Chapter 14: Web Scraping

HTML(Hyper Text Markup Language)

-웹 상의 정보를 **구조적**으로 표현하기 위한 언어

-제목, 단락, 링크 등 요소 표시를 위해 Tag를 사용

-모든 요소들은 꺾쇠 괄호 안에 둘러 쌓여 있음

-HTML도 일종의 프로그램, 페이지 생성 규칙이 있음 : 규칙을 분석하여 데이터의 추출이 가능

-HTML은 Tree 구조

HTML Parsing

-웹으로부터 데이터를 추출해 내는 행위

-대부분의 웹은 사용자 요구에 따라 동적으로 생성됨

-HTML Parsing을 위해서 1)정규식을 이용하거나 2)모듈을 활용하면 된다.

정규 표현식 (Regular Expression, regexp, regex)

-복잡한 문자열 패턴을 정의하는 문자 표현 공식

-특정한 규칙을 가진 문자열의 집합을 추출

-HTML은 tag와 같이 **일정한 형식**이 존재하여 정규식으로 추출이 용이함

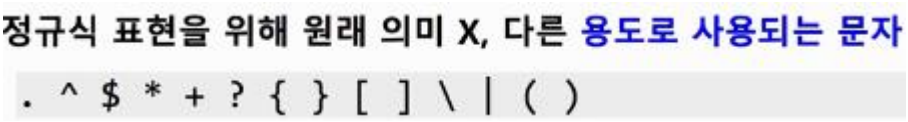
정규식 기본 문법 1#

문자 클래스 [ ]: [와] 사이의 문자들의 매치라는 의미

Ex) [abc] – 해당 글자가 a,b,c중 하나가 있다. "-"를 사용 범위를 지정할 수 있음

[a-zA-z] – 알파벳 전체, [0-9] – 숫자 전체

정규식 기본 문법 – 메타문자



. : 줄바꿈 문자인 ₩n를 제외한 모든 문자와 매치 Ex) a[.]b - **aqwexczxcb**(a에서 시작해서 b로 끝나는 모든 문자)

***:** 앞에 있는 글자를 반복 (**최소반복이 없어도 상관없다.**) Ex) tomor*ow - tomorrow

+: 앞에 있는 글자를 최소 **1회이상** 반복

Ex) [Bb]o+m -> Boom, Bom [Bb]o*m - > Bm

{m,n} : 반복 횟수를 지정 Ex) 203.252.101.40 -> [0-9]{1,3}₩.[0-9]{1,3}₩.[0-9]{1,3}₩.[0-9]{1,3}

or ₩d{1,3}₩.₩d{1,3}₩.₩d{1,3}₩.₩d{1,3}

| or (0|1){3} ^ - not

정규식 in 파이썬

- re 모듈을 import 하여 사용 : import re

- 함수 : search – 한 개만 찾기, findall – 전체 찾기

```
<dl class="blind">
    <dt>종목 시세 정보</dt>
    <dd>2019년 02월 12일 16시 12분 기준 장마감</dd>
    <dd>종목명 삼성전자</dd>
    <dd>종목코드 005930 코스피</dd>
    <dd>현재가 46,050 전일대비 상승 1,050 플러스 2.33 퍼센트</dd>
    <dd>전일가 45,000</dd>
    <dd>시가 44,650</dd>
    <dd>고가 46,250</dd>
    <dd>상한가 58,500</dd>
    <dd>저가 44,650</dd>
    <dd>하한가 31,500</dd>
    <dd>거래량 12,985,826</dd>
    <dd>거래대금 595,133백만</dd>
</dl>
```

① <dl class="blind"> ~~~~ </dl>
(₩<dl class=₩"blind₩"₩>)([₩s₩S]+?)(₩<₩/dl₩>)
<dl class에서 시작해서 / 사이에 아무 글자나 있고 / </dl> 로 끝내기

② <dd> ~~~~ </dd> 정보를 추출하면 됨
(₩<dd₩>)([₩s₩S]+?)(₩<₩/dd₩>)
<dd> 에서 시작해서 / 사이에 아무 글자나 있고 / </dl> 로 끝내기

① 를 먼저 찾고 ① 안에 ②를 차례대로 찾으면 됨