Automated Resume Evaluation with spaCy and Topic Modelling

Introduction:

Our data science project focuses on the difficulties related to hiring applicants based on resume screening. Our objective is to make the process of selecting candidates easy and quick by implementing technology to automatically evaluate resumes and position requirements. This project is important because it could change how resumes are typically screened. By using technology like machine learning, we hope to make hiring processes simpler, reduce the need for manual work, and make sure everyone has a fair chance of getting a job. To achieve this, it is architecture to employ a tool called Spacy to assist in comprehending resumes and experiment with various methods of text analysis as well. Moreover, incorporating a feature that aligns skills with job criteria shows the complexity of this project. Our project involves working with two separate datasets: one containing a resume and the other comprising skills. These datasets will be used to tailor our analysis tool accordingly.

Hypothesis / Business use:

Our working idea or rationale for this project is that if we use machine learning to automate the screening and shortlisting of resumes, we can make the candidate selection process quicker and easier. This idea shapes our approach to analyzing and modeling data, to simplify recruitment procedures and enhance the effectiveness of hiring choices.

Dataset:

For this project, we obtained the dataset from livecareer.com, which includes over 2400 resume examples. Each resume is labeled according to the job category it corresponds to, with categories such as HR, Designer, information technology, Teacher, and more.

The dataset is structured as follows:

- ID: Unique identifier and file name for each resume in PDF format.
- Resume_str: The resume text in string format.
- Resume_html: The resume data in HTML format as scraped from the website.
- Category: The job category associated with each resume.

The dataset was gathered by scraping individual resume examples from the livecareer.com website. It was then pre-processed to clean and prepare the data for analysis, including handling missing values and ensuring consistency in the format of the resume texts.

Data Exploration:

During the data exploration phase, we will thoroughly analyze the resume dataset obtained from livecareer.com. Key findings will include:

1. **Job Category Distribution:** The dataset will cover various job categories such as HR, Designer, information technology, Teacher, and more. Resumes will be unevenly distributed across these categories.
2. **Resume Length Variation:** Resumes will exhibit variation in length across different job categories, indicating varying levels of detail and emphasis on qualifications.
3. **Common Keywords Identification:** We will identify common keywords and phrases, shedding light on prevalent skills and qualifications emphasized by candidates in different fields.
4. **Visualization:** We will utilize visualizations like histograms, word clouds, and scatter plots to illustrate the distribution of job categories and common keywords, as well as to visualize relationships between resume length and job categories.

Overall, our data exploration will provide insights into the characteristics and patterns present in the resume dataset, guiding subsequent modelling and analysis tasks.

Modelling / Optimization:

For our modelling phase, we will be utilizing the spaCy library, which offers powerful natural language processing tools. We will fine-tune the spaCy model parameters and configurations to ensure accurate classification of resumes into job categories. Additionally, we may explore customizing the model by incorporating domain-specific data or creating custom entity recognition patterns. We aim to optimize the spaCy model's performance to achieve precise and efficient classification results.

Conclusion:

In this project, we will employ an entity ruler to generate extra entities and customize their presentation with unique colors. Moreover, we will display the distributions of categories and skills, allowing users to directly upload resumes with a skill match percentage feature. Additionally, we will utilize LDA for topic modeling and pyLDAvis to visualize the resultant topics. This project provides an invaluable learning experience for me to explore the capabilities of spaCy. Through this project, we expect to uncover various methods to enhance the hiring process by efficiently identifying the most suitable candidates for job positions. Overall, we are eager to apply these techniques to streamline the recruitment process and enhance candidate selection efficiency.