

Demonstration 1

Adding sheets to a workbook.

At the end of this demonstration, you should be able to:

- Add sheets to a workbook
- Run a workbook

Demonstration 1: Adding sheets to a workbook

Demonstration 1: Adding sheets to a workbook

Purpose:

You will create a new workbook based on the web crawler notebook and use sheets to extract data from web pages.


User/Password: **biadmin/biadmin**
Root/dalvm3
 Service Password: **ibm2blue**

Task 1. Background.

In the previous demonstration you created a workbook based on the results of a web crawler. The crawler was directed to extract information from a website that dealt with patents. Essentially the web crawler looked at a site that had a list of names. Each name is a hyperlink to a page that lists the patents for that person.

1. If you wish, you can review the contents of the website that was crawled. Launch **Firefox**, or open up a new tab in the existing browser. If you wish to see it again, go to the following website:
<http://www.ibm.com/software/ebusiness/jstart/bigsheets/demo/Patents.html>
2. There you see the list of names, Click on any name and it takes you to a page that lists all of the patents registered to that individual. This is to give you a frame of reference when doing this exercise.
3. You can close the newly opened tab.

Task 2. Create a Function sheet.

1. Go to the **BigSheets** Home page.
2. Open the **PatentCrawler** workbook.
 PatentCrawler is a master workbook. No modifications can be made to it. However, you can create a new workbook that is based on PatentCrawler and this new workbook can be modified.
3. Click **Build new workbook** .
 Before you start working with sheets, you will rename each header into something more meaningful.
4. Rename the columns by clicking on the drop-down button next to the column header and then select **Rename**.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

5. Name the columns as follows:

6. Header1 = URL

7. Header2 = Type

8. Header3 = Content

9. Click **Add sheets**



You are to apply a function to some of the data.

10. Select **Function** from the list of sheet types.

Since you are not aware of all of the functions at your disposal, it is best to get a list of them.

11. Click the **Categories** hyperlink.

The ultimate goal is to get patent information for each person on the patents website. You want to apply some function that works with HTML tags.

12. Click the **html** hyperlink.

For each individual's patent site, the person's name is associated with the HTML tag H1. The first thing that needs to be done is to get the content for each H1 tag.

13. Select the **HTMLEXTRACTTAG** hyperlink from the list of functions.

You have the option of giving each new sheet a descriptive name, but you will use *Sheet1* as the name.

14. Click the **content** drop-down box.

You are presented with a list of all columns in the sheet from which to make a selection.

15. Select **Content**.

You want to get the names of the individuals. They are displayed using an HTML H1 tab.

16. In the *tag* field, type in **H1**.

17. In the **occurrence** field, enter a value of **1**.

18. At the bottom of the entry dialog (you may need to scroll down), click the **Carry over (0)** tab.

You might not realize it, but you not only want the extracted information in this new sheet, but, for your purposes, you also want the Content column information as well.

19. From the **Add columns to carry over** drop-down box, select **Content**.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

20. Click **Add column to carry over** .

This adds the Content column to the carry over list.



New sheet: Function Input sheet: PatentCrawler

* Sheet Name: Sheet1

HTMLEXTRACTTAG 
Returns the specified fragment of the content matching the given tag, including the tag element itself

Add columns to carry over:

Content

Parameters Carry over (1)

21. Click **Apply settings**.
You now have a new sheet. But the heading name for column A needs some work.
22. Move the cursor to the **HTMLEXTRACTTAG** column heading, click on the drop-down indicator and then select **Rename**.
23. Highlight the name and change it to **NameWithH1Tag**.
The column name cannot have any spaces.
24. Click **Apply settings**.

Task 3. Further refine the collection.

You did not get the results that you really desired. There are HTML tags encapsulating the desired data. To keep progressing towards your goal, you must create another sheet. You want it based on *Sheet1*.

1. Ensure that the tab for *Sheet1* is selected and then click **Add sheets**.
2. You want to create another **Function** sheet.
3. Click **Categories** and then **html**.
This time, you want to get the encapsulated values
4. Select the **HTMLTAGVALUE** function.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

5. Keep the default sheet name, click the **elements** drop down box and then select **NameWithH1Tag**.
6. Click the **Carry over** tab and add content to the carry over list by clicking **Add column to carry over**.
7. Click **Apply settings**.
8. Change the name of column A to **PatentOwner**.

Task 4. The saga continues.

You will now get all of the patents for each individual.



1. Ensure **Sheet2** is selected and then click **Add sheet**.
2. Select a **Function** sheet, click **Categories** and then **html**.
3. Select **HTMLEXTRACTTAGS**. (with an S)
HTMLEXTRACTTAG lets you specify the occurrence. HTMLEXTRACTTAGS selects all.
4. Keep the default sheet name.
5. From the **Content** drop-down, select **Content**.
6. For **tag** type **H2**.

This is the tag associated with the name of each patent for an individual.

7. Click the **Carry over** tab and then add **PatentOwner** to the list.
8. Click **Apply settings**.

Now you have a row for each patent that is associated with each individual. If you notice, you also removed the blank name and the one called *Found*.

Ultimately, your goal might be to count the number of patents for each individual. However, you are going to stop here. The presentation material has not covered the additional topics that are required to complete that goal. Later, once you have this additional information, you could come back here and code the additionally required sheets.

9. Rename your workbook.
10. At the top of the spreadsheet, click **Edit workbook** .
11. Change the name to **Patent Extract** and then click **Save tag** .
12. Click **Save**.
13. Select **Save & Exit**.



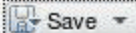



14. At this point you also have the option to rename the collection.
15. Click **Save**.

Notice that this new workbook has not been run. You have only been working with a subset of data. In order for your work to be applied to all of the data, you must run the workbook.

16. Click either one of the **Run** buttons.

There is a yellow triangle on the left side above the data that indicates that the workbook has not been run. (There is also a progress indicator on the right side.) When the processing completes, the triangle changes to a green checkmark. Also, the percentage complete on the right side goes to 100%.

The results appear as follows:

Wordcount Totals  		
 Save  X Exit  Add sheets 		
<i>fx</i>		
	A	B
	Occurrences	Num
1	1	1266
2	2	248
3	3	113
4	4	71
5	5	51
6	6	31
7	7	25
8	8	15
9	9	11
10	10	14

Results:

You created a new workbook based on the web crawler notebook and used sheets to extract data from web pages.