数据集格式规范脚本

Date: 2022.3.12 Author: Jiaxin Zhuang 👨‍🔧

1 使用方法

2.1 shell 脚本使用

为了方便别人使用我的代码，连夜写了一个 `shell` 脚本供大家使用：

```
# 在目录下直接运行
./start.sh
# 运行前请先检查是否有执行权限
chmod +x start.sh
```

⚠️ **使用注意事项**：

- 脚本默认带转换的 `txt` 文件放在 `./dataset` 目录中，如果检测该目录下没有文件不会运行
- shell 脚本和 python 脚本都默认在 `./output` 目录下（**如果初次运行没有该目录的话回自动创建一个目录，不用担心**）
- 在 `./output` 存放已经转换完成的文件（**文件名保持不变**）

2.2 python 脚本使用

直接使用 `python` 脚本也可以转换哦～

```
# 命令格式如下，
# [path] 是相对脚本 transfer.py 的目录，使用之前请确保当前目录与transfer.py目录相同
# [type] s 表示 .source 文件； t 表示 .target 文件。
# 如果文件名符合 xxx.target 或者 xxx.source 可以省去这个参数，python 脚本会自动识别类型
python3 transfer.py [path] [type]
# 示例
python3 transfer.py ./dataset/test.source
python3 transfer.py ./dataset/test.target
```

2 脚本介绍

代码由三个部分组成：

- 格式化文本
- 提取人物名
- 根据人名重组文本
- 组装模块和控制输入输出

2.1 格式化文本

`Fomrat.py`

针对原先数据集的特点，对文本进行三个处理：

- 去除标点符号前后多余的空格，句首大写处理
- 纠正缩写产生多余的空格
- 去除一些多余字符，针对 target 文件开头大写

这里采用 Python 的 `re` 模块，使用 **正则表达式** 对文本进行匹配查找，通过 `re.sub()` 函数对匹配的文本进行操作。

2.2 提取人物名

`NameToken.py`

针对文本的特点，对文本的人名进行提取：

- 识别出每行说话者的名字
- 将名字制作成一个集合
- 加入特定字符方便后续重组文本

利用正则表达式对文本进行匹配，识别 `NAME:` 部分，并提取出来形成 `NameList` 列表，消除重复元素后返回 `NameList`。在提取出 `NameList` 后对文本中的人名大小写规范，并加入特殊字符方便后续重组语句。

2.3 根据人名重组文本

`Resemble.py`

经过上述的处理，现在根据人物来聚合文本：

- 利用上一个模块添加的特殊字符对文本进行分割
- 利用 `NameList` 来对分割后的文本重组
- 对每个文本进行补充标点
- 处理完后重组成一行

这个部分考虑到对话者可能不止两个，因此采用了二维列表来分类语句，最后通过 `str.join()` 函数和在一起。

2.4 组装模块和控制输入输出

`transfer.py`

将上述模块组合在一起，并添加输入输出提示和控制：

- 从命令行提取路径和文本类型（文本类型选填，可以自动识别）
- 按顺序进行格式化、人名提取、文本重组处理
- 针对 target 进行特殊格式化

通过 `sys.argv` 读取 `path` 和 `type`。如果文件名符合要求，则 `type` 可以不写。根据类型的不同，调用不同的处理函数。并增加输入错误提示。默认输出目录为 `./output`

Have a nice day ~ ❤️