TEXT SIMILARITY SCORING APPLICATION

Introduction:

The purpose of this report is to outline the process of analyzing resumes from job applicants and comparing their content to job descriptions through text preprocessing and cosine similarity calculations. The key steps include obtaining resumes and job descriptions, extracting text from PDFs, performing text preprocessing, converting text to embeddings, and calculating cosine similarity.

1. Data Collection:

The process starts with obtaining resumes from job applicants and the relevant job descriptions. Resumes are typically provided by candidates as PDF files, and job descriptions are typically in text format.

2. Text Extraction from PDF:

To analyze the text content of resumes (PDF files), the pdfminer library can be used. It extracts the text from the PDF, making it available for further analysis.

3. Text Preprocessing:

Text preprocessing is crucial to clean and prepare the text for analysis. The following steps are performed:

Remove Punctuation: Punctuation marks are removed to focus on meaningful words.

Remove Stop Words: Common words (stop words) that don't add much meaning to the text are removed.

Lemmatization: Words are reduced to their base or root form (lemmas) to improve text consistency.

4. Text Embedding:

Text embedding involves converting text data into numerical vectors that can be used for mathematical calculations. Sentence-Transformers is a powerful natural language processing tool that allows efficient comparison of text documents using cosine similarity. This library utilises pre-trained models to transform sentences or paragraphs into dense vectors in a semantic space, enabling semantic understanding and similarity analysis. So, we have used a LLM called all-mpnet-base-v2 from HuggingFace.

5. Cosine Similarity Calculation:
Cosine similarity is used to measure the similarity between two text embeddings. It calculates the cosine of the angle between the two vectors, indicating how closely they are related. A value of 1 means they are identical, while 0 indicates no similarity.

6. Process Overview:

- Obtain resumes and job descriptions.
- Use pdfminer to extract text from resumes (PDFs).
- Preprocess text by removing punctuation, stop words, and lemmatizing.
- Convert preprocessed text to text embeddings using pre-trained models.
- Calculate cosine similarity between resume embeddings and job description embeddings.

7. Benefits and Applications:

- Efficient Resume Screening: This approach helps HR teams quickly screen and shortlist candidates based on the similarity of their skills and experiences to the job requirements.
- Tailored Job Matching: By comparing resumes with job descriptions, companies can provide more targeted job recommendations to candidates.

8. Limitations and Considerations:

- Language Dependency: The approach may be language-dependent, requiring different preprocessing for different languages.
- Contextual Understanding: The method does not consider the context of words in sentences, which may impact accuracy.

9. Conclusion:
By using techniques such as text preprocessing and cosine similarity calculations, it's possible to compare the content of job applicants' resumes to job descriptions objectively. This can lead to more efficient and accurate resume screening processes, helping companies find the best candidates for their roles.