# 编译原理实践第1课 --- 正则表达式

说明:

(1) 本次作业提交最晚时间为9月26日晚23:59。

(2) 将您的解答填写至每道题后. 如果是程序题, 代码量不大的话，请将代码放至文本框内，如下所示:

#/usr/bin/perl

print "HELLO WORLD!\n”

代码量大的话, 请提交代码文件, 并在相应地方填写，如下所示:

见\*\*\*.pl文件

(3) 如果有多个提交文件, 请打包压缩后提交. 请使用zip、tar.gz等压缩格式，请匆使用rar格式。

一、分别为以下各题写出相应的正则表达式

(1) 给定一个文件名, 验证其格式是否为jpg, jpeg, gif, bmp, 大小写均可.

答:

(2) 识别C语言代码中的标识符 (C语言的标识符是字母、数字和下划线组成的串).

答:

(3) 识别C语言代码中的无符号数，例1946,11.28,63E8,1.99E6等.

答:

二、编程题

(1) 给定一个天猫商品页面, 写程序使用正则表达式抽取关于该商品的产品参数信息 (如果该商品有产品参数信息的话). 要求获取(参数名称, 参数内容)对.

例如: <https://detail.tmall.com/item.htm?spm=a220m.1000858.1000725.1.78b13727dl7gbT&id=598724949790&skuId=4185713437361&areaId=320500&user_id=880734502&cat_id=2&is_b=1&rn=8c61077a198717d0afd1a0298a02dfaf>

获取的内容如下:

生产许可证编号：SC12433020300393

产品标准号：GB/T19855

厂名：宁波思味特食品有限公司

厂址：浙江省宁波市海曙区集士港镇卖面桥村

厂家联系方式：400-800-4900

配料表：详情见包装

储藏方法：常温下通风、避光保存，开封后请尽快食用。

保质期：60 天

食品添加剂：详情见包装

品牌: Three Squirrels/三只松鼠

系列: 这就是月饼\_月下礼盒/640g

月饼馅种类: 月下礼盒640g/8只装

产地: 中国大陆

省份: 浙江省

城市: 宁波市

月饼价格区间: 100元-150元

包装方式: 包装

答:

(2) 提及句子中的单词及词性 (选作题)

读取bracketed.txt文本中的每一行, 并为每一行得到该句子的单词/词性序列. (注，如果bracketed.txt内容是乱码, 请使用bracketed.gbk.txt)

如:

(TOP (IP (NP (NP (NR 上海) (NR 浦东)) (NP (NN 开发) (CC 与) (NN 法制) (NN 建设))) (VP (VV 同步))) )

该句子包含的单词为: 上海 浦东 开发 与 法制 建设 同步，每个单词前的标签为其词性. 如 “上海”的词性为NR, “浦东”的词性为NR等。

得到的单词/词性序列为:

上海/NR 浦东/NR 开发/NN 与/CC 法制/NN 建设/NN 同步/VV

答: