

An introduction to Web Scraping with Python and Azure Functions

with



Welcome!

26-05-2021





Hello!

I am Daniela Miranda

and I'm here because I'm passionate about Data



Agenda

- 1- Web Scraping (generating output)
- 2- Blob Storage (storing the output)
- 3- Wordcloud (convert output into a wordcloud)



Web Scrapping

- Common uses
- Benefits
- How to scrape data from a website?
- Scrapy
- Where to store the output?





Common uses



Price comparison



Email address gathering



Research and development



Benefits

From the business point of view:

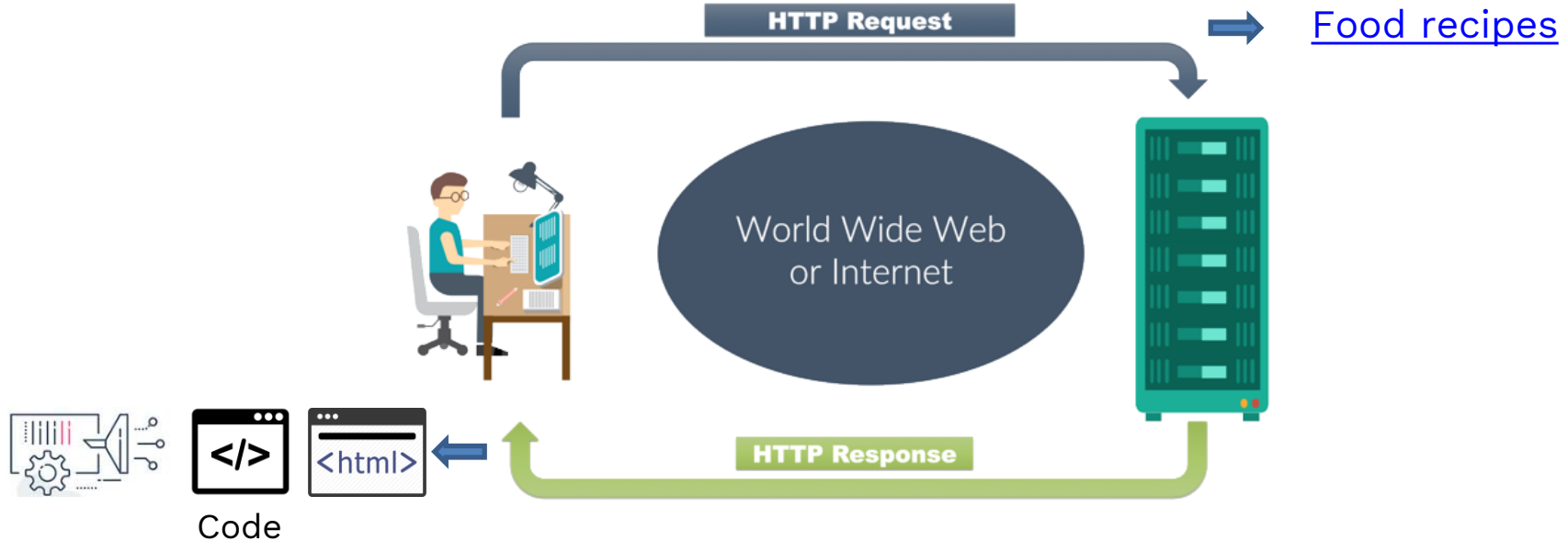
- To achieve automation
- Acquisition of insights
- Unique and rich datasets

From python point of view:

- Easy to use
- Community support
- Several options for web scraping



How to scrape data from a website?





How to scrape data from a website?

1. Find the URL that you want to scrape
2. Inspecting the Page
3. Find the data you want to extract
4. Write the code
5. Run the code and extract the data
6. Store the data in the required format



[Food recipes](#)

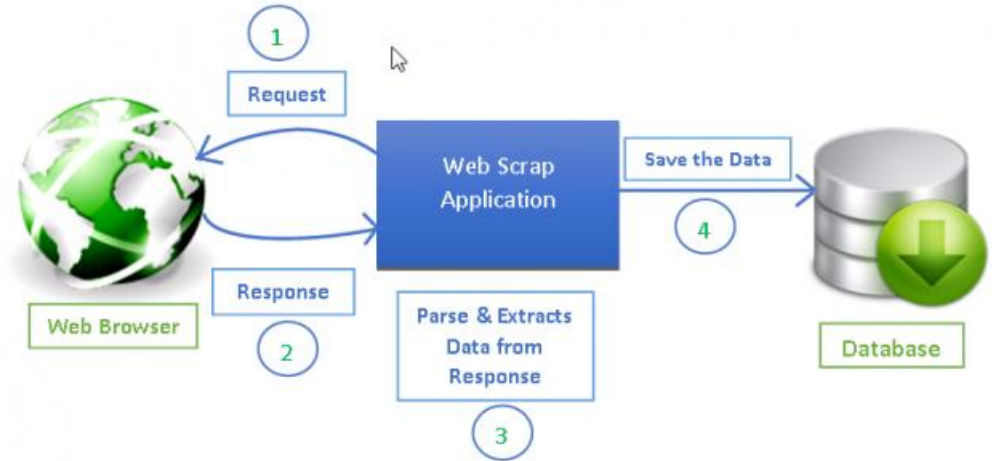


How to do it
automatically and in a
robust way?



Scrapy

- A framework for extracting, processing, and storing web data
- Initial release in June 2008





Scrapy

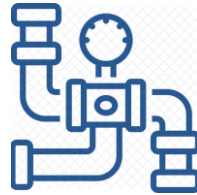
Main components:



Spiders



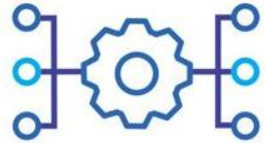
Items



Pipelines



Settings



Middlewares



Scrapy – Hands on

Pre-requisites:

1. Python installed (Check requirements.txt in [GitHub repository](#))
2. Code Editor of preference. Recommended for today: VSCode

Outline:

1. Page to scrape: [Food recipes](#)
2. Goal: Get the list of dish names of British cuisine



Demo Time (1)

Task:

Pick 5 cuisines of your preference and crawl the recipe names from all of them





Storing

- Text file on local machine
- Store it on a Database

What if you are a DS and work on a team. You want your other teammates to see your scraped data. How?





What is a Blob storage?

- Blob storage is optimized for storing massive amounts of unstructured data. Unstructured data is data that doesn't adhere to a particular data model or definition, such as text or binary data.



Blob Storage creation – Hands on

Pre-requisites:

1. Create your azure account [here](#)
2. Create a storage account
3. Create blob container

Outline:

1. Create input and output folders
2. Upload our text file to input folder



Demo Time (2)

We will create a storage account and input/output folders via Azure Portal





Wordcloud: Azure Functions

- What is a Wordcloud?
- How to generate a Wordcloud using Azure Functions?



Microsoft Azure

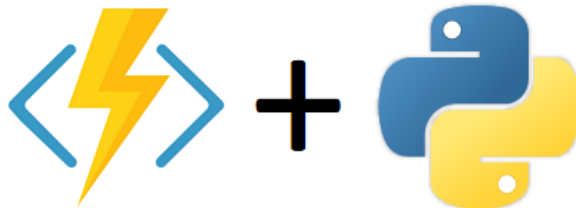




Azure functions

- serverless, lightweight, language independent

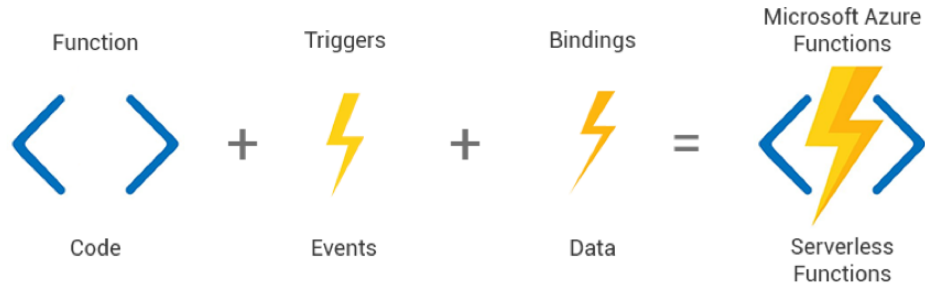
“Azure functions allows you to implement you systems logic into readily available blocks of code, these code blocks are called functions.”





2 additional components of the function: Triggers and Bindings

- Triggers: determines the way the function will execute
- Bindings: determines the input and output of my function





Triggers



- Blob storage trigger

Events

Bindings



- Text file input



Scrapy
Python Library



overly dried fruit expressive
citrus herb include palate
offering brimstone
Aromas sage brisk
broom tropical apple
unripened acidity alongside



What is a Wordcloud?

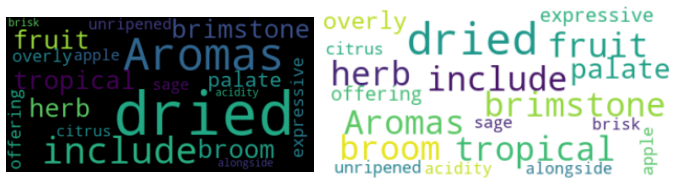
- Collection of clusters of words
- WordCloud is a technique to visualize which words are the most frequent among the given text.





How to generate a Wordcloud?

- Extract the text from file in input folder
- Create and generate a wordcloud image:
 - Split the text in tokens
 - Remove repetitive words
 - Count word frequency
 - The words with higher frequency will be shown in a bigger size
- Save the image in output folder





Wordcloud using Azure Functions– Hands on

Pre-requisites:

1. VSCode
2. Microsoft Azure Storage Explorer
3. Azure Functions Core Tools
4. Python Extension for VSCode
5. Azure Functions Extension for VSCode
6. Azurite Extension for VSCode

Outline:

1. Create local project
2. Create code for wordcloud generation
3. Run the function locally
4. Publish the project to Azure
5. Run the function in Azure



Demo Time (3)

Tasks:

- Make sure your function is triggered by .txt files and generates .png files
- Clean up your text to get more relevant results





Useful resources

- Scrapy documentation: <https://scrapy.org/>
- Azure documentation: <https://docs.microsoft.com/en-gb/learn/modules/fundamental-azure-concepts/>
- Azure functions for beginner – Pyladies: <https://github.com/pyladiesams/Azure-functions-beginner-mar2020>
- Azure function – Microsoft documentation: <https://docs.microsoft.com/en-us/azure/azure-functions/create-first-function-vs-code-python>
- Wordcloud tutorial: <https://www.datacamp.com/community/tutorials/wordcloud-python>



Questions?



Thank you!

Daniela Miranda


 [@danielamiranda](#)

 [@ElaMirandar](#)

 [LinkedIn Daniela](#)



 [@pbblnl](#)

 [@pebbledotnet](#)

 [Website Pebble](#)