## Automated Resume Analysis

**Executive Summary:**

The Automated Resume Analysis project is a comprehensive endeavor aimed at streamlining the recruitment process through the utilization of advanced Natural Language Processing (NLP) techniques. By leveraging SpaCy for entity recognition and various NLP tools for text analysis, the project endeavors to expedite candidate screening and selection while providing valuable insights into resume content and skill matching.

**Introduction:**

The project's primary objective is to assist recruiters in efficiently navigating through large volumes of resumes by automating the process of entity recognition and skill extraction. Additionally, a skills matching feature has been integrated to aid hiring managers in evaluating candidate suitability based on predefined criteria.

**Dataset:**

The [dataset](https://www.kaggle.com/datasets/snehaanbhawal/resume-dataset) comprises over 2400 resume examples sourced from livecareer.com, encompassing diverse job categories such as HR, Designer, Information Technology, Teacher, and more. Each resume entry includes the resume text in both string and HTML formats, along with the associated job category.

The data acquisition involved scraping individual resume examples from the [livecareer.com](http://www.livecareer.com) website. The [jobzilla skill](https://github.com/kingabzpro/jobzilla_ai/blob/main/jz_skill_patterns.jsonl) dataset is jsonl file containing different skills that can be used to create **spaCy** entity\_ruler. The data set contains label and pattern-> diferent words used to describe skills in various resume.

**Dependencies and Libraries:**

The project relies on various libraries and tools, including SpaCy, Gensim, Plotly Express, Matplotlib, Pandas, NumPy, and NLTK, for data processing, analysis, and visualization.

**Loading Dataset:**

The SpaCy model, resume dataset, and Jobzilla skills dataset are loaded directly into the entity ruler for streamlined processing.

**Skills Extraction:**

Custom Python functions are developed to extract skills mentioned within each resume text. These functions facilitate the creation of an array containing all unique skills across the dataset, enabling trend analysis and pattern identification.

**Cleaning Resume Text:**

Text preprocessing is carried out using NLTK, involving steps such as removal of hyperlinks, special characters, and punctuation, lowercasing text, tokenization, lemmatization, and elimination of English stopwords.

**Applying Functions:** The previously created functions are applied to the dataset, resulting in the creation of columns for cleaned resume text and extracted skills. Duplicates are removed from the skills column for accurate analysis.

**Visualization:** Various visualizations, including histograms and word clouds, are generated to provide insights into job category distributions and prevalent skills across the dataset.

**Entity Recognition:** SpaCy's displacy.render function is utilized to visualize various entities within the raw resume text, including skills and job categories. Custom entity colors and gradients are applied for enhanced visualization.

**Custom Entity Recognition:** Additional entities such as job categories are incorporated into the entity ruler, enhancing the scope of entity recognition. Custom colors and gradients are assigned to different entity types, resulting in visually appealing highlights within the text.

**Resume Analysis:**

Users can input their resumes for analysis, wherein SpaCy effectively identifies and highlights skills mentioned within the text. The analysis facilitates recruiters in evaluating resume suitability based on required skills.

**Match Score:**

A match score feature enables recruiters to input desired skills and obtain a percentage indicating the extent to which a resume matches the specified criteria. This functionality aids in streamlining the candidate selection process.

**Topic Modeling - LDA:**

Topic modeling using Latent Dirichlet Allocation (LDA) is employed to identify prevalent themes within the resume dataset. The resultant topics are visualized using PyLDAvis, offering insights into the underlying content structure.

**Conclusion:**

The project showcases the utilization of SpaCy and NLP tools for automated resume analysis, with a focus on entity recognition, skills extraction, and topic modeling. The developed system offers valuable insights to recruiters, expediting the hiring process and enhancing candidate selection efficiency. The comprehensive analysis and visualizations contribute to a deeper understanding of the resume dataset, facilitating informed decision-making in recruitment endeavors.

**Acknowledgements:**

Huge thanks to [Abid Ali Awan](https://deepnote.com/app/abid/spaCy-Resume-Analysis-81ba1e4b-7fa8-45fe-ac7a-0b7bf3da7826?Job_Category=AVIATION) for the great support and documentation to implement project.