Week 8

Name: Sri Harsha, Nimmagadda
Email: harshachaitanya27@gmail.com
College: Oregon State University
Specialisation: NLP

Problem Statement: The challenge arises when HR or hiring managers review resumes, as they often encounter an overwhelming amount of irrelevant information that requires careful examination. This tedious process makes it challenging and time-intensive for them to pinpoint the most qualified candidates. To tackle this issue, a proposed solution involves leveraging Named Entity Recognition (NER) within Natural Language Processing (NLP). This advanced technology can autonomously recognize and categorise key details in resumes, such as the candidate's name, educational history, work experience, and skills. Implementing NER streamlines the candidate shortlisting process, significantly enhancing efficiency for HR professionals and reducing their time and effort investment.

Data understanding: The dataset I work with is composed of unstructured text data presented in JSON format. Upon importing this data into a pandas DataFrame, a more comprehensive understanding of its structure has been acquired. Within the dataset, there are two principal columns: "content," housing the primary textual content from resumes, and "annotation," which provides labels for the information embedded in the content. This dataset encompasses resumes from 200 different individuals, each containing diverse details about the applicants. The information is systematically categorised into segments, covering aspects such as name, location, contact details (Indeed account), university/college name, degree, graduation year, years of experience, past companies, job designation, and skills.

Type of the Data used: The nature of the data in our dataset is primarily characterised as unstructured data, specifically in the form of text. This textual information is stored in JSON (JavaScript Object Notation) format, a versatile and widely used data interchange format. The unstructured aspect of the data implies that it doesn't adhere to a predefined data model, allowing for flexibility in representing information. This text-based content, organised within the structured framework of JSON, forms the foundation of our dataset, providing valuable insights into the varied and nuanced information present in each data entry.

Data Issues: Given that our dataset primarily consists of text data, the concept of outlier data points is not applicable, as text data does not exhibit numerical values that could deviate significantly from a norm. Additionally, many challenges associated with quantitative data distribution, including issues like skewness or the necessity for statistical normalisation, are not pertinent to our specific case. The unique nature of textual information mitigates certain concerns commonly encountered in datasets with numerical or structured data, providing a distinctive context for analysis and interpretation.

Solution Approaches:

Conduct thorough text cleaning to remove noise, special characters, and irrelevant information.

Github link:

https://github.com/harshachaitanya27/DataGlacier/tree/main/week8