Team member: Zongdao Wen (Indivisual Project)
Email: zongdaow@usc.edu
Country: United States
College: University of Southern California
Specialization: NLP
Internship Batch: LISUM21
GithubLink: https://github.com/louiswenz/dataglacier/blob/main/week7/NLPResume-Description.pdf

Problem Description:

The goal is to develop a system that can automatically extract and classify specific entities from resumes into predefined classes such as person name, college name, academics information, relevant experiences, skill set, and more.

The input to the system is a collection of resumes stored in a JSON file. Each JSON object represents a resume and contains information about the candidate's profile, including the resume content and any available annotations or tags.

The system should extract the relevant entities from each resume and classify them into predefined classes. For example, for the resume content "John Doe graduated from XYZ University with a degree in Computer Science and has 5 years of experience in software development," the extracted entities could include "John Doe" (person name), "XYZ University" (college name), "Computer Science" (academic information), and "software development" (skill set).

Business Understanding:

In today's competitive job market, companies receive a vast number of resumes from job seekers for various positions. Manually reviewing and extracting relevant information from each resume is a time-consuming and tedious task. To address this challenge, the business aims to develop a resume entity extraction system using NLP techniques. This system will automate the process of extracting and classifying key entities from resumes, providing valuable insights to hiring teams and improving overall efficiency in the recruitment process.

Project Cycle (7 weeks):

1. Week 1: Project Planning and Data Collection (Deadline: 6.19)
   - Define project scope and objectives.
   - Identify the specific entities to be extracted from resumes.
   - Determine the classification categories for the entities.
   - Data Exploration.
2. Week 2: Data Preprocessing and NER Model Selection (Deadline: 6.26)
   - Preprocess the collected resumes, removing irrelevant information and standardizing the format if needed.
   - Explore different NER models and select the most suitable one for entity extraction.
   - Split the data into training and evaluation sets.
3. Week 3: NER Model Training and Fine-tuning (Deadline: 7.2
   - Train the selected NER model using the labeled training data.
   - Fine-tune the model on the specific resume entity extraction task.

Evaluate the performance of the trained model on the evaluation set and iterate if necessary.

4. Week 4: Entity Classification Design and Implementation (Deadline: 7.9)
   Design a set of classification rules or develop a machine learning model for entity classification.
   Implement the classification logic based on the predefined entity categories.
   Test the entity classification component using a sample set of resumes.
5. Week 5: Integration and System Development (Deadline: 7.16)
   Integrate the NER model and entity classification component into a cohesive system.
   Develop the necessary modules or functions to handle resume input, entity extraction, and classification.
   Conduct thorough testing and debugging of the system.
6. Week 6: Evaluation and Refinement (Deadline: 7.23)
   Evaluate the overall performance of the resume entity extraction system.
   Measure the accuracy, precision, recall, and F1-score of the extracted entities.
   Refine the system based on evaluation results and feedback.
7. Week 7: Documentation, Deployment, and Finalization (Deadline: 7.30
   Document the entire project, including the methodology, implementation details, and usage instructions.
   Finalize the system by addressing any remaining issues or improvements.
   Prepare the system for deployment in the production environment.
   Conduct a final round of testing and validation.

Data Intake Report

Name: NLP: Resume Extraction
Report date: June 19 2023
Internship Batch: LISUM21
Version:<1.0>
Data intake by: Zongdao Wen
Data intake reviewer:<intern who reviewed the report>
Data storage location: Github

Tabular data details:

Resume

Total number of observations	200
Total number of files	1
Total number of features	2
Base format of the file	JSON
Size of the data	1.1MB

Proposed Approach:

The data is loaded using pandas.read.json with inline =True. Regex is used to clean up the data.