```
作业代号：flounder
作业布置时间：2022.10.31
作业截止提交时间：2022.11.6
作业提交文件名：flounder.txt
```

下列选择题为单选题，请选择正确的答案

一、Kettle 是”Kettle E.T.T.L. Envirnonment”只取首字母的缩写，其中，E.T.T.L.分别表示的是：

A. Extraction, Transformation, Transport, 以及Loading

B. Extraction, Transport, Transformation, 以及Loading

C. Expression, Transport, Transformation, 以及Loading

D. Expression, Transport, Treatment, 以及Loading

二、下列对分装器（Wrapper）的描述中，哪句话是不准确的：

A. 分装器是针对与数据源直接通讯的数据集成系统的重要组成, 从高层向数据源发出查询请求是它的工作之一

B. 分装器可以将查询结果转换成查询处理器可以接受或操控的数据格式

C. 只有当数据源是结构化或半结构化时，才可以使用分装器

D. 当数据源是关系型DBMS时，访问数据源时，不需要分装器与ODBC（或类似的ODBC驱动）进行交互

三、对于下面的python代码片段，使用了正则的匹配。请问其输出结果是什么?

```
s=' 192.137.1.336  192.168.1.137.123  192.168.1.138 '

print(re.findall(r'(?<![\.\d])(?:25[0-5]\.|2[0-4]\d\.|[01]?\d\d?\.){3}(?:25[0-5]|2[0-4]\d|[01]?\d\d?)(?![\.\d])',s))
```

A. 192.137.1.336

B. 192.168.1.137.123

C. 192.168.1.138

D. 输出为空

四、在使用基于学习的分装器时，下面哪句话存在问题：

A. HLRT分装器可以看成是Stalker分装器模型的子模型

B. Stalker分装器与 HLRT分装器均可看成是有限元自动机建模

C. Stalker主要应对平坦的元组模式，而HLRT可以应对嵌套的元组模式

D. Stalker分装器可以应对简单的平坦结构，也可以应对复杂的嵌套结构

五、自己尝试一下，如果要将xml文件中的标签全部删除，下面的正则表达描述哪个是对的：

A. 使用 `<(\S*?)[^>]*>.*?|<.*? />` 是对的，但是使用 `<[^>]*>` 也是对的

B. 使用 `<(\S*?)[^>]*>.*?|<.*? />` 是不对的，使用 `<[^>]*>` 也是不对的

C. 使用 `<(\S*?)[^>]*>.*?|<.*? />` 是不对的，但是使用 `<[^>]*>` 是对的

D. 使用 `<(\S*?)[^>]*>.*?|<.*? />` 是对的，但是使用 `<[^>]*>` 不对

六、在进行网页爬虫类的Wrapper设计时，往往存在很大的挑战。下面对这些挑战的总结中，哪一句是不准确的：

A. 学习源模式T_s非常困难

B. 同时满足正例和反例的正则表达是不存在的

C. 学习抽取程序E_w非常困难

D. 即使下降了对E_w的图灵完备性限制，只需学习模型参数的有限集合也是极其困难的

七、尽管分装器形式多样，我们仍然可以把它们划分为几种类型?

A. 1）手工分装器、2）基于学习的分装器、3）无模式的自学习分装器、以及采用2）与3）技术的混合分装器

B. 1）手工分装器、2）基于智能的分装器、3）基于算法的分装器、以及采用2）与3）技术的混合分装器

C. 1）手工分装器、2）基于智能的分装器、3）无模式的自学习分装器、以及采用2）与3）技术的混合分装器

D. 1）手工分装器、2）基于学习的分装器、3）基于算法的分装器、以及采用2）与3）技术的混合分装器

八、对我们讲解的网页爬虫实例（即采用shell script爬取Linux命令行经验分享网站条目），下面我们得出的结论哪一条是不准确的：

A. 该实例展示的爬取方法，如果按照形式化的描述，其T_s与T_w是不等的

B. 针对该网站的内容爬取，使用shell script不是唯一的手段，使用其它的语言书写E_w，也可以实现目标

C. 该案例中，使用了HTML-XML-Utils，它是WHATWG组织开发出来的一组工具集合，目的是对HTML/XML进行内容抽取

D. 课堂上展示的爬取结果是markdown格式的，如果需要的话，开发者也可以在该案例中输出到关系数据库（如MySQL）中

九、在手工方式构建分装器时，下面哪一句话是合理的：

A. 开发者对需要抽取的页面进行分析研究，即分析其T_w的模式，是该方式的首要任务

B. 由于C语言是公认的高效计算语言，在手工构建封装器时，往往成为首选的编程语言

C. 这种方式本质上是使用手工方式完成E_w的构建，但是构建的策略收到T_w的制约

D. 如果对页面进行内容提取，由于网页页面均是HTML格式的，因此采用对DOM树的XPath解析方式是唯一的有效手段

十、「Kettle在Windows平台上可以连接访问任何支持ODBC的数据源」，那么，以下的描述的数据源集合中，哪一句是符合上述的要求的：

A. Hypersonic, MS Access, Firebird SQL, Oracle, Hadoop

B. MySQL, Firebird SQL, Shapefile, PostgreSQL

C. Oracle, MySQL, DB2, Sybase, SAP, Redis

D. Informix, DB2, SAP, ArcEngine