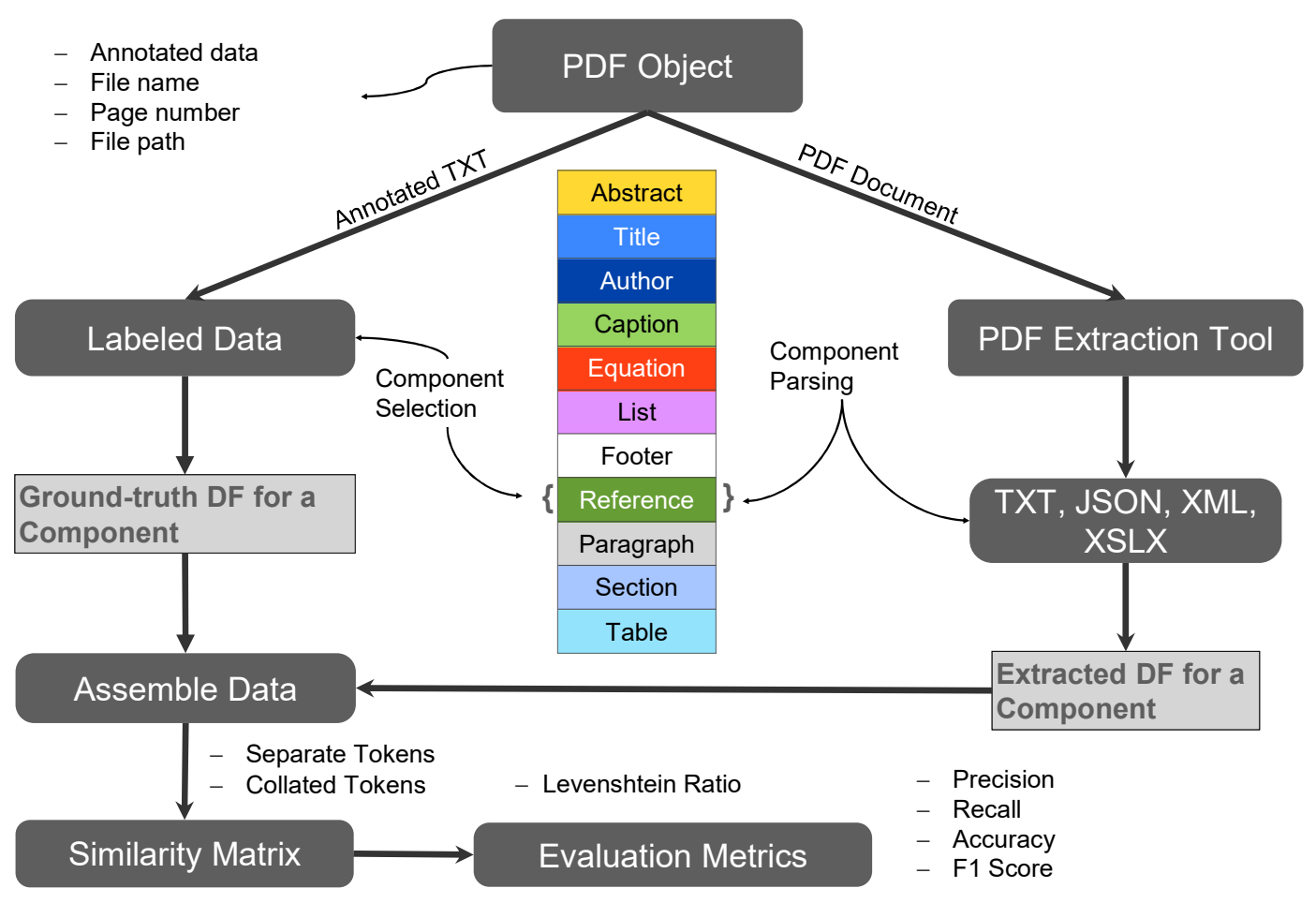
– Annotated data
– File name
– Page number
– File path
PDF Object
Annotated TXT
PDF Document
Abstract
Title
Author
Caption
Equation
List
Footer
Reference
Paragraph
Section
Table
Labeled Data
Component Selection
Component Parsing
PDF Extraction Tool
Ground-truth DF for a Component
TXT, JSON, XML, XSLX
Assemble Data
Extracted DF for a Component
– Separate Tokens
– Collated Tokens
– Levenshtein Ratio
– Precision
– Recall
– Accuracy
– F1 Score
Similarity Matrix
Evaluation Metrics