命名实体识别（NER）

班级：2016211303　　姓名：黄若鹏　　学号：2016212901

1.题目

Named entity recognition: 30 points

- Named entities: people names, organizations, locations, numerals, etc
- Your objective is to build a machine learning named entity recognition system, which when given a new previously unseen text can identify and classify the named entities in the text. This means that your system should annotate each word in the text with one of the four possible classes.
- You will be given labeled data sets to train and test your model.

2.概要

2.1.命名实体识别（NER）

命名实体识别（Named Entity Recognition，简称 NER）是信息提取、问答系统、句法分析、机器翻译等应用领域的重要基础工具，在自然语言处理技术走向实用化的过程中占有重要地位。一般来说，命名实体识别的任务就是识别出待处理文本中三大类（实体类、时间类和数字类）、七小类（人名、机构名、地名、时间、日期、货币和百分比）命名实体。

举个简单的例子，在句子“小明早上 8 点去学校上课。”中，对其进行命名实体识别，应该能提取信息

```
人名：小明，时间：早上 8 点，地点：学校。
```

2.2.实现方法

基于规则和词典的方法，用 HMM、CRF 或深度学习来实现命名实体识别等。
我这里是使用 nltk 工具包
这里我通过查博客看了许多讲解和代码，收获很多。

3.运行环境

系统：windows10

编写工具：VsCode

4.输入输出

输入：一篇小短文

输出：短文命名实体标签

5.实验过程

测试文档（介绍 FIFA，来源于维基百科）

FIFA was founded in 1904 to oversee international competition among the national associations of Belgium,
Denmark, France, Germany, the Netherlands, Spain, Sweden, and Switzerland. Headquartered in Zürich, its
membership now comprises 211 national associations. Member countries must each also be members of one of
the six regional confederations into which the world is divided: Africa, Asia, Europe, North & Central America
and the Caribbean, Oceania, and South America.

实现 NER 的 Python 代码

```
import re
import pandas as pd
import nltk

def parse_document(document):
   document = re.sub('\n', ' ', document)
   if isinstance(document, str):
       document = document
   else:
       raise ValueError('Document is not string!')
   document = document.strip()
```

```
   sentences = nltk.sent_tokenize(document)
   sentences = [sentence.strip() for sentence in sentences]
   return sentences


# sample document
text = """
FIFA was founded in 1904 to oversee international competition among the
national associations of Belgium,
Denmark, France, Germany, the Netherlands, Spain, Sweden, and
Switzerland. Headquartered in Zürich, its
membership now comprises 211 national associations. Member countries
must each also be members of one of
the six regional confederations into which the world is divided:
Africa, Asia, Europe, North & Central America
and the Caribbean, Oceania, and South America.
"""

# tokenize sentences
sentences = parse_document(text)
tokenized_sentences = [nltk.word_tokenize(sentence) for sentence in
sentences]
# tag sentences and use nltk's Named Entity Chunker
tagged_sentences = [nltk.pos_tag(sentence) for sentence in
tokenized_sentences]
ne_chunked_sents = [nltk.ne_chunk(tagged) for tagged in
tagged_sentences]
# extract all named entities
named_entities = []
for ne_tagged_sentence in ne_chunked_sents:
   for tagged_tree in ne_tagged_sentence:
       # extract only chunks having NE labels
       if hasattr(tagged_tree, 'label'):
           entity_name = ' '.join(c[0] for c in tagged_tree.leaves())
#get NE name
           entity_type = tagged_tree.label() # get NE category
           named_entities.append((entity_name, entity_type))
           # get unique named entities
           named_entities = list(set(named_entities))
# store named entities in a data frame
entity_frame = pd.DataFrame(named_entities, columns=['Entity Name',
'Entity Type'])
# display results
print(entity_frame)
```

输出结果

M-Viterbi-master\HMM-Viterbi-maste

Entity Name	Entity Type
FIFA	ORGANIZATION
Central America	ORGANIZATION
Belgium	GPE
Caribbean	LOCATION
Asia	GPE
France	GPE
Oceania	GPE
Germany	GPE
South America	GPE
Denmark	GPE
Zürich	GPE
Africa	PERSON
Sweden	GPE
Netherlands	GPE
Spain	GPE
Switzerland	GPE
North	GPE
Europe	GPE

注：上图中 LOCATION 和 GPE 有重合。GPE 通常表示地理—政治条目，比如城市，州，国家，洲等。LOCATION 除了上述内容外，还能表示名山大川等。FACILITY 通常表示知名的纪念碑或人工制品等。

可以看到，NLTK 中的 NER 任务大体上完成得还是不错的，能够识别 FIFA 为组织（ORGANIZATION），Belgium,Asia 为 GPE，但是也有一些不太如人意的地方，比如，它将 Central America 识别为 ORGANIZATION，而实际上它应该为 GPE；将 Africa 识别为 PERSON，实际上应该为 GPE。

这个实验由于是最后一个做，正处于考试周，因此只是简单的测试了一篇小短文，并且是直接调用工具包，并未像其他实验一样实现算法。