Python Installations required to make sure the python parser microservice to run properly

Scope of this Microservice:

Text extraction

Section parsing

Structured JSON output

Ready for Java + React flow

1️⃣ System-level installations (MANDATORY)

Linux (Ubuntu / Debian)

sudo apt-get update

sudo apt-get install -y tesseract-ocr libtesseract-dev

Mac (local development)

brew install tesseract

Needed for **OCR on JPG / scanned PDFs**

2️⃣ Python runtime

- **Python 3.11+** (required)

Verify:

python3 --version

3️⃣ Python dependencies (from requirements.txt)

Installed via:

pip install -r requirements.txt

Includes:

- fastapi – API framework

- uvicorn – ASGI server
- boto3 – S3 access
- PyMuPDF – PDF text extraction
- python-docx – DOCX parsing
- pytesseract – OCR engine binding
- Pillow – image handling
- python-dateutil – date parsing
- pydantic – request/response models

4️⃣ Docker (RECOMMENDED for prod)

- Docker Engine
- Docker Compose (optional)

Verify:

```
docker --version
```

5️⃣ AWS requirements

IAM Role / Credentials

Python service needs **read-only access** to resume files:

Required permission:

```
s3:GetObject
```

If running locally:

```
export AWS_ACCESS_KEY_ID=...
export AWS_SECRET_ACCESS_KEY=...
export AWS_REGION=us-east-1
```

If running on ECS/EKS:

- ❌ No keys needed
- ✅ Use IAM role

6️⃣ Network / Infra

- Internal access to:
    - Python service (/v1/parse)
- No DB connectivity required
- No Redis connectivity required

7️⃣ Optional (but recommended)

- curl or Postman (API testing)
- make (automation)
- git (version control)

8️⃣ Verification commands

uvicorn app.main:app --port 8080

curl http://localhost:8080/health

curl http://localhost:8080/ready