Report: Matching Resumes to Job Descriptions

❖ Introduction:

The goal of this project is to match resumes (CVs) to job descriptions by calculating the similarity between the content of each resume and each job description. This automated matching process aims to identify the top 5 candidates for each job description based on their similarity scores.
Approach:

❖ Data Collection:

We obtained two key datasets for this task: job descriptions and resumes (CVs).
Job descriptions were sourced from a CSV file containing "job_description" data.
Resumes were sourced from another CSV file containing "Resume_str" data, along with corresponding job categories ("Category").

❖ Text Preprocessing:

The text from both job descriptions and resumes underwent preprocessing:
Tokenization: Text was tokenized into smaller units, usually words or subwords, using the DistilBERT tokenizer.
Padding and Truncation: Text sequences were adjusted to a maximum length of 512 tokens to match the model's input requirements.

❖ Embedding Calculation:

DistilBERT, a pre-trained transformer model, was utilized to calculate embeddings (numerical representations) of the text data.
These embeddings capture the semantic meaning of the text.

❖ Similarity Calculation:

Cosine similarity was computed between the embeddings of job descriptions and resumes.
Cosine similarity is a common metric for measuring the similarity between vectors and ranges from -1 (dissimilar) to 1 (similar).

❖ Top Candidate Selection:

The top 5 candidates for each job description were determined by selecting resumes with the highest cosine similarity scores.
Challenges Faced and Solutions:

❖ PDF Text Extraction:

Extracting text from PDF files posed a challenge due to variations in PDF formats.
We used the PyPDF2 library to extract text, but some PDFs with complex layouts may not be fully supported.

❖ Performance Optimization:

Calculating embeddings and similarity scores for a large number of job descriptions and resumes can be computationally intensive.
We utilized GPU acceleration (if available) to speed up processing and optimize code for efficiency
Report: Matching Resumes to Job Descriptions
Introduction:
The goal of this project is to match resumes (CVs) to job descriptions by calculating the similarity between the content of each resume and each job description. This automated matching process aims to identify the top 5 candidates for each job description based on their similarity scores.
Approach:

❖ Data Collection:

We obtained two key datasets for this task: job descriptions and resumes (CVs).
Job descriptions were sourced from a CSV file containing "job_description" data.
Resumes were sourced from another CSV file containing "Resume_str" data, along with corresponding job categories ("Category").

❖ Text Preprocessing:

The text from both job descriptions and resumes underwent preprocessing:
Tokenization: Text was tokenized into smaller units, usually words or subwords, using the DistilBERT tokenizer.
Padding and Truncation: Text sequences were adjusted to a maximum length of 512 tokens to match the model's input requirements.

❖ Embedding Calculation:

DistilBERT, a pre-trained transformer model, was utilized to calculate embeddings (numerical representations) of the text data.
These embeddings capture the semantic meaning of the text.

❖ Similarity Calculation:

Cosine similarity was computed between the embeddings of job descriptions and resumes.
Cosine similarity is a common metric for measuring the similarity between vectors and ranges from -1 (dissimilar) to 1 (similar).

❖ Top Candidate Selection:

The top 5 candidates for each job description were determined by selecting resumes with the highest cosine similarity scores.

Job Description 1:
Top CV 1: CV 299, Similarity Score: 0.9613177180290222, Category: INFORMATION-TECHNOLOGY
Top CV 2: CV 2225, Similarity Score: 0.9565382599830627, Category: BANKING
Top CV 3: CV 1305, Similarity Score: 0.9551481008529663, Category: DIGITAL-MEDIA
Top CV 4: CV 1235, Similarity Score: 0.9536933302879333, Category: DIGITAL-MEDIA
Top CV 5: CV 577, Similarity Score: 0.9527629613876343, Category: BUSINESS-DEVELOPMENT

❖ Challenges Faced and Solutions:

✓ PDF Text Extraction:

Extracting text from PDF files posed a challenge due to variations in PDF formats.
We used the PyPDF2 library to extract text, but some PDFs with complex layouts may not be fully supported.

✓ Performance Optimization:

Calculating embeddings and similarity scores for a large number of job descriptions and resumes can be computationally intensive.
We utilized GPU acceleration (if available) to speed up processing and optimize code for efficiency

❖ Recommendations and Insights:

The automated matching process effectively identifies potential candidates for each job description based on the similarity between their resumes and the job requirements.
Further fine-tuning of the model and hyperparameters may enhance matching accuracy.
It's essential to ensure that the PDFs used for job descriptions and resumes are well-structured for optimal text extraction.
The top 5 candidates can be reviewed for each job role, and the selection can consider other factors beyond similarity scores, such as relevant experience and skills.