Week 9

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.

Github link - https://github.com/harshachaitanya27/DataGlacier/tree/main/week9

Data Cleaning and transformation done:

I applied a thorough set of data cleaning and transformation operations to text data stored in a pandas DataFrame. Firstly, I used a custom cleaning function to expand contractions, remove digits, convert text to lowercase, eliminate unnecessary characters, tokenize, remove stopwords, and lemmatize words in the 'content' column. Afterward, I employed scikit-learn's CountVectorizer to convert the cleaned text into a count matrix, representing word frequencies in the corpus. Additionally, I utilized Gensim's Word2Vec model to capture semantic relationships among words in the tokenized and cleaned text. These steps collectively provided a structured and numerical representation of the original textual data, facilitating subsequent analysis, modeling, or machine learning tasks on the cleaned and processed information.