基于文本内容的电影检索与推荐（第一部分）实验报告

吴佳龙 2018013418

一、 实验目标

本学期实验的最终目标是实现一个基于文本内容的电影检索与推荐系统，可以对电影网页进行信息提取和分词，并以此为基础建立倒排文档，实现电影查询及简单的推荐功能。

本实验的第一部分将使用栈结构解析网页文档，提取电影信息并对剧情简介进行中文分词。电影信息和分词结果将被输出到指定文件，并且代码中预留 `extractInfo`, `initDictionary`, `divideWords` 接口，为第二部分做好充分准备。

二、 实验环境

开发环境：64 位 Win10 系统，Visual Studio Community 2017，C++ 语言

三、 抽象数据结构说明

本次实验实现的数据结构有字符串 `CharString`、字符串链表 `CharStringLink`、栈 `Stack`、向量 `Vector`、哈希表 `HashMap`。

1、 字符串 CharString

`CharString` 类的成员变量有：`int` 类型的 `_len` 和 `_capacity` 分别表示当前长度和容量，以及 `wchar_t*` 类型的 `_str` 存储字符数组。

该类的实现采用动态分配内存的方式，即：若当前容量不足时，重新申请分配一段两倍原长的内存，将字符串拷贝到新内存，并释放原内存。这种实现方式保证了字符串的 `+=` 操作（在末尾连接字符串）的均摊复杂度是O(m)，m 为被合并的字符串的长度；而空间是 O(length) 的。

`CharString` 实现的成员函数（运算符）有：

- 查询：`length`, `capacity`, `empty`, `indexOf`（子串第一次出现的位置，使用 KMP 算法）
- 修改：`clear`, `assign`（`operator =` ）, `operator +=` , `reserve`（扩充容量）
- 索引：`operator []`, `substring`

友元函数（运算符）有：

- 连接：`concat`（`operator +`）
- 输入输出：`operator <<` , `operator >>`

2、 字符串链表 CharStringLink

`CharStringLink` 类采用双向链表的形式实现，链表的每个节点存有 `CharString` 类型的数据。该类在执行插入和删除等操作时动态申请（销毁）对应节点的内存。

`CharStringLink` 类实现的成员函数（运算符）有：

- 查询：`empty`, `search`（查找某元素第一次出现的下标）

- 修改：operator = , push_back（add）, push_front, remove, concat
- 遍历：begin, end（返回迭代器类型，同时还实现了迭代器的 operator ++ 和 operator *）

还重载了友元运算符 operator << 用于输出。

3、 栈 Stack

Stack 是一个模板类，它的模板参数是 value_t，表示元素的类型。它的成员变量有：int 类型的 _top, _capacity 分别表示栈顶的下标和当前容量，以及 value_t* 类型的 _stack 存储栈中元素。

该模板类的实现也采用了动态分配内存的方式，即容量不足时重新分配一段两倍的内存。这样保证了栈的 push 操作的均摊复杂度是 O(1) 的；而空间是 O(size) 的。

Stack 模板类实现的成员函数（运算符）有：

- 查询：size, capacity, top, empty
- 修改：operator = , push, pop, reserve

4、 向量 Vector

Vector 是一个模板类，实现与 Stack 基本类似。它的模板参数是 value_t；它的成员变量有：int 类型的 _size, _capacity 和 value_t * 类型的 _vector。

该模板类也采用动态分配内存的方式，保证了 push_back 的均摊复杂度是 O(1) 的，而空间是 O(size) 的。

Vector 模板类实现的成员函数（运算符）有：

- 查询：size, capacity, back, front, empty
- 修改：push_back, pop_back
- 索引：operator[], at

5、 哈希表 HashMap

HashMap 是一个模板类，它的模板参数是 key_t, value_t 分别表示键和值的类型，以及函数指针类型的 hash_t hashFunc(const key_t&)，表示 hash 函数。使用该类时需要保证每一个元素的键都是唯一的。

该类采用链表的方式实现。成员变量 unsigned hashSize 是一个质数，作为 hash 的模数，也就是 hash 表表头数组 List **head 的大小。List 是链表节点的类型，它的成员有 key_t key，value_t value，List* next。当往 hash 表中插入新元素时，将键的 hash 值对 hashSize 取模后，加入到对应链表的表头，查询时类似。

该类采用动态分配内存的方式：当前的节点总数接近 hashSize 时，重新分配两倍的内存。这样保证了在 hash 函数理想时，查询和索引的复杂度是期望 O(1)的，插入的复杂度是均摊 O(1) 的，空间复杂度是 O(size) 的。

HashMap 模板类实现的成员函数有：

- 查询：size, find（是否给定的键是否存在）
- 修改：reserve, insert
- 索引：operator []

四、 算法说明

1、 Html 解析

`HtmlParser` 类采用栈标签全遍历的形式解析 html 文档的结构。具体地，维护一个 `Stack<HtmlTag>` 类型的栈。

`HtmlTag` 是自定义的数据结构，它的成员：`CharString _type` 表示标签的类型（如 p, span, div 等），`Vector<TagAttribute> _attr` 存储标签属性（`TagAttribute` 类型本质上是一个 pair，有成员 `CharString _key`，`_value`），`CharString _content` 存储标签中的文本内容。

Html 的解析算法如下：从前往后扫描一遍 html 文本，同时维护标签栈。扫描过程中如果遇到：

1 开始标签（如 ）：新建一个 `HtmlTag`，提取标签类型和标签属性，并入栈。

2 文本内容：连接到栈顶标签的 content 的末尾。

3 结束标签（如 ）：如果结束标签的类型和栈顶一致，则弹出栈顶，并将栈顶的 content 连接到新栈顶的 content 末尾（嵌套内层标签的文本内容也属于外层标签）；否则（标签未正常关闭），不断退栈直到栈顶与当前结束标签匹配。

为了提取电影信息，在标签被弹出栈时，对标签进行判断，如果：

1 标签类型是 title，则执行相应操作提取电影名称。

2 标签类型是 div，且有属性 id="info"，则提取电影的导演、编剧等信息。

3 标签有属性 property="v:summary" 或 class="all hidden"，则提取剧情简介。

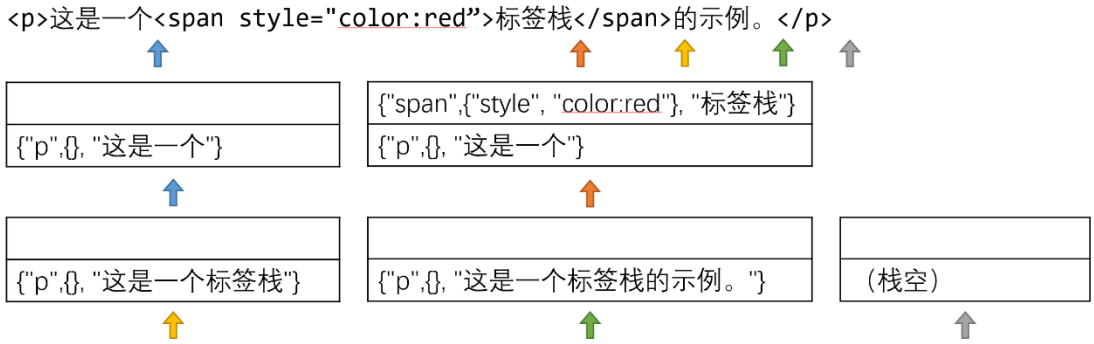

图 1 标签栈的示例（从下到上表示栈底到栈顶）

2、 中文分词算法

`WordSegmentor` 类的中文分词算法参考了 Python 中文分词组件 "Jieba" [1]，但是使用 C++语言自己实现了其算法，并对细节进行了一些针对性的调整。

本次实验的词典和 HMM 模型参数来自 Jieba [1]，停用词表来自 [2]。其中词典中有词频和词性信息。

算法框架如下：

1 利用词频求解最大概率分割，采用动态规划算法。

2 使用 HMM（隐马尔可夫模型）来处理连续的单字序列，识别未登录词。

3 将数字与量词拼接成词。

4　去除停用词。

1） 动态规划：词频和最大概率分割

将每个词的词频除以总词频，视作该词出现的伪概率。然后使用动态规划算法求解最大概率的分割方案，状态转移方程为：

$$f[i] = \max_{j<i} \{ f[j-1] \times \frac{freq(s[j:i])}{total\_freq} \}$$

其中 f[i] 表示 s[:i] 的最大概率分割的概率。具体实现时，为了保持精度，所有数值在对数意义下运算。

2） HMM：未登录词

由于以上 DP 算法倾向于将未登录词拆分成单字，因此使用 HMM 来处理连续的单字序列。

具体地，HMM 中每个观测（单字）分别有四种可能的隐状态（B M E S，分别表示词的开始、词中、词的结尾以及单字成词）。用 Viterbi 算法计算出最大概率的隐状态序列，并根据隐状态将其分词。

3） 数字与量词

将分词结果中的数字和量词连接，如‘2019’和‘年’应为‘2019 年’，量词的判断依据词典中的词频和词性，如‘个’在词典作为量词结尾的频率极高。

其余细节，如中文和西文字符（字母、数字）、汉字和标点、空白字符等的处理详见代码。

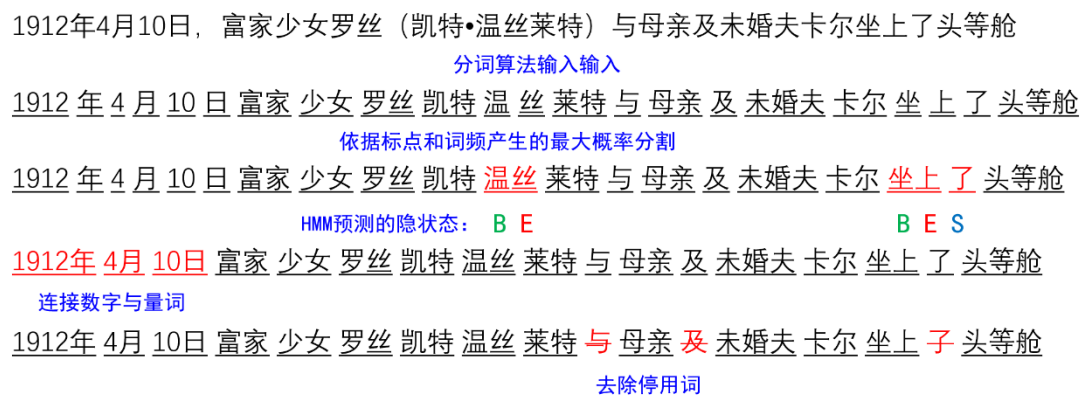

图 2 中文分词算法示例（仅供示例，实际句子有所不同）

五、 输入输出及操作相关说明

在可执行文件同名文件夹下应有 FilmContentSystem.config 文件，程序将从中读取相应配置信息，否则将采用默认配置。

```
DICT_PATH = "dict/dict.txt"
HMM_PATH = "dict/HMM.txt"
STOP_PATH = "dict/stopwords.txt"
USE_HMM = true
USE_STOP = true
INPUT_DIR = "input"
OUTPUT_DIR = "output"
```

config 文件的示例如右。其中 dict.txt，HMM.txt，stopwords.txt 都应以 UTF-8 编码存储，USE_HMM 和 USE_STOP 分别表示分词时是否使用 HMM 和停用词表。只有配置以上 7 个键值是有效的，请保证 config 文件的正确性，以及 config 文件应存储为 UTF-8 编码，否则将不保证程序能正常运行。

在 INPUT_DIR 下应放置 html 文件，请保证它们以 UTF-8 编码存储。在 OUTPUT_DIR 下将输出每个 html 文件同名的 info 和 txt 文件，分别存储电影信息和分词结果，编码也为 UTF-8。

在以上都确保无误的情况下，点击可执行文件运行即可。在屏幕上可能会打印一些日志信息，如读取、解析、分词等的运行时间，这是正常的。

六、 流程概述

程序的运行流程由 `FilmContentSystemApplication` 类的成员函数 `run` 实现。

具体的流程为：

1 读取 config 文件（`loadConfig`）
2 载入词典、停用词典、HMM 参数（`initDictionary`）
3 从 INPUT_DIR 中查找所有 html，并执行以下流程
   a) 读入文件
   b) 解析 html 并提取电影信息（`extractInfo`）
   c) 中文分词（`divideWords`）
   d) 将电影信息和分词结果输出到指定文件

七、 实验结果

示例的输出结果截图如下，输出结果符合预期。

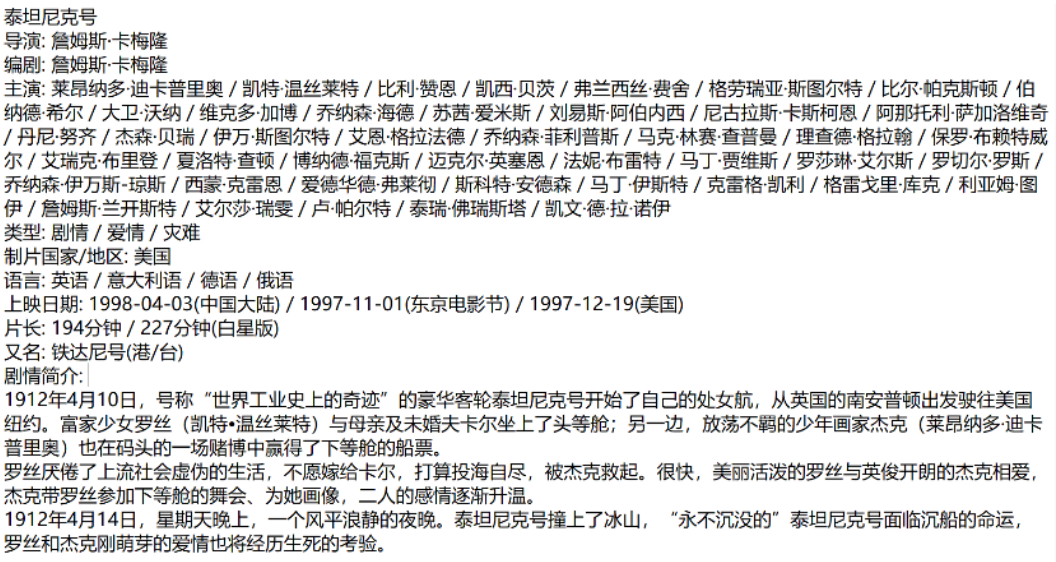
泰坦尼克号
导演: 詹姆斯·卡梅隆
编剧: 詹姆斯·卡梅隆
主演: 莱昂纳多·迪卡普里奥 / 凯特·温丝莱特 / 比利·赞恩 / 凯西·贝茨 / 弗兰西丝·费舍 / 格劳瑞亚·斯图尔特 / 比尔·帕克斯顿 / 伯纳德·希尔 / 大卫·沃纳 / 维克多·加博 / 乔纳森·海德 / 苏茜·爱米斯 / 刘易斯·阿伯内西 / 尼古拉斯·卡斯柯恩 / 阿那托利·萨加洛维奇 / 丹尼·努齐 / 杰森·贝瑞 / 伊万·斯图尔特 / 艾恩·格拉法德 / 乔纳森·菲利普斯 / 马克·林赛·查普曼 / 理查德·格拉翰 / 保罗·布赖特威尔 / 艾瑞克·布里登 / 夏洛特·查顿 / 博纳德·福克斯 / 迈克尔·英塞恩 / 法妮·布雷特 / 马丁·贾维斯 / 罗莎琳·艾尔斯 / 罗切尔·罗斯 / 乔纳森·伊万斯-琼斯 / 西蒙·克雷恩 / 爱德华德·弗莱彻 / 斯科特·安德森 / 马丁·伊斯特 / 克雷格·凯利 / 格雷戈里·库克 / 利亚姆·图伊 / 詹姆斯·兰开斯特 / 艾尔莎·瑞雯 / 卢·帕尔特 / 泰瑞·佛瑞斯塔 / 凯文·德·拉·诺伊
类型: 剧情 / 爱情 / 灾难
制片国家/地区: 美国
语言: 英语 / 意大利语 / 德语 / 俄语
上映日期: 1998-04-03(中国大陆) / 1997-11-01(东京电影节) / 1997-12-19(美国)
片长: 194分钟 / 227分钟(白星版)
又名: 铁达尼号(港/台)
剧情简介:
1912年4月10日，号称“世界工业史上的奇迹”的豪华客轮泰坦尼克号开始了自己的处女航，从英国的南安普顿出发驶往美国纽约。富家少女罗丝（凯特•温丝莱特）与母亲及未婚夫卡尔坐上了头等舱；另一边，放荡不羁的少年画家杰克（莱昂纳多·迪卡普里奥）也在码头的一场赌博中赢得了下等舱的船票。
罗丝厌倦了上流社会虚伪的生活，不愿嫁给卡尔，打算投海自尽，被杰克救起。很快，美丽活泼的罗丝与英俊开朗的杰克相爱，杰克带罗丝参加下等舱的舞会、为她画像，二人的感情逐渐升温。
1912年4月14日，星期天晚上，一个风平浪静的夜晚。泰坦尼克号撞上了冰山，“永不沉没的”泰坦尼克号面临沉船的命运，罗丝和杰克刚萌芽的爱情也将经历生死的考验。

图 3 电影信息提取结果

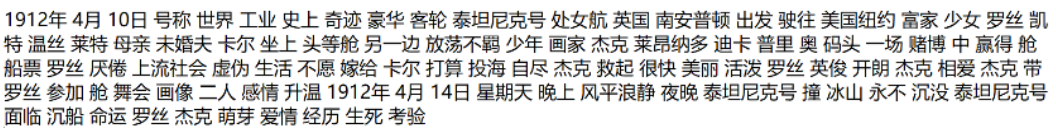
1912年 4月 10日 号称 世界 工业 史上 奇迹 豪华 客轮 泰坦尼克号 处女航 英国 南安普顿 出发 驶往 美国纽约 富家 少女 罗丝 凯特 温丝 莱特 母亲 未婚夫 卡尔 坐上 头等舱 另一边 放荡不羁 少年 画家 杰克 莱昂纳多 迪卡 普里 奥 码头 一场 赌博 中 赢得 舱 船票 罗丝 厌倦 上流社会 虚伪 生活 不愿 嫁给 卡尔 打算 投海 自尽 杰克 救起 很快 美丽 活泼 罗丝 英俊 开朗 杰克 相爱 杰克 带 罗丝 参加 舱 舞会 画像 二人 感情 升温 1912年 4月 14日 星期天 晚上 风平浪静 夜晚 泰坦尼克号 撞 冰山 永不 沉没 泰坦尼克号 面临 沉船 命运 罗丝 杰克 萌芽 爱情 经历 生死 考验

图 4 中文分词结果（为了显示效果，将实际输出文件中换行被替换为空格显示）

八、 功能亮点说明

1、自行实现的哈希表 `HashMap` ，详见“抽象数据结构说明”部分及代码注

释。

2、细致的中文分词算法：运用词频的动态规划算法；HMM 模型处理未登录词；停用词去除；数字和量词的连接。

九、 实验体会

由于实现的内容较丰富，文档篇幅略微超出限制，敬请谅解。

参考资料

[1] fxsjy/jieba: 结巴中文分词, https://github.com/fxsjy/jieba/

[2] 最全中文停用词表整理（1893 个） - 以家为家，以乡为乡，以国为国，以天下为天下 - CSDN 博客, https://blog.csdn.net/shijiebei2009/article/details/39696571

[3] c++ - Read Unicode UTF-8 file into wstring - Stack Overflow, https://stackoverflow.com/questions/4775437/read-unicode-utf-8-file-into-wstring)