Solution for Data Parsing

A data parsing tool which culls the loan level data from an electronic PDF and provides you periodic reports at tranche level along with the delinquency details and mortgage indices

Our client provides on demand web based data, analytics, risk management and trade management services spanning all fixed income products. The client has built a robust analytics and risk management platform designed for structured financial market. The client's data service collects remittance and loan level data for thousands of deals. Utilizing our client's server farm, several of the top CDO Managers run loan level analysis of complex ABS and CDO structures.

Problem

The client needed to extract meaningful data from 80 ABS Index deals in PDF files and create comparative data across deals for several parameters such as delinquency at tranche level. They also needed to compare the data with that of previous months and publish the information on the web.

It took over 800 minutes to extract data from 80 deals. In addition, since it was a manual process, several errors were committed which took additional effort and time to rectify before the data could be published in a usable format. If 1000 such deals were to be processed, it would take approximately 167 hours or 7 days for a team of 3 experts with a possibility of several errors.

Solution

Technically the problem essentially was to read data from PDF files and save it in a properly referenced manner. However, the PDFs contain string literals and such other parameters as height and weight. By defining the string literals, PDF achieves device independence. The size of the rectangles has significance and there can be rectangles inside another larger rectangle.

To be able to read the string literals with their properties and rectangles and convert them into digital data, our team came up with a parsing algorithm. The algorithm used transformation matrices to receive coordinate space of the data. The data region or table area was detected based on the theory of connected lines. If a rectangle lies inside another rectangle then the latter is called the parent rectangle. The region growing principle was applied to the parent rectangle.

Every data region in the relevant PDF page is scanned for the literals available, which help in formation of rows. These rows constitute a table. All these rows are then separated into various columns, which help retrieving the relevant headers and the respective data. The column headers are then compared to a mapping wherein the required fields are mentioned. This mapping can be converted and saved onto electronic media. With the back propagation approach the data fetched from the PDF file is mapped to appropriate fields. Finally, the properly mapped data is saved into relational databases.

The algorithm was also extended to read excel files. The challenge in reading excel files was that the files are of a different format and a generic engine is required to read all types of formats. This was resolved by introducing the concept of templates which can guide the engine on how to read an excel file of a specific format. When the engine is made available with the file for processing, the target file routes the engine to one particular template. The template subsumes the information for reading the imperative data.

This algorithm would take less than three hours on one machine or less than one hour on three machines for a team of three people to cull out, process and convert the data from a 1000 PDF files into database files. The algorithm parses the electronic files by using the concept of indexing. Indexing refers to the act of finding the key data from the file format and locating the related data based on the index.

A new solution is born

With the development of the parsing algorithm a new patent pending solution was born. The uniqueness of the solution is the conversion of electronic data to digital format. It enables taking data in a format where only a human can interpret it and converting it to a format where a machine can interpret it. The significance is that once data is available in a machine interpretable manner or in a digital format, it can be analysed using machines to generate business intelligence.

Practical applications of the solution

The solution proved best during the creation of monthly deal data for trading purposes. When hundreds of deals were downloaded the data required by the dealers and brokers are fetched by the converter and viewed online, based on which the advisory body takes a decision. The solution also provides the comparison rates of the previous and subsequent months for a better insight of the situation. The deals and individual reports for selected months can also viewed. The ratings are shown with their change values with respect to the previous month rating. Some of the data is available at www.dataspan.in