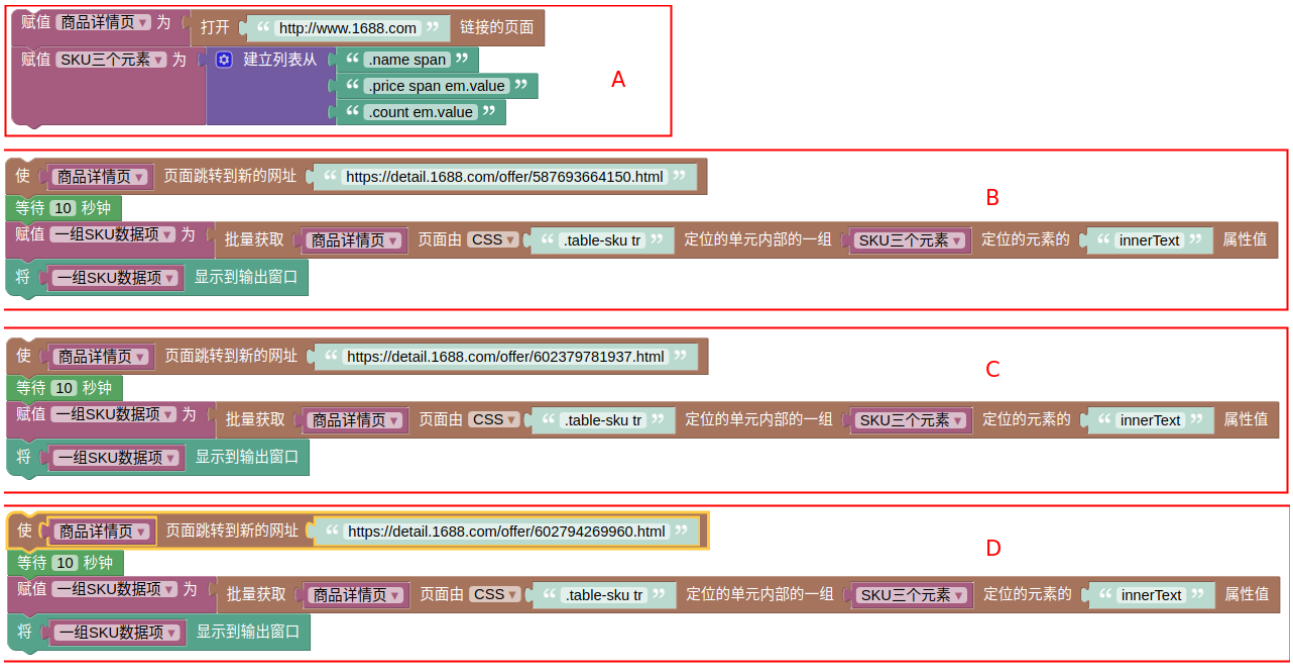
第五课. 循环提取多个网页内的同类数据

首先让我们整理一下现在面临的问题。

如果要采集一个商品详情页的 SKU 数据，我们需要操作 4 行积木块。如果我们想要获得一批商品详情页的 SKU 数据，例如 3 个竞品，那么至少要操作 3 次第四课的步骤。老老实实的一行一行搭建的话，我们需要拖入 2+3*4=14 行积木块，也就是像下面这样的一个流程。



显然，我们不可能用这样一个流程去采集成千上万的商品数据。

注意，流程开始的 A 组积木实现两个步骤：一是打开一个新的网页，这个网页将在后面的步骤中反复使用;二是定义了一个数组，用于存放 SKU 三个元素的 CSS 定位字符串。在 A 组积木后面的 B、C、D 组积木，都包括 4 个积木块，步骤基本上相同，唯一不同之处在于网址，这是因为流程分别要访问三个商品详情页：

https://detail.1688.com/offer/587693664150.html https://detail.1688.com/offer/602379781937.html https://detail.1688.com/offer/602794269960.html

每次打开网页后，为了等待 SKU 数据显示在网页上，需要等待 10 秒钟，然后提取 SKU 数据项，最后输出到窗口。

试想一下，要一模一样搭建 1000 个商品详情页的采集流程，会耗费多少时间，过程中如果操作错误，那么要找出错误之处，又要耗费多少精力。况且，这样搭建流程，根本不能称为自动化流程，所需的人力成本并不会比直接用鼠标从每个网页拷贝所需信息要少。

所以，既然我们要实现机器人自动化流程，那么接下来的学习任务也很明确了，就是用尽可能少的积木块，实现大量的步骤，特别是这种重复性非常高的“机械式”流程。

本课将介绍一个新的积木块——“循环”。我们将用它实现类似“把这几个步骤重复 1000 遍”这样的流程，从而避免人工重复搭建几千个积木块的问题。

从这一课开始，我们将真正进入自动化流程的搭建实验中。

绝大多数网页数据采集流程，都会用到循环类积木块，它是帮助我们节省人力成本的关键。常见于采集搜索结果、商品目录、订单列表、用户评论等具有多页内容的数据。请读者跟着步骤试着搭建好这个流程，通过指定数量不同的商品详情页网址，反复运行，掌握循环类积木块的使用方法。

搭建步骤详解

本课内容讲解如何用机器人流程自动化采集多个网页上一组指定元素的数据。

虽然用于演示的例子是 1688 网站的商品详情页的 SKU 数据，图示是循环采集三个网页的数据。读者学会后可自行开发流程批量采集任何网页的多个指定元素的数据。

此处以使用 1688 平台为例，选择三个商品详情页，运行机器人流程，无人工干预的提取各项 SKU 数据，所产生的最终结果类似于一张表格。（注：步骤中的图片只用来介绍方法中的主要步骤，并不连续。）

1. 第一步，运行电商记桌面版客户端的浏览器，或者安装了电商记插件的 chrome、360、QQ 等浏览器，打开“机器人流程自动化”页面（https://m.dianshangji.com/rpa.html）

2. 第二步，在“机器人流程自动化”页面 中搭建第一个**浏览类**积木块 A，用于打开一个新的页面，这个页面在后面的步骤中将被反复用于访问每个商品详情页。为了代表这个页面，我们会定义一个**变量**。出于演示的目的，图中将变量命名为“商品详情页”。

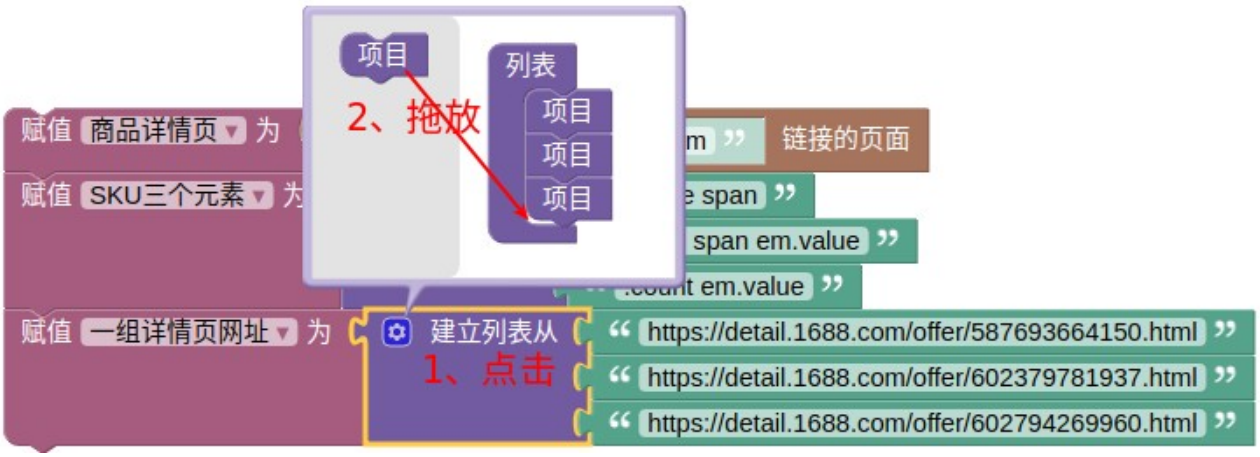
3. 第三步，在流程中添加**数组类**积木块 B，建立一个列表，表示商品详情页的 SKU 单元内三个元素各自对应的 CSS 定位字符串。分配新的变量“SKU 三个元素”，代表这个列表。由于这个列表包含三个元素的数据，因此需从**文本类**积木块中选三个空白文本块放入列表，并将第四课中取得的三个 CSS 定位字符串复制到列表中。

这组 CSS 定位字符串用于提取商品详情页的 SKU 数据（如图所示）。SKU 数据包含多个数据项，每个数据项由 SKU 标题、价格和库存构成，在网页上对应一个单元，包含三个文本元素。

SKU 数据项

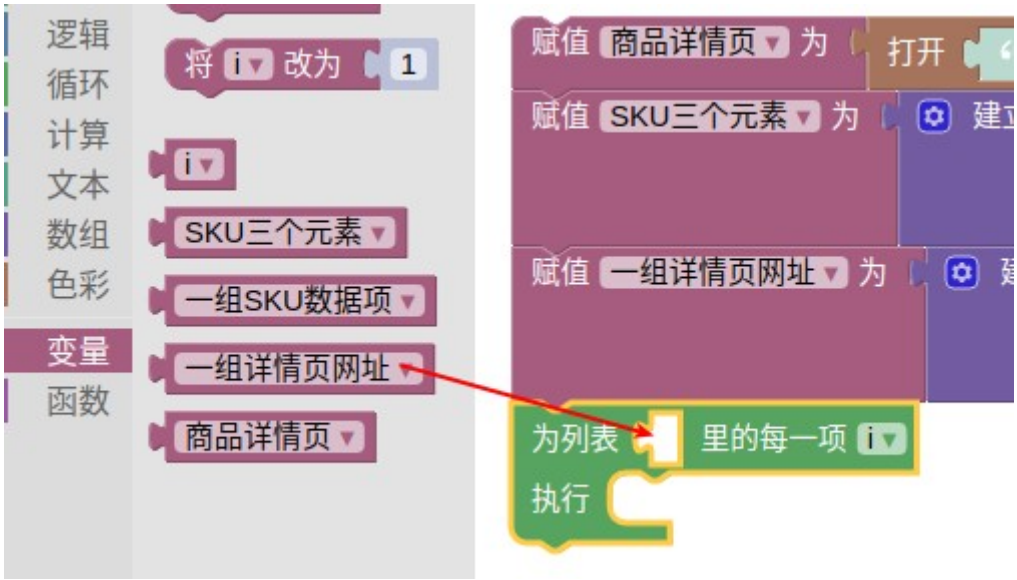
4. 第四步，再添加一个**数组类**积木块 C，建立一个列表，表示三个商品详情页的网址。分配新的变量“一组详情页网址”，代表这个列表。由于这个列表包含三个网址，因此需从**文本类**积木块中选三个空白文本块放入列表，并将三个网址的文本分别复制到列表中。

如果我们需要采集四个网址，可以按图示步骤，为列表增加一个项目：1、点击“建立列表”左侧图标;2、将“项目”从左侧拖放到右侧下方。要增加更多项目，只需按此步骤多次操作即可。

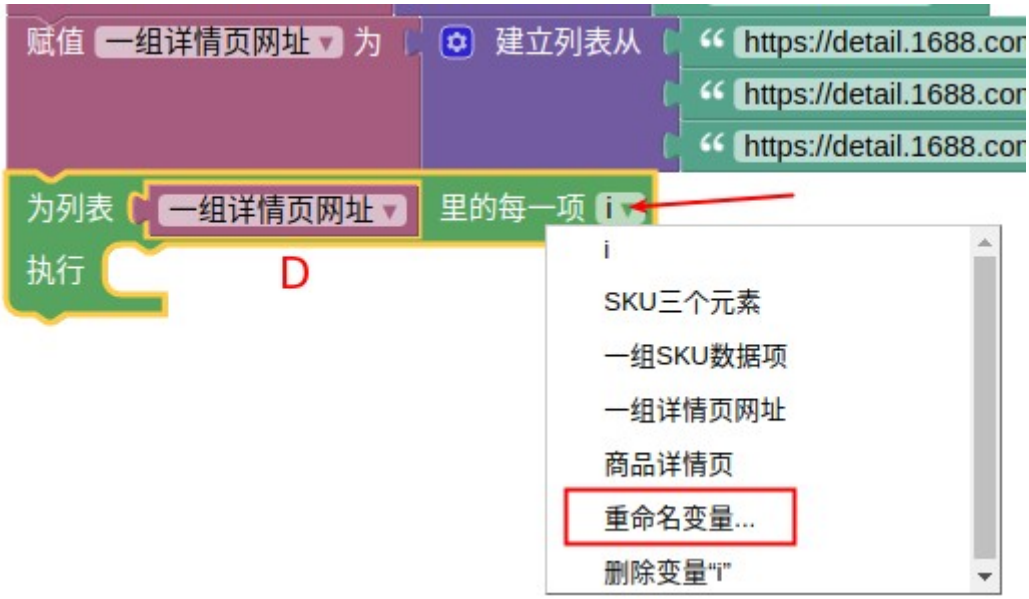


5. 第五步，添加**循环类**积木块 D，它的作用是为列表“一组详情页网址”中的每一项网址，重复运行采集 SKU 数据的四个积木块。

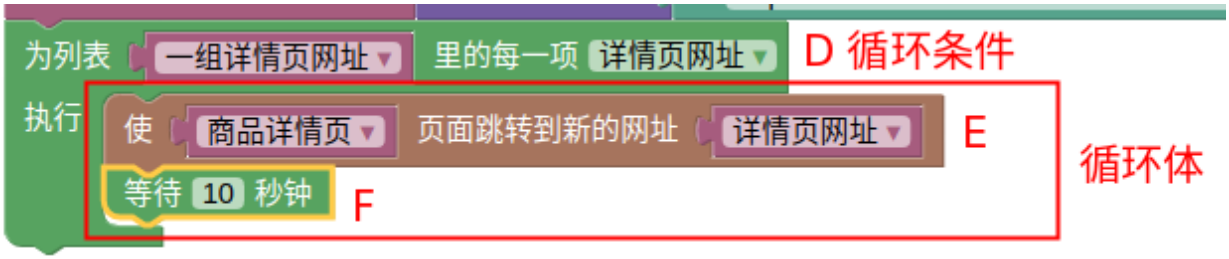
从**变量类**积木块中，将“一组详情页网址”这个变量拖放到积木块 D 的“列表”名称中。



然后将列表中的每一项“i”变量，重命名为“详情页网址”。



6. 第六步，添加**浏览类**积木块 E 到积木块 D 的空白区，然后添加**时间类**积木块 F。虽然这段流程没有什么实际意义，但是可以正常运行。请自己输入并尝试运行一下。



实际运行时，浏览器会每隔 10 秒钟，在“商品详情页”变量所代表的页面内自动跳转到一个网址，这个网址由“详情页网址”变量所指定，按先后顺序分别是“一组详情页网址”列表中的三项网址。

在图示这些积木块中，积木块 D 所在这一行，我们称为**循环条件**。D 所包含的一组积木块（如 E 和 F），我们称为**循环体**。

循环条件，表明了运行循环体内一组积木块时所需要的条件。在循环类积木块中，你可以根据需要使用各种条件来控制循环体。我们这里使用了最简单的一种，即将变量“详情页

网址”依次赋值为列表“一组详情页网址”的第 1 项、第 2 项、第 3 项，运行循环体内的积木块。也就是说，循环体总共会运行三次：

(1) 第一次运行时，变量“详情页网址”的值是“https://detail.1688.com/offer/587693664150.html”；

(2) 第二次运行时，变量“详情页网址”的值是“https://detail.1688.com/offer/602379781937.html”；

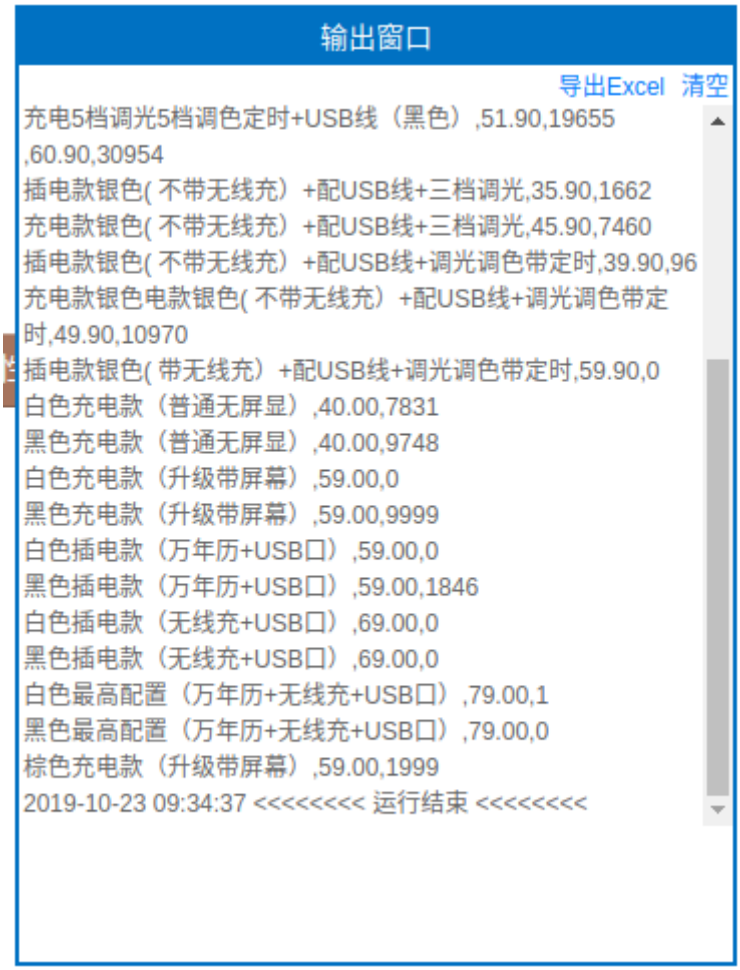
(3) 第三次运行时，变量“详情页网址”的值是“https://detail.1688.com/offer/602794269960.html”。

依此类推，如果我们在列表“一组详情页网址”中添加了第四个网址“https://detail.1688.com/offer/601321744700.html”，那么循环体总共运行 4 次，并且第 4 次时，变量“详情页网址”的值是“https://detail.1688.com/offer/601321744700.html”。

至此，我们初步介绍了**循环类**积木块的使用方法。关键的知识点在于，**循环条件**会创建一个变量，每次运行循环体之前，这个**循环条件变量**（即详情页网址）的值是自动变化的。尽管循环体中的积木块在每一次循环时可能运行相同的步骤，但是由于**循环条件变量**的值改变了，所以循环体内积木块操作的数据有所不同，最终结果也就不同了。

7. 最后一步，在循环体内添加**浏览类**积木块 G 和**文本类**积木块 H。积木块 G 会提取页面的 SKU 数据，存放在变量“一组 SKU 数据项”中。积木块 H 则将 SKU 数据显示到窗口。 这两个步骤在第四课中已经详细解释，在此不再赘述。

本课的机器人流程到此已经可以运行了。你可以多次运行这个流程，验证它总是能准确采集到全部的 SKU 数据。如果你采集的 SKU 数据项很少，请检查三个 CSS 定位字符串是否有误。如果运行无误，请再尝试多添加几个详情页网址，理解循环条件和循环体的作用。



本课总结

总结一下本课所传达的概念。

- **循环条件**可用于多次重复运行一组积木块，通过自动变化的**循环条件变量**来改变每次循环所操作的数据。

- **循环体**是循环条件成立时所运行的一组积木块，重复可运行的次数由循环条件所确定，而每一次运行时，访问的循环条件变量的值是不同的，因此运行的结果也会不同。

本课内容是对第四课的进一步深化，用于批量采集一批网页上类似表格结构的一组数据项。提取的结果是一批列表，且每一个列表中每一项也是一个列表。

第四课的流程只是单纯的顺序步骤，让流程从上到下一路运行下去就可以了。然而在学过本课后，你已经可以改变流程的运行顺序，从上到下，然后回到上面，不断往复。

本课最重要的一点在于，你学会了让流程按照预期设计有条件的运行，流程的步骤之间存在关联结构，而不是的按部就班的简单顺序。这是编程的思维方式。只要你理解了本课的内容，你就不算是编程的门外汉了。

本课附带了一个课后练习，如果你希望继续第六课的学习，请务必用上你的鼠标和键盘，在自己的浏览器上实践几次。如果中途出错，没有关系，只要刷新一下页面就可以了， 直到获得预期的结果。

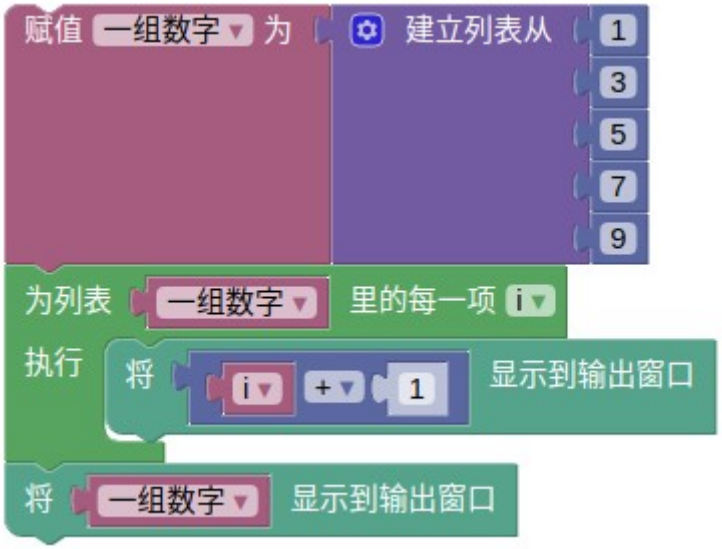
课后练习

请你尝试在图示循环体中添加一个**计算类**积木块，通过循环，从一组数字（1,3,5,7,9）经简单计算运行 5 次，依次显示 2,4,6,8,10，在循环体运行后再添加一个文本类积木块，显示列表“一组数字”的值。

列表中的数字可以从计算类积木块中拖放，具有显示功能的积木块可以从文本类积木块中拖放。

计算类积木块可以实现一个变量的算数加减。

请你先不要看下一页的答案，创建新的流程，从零开始搭建这个流程。流程搭建完毕后，运行流程，然后再添加数字，这样反复实验多次，掌握重复运行的关键——**循环条件变量**。


赋值 一组数字 为
建立列表从
1
3
5
7
9
为列表 一组数字 里的每一项 i
执行
将 i + 1 显示到输出窗口
将 一组数字 显示到输出窗口


输出窗口
导出Excel 清空
2019-10-23 12:03:18 >>>>>>>> 运行开始 >>>>>>>>
2
4
6
8
10
1
3
5
7
9
2019-10-23 12:03:18 <<<<<<<< 运行结束 <<<<<<<<