**InfoTrie**

**Exercise**

**Lin, Jung-Yen**

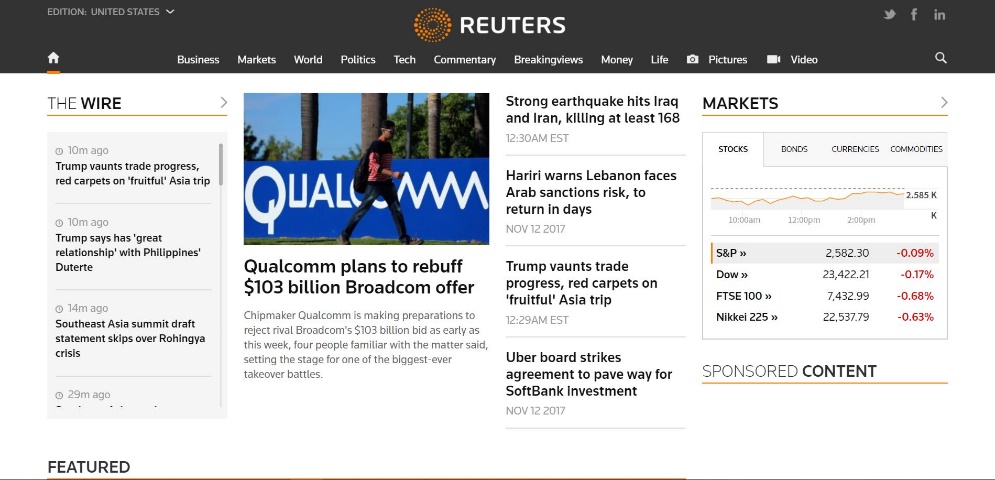
The news portal: Reuters (<https://www.reuters.com/>)

I choose this news website to do the exercise one. I crawl two parts of news in the home page.

First part is a dynamic news which is keep updating (Figure 1). I check the source code first. Those titles and links of the news are under a tag called “article”, however, the whole page contains many “article” tags. Therefore, I check the class name of each article tag. For the dynamic news is under the class called “article USKBNXXXXXXXX” (Figure 2). Each article has its own certain identical number. In the python, I use regular expression to catch the tag and class name which is start with “article” (Figure 3). After catching the correct part of data, I parse the time, title, and link in different tag. For the time is under the “span” and its class name is “timestamp”. For the title and link are under the tag, “a”.

Second part is the main news which won’t be updated frequently (Figure 1). In the source code, this part of news is under the tag, “div” which class name is “story-content” (Figure 4). In the python code, I parse the title and link by finding the div tag and its class name, “story-content” (Figure 5).

Moreover, I define a function to make the code run periodically. It will run the code every 60 second (1 minutes) automatically (Figure 6).



**Second**

**First**

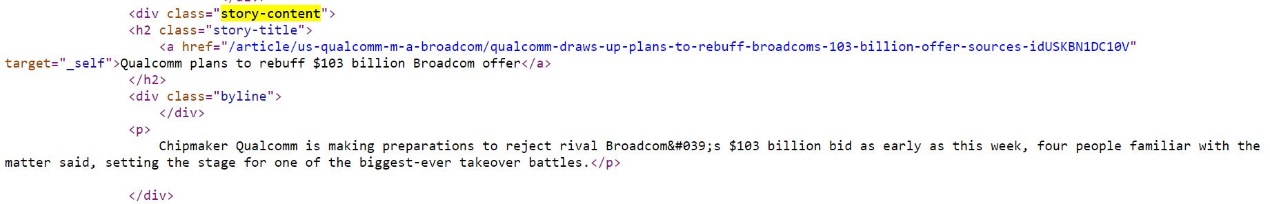
Figure



Figure



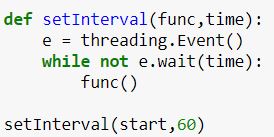
Figure



Figure



Figure



Figure