Round 1B: Persona-Driven Document Intelligence - Approach Explanation

Methodology Overview

Our solution implements a multi-stage intelligent document analysis system that adapts to different personas and their specific job requirements. The approach combines traditional text processing with relevance scoring algorithms to extract and prioritize the most pertinent information.

Core Architecture

1. Document Preprocessing

The system begins by extracting structured content from PDFs using PyMuPDF, capturing not just text but also metadata like font sizes, formatting, and positional information. This rich extraction allows us to identify logical document sections and understand content hierarchy.

2. Section Detection and Extraction

We implement a sophisticated heading detection algorithm that combines multiple signals:

- **Font Size Analysis**: Identifies headings by comparing font sizes against document averages
- **Formatting Cues**: Detects bold text, capitalization patterns, and numbering schemes
- **Pattern Recognition**: Uses regex patterns to identify common heading structures across domains
- **Contextual Analysis**: Considers section length and content density for validation

3. Relevance Scoring Engine

The heart of our system is a multi-factor relevance scoring algorithm that evaluates each section against the persona and job requirements:

Persona Matching (30% weight): Extracts keywords from persona description and measures their frequency in section content. This ensures content alignment with the user's expertise and background.

Job Alignment (40% weight): Analyzes job-to-be-done requirements and prioritizes sections containing related concepts and terminology. This receives the highest weight as it directly addresses the user's immediate needs.

Content Quality (20% weight): Evaluates section length and substance, favoring comprehensive content over fragmented text snippets.

Structural Importance (10% weight): Provides bonus scoring for identified headings and key sections, recognizing that well-structured content often contains crucial information.

4. Subsection Analysis

For top-ranking sections, we perform granular analysis by:

- Segmenting content into logical paragraphs
- Applying relevance scoring at the paragraph level
- Refining and cleaning text for optimal readability
- Ranking subsections by their specific relevance to the persona-job combination

Key Innovations

Adaptive Keyword Extraction

Our system dynamically extracts domain-specific keywords from persona descriptions, allowing it to adapt to diverse professional backgrounds without manual configuration. This ensures relevance across academic, business, and technical domains.

Multi-Domain Generalization

The solution handles varied document types (research papers, financial reports, textbooks) by using flexible content patterns and avoiding domain-specific assumptions. This generalization is crucial for the hackathon's diverse test cases.

Hierarchical Content Understanding

By analyzing document structure at multiple levels (sections, subsections, paragraphs), the system provides both broad overview and detailed insights, matching how professionals actually consume information.

Technical Implementation

Performance Optimization

- **Efficient Text Processing**: Single-pass document analysis minimizes I/O operations
- **Vectorized Operations**: NumPy-based calculations for fast relevance scoring
- **Memory Management**: Processes documents sequentially to maintain memory efficiency
- **Selective Analysis**: Focuses computational resources on highest-ranked content

Scalability Design

The architecture supports processing multiple documents simultaneously while maintaining consistent performance. The scoring algorithm scales linearly with document count, ensuring reliable performance within the 60-second constraint.

Robustness Features

- **Error Handling**: Graceful degradation for corrupted or unusual PDF formats

- **Content Validation**: Filters out noise and irrelevant content automatically
- **Format Flexibility**: Adapts to various document layouts and structures

Expected Outcomes

This methodology consistently identifies the most relevant content for diverse persona-job combinations while maintaining high processing speed and accuracy. The system's ability to understand context and prioritize information makes it particularly effective for professionals who need to quickly extract actionable insights from large document collections.

The approach balances computational efficiency with analytical depth, ensuring it meets the hackathon's technical constraints while delivering meaningful results across the varied test scenarios.