Homework_Data_Pre-Processing_Neural_POS-tagging

October 22, 2018

Homework: Pre-process data for Neural POS-tagging

Task (due Tuesday, 30 October, at 10:00)

1. Download the folder with CONLL2003 dataset:
https://github.com/Franck-Dernoncourt/NeuroNER/tree/master/data/conll2003
(CONLL2003 stands for Conference on Natural Language Learning, 2003)

Take a look at the data. The data are already divided into train, test, and validation (development) sets (please ignore the metadata file).

Each file has 4 columns:

- token (word or punctuation mark)
- part of speech tag (POS tag)
- information for chunking
- information for Named Entity Recognition (ner)

a) Load the data and give the set of values for each column (except the first, three sets in total).

b) How many items are in each of the three sets?

c) What do the values in those three sets stand for?

2. Lines between blank lines represent sentences, e.g.:

```
It PRP B-NP O
was VBD B-VP O
the DT B-NP O
second JJ I-NP O
costly JJ I-NP O
blunder NN I-NP O
by IN B-PP O
Syria NNP B-NP B-LOC
in IN B-PP O
four CD B-NP O
minutes NNS I-NP O
. . O O

It was the second costly blunder by Syria in four minutes .
```

Write a function readFile(path), which takes a file path as an input, loads data from this file and returns a list of sentences as an input.

Please ignore the first line '-DOCSTART- -X- -X- O'

```
In [25]: def readFile(path):
             list_of_sentences = []
             sentence = []
             # <your code here>
             return list_of_sentences
```

Applied to a path in the 'conll2003/en/' folder, you should get the following output:

```
In [31]: test_sentences = readFile('conll2003/en/test.txt')

         print("number of sents: ", len(test_sentences))
         print()
         print(test_sentences[0:3])
```

```
number of sents:  3453

[[['SOCCER', 'NN', 'B-NP', 'O'], ['-', ':', 'O', 'O'], ['JAPAN', 'NNP', 'B-NP',
↪  'B-LOC'],
    ['GET', 'VB', 'B-VP', 'O'], ['LUCKY', 'NNP', 'B-NP', 'O'], ['WIN', 'NNP',
↪  'I-NP', 'O'], [',', ',', 'O', 'O'], ['CHINA', 'NNP', 'B-NP', 'B-PER'], ['IN',
↪  'IN', 'B-PP', 'O'], ['SURPRISE', 'DT', 'B-NP', 'O'], ['DEFEAT', 'NN', 'I-NP',
↪  'O'], ['.', '.', 'O', 'O']], [['Nadim', 'NNP', 'B-NP', 'B-PER'], ['Ladki',
↪  'NNP', 'I-NP', 'I-PER']], [['AL-AIN', 'NNP', 'B-NP', 'B-LOC'], [',', ',', 'O',
↪  'O'], ['United', 'NNP', 'B-NP', 'B-LOC'], ['Arab', 'NNP', 'I-NP', 'I-LOC'],
↪  ['Emirates', 'NNPS', 'I-NP', 'I-LOC'], ['1996-12-06', 'CD', 'I-NP', 'O']]]
```

3. Please send me your solutions by next Tuesday, 30 October, at 10:00. Please send me **just a .ipynb or .py file, not a .zip file**, since I have the data.