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/week9

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).

Data Cleaning:

- The labels are Companies worked at, Location, Name, Designation, Email Address, UNKNOWN, Graduation Year, Years of Experience, Skills, College Name, Degree. UNKNOWN only occurred twice in the data, it is considered as a NA value and will be removed.
- Leading and trailing white space and punctuations affects the training of NER. Therefore they are also removed, and ‘start’ and ‘end’ are adjusted accordingly.
- Indices also changed from [start, end] to [start, end].
- Some entities are overlapping. For the overlapping ones, only the longest is preserved.