IST 510 - Team 10 Report - PDF parsing

Hongjian Wang [hxw186@ist.psu.edu](mailto:hxw186@ist.psu.edu)

Dongpeng Xu [dux103@ist.psu.edu](mailto:dux103@ist.psu.edu)

Mohammad H. Habibi [mzh191@ist.psu.edu](mailto:mzh191@ist.psu.edu)

April 12th, 2014

# 1. Progresses

1. We have implemented a pdf extractor using perl. First, a tool named pdftotext is used to generate formatted text file from the given pdf document. Then we feed the text file into our extractor, which matches and extracts specific character sequences based on regular expression. Currently, it can get information like company name, phone number, price and time limitation. Below is the screenshot of a snippet of the output.



1. We have specified the interface between the extractor and the database server. The table will contain columns as: <id>, <comany\_name>, <tel\_num>, <price>, <lasting\_time>

# 2. Future Work

1. Our extractor now can handle around 70% of the information in the pdf file. Also, there are some redundant information in the output. Next, we will improve the extractor so as to increase the accuracy and cover rate.
2. Once the output of the extractor is fixed, we will begin build the web server and database.