Project 2: Named Entity Recognition

Description

For Project-2, you will implement a feature-driven model that extracts named entities (NEs) from a document. The goal is to extract relevant concepts from a document such as names of persons, locations, geo-political entities, etc. To train your NER model, you are provided with the CoNLL-2003 English dataset. This corpus describes four kinds of NE's: Locations (LOC), Organizations (ORG), Persons (PER) and Miscellaneous (MISC). This corpus follows the BIO notation for representing labels where B marks the beginning of a tag, I marks the continuation of a previous tag and O marks that the given token is a named entity.

Instructions

1. Download the `modified_train.txt`, `modified_test.txt` and `proj_2.py` files. Put these files in the same folder.
2. Install dependencies:

   Linux or macOS

   ```
   pip3 install -r requirements.txt
   ```

   Windows

   ```
   pip install -r requirements.txt
   ```

3. To run, type in the command line interpreter:

   Linux or macOS

   ```
   python3 proj_2.py
   ```

   Windows

   ```
   python proj_2.py
   ```

NOTE: This application requires the use of a large amount of RAM, at least approx. 80 GB, to run properly. You have been warned.