

专利机翻检测工具 - 安装文档

线上地址

包括检测工具和可视化工具两部分，

- 检测工具需要做安装配置，用于生成待分析的*.json文件目录和*.csv文件前缀
- 可视化工具不需要，仅需拷贝views/目录、待分析的*.json文件

技术路线

- 定义翻译锚点，锚点选择方案：
 1. 对照编码表，定义标准锚点字符集；
 2. 归纳总结，数字块、HTML实体、缩写、序号、等式、非规范引用等；
 3. 对照符号串，在原文和译文中对这些锚点符号串进行对照；
 4. 如果未对照出现，则认为有错误
- 原文筛选锚点；
- 译文对照锚点；
- 统计排序；

环境安装

win和linux都可以，如果没安装python的话需要先安装

- Windows安装python3
 1. 打开 WEB 浏览器访问<https://www.python.org/downloads/windows/>
 2. 在下载列表中选择Window平台安装包，版本选择python3.5+，可以是3.9
 3. 下载后，双击下载包，进入 Python 安装向导
- Linux安装python3和pip
 1. centos: yum install python3
 2. ubuntu: apt install python3

详细教程可以参见：[安装教程](#)

安装完python3之后，检查python：

```
python -V
# 或
python3 -V
```

检查pip版本：

```
pip -V
# 或
pip3 -V
# 检查pip版本和对应的python，没问题：
pip3 install -r requirements.txt
```

安装环境和依赖包

```
pip3 install -r requirements.txt
# 或
pip install -r requirements.txt

pip install tqdm lxml regex xlrd
```

如果有其他缺少的模块，继续使用pip install进行安装

```
pip install <缺少的模块名>
# 或
pip3 install <缺少的模块名>
```

使用和部署

包括检测工具和可视化工具两部分：

- 检测工具是.py文件，run.py，需要的目录结构如下：

```
|—run.py
|—run.excel.py
|—run.txt.py
|—input.list
|—config.ini
|—data
```

使用方式：

```
python3 run.py -c config.ini -i input.list -o output
# 或
python run.py -c config.ini -i input.list -o output
```

run.excel.py和run.txt.py分别是输入excel和输入txt的版本。

全部参数说明

```
usage: run.py [-h] [-i INPUT_FOLDER] [-o OUTPUT_FOLDER] [-c CONFIG] [-j JOBS]
```

Data process tool.

optional arguments:

- h, --help show this help message and exit
- i INPUT_FOLDER, --input_folder INPUT_FOLDER
输入文件夹名
- o OUTPUT_FOLDER, --output_folder OUTPUT_FOLDER
输出文件夹名称前缀
- c CONFIG, --config CONFIG
配置文件
- j JOBS, --jobs JOBS 进程数，一般默认即可

- 可视化工具是一个HTML页面，以及配套的css和js文件，目录结构如下：

```
|—index.html
|—utils.js
|—css
|—fonts
|—js
```

- 使用方式：

1. 选择检测工具输出的文件，如output_folder
2. 点击“开始分析”