实体识别项目说明

## 一、数据集

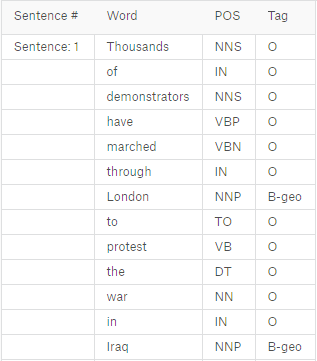
### 1. 实验数据来源：CoNLL 2002

共含47959个句子，1048575个词，已标注好；

下载地址：<https://www.kaggle.com/abhinavwalia95/entity-annotated-corpus/data>

（ner\_dataset.csv文件）。

**文件格式：**

****

**标签含义**：

geo = Geographical Entity

org = Organization

per = Person

gpe = Geopolitical Entity

tim = Time indicator

art = Artifact

eve = Event

nat = Natural Phenomenon

Using a tagging scheme to distinguish between the beginning (tag B-...), or the inside of an entity (tag I-...)

### 2. 数据集设置（保存在\data\input\目录下）

训练集train.txt：60%

验证集dev.txt：20%

测试集test.txt：20%

## 二、预训练词向量：Glove

下载地址：<https://nlp.stanford.edu/projects/glove/> （glove.6B.zip文件）

使用向量维度为300的词向量，由于源文件过大（不上传），所以从预训练的词向量中获取有用的部分进行存储：保存在\data\reduced\_glove\_300dimension.npz文件中。

## 三、项目运行说明

**step 1（可省）:** 运行build\_vocab.py文件，字典与词向量文件已创建好，存放在\data目录下：words.txt, tags.txt, chars.txt, reduced\_glove\_300dimension.npz

**step 2 （可省）:** 运行train.py文件，模型已训练好，存放在\results\model\目录下，运行日志存放在\results\ log.txt文件中

**step 3:** 运行test.py文件，模型测试，测试日志文件存放在\results\log.txt文件中

step 1. python build\_vocab.py #创建模型字典，加载预训练的词向量

step 2. python train.py # 模型训练

step 3. python test.py # 模型测试

## 四、性能说明

**模型训练:**

迭代轮数：13

训练数据集：\data\input\train.txt 句子数：28775 词数：629421

验证数据集：\data\input\dev.txt 句子数：9592 词数：210010

训练时长：约 6 小时

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 迭代次数 | Accuracy | Precious | Recall | **F1** |
| 1 | 95.75 | 76.5 | 75.46 | **75.98** |
| 2 | 96.35 | 80.91 | 78.36 | **79.62** |
| 3 | 96.56 | 82.25 | 79.61 | **80.91** |
| 4 | 96.64 | 82.66 | 80.46 | **81.55** |
| 5 | 96.74 | 83.40 | 80.85 | **82.11** |
| 6 | 96.79 | 83.53 | 81.24 | **82.37** |
| 7 | 96.82 | 83.60 | 81.49 | **82.53** |
| 8 | 96.80 | 83.66 | 81.55 | **82.59** |
| 9 | 96.86 | 83.90 | 81.80 | **82.83** |
| 10 | 96.87 | 83.68 | 82.11 | **82.89** |
| 11 | 96.87 | 83.65 | 82.06 | **82.85** |
| 12 | 96.89 | 83.78 | 82.18 | **82.97** |
| 13 | 96.86 | 83.59 | 82.26 | **82.92** |

**模型测试:**

测试数据集：\data\input\test.txt 句子数：9592 词数：206669

测试所需时间：3.6 分钟

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precious | Recall | **F1** |
| 97.12 | 84.77 | 83.64 | **84.20** |