Contents

1 Current Extraction Mode Data Flow

1.1 All Three Modes (After Recent Unification)

1.1.1 Entry Point: `_load_annotated_from_couchdb()`

All three extraction modes now use the same entry point in classifier_v2.py:888.

```python
def _load_annotated_from_couchdb(self) -> DataFrame:
    database = self.couchdb_training_database or self.couchdb_database

    # Branch based on extraction_mode
    if self.extraction_mode == 'section':
        return self._load_sections_from_couchdb(database=database)
    else:
        # Line and paragraph modes
        conn = CouchDBConnection(...)
        df = conn.load_distributed(self.spark, pattern)
        parser = AnnotatedTextParser(line_level=self.line_level)
        return parser.parse(df)
```

1.2 Line/Paragraph Mode Flow

1. **Load**: `CouchDBConnection.load_distributed()`
   - Returns: DataFrame(doc_id, attachment_name, value=full_annotated_text)
2. **Parse**: `AnnotatedTextParser.parse()`
   - Extracts YEDDA annotations from full text
   - Splits into lines (line_level=True) or paragraphs (line_level=False)
   - Returns: DataFrame(doc_id, attachment_name, value, label, section_name, [line_number])

1.3 Section Mode Flow

1. **Extract**: `PDFSectionExtractor.extract_from_multiple_documents()`
   - Loads PDF/text attachments from CouchDB
   - Extracts text using PyMuPDF or form-feed parsing
   - Parses into sections/paragraphs
   - Extracts YEDDA annotations
   - Detects section headers
   - Returns: DataFrame(doc_id, attachment_name, value, line_number, page_number, section_name, label, ...)

1.4 Key Differences

1.4.1 Line/Paragraph Modes

- **Input**: Pre-annotated text files (.txt.ann)
- **Processing**: Textual parsing only
- **Uses FileObject**: Yes (indirectly through CouchDBConnection internals)
- **YEDDA parsing**: AnnotatedTextParser

1.4.2 Section Mode

- **Input**: PDF files or text with form feeds
- **Processing**: Structural extraction (sections, pages) + textual parsing
- **Uses FileObject**: No (uses PyMuPDF and custom parsing)
- **YEDDA parsing**: PDFSectionExtractor (during extraction)

1.5 Current Status

After recent fixes: ✅ All modes enter through same method (`_load_annotated_from_couchdb`) ✅ All modes return similar DataFrame schemas ✅ Line numbers are preserved in all modes ✅ YEDDA labels are extracted in all modes

1.6 Remaining Inconsistency

The FileObject interface (`fileobj.py`, `couchdb_file.py`, `line.py`) is used for line/paragraph modes but NOT for section mode.

Section mode uses its own extraction logic in PDFSectionExtractor which: - Directly processes PDF bytes via PyMuPDF - Parses text structure (sections, paragraphs, pages) - Handles YEDDA annotations during extraction

This is functionally correct but architecturally inconsistent with line/paragraph modes.