Team Oxidized

Sidharth Kumar Singh **Pranay Buradkar** **Maneesh Shukla**

Technical Solution Brief

1A] PDF Heading Extraction

- **What We Did:**
  Developed a tool extracting and classifying PDF headings using **pdfminer.six** and **DecisionTreeClassifier**.
- **How We Solved It:**
  1. Extracted styled **text spans with metadata**
  2. Engineered features: **font size, styling, numbering, casing**
  3. Trained on **1,500+** labeled spans
  4. Built hierarchical JSON outlines
- **Accuracy:**
  Achieved **97.3%** classification accuracy, surpassing prior **research**.
- **Conclusion:**
  A robust, Dockerized solution for high-accuracy PDF heading extraction.

Example Output:

```
"title": "COP 3330 Object Oriented Programming",
 "outline": [
    {
        "level": "H2",
        "text": "SYLLABUS",
        "page": 1
    },
    {
        "level": "H1",
        "text": "COP 3330 Object Oriented Programming",
        "page": 1
    },
    {
        "level": "H2",
        "text": "Summer 2012",
        "page": 1
    },
    {
        "level": "H3",
        "text": "Office Hours:",
        "page": 1
    },
    {
        "level": "H3",
        "text": "Course Description:",
        "page": 1
    },
    {
        "level": "H3",
        "text": "Prerequisite: COP 3223 or EGN 3211",
        "page": 1
    },
    {
        "level": "H3",
        "text": "Corequisites: NONE",
        "page": 1
    },
    {
        "level": "H3",
        "text": "Course Objectives:",
        "page": 1
    }
]
```

1B] PDF Section Ranking & Analysis

- **What We Did:**
  Built a **semantic analysis** tool ranking PDF sections by relevance to persona/job using **SentenceTransformers**.
- **How We Solved It:**
  1. Used 1A module for structure extraction
  2. Flattened content and **aggregated paragraphs**
  3. Generated embeddings via **MiniLM-L6-v2 (384d)**
  4. Computed **cosine similarity** against persona-job query
  5. Ranked **top 10** relevant sections in JSON
- **Accuracy & Performance:**
  Embedding benchmark: 58.8/100 average
  Speed: 14,200 sentences/sec (80 MB model)
- **Conclusion:**
  Efficient Dockerized pipeline for targeted semantic PDF analysis, reducing manual review.

Example Output:

```
{
  "metadata": {
    "input_documents": ["file01.pdf", "file05.pdf", "201.pdf", "..."],
    "persona": "A software engineer with expertise in agile methodologies",
    "job_to_be_done": "Comprehensive analysis on agile testing methodologies",
    "processing_timestamp": "2025-07-27T09:12:36Z"
  },
  "extracted_sections": [
    {
      "document": "file02.pdf",
      "section_title": "2. Introduction to Foundation Level Agile Tester Extension",
      "importance_rank": 1,
      "page_number": 7
    },
    {
      "document": "file02.pdf",
      "section_title": "Introduction to Foundation Level Agile Tester Extension",
      "importance_rank": 2,
      "page_number": 4
    },
    { "...": "..." }
  ],
  "subsection_analysis": [
    {
      "document": "file02.pdf",
      "refined_text": "The certification for Foundation Level Extension Agile Tester is designed for professionals who are working within Agile environments. It is also for professionals who are planning to start implementing Agile methods in the near future, or are working within companies that plan to do so, The certification provides an advantage for those who would like to know the required Agile activities, roles, methods, and methodologies specific to their role.
",
      "page_number": 7
    },
    { "...": "..." }
  ]
}
```