Introduction

Create Python code to handle one of the following scenarios:

1. Train new NER entity for position (for both English and Chinese dataset)
2. Train new relation extraction (for **Chinese dataset only**) for various relationship types given in the sample data

For each of scenario, part of the supplied dataset must be used as the basis to train and test.

For additional points, you can also find ways to supplement more test data via other means to prove or further enhance your results. The description of the dataset are provided at the end of this document.

Below are general description for the two scenarios.

Scenario 1 Description

As shown in the English dataset and as explained in the dataset description below, the relationship is extracted using template filling method, so accuracy and exact extraction is not guarantee. To improve on such result, training a model to recognize "Position" entity is necessary.

So that the sample position relation below:

```
"position": "Vice president of the Board and Chief Executive Officer of the Group"
```

can further be extracted by NER as list of common position name like:

```
"position": ["Vice President", "Chief Executive Officer"]
```

Furthermore, when training position entities, please account for various way the same position can be written, such as: **"Vice President", "VP", "V.P." "Vice-President"**

Scenario 2 Description

For relation extraction, most common approach is to take a two steps process:

1. Extract known NER entities as shown in below sample (PERSON, DATE, ORG, etc)


```
{
        "text": "許立榮先生，63 歲，現任中國遠洋海運集團有限公司董事長、黨組書記；本公司執行董事、董事長，東方海外（國際）有限公司（一家本公司之非全資附屬公司及於聯交所上市之公司（股份代號：316））執行董事、董事會主席。許先生於一九七五年三月參加工作。許先生過往曾任上海遠洋運輸有限公司船舶管理一處副處長、總經理助理、副總經理及總經理；上遠貨運公司副經理、經理兼黨委書記；上海航運交易所的總裁及黨委書記；中遠海運集裝箱運輸有限公司的總經理、黨委委員及黨委副書記；本公司副總經理、黨委委員、副書記；中國遠洋運輸（集團）總公司（現稱為中國遠洋運輸有限公司，為直接控股股東）副總裁、工會主席及黨組成員；中國海運（集團）總公司（現稱為中國海運集團有限公司，中國遠洋海運的附屬公司）董事、總經理、黨組成員、董事長、黨組書記；東方海外貨櫃航運有限公司董事會主席、執行委員會委員等職。許先生取得上海海事大學工商管理碩士學位，為高級工程師。",
        "ner": "[{\"text\": \"許立榮\", \"label\": \"PERSON\", \"start\": 0, \"end\": 3}, {\"text\": \"63 歲\", \"label\": \"DATE\", \"start\": 6, \"end\": 9}, {\"text\": \"中國遠洋海運集團有限公司\", \"label\": \"ORG\", \"start\": 12, \"end\": 24}, {\"text\": \"東方海外（國際）有限公司\", \"label\": \"ORG\", \"start\": 45, \"end\": 57}, {\"text\": \"交所\", \"label\": \"ORG\", \"start\": 74, \"end\": 76}, {\"text\": \"許\", \"label\": \"PERSON\", \"start\": 103, \"end\": 104}, {\"text\": \"一九七五年三月\", \"label\": \"DATE\", \"start\": 107, \"end\": 114}, {\"text\": \"許\", \"label\": \"PERSON\", \"start\": 119, \"end\": 120}, {\"text\": \"上海遠洋運輸有限公司船舶管理一處\", \"label\": \"ORG\", \"start\": 126, \"end\": 142}, {\"text\": \"上遠貨運公司\", \"label\": \"ORG\", \"start\": 161, \"end\": 167}, {\"text\": \"上海航運交易所\", \"label\": \"ORG\", \"start\": 179, \"end\": 186}, {\"text\": \"中遠海運集裝箱運輸有限公司\", \"label\": \"ORG\", \"start\": 195, \"end\": 208}, {\"text\": \"中國遠洋運輸（集團）總公司\", \"label\": \"ORG\", \"start\": 241, \"end\": 254}, {\"text\": \"中國遠洋運輸有限公司\", \"label\": \"ORG\", \"start\": 258, \"end\": 268}, {\"text\": \"中國海運（集團）總公司\", \"label\": \"ORG\", \"start\": 291, \"end\": 302}, {\"text\": \"中國海運集團有限公司\", \"label\": \"ORG\", \"start\": 306, \"end\": 316}, {\"text\": \"中國遠洋海運\", \"label\": \"ORG\", \"start\": 317, \"end\": 323}, {\"text\": \"東方海外貨櫃航運有限公司\", \"label\": \"ORG\", \"start\": 350, \"end\": 362}, {\"text\": \"許\", \"label\": \"PERSON\", \"start\": 378, \"end\": 379}, {\"text\": \"上海海事大學\", \"label\": \"ORG\", \"start\": 383, \"end\": 389}]"
    },
```


2. Build linkage among entities, like: PERSON→ORG = WorkAt or StudyAt relation
   A sample output can be like this:


```
{"text": "許立榮先生，63 歲，現任中國遠洋海運集團有限公司董事長、黨組書記；本公司執行董事、董事長，東方海外（國際）有限公司（一家本公司之非全資附屬公司及於聯交所上市之公司（股份代號：316））執行董事、董事會主席。", "relation_list": [{"predicate": "Age", "object_type": "DATE", "subject_type": "PERSON", "object": "63 歲", "subject": "許立榮"}]}, {"predicate": "WorkAt", "object_type": "ORG", "subject_type": "PERSON", "object": "中國遠洋海運集團有限公司", "subject": "許立榮"}]}
```

Submission Requirements

- Code written in Python
- For data preparation & tagging, you can use any existing tools but please state what tools are used.
    - For reference, doccano is a good entities annotation library but feel free to explore others.
- For machine learning framework, you can choose one of the following: Pytorch or Spacy. Please state the versions used.
- Your code should come with setup instruction in case you utilize additional modules (e.g., jieba, pkuseg, docker setup, etc) to get it working.
- For your result, you can provide your accuracy with any supporting that you like.
- Bonus points will be rewarded if you can highlight any limitations or conditions not considered in your submission.

Dataset Descriptions

The dataset contains texts and pre-annotated named entities, relations from specific sections of 23 annual reports. All the data are organised in the file **annual_report_23.json**, including the following attributes:

- all_data: List[single]
- single: Dict
    - stock_code: str, the stock code of the company
    - lang: Union[en,zh], e.g., en: English, zh: Chinese
    - document_type: str
    - text: List[ner]
    - relation: List[rel]
- ner: Dict
    - text: str, the raw text, corresponding to one paragraph in the document
    - ner: List[ner_]
- ner_: Dict
    - text: str, named entity found in the text
    - label: str, entity label
    - start: str, starting index for the entity
    - end: str, end index for the entity
- relation: Dict (For English data set only)
    - name: str, personal names found in the whole section
    - age: str, age for that person
    - education: List[str], university that the person went to
    - join_date: str, the date the person joined the company
    - past_experience: the person's past experience

Note, the entities are identified by Spacy, and the relations are generated using template filling method. Therefore, the accuracy can not be guaranteed. They act as reference only. Besides, Spacy can only recognise entity labels such as person, location, others for example, job position does not work out. So it should be additionally labelled.

Here is an example sentence:

Dr Auyeung is an independent non-executive director of China Construction Bank (Asia) Corporation Limited and C-MER Eye Care Holdings Limited.

where, Auyeung is person name, an independent non-executive director is job position, China Construction Bank (Asia) Corporation Limited and C-MER Eye Care Holdings Limited are organizations. the relation between Auyeung and China Construction Bank (Asia) Corporation Limited (or C-MER Eye Care Holdings Limited) can be defined as per:employee_of.