基于文本内容的销售线索检索

*实验目的：*对网页进行信息提取；对字符串进行中文分词

*实验环境：*Mac 的 Vmare 虚拟机下， Windows10 Visual Studio 2012

抽象数据结构说明：

数据结构	函数名称	函数功能
栈 Stack	push	压栈
	pop	退栈
	top	返回栈顶元素
	empty	判断栈是否为空
字符串 CharString	indexOf	查找子串的位置
	subString	截取字符串，返回一个新串
	concat	将另外一个字符串接到当前字符串的末尾
	operator=	重载赋值操作符
	insert	在字符串末尾插入字符
	remove	删除第 i 个字符
	operator==	判断操作符
	operator[]	返回第 i 个位置元素的引用
	operator<<, >>	输出流、文件输入流
	removeSpace	删除字符中的所有空格
	size	返回字符串的规模
	data	将字符串转化为字符数组
字符串链表 CharStringLink	add	添加元素（字符串、字符数组、另外一个字符串链表）
	remove	删除元素（字符串、字符数组），删除第 i 个位置的元素
	search	查找某元素的位置（字符串、字符数组）
	size	字符串链表的规模
	operator<<	重载输出流（逆序输出）
字典 Dictionary	init	字典的初始化
	put	插入词条
	search	查询某个词是否在词典中
	size	查询词条的数目
	dividePhrase	对短语进行分词（短语中只有中文字符）
	divideSentence	对句子进行分词（句子中可以有标点、数字、英文）
网页处理 WebsiteProcessor	extractInfo	处理文件中存储的网页 url，将处理结果存储到另外一个文件
	initDictionary	初始化词库
	divideWords	对某个句子进行分词

编码转换 Converter	UnicodeToChinese	将 unicode 编码转换为中文

算法说明

网页文件的解析和提取

1. 首先从头扫描整个网页源码，转 1。
2. 如果找到了<***>，</***>，<***/>形式的子串，则生成一个节点，节点分别标记为左界、右界、自匹配，转 3。如果没有找到，则转 4。
   Note ： 左界、右界、自匹配的说明例子
3. 如果当前节点是左界，则将结点压入栈中；

```
<a href="http://dealer.cehome.com/" target="_blank">代理商</a>

<meta name="keywords" content="" />
```

   如果当前节点是右界，则将结点弹出，同时根据节点的标签分别处理；
   如果当前节点是自匹配，则不作处理。转 2。
4. 如果已经搜集到所有的信息，则将信息写入文件；否则，则向文件输出空。
   Note 1: 此处采取的策略是搜集所有可能满足如下条件的模式串，然后在搜集完成后，挑选其中特定位置的模式串中的内容，即为所要提取的信息。

模式串	包含的文本类型
<a>***	发帖大类、小类、标题
<td class=”t_f”>***</td>	发帖内容
<a>***	发帖人
<a>***	发帖日期
<div class=”ts z hl”><a>***	发帖分类、标题

Note 2: 由于<td></td>中除了发帖内容外，还可能出现其他的标签，使用 getText(“WebsiteProcessor.h”)对一个字符串去除标签内容，保留文本内容。

分词算法

1. 首先正向扫描一个句子 sentence，根据其中的英文标点符号（单字节）、数字（单字节）和中文标点符号（双字节）将其分成多个短语（phrase）。
2. 对每个短语分别进行分词，采用逆向最大匹配法。

a) 从后往前取一个固定长度的子串。若子串长度大于 0，转 b；否则，结束。
b) 如果子串在词库中，则将其加入分词结果，将子串从短语中删除，转 a。
c) 如果子串不在词库中，则长度减少 1，若子串长度大于 0，则转 b；否则，取出末尾的汉字，加入分词结果，短语中删除该汉字，转 a。

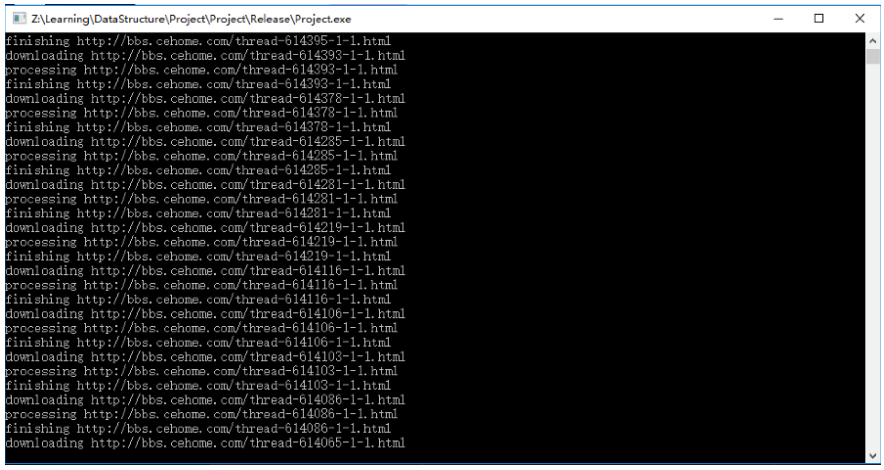
Note 1 : 分词算法可以选择删除还是保留无用词。无用词词库从/Dictionary/uselessWord.dic 加载，其中包括中文中的连词、感叹词、谓词等。

Note 2 : 此处采用了 hash 算法来实现对词库中词语的快速查找，每个桶都是一个字符串链表，假如两个字符的 hash 码相同，则将它们放入同一个桶。桶容量为素数 10133， 总共的词条数为 275909 条，平均意义下的每个字符串链表的长度为 27， 实际测得的字符串链表最长为 48。因此这种策略实现了空间和时间之间较好的折中，如果希望更加节省空间或者更加节省时间，只需要调整桶容量即可。

流程： [载入词库]->[读取网页到本地文件]->[从本地文件中提取关键信息]->[中文分词]->[输入结果]

遇到的问题及解决： 下载下来的网页源码中，有形如“我”的字符串，没能找到相关的函数去解析这些字符串。于是自行写了一个函数 UnicodeToChinese(“Convert.h”)，通过将“我”转化为整数 25105，然后再转化为 wchar_t，然后调用 wstring_convert，转化为对应的中文字符串。

操作说明： 打开 Project.exe，出现如图所示界面，显示当前正在下载或者处理的网页信息。



实验结果： 100 个网页中有 3 个无法获取信息（网页本身无法访问）。

序号	网址	发帖大类	发帖小类	发帖标题	发帖内容	发帖人	发帖日期	发帖类型	分词结果
1	http://bbs.	挖掘机专区	我爱我挖	求购	精品/打算最近买台小挖，在农村做事。	青春的会是	2016-1-3	提车作业	求购 精品 小挖 到 进口 合资 机 优先 打算 最近 买 台 小挖 农村 做事 邮号 不能 小于 车况
2	http://bbs.	挖掘机专区	杂谈图库	本人	系最近本人最近想求购一台200型挖掘机	哇哇咖	2016-1-3	生活杂谈	本人 系 最近 想 求购 一台 型 挖掘机 本人 最近 想 求购 一台 型 挖掘机 万 左右 请 大家 推
3	http://bbs.	挖掘机专区	选购询价	乌合	大养才本人真心求购一台个人二手小挖机	成勇	2016-1-3	6-12吨	乌合 大养 求购 二手 挖机 一台 本人 真心 求购 一台 个人 二手 小 挖机 三一 品牌 也行 时间
4	http://bbs.	挖掘机专区	选购询价	求购	三一6 最好是法勿机求浙江地区或者安徽	铁甲王子多	2016-1-3	6-12吨	求购 三一 最好 法 勿 机 求 浙江 地区 安徽 三一 代理
5	http://bbs.	挖掘机专区	选购询价	求购	台3 翻新的潲油的勿扰	和v悦	2016-1-3	13-20吨	求购 一台 万 左右 小松 翻新 潲油 勿扰
6	http://bbs.	挖掘机专区	选购询价	求购	55或6 买来自己用，车况要好，最好是山	电话15169	半支烟的角	2016-1-3	5吨以下 求购 小挖 买 来自 车况 要好 最好 山东 电话
7	http://bbs.	挖掘机专区	选购询价	求购	小挖掘车况好点，30-55的都行，价格3-	大佛羚兹	2016-1-3	5吨以下	求购 小 挖掘机 车况 好点 都行 价格 万元
8	http://bbs.cehome.com/thread-614285-1-1.html								
9	http://bbs.	挖掘机专区	选购询价	求购	有没有卡特机子11年的，价格35w	半世	2016-1-2	21-30吨	求购 有没有 卡特 机子 年 价格
10	http://bbs.	挖掘机专区	选购询价	求购	求购卡特320，住友200a3	soudwe42	2016-1-2	21-30吨	求购 求购 卡特 住友
11	http://bbs.	挖掘机专区	我爱我挖	求购小挖	10年以后的小挖机	小甜86971	2016-1-2	横作技巧	求购 小挖 年 以后 小 挖机
12	http://bbs.	挖掘机专区	选购询价	求购久保田	本来想去深圳看机，水大深呢！！	特大近	2016-1-2	5吨以下	求购 久保田 本来 想去 深圳 看 机 水 大深 还 铁甲 里 投 广 人机 较好 微 号
13	http://bbs.	挖掘机专区	选购询价	安徽求购	一手机，，，欢迎机主来电，1381	1.39E+10	2016-1-2	6-12吨	安徽 求购 精品 小挖 一台 小松 卡特 日 力 手机 欢迎 机主 来电
14	http://bbs.	挖掘机专区	选购询价	广东省内	求购小挖，性能可靠，无暗病，价	我风我发	2016-1-2	6-12吨	广东 省内 求 台 机 牌子 不限 性能 可靠 无 暗病 价格合理 最好 已 改好 装 插 木棚 话 上图
15	http://bbs.	挖掘机专区	选购询价	求购	想买台100 150的机子	饮州小吴	2016-1-1	13-20吨	求购 想买 台 机子
16	http://bbs.	挖掘机专区	选购询价	求购	12-16要有合格证发票，原版漆，土方机	修不完的农	2016-1-1	6-12吨	求购 之间 一手 合资 小挖 要有 合格证 发票 原版 漆 土方 机 年限 年前 后 求购 地点 江西
17	http://bbs.	挖掘机专区	选购询价	求购	一台割求购一台新源90轮挖，带灭木器的	Chuanjie	2016-1-1	机型对比	求购 一台 新源 轮 挖 带 灭 木器 求购 一台 新源 轮 挖 带 灭 木器 手机号
18	http://bbs.	挖掘机专区	选购询价	求购	小挖 60机价格在17万左右年限不超过2	开山2013	2016-1-1	6-12吨	求购 小挖 一台 机 价格 万 左右 年限 不 超过 年 国产 小 牌子 免谈
19	http://bbs.	挖掘机专区	选购询价	求购	小挖1 微信13561085566	广饶求购	2016-1-1	5吨以下	求购 小挖 年 下半年 以后 系列 微 信
20	http://bbs.	挖掘机专区	杂谈图库	求60	福定地区求，口力60，70 神刚7	lby123	2016-1-1	美图转发	求 福建 地区 求 口 力 神 刚 三一 二十 上 下
21	http://bbs.	挖掘机专区	选购询价	求购	20与台求购一台15万左右的日历200挖机	山东大	2016-1-1	21-30吨	求购 吨 挖机 求购 一台 万 左右 日历 挖机 要求 日本 原装 进口 报 过关 无弹 史 大件 没 动
22	http://bbs.	挖掘机专区	选购询价	求购	小挖时河北境内求购小挖机一台价格十五	liujinliang	2016-1-1	机械对比	求购 小 挖机 河北 境内 求购 小 挖机 一台 价格 十五 左右 电话
23	http://bbs.	挖掘机专区	选购询价	吉海附近	求购一台60 吉A8小聪	吉A8小聪	2016-1-1	6-12吨	吉海 附近 求购 吉海 附近 求购 一台 微 信
24	http://bbs.	挖掘机专区	选购询价	买台二手马中友们，现代150车怎么样？我想	唐纲能	2016-1-1	13-20吨	买 台 二手 现代 中 友 现代 车 样 想买 台 二手 车 年 值 多少钱 土方 机
25	http://bbs.	挖掘机专区	选购询价	求购	台2 个人和我联系30万以内微信15252	徐州苏刚	2016-1-1	机型对比	求购 一台 以上 挖机 一台 个人 和我 联系 万 以内 微 信
26	http://bbs.	挖掘机专区	选购询价	广西求购	一台可以转分期的小松3k潮大鸡		2016-1-4	30吨以上	广西 求购 一台 转 分期 挖掘机 广西 求购 一台 转 分期 小松 非 诚 勿扰
27	http://bbs.	挖掘机专区	选购询价	求购卡特3	求购卡特323挖机，要12年以后的	求购卡特3	2016-1-4	21-30吨	求购 卡特 挖机 求购 卡特 挖机 年 以后 原厂时间 发过 求购 贴 资金 问题 谈 买成 初步 打算

功能亮点

1. 自行设计和改进了 hash 函数。具体参见算法说明。
2. 分词算法能够过滤掉英文、标点、数字等，同时还能去除中文中的无用词，保留信息量更大的词语，为后续的搜索做好准备。具体算法参加算法说明。

实验体会

在建立一个较大的系统时，选取效率较高的数据结构就显得至关重要，比如这次实验中使用的——字符串的 KMP 算法、使用 hash 的字典。

此外，自己在不同的编码之间的转换问题上还需要更多的学习，这次在编码转换问题上花费了较多时间，采用的策略也不够优雅。