# 实体识别系统

1. **需求分析**

在地理高考问答系统中，理解试题文本是系统的重要部分。实体识别是为试题文本做后续处理做支撑，也是高层抽象语义理解的基础。实体识别只对地理试题中出现频次高且非常重要的时间、地点这两类实体做识别，对其他不重要或是出现频次低的实体（如货币、距离等）不做识别。

1. **实体定义**

时间TIME：具体的、能在时间轴上找到位置的时间

例：2005年、8月12日、15时等

地点LOC：能够在地图中找到的实际存在的点

例：美国、北美洲、黄土高原、亚欧板块等

## 数据格式

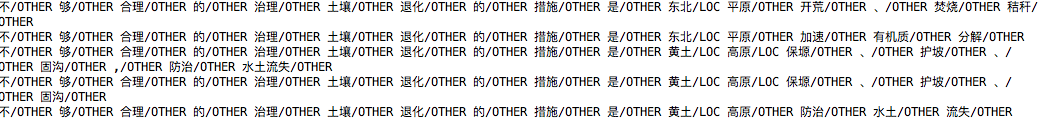
### 输入文件格式

以utf-8格式编码的文本文件，文件中每一行都是是分好词的试题。附带的测试文本test.txt是2003-2015年的北京高考选择题文本，共649句话，每句话由选择题题干+选项构成。

../Documents/OneDrive/NER/NER/test.png

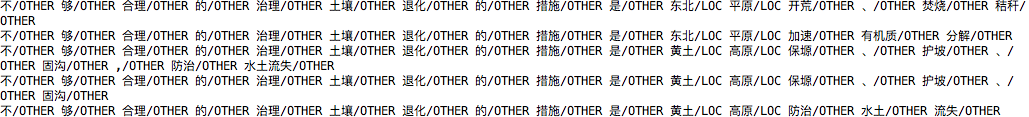
### 输出文件格式

输出的文件包含测试句子每个词及其预测出来的实体标记。每个词与其标签间用“/”隔开，不同词之间用空格隔开。



### 评测文件格式

与输出文件一致，包含句子及其正确实体标记。



## 软件运行环境

环境：JDK1.8