

# 文本核心地理位置提取相关代码文件及说明

---

本实验中主要包含两大部分：

1. 文本核心地理位置提取部分
2. 展示平台

## 文本核心地理位置提取部分

在这一部分实验中主要分为3块

1. Mordecai库相关实验
2. 动词地名树实验
3. 机器学习实验

### Mordecai库相关实验

#### 相关文件

| mordecaiProcessing

| txtExtracting.py

| txtProcess.py

#### 相关说明

由于Mordecai库本身是对于英语文本的信息提取，因此在进行处理之前通过有道接口帮助中英文转换。

在有道接口使用过程中，由于需要使用大量的访问，为了避免ip被封，我使用了代理ip等方式。

Mordecai库本身是一个相对比较成熟的核心地理名词提取库，其相关论文为Geolocating Political Events in Text.pdf

#### 使用方法

安装mordecai库直接运行即可，其安装方法：<https://github.com/openeventdata/mordecai>

遇到的问题在20200305本周报告.pdf中有详细的介绍

### 动词地名树相关实验

#### 相关文件

| verb-location tree

| dataCreating.py

| judgeLocation.py

#### 相关说明

动词地名树方法核心是基于规则的树结构搭建，通过stanfordnlp工具获取相关文本信息并形成相关信息数组，使用树结构对于节点信息进行合并操作后，得到最有可能是核心地理位置的节点。核心代码文件judgeLocation.py

### 使用方法

下载安装stanfordnlp相关工具，在Mac上由于存在一定的权限问题，在执行代码的时候使用的基本都是命令行 `sudo python xxx.py`

遇到的问题及详细说明在**20200327本周报告.pdf**中有所介绍

## 机器学习相关实验

### 相关文件

| machinelearning

| model\_train.py

| model\_predict.py

### 相关说明

机器学习这一块主要使用的Kashgari库帮助完成。

### 使用方法

下载安装Kashgari相关工具，直接运行即可。相关地址：<https://kashgari-zh.bmio.net/tutorial/text-labeling/>

遇到的问题及详细说明在**20200410 本周报告.pdf**，**20200424 本周报告.pdf**，**20200507 本周报告.pdf**中有所介绍。

## 展示平台代码

### 使用方法

直接用python打开文件即可。使用的是Django框架。前段在其中frontend文件夹。