Deloitte.

Deloitte's table extraction tool: TableMiner

Saving Time for Deeper Analysis

Deloitte's table extraction tool "TableMiner" reproduces tables from unstructured (pdf) documents into spreadsheets, taking all-too-common dirty work out of daily life of the analyst.

The Need

Sound analysis is based on data... generally, the more, the better. Modern organizations increasingly rely on machines to analyze large volumes of data. This data must be structured, i.e. in the form of tables and databases that can be programmatically queried. The data age has ushered in widespread availability of structured data. Often, but not always. Some data remains "unstructured" – buried within narrative of reports, or inserted as tables within published (digital) documents. The data may be available, yet it is not easily accessible for machine-enabled analysis.

The ubiquitous Portable Document Format (PDF) guarantees formatting consistency and in a generally compact filesize. It is also notoriously unhelpful to those seeking to extract tabular data from its contents. This difficulty lies in the fundamental design of PDFs to be easy on the eyes. Unlike other formats (MS or other Office formats), which store tabular data explicitly as embedded tables, PDFs store tables and text as vector graphics. Converting content to graphics preserves formatting at the cost of removing context: any formatting and structure is lost when copying and pasting text out of a PDF document. Already a problem with e-documents (Office documents) saved as PDFs, scans saved as PDFs without embedded OCR (optical character recognition) are even more unwieldy.

The result: analysts are left with few options other than to manually transfer data to editable formats (spreadsheets) – a labor intensive and error-prone process. This binds qualified resources to menial tasks, representing a costly productivity drain, inviting fatigue-related manual errors, and leaving less time for value-added analytical work.

Our Solution: TableMiner

Deloitte's table extraction tool "TableMiner" addresses this very issue, joining multiple Computer Vision and Natural Language Processing methods to provide an easy solution to an all too common problem.

TableMiner's neural networks scan each page for tabular data – irrespective of whether the document contains only a single or hundreds of tables in various formats and styles, even multiple per page. Once identified, tables are then automatically extracted and converted into a specified format, directly viewable in the TableMiner application or downloaded and viewed in a separate (MS or other) spreadsheet application.

It deftly handles so-called "dirty" scans without OCR – meaning: only a picture, no associated text meaning. TableMiner can automatically distinguish between e-documents saved as PDF, "clean" scans (with OCR) and "dirty" scans (without OCR). Finding a "dirty" scan, TableMiner first applies state-of-the-art OCR techniques: scanned tables are partitioned into smaller sub-boxes and characters are digitized. In other words, TableMiner "reads" the document and saves its meaning. TableMiner then summarily reconstructs the extracted information to form a text version of image.

TableMiner offers a convenient graphical user interface for the user to selectively search for and extract targeted tables. For larger jobs, TableMiner's batch processing feature saves valuable time, allowing the user to upload multiple documents, determine output format and let TableMiner get to work, automatically identifying and extracting all tables within the uploaded documents.

Fig. 1 – Proper recognition of columns

Original PDF

This is a sample unstructured document ...

	A	B	C	D
Rev	5	6	7	8
Cost	2	3	4	4
Profit	3	3	3	4

... with a table inserted within the text.

Without TableMiner

Manual copy-pasting from the PDF results in all the columns being concatenated ...

ABCD
Rev5678
Cost2344
Profit3334

... requiring as much, if not more work to clean than re-typing.

With TableMiner

The grid positioning of the values within the table remains intact ...

	A	B	C	D
Rev	5	6	7	8
Cost	2	3	4	4
Profit	3	3	3	4

... transferred to a spreadsheet for further analysis.

Advantages/Benefits

- Shifts the analyst focus to what really counts: analysis vs data collection and aggregation.
- Reduced transmission error
- Automatically extracts tables from hundreds of documents via batch-processing.
- Reliably handles different table formats and types of PDF documents.
- Scanning throughout entire document
- Easy integration with existing applications and workflows via the TableMiner API.
- Can be hosted on the cloud for subscription service or implemented locally with client firewall.

Example Use Cases

- Facilitating balance sheet analysis (e.g. for underwriting SME/corporates).
- Various audit functions.
- Technical accounting/extraction of terms form contracts for input to systems.
- Extension of RPA capabilities.
- Exhaustive audit.
- Creating new and perfecting existing workflows: For example, a setup that directly forwards scanned documents to TableMiner via the API and stores a copy of the extracted tables.

Contacts

David Thogmartin
Leader aiStudio
dthogmartin@deloitte.de

Issue 05/2020