Martech Invoice Scraper

ISYS540
Dr. Gove Allen
By: Jorel Perez

Situation

The Martech invoice scraper is a macro enabled word document that programmatically scrapes invoices and inserts all the scraped information into an Access database. Martech Computers is a small IT and computer shop in the metro Atlanta area. The company has been in business since the early 90s and has invoices written in standard Microsoft Word Documents. Unfortunately for the company, the rate of business has not allowed them to implement a more sophisticated method of building invoices or the ability to look at customer and service history.

Solution

The Martech invoice scraper hopes to solve that by allowing the user to specify the directory where older invoices are located and programmatically going through each document and scraping the invoices for customer, invoice, and sale items information and putting them into an Access database.

Process

The user opens the word document and clicks on the "Scrape Invoices" button located near the top.

Figure 1: Scraper at Start

The scraper then presents the user with an input box for entering the directory path where the invoices are saved.

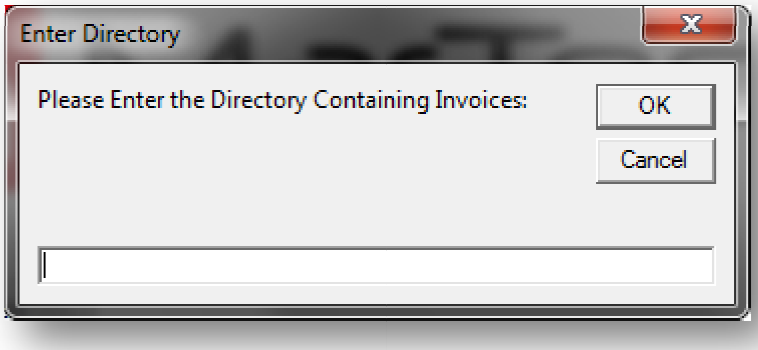

Figure 2: Directory Prompt

Once the user has entered the directory and clicked "OK" the program goes through the directory and creates a collection of all the files in that directory. The program then begins a loop to go through each invoice and scrapes all the information on the document marked with a percent (%) character. This percent character is used to manually show where the information is on the document due to changes in invoice formatting during the time the company has been in business.

The database the program connects to is called Martech and is in the same file path as the scraper in order to avoid complexities. The program connects to the database using the ADO connection method.

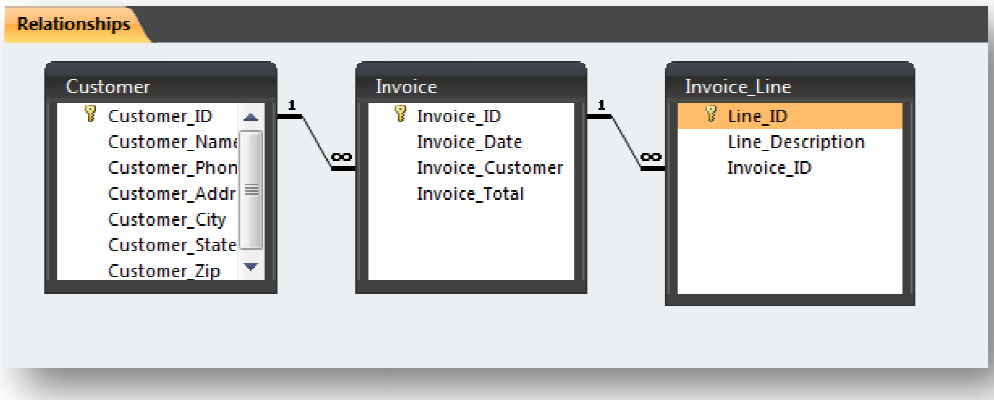

Figure 3: Database Structure

After gleaning the information off of the invoice the program then goes through the process of adding customers that do not currently exist in the Access database. If the customer does not exist then the program goes ahead and creates the customer in the database then proceeds to enter in the invoice information into the database. Unfortunately for the accuracy of this program the company to this day has not implemented an invoice numbering system, so the

invoices cannot truly be tracked for uniqueness unless a full analysis of the sale items is performed.

Within a short while the program finishes entering in the invoice information and then starts entering all the sale items that make up the invoice with their description. This finishes the loop and continues to update any statistics that are gathered by the program before moving on to the next invoice.

The final results of the program are the entries in the database and a small report on the statistics gathered by the program such as the time it took to process all the invoices, the number of invoices processed, the number of new customers added to the database and the total revenue from the invoices added. This report allows the user verify that the correct number of invoices was processed, that the invoices were scraped correctly, and to be able to estimate how long jobs with numerous invoices will take.

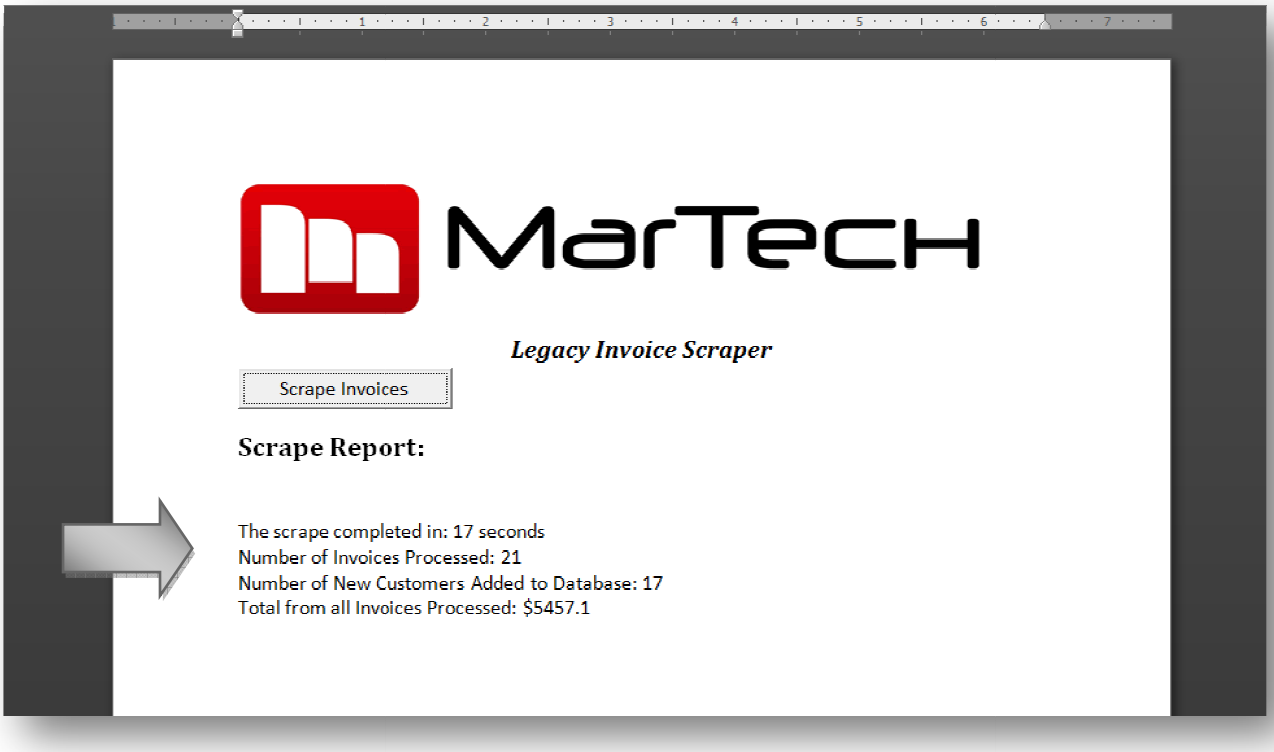

Figure 4: Results View

Conclusion

Even though the ADO connection to Access, the inconsistent invoice formats, and the difference in typing out information using Word instead of Excel made this project challenging, I believe that this program will be a great asset to Martech Computers and be able to give the company a new edge to their market strategy. Information is key when determining how a company should move forward and this will give the company all the information they need.