作业一补充说明

2024.10.04 曾泉胜

1. 程序补充说明

1.1 分词方式

- 每个词项只包含 `[a-zA-Z+]` （仅字母）。如 `don't` 被分成 `don` 和 `t`， `state-of-the-art` 被分成 `state`, `of`, `the` 和 `art`.
- 所有词项全部小写处理。

1.2 程序输入输出

- 程序应当接受一个输入文件路径列表作为命令行参数，至少有一个输入文件，且均为txt格式；输入文件的每行结束都有换行符 `\n`。

```
./main in1.txt in2.txt ...
```

- 输出文件数量应当与输入文件相同。
- - 第 `1` 个输出文件命名为 `out1.txt`，反应的是在完成 插入 `in1.txt` 之后的词表与为位置表的状态；
- - 第 `2` 个输出文件命名为 `out2.txt`，反应的是在完成 插入 `in1.txt` 和 `in2.txt` 之后的词表与为位置表的状态；
- 格式如下：

```
<word>;<count>;(<fid>,<row>,<col>);...
<word>;<count>;(<fid>,<row>,<col>);(<fid>,<row>,<col>)...
...
```

- - `<word>` 是词本身，也是词表排序的主键；
- - `<count>` 是该词在已经输入的文件中出现的总次数；
- - `(<fid>,<row>,<col>);...` 是这个词的位置索引列表，其中每一项由 （出现文件位置，行数，列数）组成，均从 `1` 开始计数。

可参考 `exemplar` 中的示例。

1.3 注意事项

- C++ 标准模板库 （STL）中可能提供了大量现成的容器和算法。同学们在作业中的关节数据结构和算法部分应当尽量自己实现。一些线性容器的使用是可以的 （`string`, `array`, `vector`）。

2. 编译模板统一

为了便于统一评测，统一使用 `cmake` 构建项目。下面以 `Windows-x64` 系统为例，给出一种默认的配置：

2.1 Step1 安装编译器

推荐使用安装简单的 TDM-GCC。下载链接。

按照默认选项安装完成后，在终端输入 `g++ -v`，应当看到类似这样的输出。

```
COLLECT_GCC=<你的安装路径>\bin\g++.exe
<... 此处省略 ...>
gcc  10.3.0 (tdm64-1)
```

如果显示找不到命令，请检查环境变量并重启终端。

2.2 Step2 安装 cmake

在 https://cmake.org/download/ 获取 cmake 的安装包。

安装时添加环境变量。完成后，同样在终端输入 `cmake --version`，可以看到类似这样的输出（你的版本可能更高，只要别太低就行）。

```
cmake version 3.26.0-rc2

CMake suite maintained and supported by Kitware (kitware.com/cmake).
```

2.3 Step3 修改 CMakeLists.txt

这里给了一个简单的 hello world 的项目。

```
├─CMakeLists.txt
├─include
└─src
```

在 `CMakeLists.txt` 修改编译器的绝对路径，以及头文件，源文件等信息。（请让最后得到的可执行文件就叫 `main`）

之后 在 CMakeLists.txt 所在目录下执行

```
cmake -G "MinGW Makefiles" -B build
```

可以发现，创建了一个 `build` 目录，其中包含 `Makefile`，这时进入 `build` 目录并执行

```
mingw32-make
```

之后就可以看到 `build` 下的可执行文件 `main.exe`。

2.4 注意事项

- 如果源代码有修改，建议 `mingw32-make clean` 清除可执行文件后重新编译；
- 如果 `CMakeLists.txt` 有修改，建议 删除 `build` 文件夹后重新构建；
- 为了保证能够统一评测，请只提交源代码以及 `CMakeLists.txt`，且按照上述流程构建完成后，应当在 `build` 目录下有一个 `main.exe`，并且执行 `main.exe` 后，输出文件同样在 `build` 中。
- 提交前请检查编译出的可执行文件能够正常运行。

3. 提交时间及其方式

1. 请在 `2024.10.09` 实验课下课之前将代码按照上述要求打包发送至邮箱 `jw_assist@163.com`。邮件主题与压缩包名称均为 `姓名_学号_第一次作业`。
2. 上述评测只是对正确性的一个简单考察，很多流程是摸索着来的。有任何建议或者问题在飞书上call我，take easy。
3. 上课时每位同学主要准备在5min内讲清楚两个问题：
    - 你是如何设计词表和索引表的数据结构的？
    - 你是如何在插入新的单词时调整字典序和频次的？
    - 此外，你可以讲一些额外的实现。