**CRAWL A WEBSITE**

1. **Sampath Kumar**

**19131a0511**

**GVP College of Engineering**

* The first step in extracting any data from a website is to inspect its html properties.
* This can be done in browser through the inspect option.
* After inspecting the site, we can choose a data scraping tool of our choice (Selenium for example).

**Extracting text:**

* If we want to extract all the text, then we can write code like driver.find\_element(By.XPATH,//hody’) using selenium framework in python.
* If we want to extract text from a specific tag, then we can give the XPATH of that html element that we can see from our browser inspection.

**Extracting images:**

* To extract all the images in the page, we can use driver.find\_elements\_by\_tag\_name(‘img’) using selenium framework in python.
* If we want to extract a specific image, then we can g give the XPATH of that image that we can see from our browser inspection.
* We can also select the image number from the images available. That is we can write code like driver.find\_elements\_by\_tag\_name(‘img’)[0] will provide us the selenium tag of that image.
* After finding the image tags, we can get the source attribute by using getAttribute method. Then we can save the image from that source attribute.