Resume Matching

Objective

To build a PDF extractor to pull relevant details from CVs in PDF format and match them against job descriptions from the Hugging Face dataset.

Approach

I divided the task into three sub-divisions:

- **PDF Extraction**
- **Job Description**
- **Candidate-Job Matching**

PDF Extraction

- Used the Resume Dataset from Kaggle to extract data from PDFs.
- Extracted all text from PDFs and created a CSV file for further use.
- Used various PDF extracting tools to extract details:
  - PyMuPDF: Used to extract images and text from PDFs.
  - PyPDF2: Similar to PyMuPDF but can only extract text.
  - pdfquery: Converts PDFs to XML files and can query text from XML files.

Important insights from data:

- Total Number of Resumes: 2484
- Total Number of Categories: 24
- There are no images in the resumes.
- The average number of pages is 2.

Problems:

- I tried to extract only specific sub-headings like education, summary, etc., using `pdfquery` but couldn't achieve it.
- The PDFs aren't representative samples of real-world resumes, as they don't contain any images and table structures.

Recommendations or Future Work:

- Assign a score to readability and structure in PDFs.
- Use document segmentation tools to extract only certain parts of the resume rather than the entire document.
- Consider using tools like SwinDocSegmenter.

Job Description

- Used the Hugging Face Job Description dataset to fetch job descriptions.
- Selected 10 random samples to work on.
- The dataset contains a column named `model_response`, which is a summary of the job description. I chose to work with these summaries.

Recommendations or Future Work:

- Since real-world data contains only job descriptions, there is a need to build a model that summarizes the job descriptions.

Candidate-Job Matching

- The task boils down to finding semantic similarity between resumes and job descriptions.

Pre-Processing and its problems:

- Started with basic pre-processing steps like checking for NA values and duplicate values in the resume text data and job description dataset.
- Wrote a function to preprocess text, which removes hyperlinks, web links, punctuation, stop words, and converts text to lowercase.
- Applied this function to both resume text and job descriptions.
- There is a problem with the processed resume text; it's lengthy and not in an order that makes it easy to compare with job descriptions.
- So, I came up with a new text preprocessing method for resume text. It processes the text only if it contains any of the keywords: skills, education, responsibilities, and experience. This approach significantly reduced the text length compared to the previous method.

Sentence Transformer:

- Used the Sentence Transformers library to find embeddings.
- Employed three different models to generate embeddings:
  - gtr-t5-large
  - all-mpnet-base-v2
  - all-MiniLM-L12-v2
- Computed cosine similarity to find the similarity between resumes and job descriptions, thereby identifying top-matched resumes for a given job description.

Note:

- I chose not to lemmatize the text as it might remove action words from the resume, which are equally important.
- I initially thought of reducing the search space by filtering resumes based on categories with the help of job positions. However, it would have been a blunder if I had implemented it.
- For example, in the case of Company: Lear Corporation and Position: Customer Service Representative, the top resumes come from healthcare, automobile, and fitness backgrounds.

Recommendations or Future Work:

- For extensive research on a resume, it's equally important to consider hyperlinks and weblinks.
- Training the BERT model on multiple resume and job datasets and using that model to find embeddings could yield much better results.

Experiment Results

Model	Wall Time	Memory
gtr-t5-large	2 min 48 sec	640 MB
all-mpnet-base-v2	1 min 11 sec	420 MB
all-MiniLM-L12-v2	43 sec	120 MB

Note:

- All these experiments were conducted on my local computer.

Top 5 candidates

Company : Volt
Position : Talent Acquisition Specialist / Recruiter

gtr-t5-large:

Category	Id	Similarity
PUBLIC-RELATIONS	13727873	0.854944
HR	30862904	0.848326
HR	73077810	0.847029
HR	25676643	0.839240
HR	11480899	0.838585

all-mpnet-base-v2:

Category	Id	Similarity
HR	30862904	0.809733
HEALTHCARE	17864043	0.777453
HR	19179079	0.765547
HR	18297650	0.763960
HR	17412079	0.744991

all-MiniLM-L12-v2:

Category	Id	Similarity
HR	30862904	0.741534
HEALTHCARE	17864043	0.710526
HR	19179079	0.703621
HR	46258701	0.703336
AUTOMOBILE	23522150	0.701682

The rest can be found in the Jupyter notebook.

References

- https://www.pinecone.io/learn/semantic-search/
- https://www.sbert.net/index.html

Name : Rahulram P
Phone No : +91 9150873896