Computer Science & Computer Engineering

PDF Extraction and Clean-Up

Sarah Bondurant, Nathan Davis, Richard Mays, Keegan Riley, Hayden Willeford
University of Arkansas, College of Engineering: Computer Science, Computer Engineering

Introduction

This was a project given to us by Sorcero, an enterprise NLP suite created to support the Life Sciences & Insurance industries.

The arbitrary structure of a PDF is not useful for data extraction, and they would like to use Natural Language Processing, Computer Vision, and human intervention to train a model that would convert PDFs to document files.

Purpose

The purpose of our project was to gain technical knowledge of PDFs, research possible conversion methods, and inspect existing implementations in order to find the best solution for Sorcero to pursue.

Figure 2 (first image left)

Markdown output

Figure 3 (second image left)

Raw text output

Best Solution: Parsr

The best solution we found was Parsr, an open source document cleaning, parsing and extraction toolchain. Its strongest suit is the fact that it takes many already existing PDF parsers and combines them. However, Parsr is only available to run in a Linux Ubuntu environment with Docker installed. The program launches a GUI that can import PDF files for conversion.

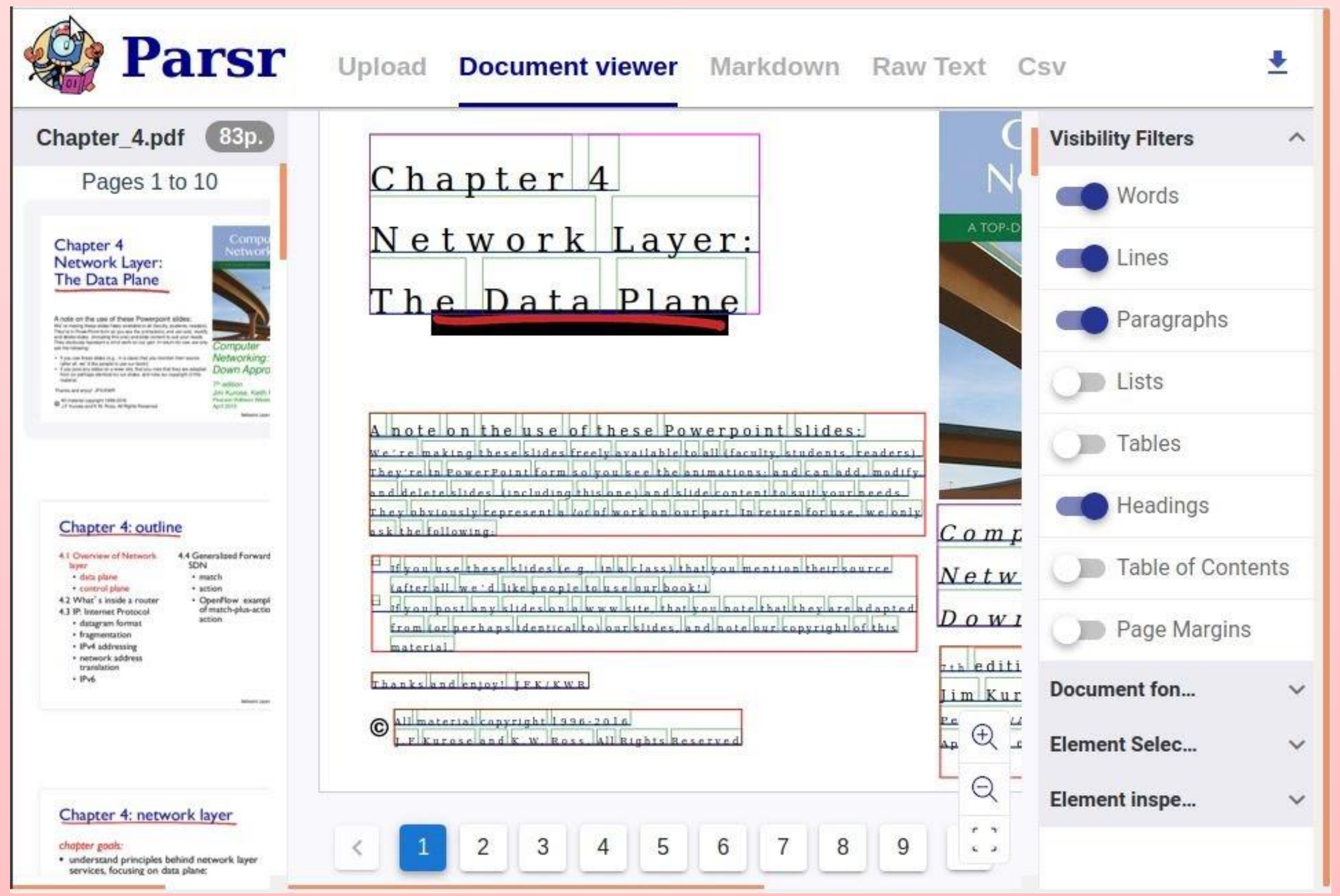
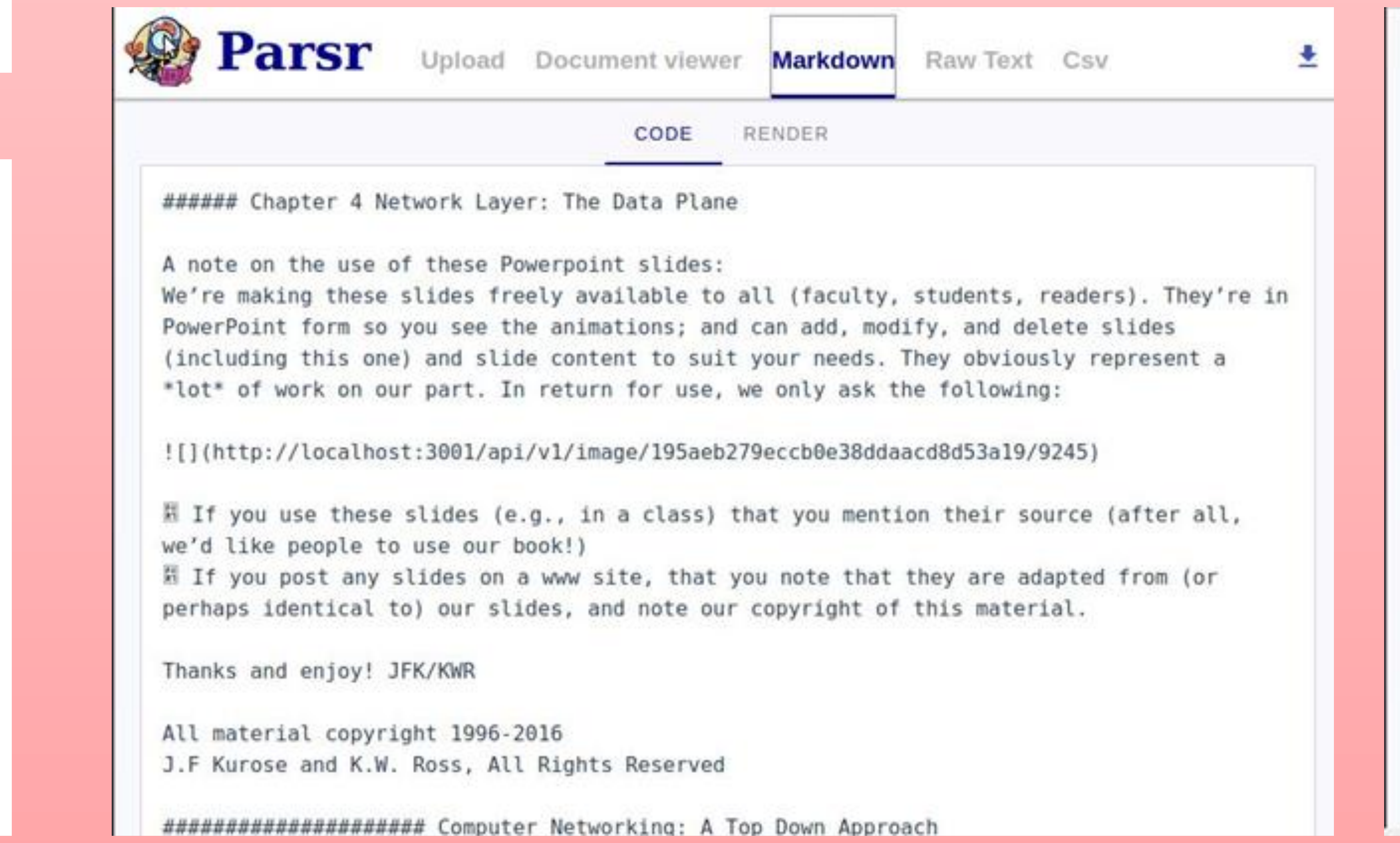
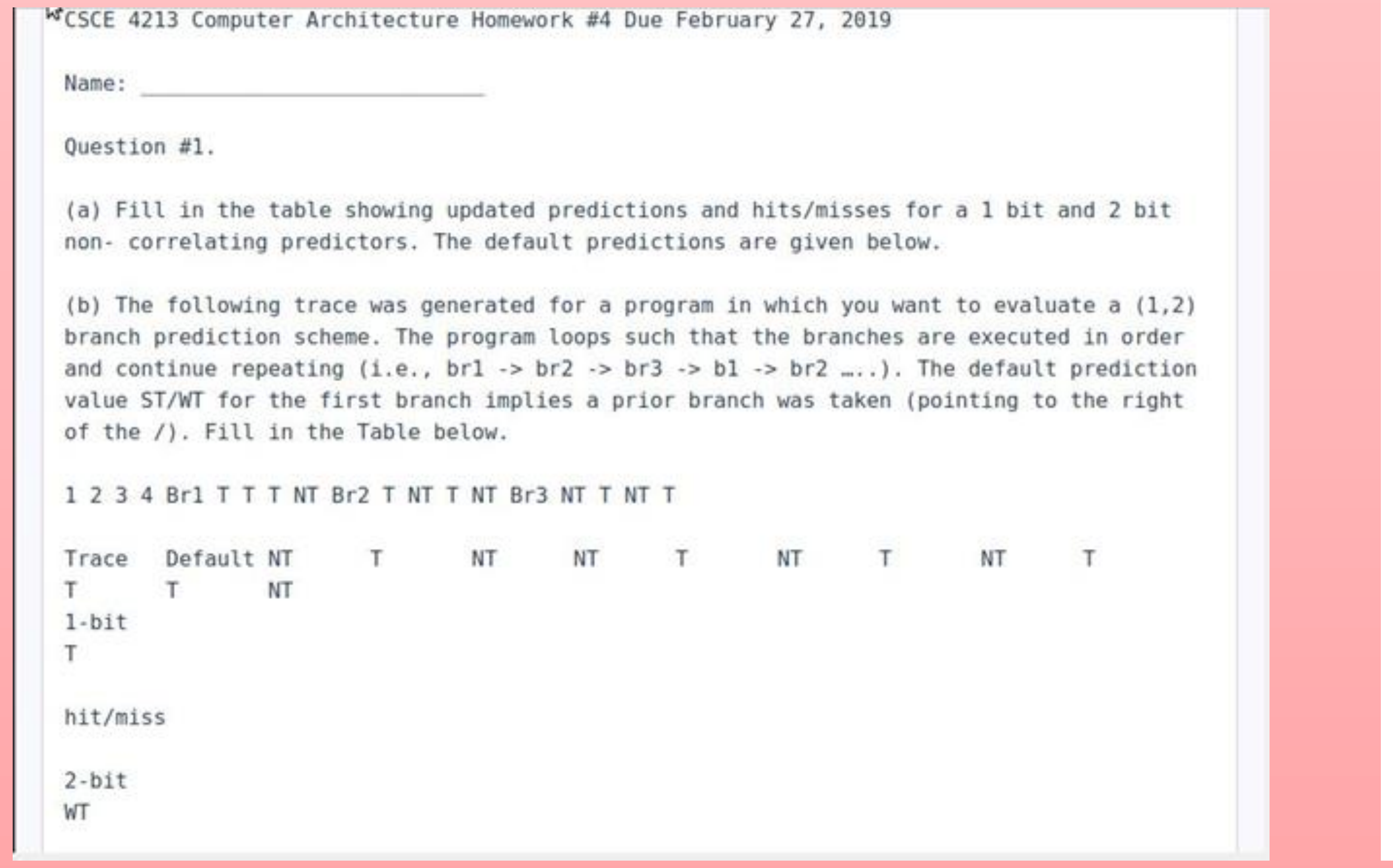


Figure 1 (image right)

The main output of Parsr includes the original document to the left, Parsr's output in the center, and the filters and element information to the right. Words are outlined in green, each line in blue, paragraphs in orange, and headers in pink.



```
###### Chapter 4 Network Layer: The Data Plane

A note on the use of these Powerpoint slides:
We're making these slides freely available to all (faculty, students, readers). They're in
PowerPoint form so you see the animations; and can add, modify, and delete slides
(including this one) and slide content to suit your needs. They obviously represent a
*lot* of work on our part. In return for use, we only ask the following:

![](http://localhost:3001/api/v1/image/195aeb279eccb0e38ddaacd8d53a19/9245)

 If you use these slides (e.g., in a class) that you mention their source (after all,
we'd like people to use our book!)
 If you post any slides on a www site, that you note that they are adapted from (or
perhaps identical to) our slides, and note our copyright of this material.

Thanks and enjoy! JFK/KWR

All material copyright 1996-2016
J.F Kurose and K.W. Ross, All Rights Reserved

##################### Computer Networking: A Top Down Approach
```



```
CSCE 4213 Computer Architecture Homework #4 Due February 27, 2019

Name: ____________________________

Question #1.

(a) Fill in the table showing updated predictions and hits/misses for a 1 bit and 2 bit
non- correlating predictors. The default predictions are given below.

(b) The following trace was generated for a program in which you want to evaluate a (1,2)
branch prediction scheme. The program loops such that the branches are executed in order
and continue repeating (i.e., br1 -> br2 -> br3 -> b1 -> br2 …..). The default prediction
value ST/WT for the first branch implies a prior branch was taken (pointing to the right
of the /). Fill in the Table below.

1 2 3 4 Br1 T T T NT Br2 T NT T NT Br3 NT T NT T

Trace   Default NT     T      NT     NT     T      NT     T      NT     T
T       T       NT
1-bit
T

hit/miss

2-bit
WT
```

Other Solutions

XpdfReader is another open source PDF reader.

PyPDF4 is a python based PDF library.

Textricator is a tool that extracts text from documents.

Apache PDFBox is an open source Java tool for PDF extraction.

Conclusion

When it comes to PDF translation, there are many methods that can be used in conjunction to get the best output. More research is needed to develop a solution for Sorcero's specific needs. However, Parsr fills very many of those needs. It combines the parsing capabilities of multiple different parsing programs to create a more powerful application.

The programs and methods we looked at in our research have their own pros and cons. Overall, the area of easy pdf translation that can accurately represent the original data is one that needs more research and development.