Round 1B: Persona-Driven Document Intelligence

Theme: "Connect What Matters — For the User Who Matters"

Challenge Brief (For Participants)

You will build a system that acts as an intelligent document analyst, extracting and prioritizing the most relevant sections from a collection of documents based on a specific persona and their job-to-be-done.

Input Specification

2. **Document Collection**: 3-10 related PDFs **Persona Definition**: Role description with specific expertise and focus areas
3. **Job-to-be-Done**: Concrete task the persona needs to accomplish

Document collection, persona and job-to-be-done can be very diverse. So, the solution that teams need to build needs to be generic to generalize to this variety.

- Documents can be from any domain (Example: Research papers, school/college books, financial reports, news articles etc.)
- Persona can again be very diverse (Example: Researcher, Student, Salesperson, Journalist, Entrepreneur etc)
- Job-to-be-done: This will be related to the persona (Example: Provide a literature review for a given topic and available research papers, What should I study for Organic Chemistry given the chemistry documents, Summarize the financials of corporation xyz given the detailed year end financial reports etc.)

Sample Test Cases

Test Case 1: Academic Research

- **Documents**: 4 research papers on "Graph Neural Networks for Drug Discovery"
- **Persona**: PhD Researcher in Computational Biology
- **Job**: "Prepare a comprehensive literature review focusing on methodologies, datasets, and performance benchmarks"

Test Case 2: Business Analysis

- **Documents**: 3 annual reports from competing tech companies (2022-2024)
- **Persona**: Investment Analyst
- **Job**: "Analyze revenue trends, R&D investments, and market positioning strategies"

Test Case 3: Educational Content

- **Documents**: 5 chapters from organic chemistry textbooks
- **Persona**: Undergraduate Chemistry Student
- **Job**: "Identify key concepts and mechanisms for exam preparation on reaction kinetics"

Required Output

- Output JSON format: Refer *challenge1b_output.json*

The output should contain:

1. Metadata:
    a. Input documents
    b. Persona
    c. Job to be done
    d. Processing timestamp
2. Extracted Section:
    a. Document
    b. Page number
    c. Section title
    d. Importance_rank
3. Sub-section Analysis:
    a. Document
    b.
    c. Refined Text
    d. Page Number **Constraints**

- Must run **on CPU** only
- Model size **≤ 1GB**
- Processing time ≤ 60 **seconds** for document collection (3-5 documents)
- **No internet access** allowed during execution

Deliverables

- approach_explanation.md (300-500 words explaining methodology)
- Dockerfile and execution instructions

Sample input/output for testing

-

Scoring Criteria

Criteria	Max Points	Description
Section Relevance	60	How well selected sections match persona + job requirements with proper stack ranking
Sub-Section Relevance	40	Quality of granular subsection extraction and ranking

Appendix:

https://github.com/jhaaj08/Adobe-India-Hackathon25.git