Named entity recognition for Fintech

`Group ID` 2

`Group members` Yixi Zhou, Jingyi Zeng, Qianru Li

`Readme` written on: June, 9, 2024

`Author` Yixi Zhou

Project Structure

```
|  README.md
|  README.pdf
|  requirement.txt
|  NER_main.ipynb
├─ NER_data
|      origin.txt
|      output.txt
|      outputli.txt
|      outputzeng.txt
|      test.char.bmes
|      train.char.bmes
|  evaluating.py
|  utils.py
├─ pic
|      Confusion Matrix HMM.png
|      Confusion Matrix CRF.png
|      Score for CRF.png
|      Score for HMM.png
|      transition.png
└─.ipynb_checkpoints
       NER_main-checkpoint.ipynb
```

Environment Configuration

Configuring a Virtual Environment

```
conda create --name nerproject python=3.9
```

```
conda activate nerproject
```

Installing third-party libraries

```
numpy
sklearn-crfsuite
torch
matplotlib
seaborn
```

- from **sklearn_crfsuite** import CRF, we use this because we don't have the basic knowledge of CRF, we just import it for the comparison with HMM model.
- import **torch**, PyTorch provides powerful tensor manipulation capabilities, which makes it convenient to represent and compute the probability matrices (transition probability matrix, observation probability matrix) and the initial state probability vector： the core parameters of the HMM model. PyTorch allows for easy matrix and vector operations.
- import **matplotlib.pyplot** as plt, import **seaborn** as sns, for visualization.

```
pip install -r requirement.txt
```

Run

After completing the environment configuration, please run `NER_main.ipynb` directly and run it normally for two minutes.

Announcement of the LLMs

To reduce some of the repetitive mechanical work, we have used Large Language Model to assist with a portion of the content in this project. However, **the core HMM code section will not involve LLM assistance**. We will declare the areas where we have used LLMs:

- In the data annotation part of named entity recognition, we used LLM for preliminary labeling, and completed the annotation of financial news using human verification.
- In the evaluation function part, we used LLM to assist with the visualization work and guide us which index should be evaluated.
- During the paper writing process, we used LLM for partial paragraph translation and grammar checking.
- In the principle analysis process, we used LLM to write pseudocode.
- For the tables in the LaTeX part, we used LLM for typesetting.

Reference

The reference can be found in the paper `reference` part for more details.