies of a pathology report in each received fax.

The PDF to text process is excruciatingly slow. In the example video, showing the extraction of 30 reports from 28 PDF files, the conversion took over 5 minutes.

4.2 Indexing

The resultant documents were very short and the indexing step executed very quickly. As we see in the search performance, indexing of these concise and technical documents allowed for highly accurate search results.

4.2 Searching

The BM25 ranking algorithm performed very well on this data set. Keyword searches showed exceptional precision and recall, as in near perfect precision and recall in the test cases examined.

4.3 Optimization

Efforts at optimization did not reliably improve over the basic BM25 text retrieval function. Stemming increased recall, but reduced precision. Query expansion, by incorporating words that often are found along side the search terms, dramatically reduced precision. Both of these phenomenon are not entirely unexpected. These documents are incredibly short and use highly technical language. Not many inflectional forms of words are used (reducing the efficacy of stemming). Likewise, given the concise nature of the documents, query expansion was found to include words not associated with a search. An example is adenocarcinoma. Query expansion could include the word “malignancy,” which is often associated with the query term. However, query expansion brought in results that included “no malignancy” and drastically reduced precision.

5. Summary

Overall, the software works well and as designed. Of course, with additional time further refinements could be made. The major runtime performance barrier is the conversion of the reports from images to text. Work with the reference lab to obtain the reports in a different manner – other than fax – would likely dramatically improve the performance. The text retrieval itself performs adequately and, when compared to browsing PDFs for diagnoses, dramatically improves the workflow associated with finding information in these documents.