**Web数据挖掘项目报告**

——钱昕扬 2016118080

1. **项目概要**

本项目（Pricer）是基于windows环境运用python语言进行开发的csgo饰品价格查询软件，可用于在igxe网站上进行csgo游戏饰品的价格查询与比对，对于igxe.cn网站进行文本爬取，信息检索以及数据分析，通过一定算法来对于网页源代码进行特征标记，从而从纷繁复杂的源代码中检索到需要的信息，并反馈给用户。同时，本软件还支持精确查询与模糊查询，并可将查询到的结果按照某种规则返回给用户，并且将查询到的数据进一步存储，以便日后对于某类数据进行研究分析。

1. **用户需求**

为了了解网站上某件csgo游戏饰品的价格，多次打开网站以及人工检索是非常繁琐的。用户希望在不用打开网站的情况下，就能得知某样游戏饰品的价格以及同类饰品的价格状况，从而对于游戏饰品进行选择和购买。例如，某用户想要购买一把蝴蝶刀，通过本项目，用户就可以实现一键查询具体某种蝴蝶刀的价格信息，并还可以通过模糊查询其他相应类型游戏饰品的价格，并结合得到文本数据进行进一步分析，从而帮助用户完成蝴蝶刀的选择与购买，而不用再次打开网站进行查询。

1. **功能描述**

主要功能如下：

1. **精确查询**

该功能可支持用户对于某个具体csgo饰品的价格进行查询。用户通过键入具体csgo饰品名称进行查询，输入的饰品名称一定要完全正确，否则会导致查询不成功。成功查询后将返回相应csgo饰品的价格结果；若出错，则会有相应错误提示（网络连接错误/饰品名称输入不准确/网站更新变化等）。并且用户可以同时对于多个csgo饰品的价格进行查询，输入的时候用分号将不同的饰品名称隔开，成功查询时会按次序返回相应饰品的价格。

1. **模糊查询**

该功能可支持用户对于某一类的csgo饰品的价格进行查询。用户通过键入相关关键词进行查询，输入名称不需要完全正确也可以成功查询。查询成功时，程序将会返回与搜索关键词相关的数条结果，并按顺序显示；查询失败时，则会有相应错误提示（网络连接错误/饰品名称输入不准确/网站更新变化等）。考虑到运行时间，程序默认返回10条最相关结果，用户可以通过加载更多来获取更多结果。通过模糊查询，可以对于同类游戏饰品价格进行比较分析，得出结论。

1. **规则排列**

该功能对于获得的结果按照一定规则的进行排序，从而使得数据关系更加清晰明确，方便用户进行进一步的分析。程序初步支持价格升序排列与价格降序排列，用户通过相应按钮的点击，可以使得结果按照价格升序或是降序呈现。

1. **数据分析**

该功能可以对于某种具体饰品历史数据进行分析比对，从而决定什么时候进行饰品的买入与卖出。在每次查询后，程序将会对于查询结果进行文件保存；随后，可供用户在历史数据中进行具体类型的数据查阅，并描述数据之间关系。

1. **具体实现**

主要功能的具体实现方法如下：

1. **精确查询**

整体上使用requests库中的相应函数对于igxe网站的网页源代码进行抓取。首先，通过对于相关URL的不断遍历获取相应网页页面信息；在得到网页页面信息后，对于源代码进行观察，针对想要检索的数据的前后特征来进行特征标记，从而获取标记之间的有用信息，并将运用字典类型对于信息进行存储记录，字典中的键是饰品名称，值是饰品的价格数据；最后，将记录的有用信息筛选整理后呈现给用户。

1. **模糊查询**

针对用户输入的关键词，通过相关URL并结合requests库函数进行网页遍历，并分析网页源代码及进行信息检索，从而获取到与关键词最相关的一个定长的名称列表。将该列表中的饰品名称进行精确查询，并将查询后的结果按照相关性顺序呈现给用户。为了程序运行时间不会过长，设定了查询上限——20个结果，用户可以通过“加载更多”按钮来进行更改，每次多加载五个结果。

1. **规则排序**

对于临时存储的字典中的信息进行处理，借助operator相关函数与迭代器，从而根据值的大小对于字典中的键值对进行排序，结果转化成一个元素为元组的列表，整理信息后呈现给用户。

1. **信息存储**

查询数据的保存通过文件存储来完成，每一个饰品的信息用以饰品名称命名的txt文件来存储，存储的信息格式为“[时间]： 价格”。因为文件命名时，不能出现某些具体的字符，所以要对于特殊文件名称进行转码和解码。

1. **关键函数**

1.def inquiry\_price\_sure(self):

进行精确查询功能。

调用函数4，5，6，10。

2. def inquiry\_price\_fuzzy(self):

进行模糊查询功能。

调用函数3，4，5，6，10。

3.def get\_names(self, search):

获取模糊查询时需要的相关饰品名称列表。

输入类型为string，返回类型为list。

调用函数10。

5.def inquiry\_process(self, searchList):

进行源代码的抓取以及标记过程，从而获取有用信息。

输入类型为list。

5.def print\_result(self):

对于检索信息结果进行整理并打印。

6.def save\_results(self):

对于得到数据进行存储

7.def more\_results(self):

加载更多模糊查询的结果。

调用函数2。

8.def up\_sort\_process(self):

将所得结果按照价格升序排列。

9.def down\_sort\_process(self):

将所得结果按照价格降序排列。

10. def remove\_duplication(self, input: list):

去除列表中的相同元素并对于列表进行排序。

输入类型为list，返回类型为list。

1. **运行结果**



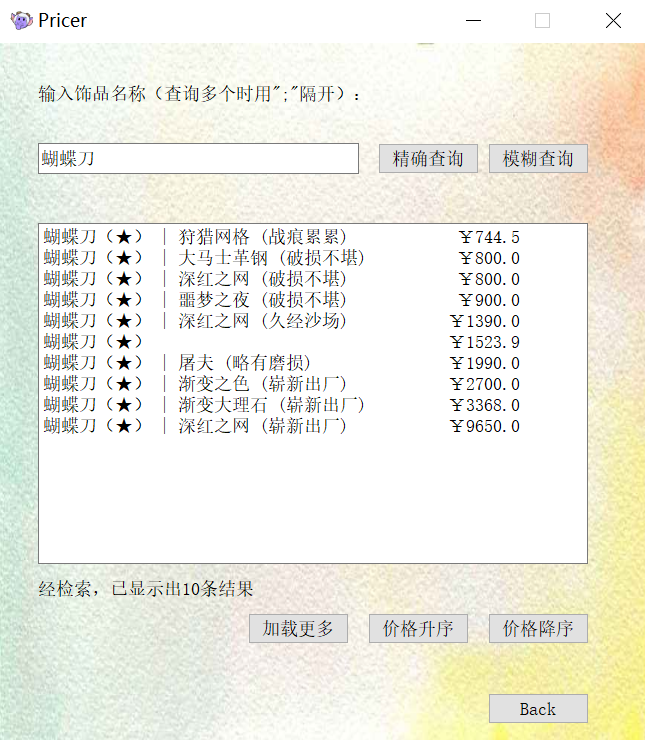
菜单界面



精确查询



模糊查询



价格升序



加载更多