拼音输入法 运行指南

致理书院 郭士尧 2022012406

1 基于字的二元模型

该模型可以直接运行。

```
make run
# 或者：make main && ./main < data/input.txt > data/output.txt
```

2 基于词的模型

在运行基于词的模型前，需要下载必要的数据文件（`extra` 部分数据）。见 附录 小节 1.2，或直接通过下面的命令下载：

```
wget 'https://cloud.tsinghua.edu.cn/d/11977280567f4609a3bf/files/?p=%2Fextra.
zip&dl=1' -O extra.zip
unzip extra.zip -d extra && rm extra.zip
```

在此之上还需要下载词典文件。词典文件通过语料预处理生成，具体见 附录 小节 1.3。也可以通过下面的命令下载：

```
# 二元模型词典，新浪新闻数据集
wget 'https://cloud.tsinghua.edu.cn/d/11977280567f4609a3bf/files/?p=%2Fdict_
sina.zip&dl=1' -O dict_sina.zip
unzip dict_sina.zip -d extra && rm dict_sina.zip

# 二元模型词典，社区问答数据集
wget 'https://cloud.tsinghua.edu.cn/d/11977280567f4609a3bf/files/?p=%2Fdict_web.
zip&dl=1' -O dict_web.zip
unzip dict_web.zip -d extra && rm dict_web.zip

# 三元模型词典，新浪新闻数据集
wget 'https://cloud.tsinghua.edu.cn/d/11977280567f4609a3bf/files/?p=%2Fdict_tri_
sina.zip&dl=1' -O dict_tri_sina.zip
unzip dict_tri_sina.zip -d extra && rm dict_tri_sina.zip

# 三元模型词典，社区问答数据集
wget 'https://cloud.tsinghua.edu.cn/d/11977280567f4609a3bf/files/?p=%2Fdict_tri_
web.zip&dl=1' -O dict_tri_web.zip
unzip dict_tri_web.zip -d extra && rm dict_tri_web.zip
```

下载好词典后，可以通过下面的命令运行二元词模型或三元词模型：

```
# 基于词的二元模型
make main_word && ./main_word run [dataset] < data/input.txt > data/output.txt
```

```
# 基于词的三元模型
make main_word_tri && ./main_word_tri run [dataset] < data/input.txt > data/
output.txt
```

这里 `[dataset]` 可选 `web`（社区问答语料）或 `sina`（下发的新浪新闻数据集）。实际表现中，`web` 数据集的准确率更高。

附 1 数据来源与许可协议

所有可下载数据统一放在清华网盘：https://cloud.tsinghua.edu.cn/d/11977280567f4609a3bf/

附 1.1 社区问答语料（`webtext2019zh`）

该语料只有需要生成对应社区语料词典时才需要下载。

- 来源：https://github.com/brightmart/nlp_chinese_corpus
- 清华网盘：`webtext2019zh.zip`
- 文件位置：`corpus/webtext2019zh`
- 许可协议：MIT License

附 1.2 `extra` 部分数据

该部分数据统一打包在清华网盘 `extra.zip`，下载后解压到 `extra` 目录下即可。

附 1.2.1 cppjieba 所需分词数据

- 来源：https://github.com/yanyiwu/cppjieba/tree/b11fd29697c0a6d8e5bef8eab62bae4221e0eda6/dict
- 文件位置：`extra/{jieba.dict.utf8, hmm_model.utf8}`
- 许可协议：MIT License

附 1.2.2 标点符号列表

- 来源：https://github.com/yanyiwu/cppjieba/blob/b11fd29697c0a6d8e5bef8eab62bae4221e0eda6/dict/stop_words.utf8，经过手工过滤
- 文件位置：`extra/punctuations.txt`
- 许可协议：MIT License

附 1.2.3 汉字词语拼音表

- 来源：https://github.com/wolfgitpr/cpp-pinyin/tree/b05278f14ace213d85777ffa55de6967585a6a93/res/dict/mandarin
- 文件位置：`extra/{word.txt, phrases_dict.txt, user_dict.txt, License.txt}`
- 许可协议：CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0/)

附 1.2.4 基础词语拼音表

- 来源：基于 汉字词语拼音表 生成，方法为 `python3 src/gen_words.txt`
- 文件位置：`extra/word.txt`

附 1.3 词典数据

该部分可以通过 `make-dict` 命令生成；但生成时间可能较长，可以选择直接下载。

生成方式：

```
# 二元模型词典
make main_word && ./main_word make-dict [dataset]

# 三元模型词典
make main_word_tri && ./main_word_tri make-dict [dataset]
```

当 `[dataset]` 为 `web` 时，需要先下载社区问答语料，见 附录 小节 1.1。

生成的词典文件会放在 `extra` 目录下，命名为 `dict_[dataset].bin` 和 `dict_[dataset]_words.txt`。

- 二元模型词典（`main_word`）
  - `sina` 数据集：`dict_sina.zip`
  - `web` 数据集：`dict_web.zip`
- 三元模型词典（`main_word_tri`）
  - `sina` 数据集：`dict_tri_sina.zip`
  - `web` 数据集：`dict_tri_web.zip`