Semi-Structured Data Parsing

IDENTIFY, EXTRACT AND ANALYZE DATA FROM MEDICAL, FINANCIAL, AND LEGAL DOCUMENTS

Text documents generally contain structured and unstructured data. **Structured data** points “live” at defined locations that analytics tools use to locate them. **Unstructured data** requires natural language processing to identify and structure for further analysis.

But not all documents are cleanly structured or unstructured. Some contain **unstructured text inside structured elements**, including tables and headers, which add important context and nuance to the data. Meanwhile, other documents have **structured elements hidden in unstructured text**, such as lists written as sentences.

Think about legal contracts, financial documents and medical records. These are written in text but have structured sections that add context to the words within. For example, it’s useful to know that a land deed mentions a company, a person and a bank. But it’s even more valuable to know that these are brought up in the context of “Lender,” “Borrower” and “Trustee.”

Natural language processing (NLP) systems excel at analyzing unstructured text but don’t account for how structure of a document influences the data within. **Business intelligence** (BI) tools are built to analyze structured data but fall short with unstructured text because each sentence looks like one very large datum.

**Semi-structured documents** are left in a “dead zone” where both BI and NLP tools fall short. Lexalytics solves this problem by combining a semi-structured data parser with natural language processing.

**Figure 1** | We use semi-structured data parsing to identify, extract and structure data from financial documents

**Figure 2** | Example of a semi-structured document: real estate purchase agreement
Applications of Semi-Structured Data Parsing

- Extract and structure data from financial Statement of Advice documents for auditors to review
- Extract relevant information from EHRs to improve clinical decision making and revenue cycle management
- Flag input errors and suspicious financial recommendations for an auditor to review
- Stay up-to-date with regulatory updates and changes in healthcare diagnostic and billing codes

DATA TO EXTRACT:

- Medical codes
- Contract roles
- Stock ticker symbols
- Illnesses
- Disclaimers
- Subscription details
- Deadlines
- Age ranges
- Products
- Addresses
- Order numbers
- Disclosures

SEMI-STRUCTURED DOCUMENTS:

- Contracts
- Regulatory updates
- Research papers
- Financial documents
- Market reports
- News articles
- SEC filings
- Requests for Proposal (RFPs)

SEMI-STRUCTURED DATA PARSING PROCESS

1. Evaluate structure
   We use our semi-structured data parser to evaluate the underlying structure of your medical, financial or legal documents.

2. Extract structured data
   Our parser identifies and extracts already-structured textual and alphanumeric data.

3. Process unstructured data
   Our parser extracts unstructured data contained within unstructured elements. Then we use NLP to process and structure it.

4. Analyze or export
   You use our NLP on this structured data for further analysis, export it to another business intelligence tool, or work with us to build some sort of custom output based on what you want to do with the data.

Our semi-structured data parser is a set of python code that we can “teach” to understand the structure of a document, such as land deeds, Medicaid updates, or SEC filings. Sometimes this is as simple as writing a few software rules. Other times, we train machine learning models and combine them with rules. Either way, the parser learns how to break down each document down into its component sub-structures, including tables, headers and lists. Then we use API calls to our Salience text analytics engine to identify, extract and structure all of the textual and alphanumeric data within those elements into a usable format.

Once the data is prepared, you can do whatever you want with it: use our NLP engine for further analysis, export it into another business intelligence tool, or work with us to build some sort of custom output. By combining our technologies in this way, we account how the structure of a document influences the data within it. This approach unlocks the full value of semi-structured documents and opens up new possibilities in regulatory compliance and other fields.

FURTHER READING:

- Semi-Custom Applications
- AI for Regulatory Compliance
- lexalytics.com/resources

Figure 3 | Lexalytics uses semi-structured data parsing and NLP to transform previously-inaccessible information into usable data