Data Scientist - Technical Assignment

Please complete the assignment below. Organize your code in a Jupyter notebook where your code is readable and sufficient documentation is added. Finally, upload your final solution to a public GitHub repository and share the link with us.

Note that you are allowed to use available AI tools or search engines to complete the assignments; however, you are responsible for the submitted solutions and are expected to explain your technical choices and answer any follow-up questions.

Assignment

We are developing a generic **Resume Parser** solution that supports ***two file formats (PDF and Word)***. Given a resume, the parser is expected to extract ***name***, ***email*** and ***skills*** into a structured JSON object:

```
{
        “name”: string,
        “email”: string,
        “skills”: List[string]
}
```

For example:

```
{
        “name”: “Jane Doe”,
        “email”: “jane.doe@gmail.com”,
        “skills”: [“Machine Learning”, “Python”, “LLM”]
}
```

Create a Jupyter Notebook to demonstrate your development process for this solution. You should cover key aspects such as data gathering and processing, algorithm design, and evaluation. Ensure the notebook is highly readable and that all cells run successfully. Incorporate data science best practices wherever possible, as this exercise will be used to evaluate how you would work as a data scientist within the team.

Notes:

- Feel free to gather resume samples from the internet. For this exercise, 4~10 resumes would be fine.
- Feel free to use LLMs. You are not required to fine-tune any models. Gemini provides free API tiers for low usage (link), but you are not restricted to that – be creative. **Do not include your API key into the public GitHub repo**.