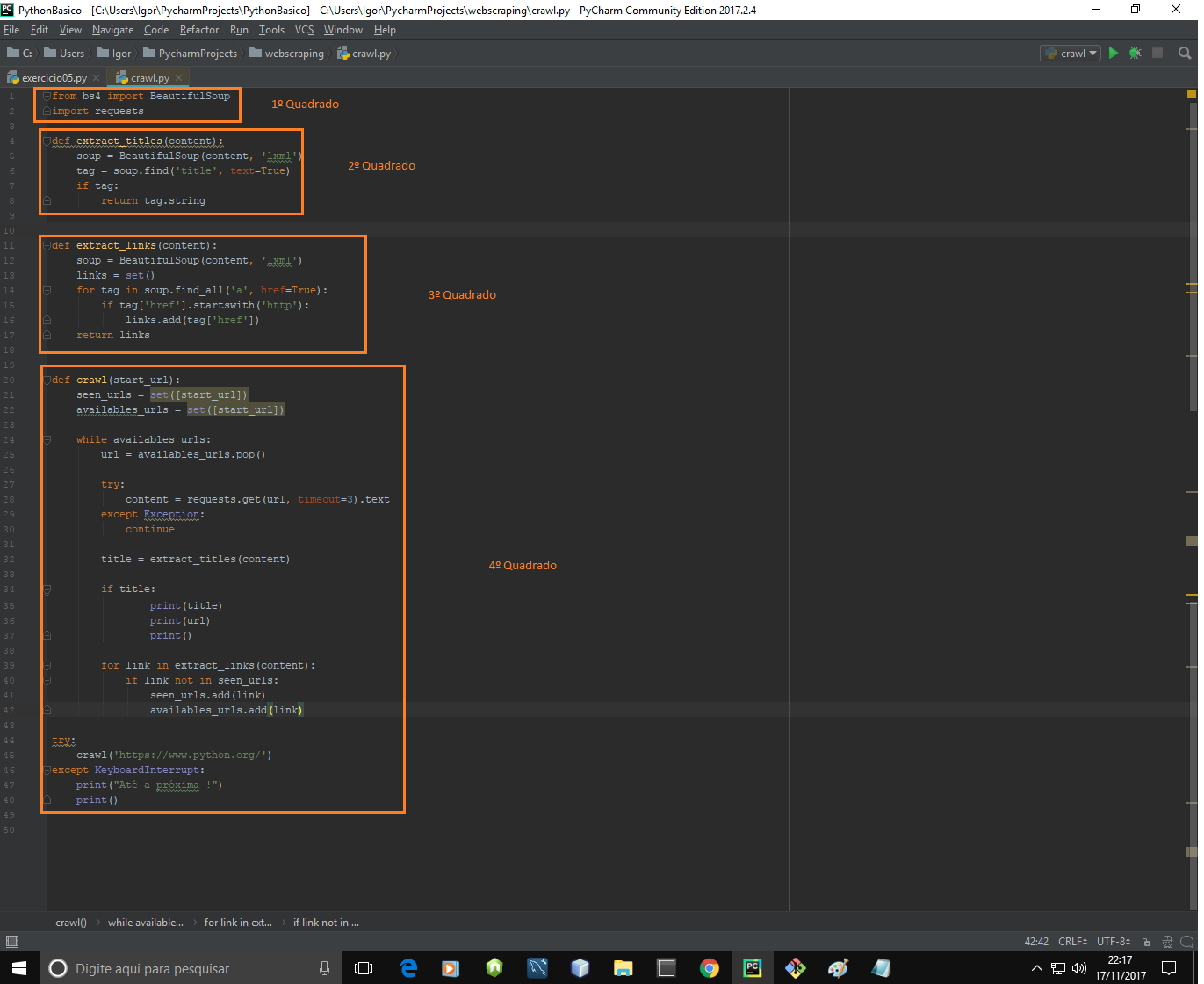
**Crawler - c01**

Este documento apresenta informações de como o Crawler foi desenvolvido, e uma explicação simples das informações dentro dele.

Objetivo do Crawler: Capturar todos os links http presentes em uma página HTML, percorrer todos eles, imprimido o título e URL da página visitada.

****

**Informações referentes ao Crawler – c01**

**1º Quadrado**

Refere-se a importação das bibliotecas que serão utilizadas no Crawler, sendo elas a BeatifulSoup e a Requests.

BeatifulSoup para analisarr o documento HTML, como por exemplo realizar buscas dentro dele.

Requests para fazer a chamada da Página HTML.

**2º Quadrado**

Refere-se a uma função de extração de título da página HTML.

A variável soup está recebendo BeautifulSoup analisando o documento HTML como arquivo XML.

Observação: A variável content receberá o arquivo HTML a ser analisado.

A variável tag está recebendo o resultado de uma busca pelo título da página, se caso houver texto no título da página ele será retornado como uma string, sem suas tags .

**3º Quadrado**

Refere-se a uma função de extração de links presentes na página HTML.

A variável soup está recebendo BeautifulSoup analisando o documento HTML como arquivo XML.

Observação: A variável content receberá o arquivo HTML a ser analisado.

A variável links definida como set() está recebendo de forma desordenada todos os links encontrados na página, porém que não se repetem.

No laço for a variável tag está recebendo todos os links encontrados no HTML e que possuam o seu atributo href preenchido e que tenha http no começo de seu link, e se caso o link possua essas características ele será adicionado ao Set de links. Por fim o loop está retornando o Set **links.**

**4º Quadrado**

Refere-se a uma função de inteligência para o Crawler em relação as urls visitadas e não visitadas, para que ele não venha a visitar a mesma página duas vezes.

No começo do código estão sendo definidas 2 Sets, um para as páginas já visitadas seen\_urls e outro para páginas ainda não visitadas, porém que ele sabe que existe available\_urls.

O loop while, Enquanto houver páginas para serem visitadas uma variável url será criada para receber o método pop() que está retirando um url do Set availables\_urls.   
Em seguida o laço try, except está fazendo com que a na variável contente receba a url retirada do availables\_urls e tente baixar seu conteúdo em 3 segundos caso contrário, a url é descartada e outra entra em ação retirada novamente da available\_urls.

Caso o conteúdo da página seja capturado a função title é chamada em seguida para realizar a extração do título da página e se o título da página obedecer às características descritas o laço If irá imprimir o título e a Url da página que está sendo analisada. Logo em seguida temos um laço For que verifica se um link está presente na Set seen\_urls, caso ele não esteja ele será adicionado. Por fim o último laço try except, inicia o Crawler a partir de uma página especificada e só irá parar quando o método KeyboardInterrup for acionada que seria o conjunto de teclas para pará-lo.