Web Crawling Review

Dart : 전자공시 사이트

전자공시시스템

- 주소 : http://dart.fss.or.kr/
- DART : Data Analysis, Retrieval and Transfer System (전자공시시스템)
- 기능 : 상장법인 등 공시서류를 인터넷으로 제출하고, 추자자 등 이용자는 제출 즉시 인터넷을 통해 조회할 수 있는 종합적인 기업공시 시스템임.
- 최근에는 API를 만들어서 제공하기도 함.
- 여러 증권관련 어플이 이를 기반으로 하는 것이 많음.
- 여기서는 정식 API가 아닌, 웹 상에서 페이지정보를 유추해서 하는 방법사용

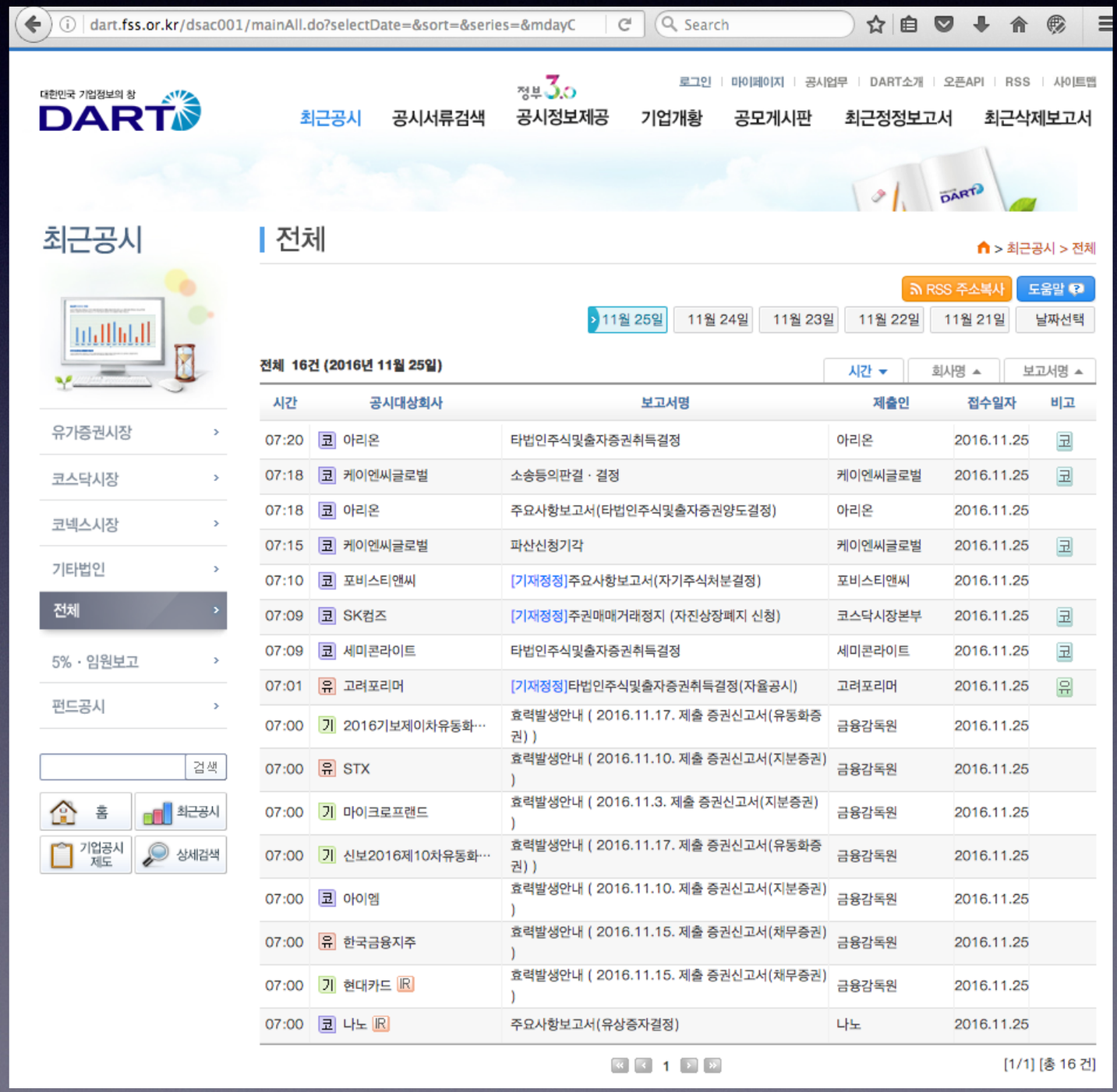
실습

- 실습 목표 : 원하는 날짜에 대한 공시정보를 아래 그림과 같이, “속한시장종류/회사이름/주요공시내용/회사id/공시날짜/공시시간/공시id/요청일자”에 대한 정보를 DataFrame으로 받는 것이 목적임.

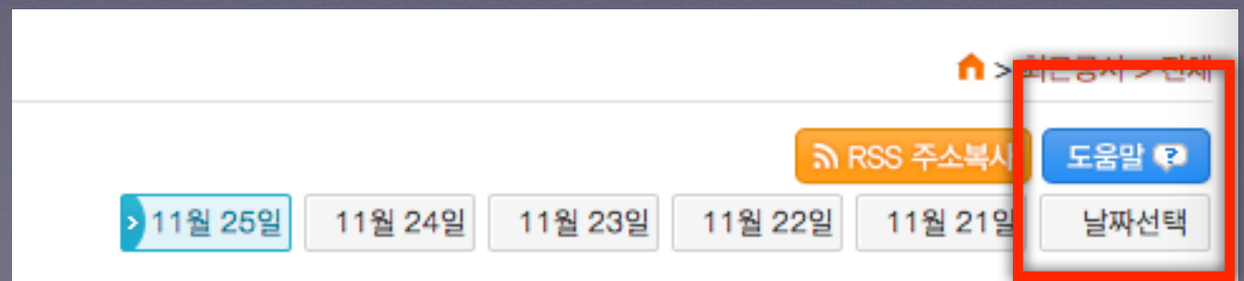
- (참고) 해당 공시의 상세내용 요청 및 공시내용에 대한 분석은 여기서는 수행하지 않음!

Step1)
내가 원하는 정보는 어디에 있나?

- 최근 공시 - 전체



- **Q) 그럼 주소는???**
  - 앞의 해당하는 내용에 대한 url을 확인하니 아래와 같이 나타남.
  - http://dart.fss.or.kr/dsac001/mainAll.do?selectDate=&sort=&series=&mdayCnt=0
  - 이 주소를 보니, 원하는 날짜와 정렬 등의 옵션이 존재하는 것을 알 수 있음(selectDate / sort / Series / mdatCnt)
- Try) 그럼 좌측의 “날짜선택”을 이용해서 한 번 해보자.



dart.fss.or.kr/dsac001/mainAll.do?selectDate=2016.07.13&sort=&seri

Search

대한민국 기업정보의 창 DART

정부3.0

로그인 | 마이페이지 | 공시업무 | DART소개 | 오픈API | RSS | 사이트맵

최근공시 공시검색 공시정보제공 기업개황 공모게시판 최근정정보고서 최근삭제보고서

"아~이 부분에 날짜 관련 정보를 요청하면 되겠다!"

최근공시 | 전체

> 최근공시 > 전체

RSS 주소복사 도움말

07월 13일 | 07월 12일 | 07월 11일 | 07월 08일 | 07월 07일 | 날짜선택

유가증권시장
코스닥시장
코넥스시장
기타법인
전체
5% · 임원보고
펀드공시

검색

홈 | 최근공시 | 기업공시 제도 | 상세검색

전체 364건 (2016년 07월 13일)

시간 ▼ | 회사명 ▲ | 보고서명 ▲

시간	공시대상회사	보고서명	제출인	접수일자	비고
18:40	코 케이피엠테크	주요사항보고서(유상증자결정)	케이피엠테크	2016.07.14	정
18:32	유 현대상선	유상증자신주발행가액(안내공시)	현대상선	2016.07.13	유
18:18	유 미원상사	주요사항보고서(자기주식취득결정)	미원상사	2016.07.14	
18:13	유 미원상사	영업(잠정)실적(공정공시)	미원상사	2016.07.13	유
18:12	유 미원상사	주식소각결정	미원상사	2016.07.13	유
18:08	코 텔콘	조회공시요구(현저한시황변동)에대한답변(미확정)	텔콘	2016.07.13	코
18:04	유 보루네오가구	[기재정정]주주총회소집결의	보루네오가구	2016.07.13	유 정
18:01	코 텔콘	타법인주식및출자증권취득결정	텔콘	2016.07.13	코 정
18:00	기 에이블디파이시리즈1	감사보고서 (2016.03)	이촌회계법인	2016.07.13	
17:59	기 브라이튼제이차	감사보고서 (2016.03)	이촌회계법인	2016.07.13	
17:58	기 아이비에스제십이차	감사보고서 (2016.03)	이촌회계법인	2016.07.13	
17:57	기 에이비에프티제이차	감사보고서 (2016.03)	이촌회계법인	2016.07.13	
17:56	기 한국투자증권	[기재정정]투자설명서(일괄신고)	한국투자증권	2016.07.13	

- 원하는 요청 url 구조 간략히 파악
  - http://dart.fss.or.kr/dsac001/search.ax?selectDate=“
  - 원하는 날짜(단, 양식은 YYYY.MM.DD 2016.07.13)
  - &sort=&series=&mdayCnt=0&currentPage=
  - 뒤의 옵션을 사용할 필요가 있으면 할 것!(선택)

Step2)
원하는 정보를 가져오기 위해 필요한 것은??

- 우리가 원하는 정보를 가져오기 위해서는 로컬에 있는 파일이 아니라, **http통신을 이용해서 웹에 있는 자료**를 요청하고 받와와야 한다.

- Review) urllib2 / BeautifulSoup 패키지 사용?
  - urllib2 : http통신용 패키지
  - BeautifulSoup : Parsing용 패키지

- 참고) 여기서는 혹시 urllib2에서 요청하고 안 되는 경우도 있을 수 있어서, 여기서는 **requests**의 패키지를 이용해서 요청할 것임. [이미 Anaconda로 설치하였을 경우에는 설치되어 있음]

- Q) request 패키지를 이용해서 원하는 url에 대한 정보를 어떻게 접속하고 받아올 수 있는가?

```
from bs4 import BeautifulSoup as bs
import requests

date = "2016.07.13"
url_part1 = """http://dart.fss.or.kr/dsac001/search.ax?selectDate="""
url_part2 = """&sort=&series=&mdayCnt=0&currentPage="""
url = url_part1 + date + url_part2

res = requests.get(url)
soup=bs(res.text, 'html.parser')
```

- urllib2와 비교해서 크게 다른 부분은 없음! 메소드 부분만 조금 다른 점이 있을 뿐!

Step3)
이제 그러면 원하는 페이지 정보 가져오자!!

- Step3-1) 필요 패키지

```
from bs4 import BeautifulSoup as bs
import requests
import pandas as pd
import re
```

- 참고) pandas는 dataframe의 형식으로 정리하기 위해서 필요
- 참고) re의 경우에는 문자열에서 정규식을 이용하기 위한 것!(실제 페이지 상에서 불필요한 문자들이 많아서 이에 대한 처리가 필요함!)

- Step3-2) Step1에서 찾은 url 구성

```
# 여기서 날짜만 yyyy.mm.dd 형식으로 지정!
date = "2016.07.13"
url_part1 = """http://dart.fss.or.kr/dsac001/search.ax?selectDate="""
url_part2 = """&sort=&series=&mdayCnt=0&currentPage="""
url = url_part1 + date + url_part2
```

-

- Step3-3) 해당 url요청하고, BeautilfulSoup로 받기!

```
# 일단 전체 페이지 가지고 와서 전체 데이터 확인용...
res = requests.get(url)
soup=bs(res.text, 'html.parser')
```

- Step3-4) 받은 정보 확인!

```
print soup.prettify()
```

```
<!--검색건수-->
<div class="table_list">
 <p class="table_tit">
  <b>
   전체
            364건
                                (2016년 07월 13일)
  </b>
 </p>
<!--검색 sort -->
<div class="sort">
 <a href="#time1" id="time1" onclick="setOrder(time); return false;">
  <img alt="시간 내림차순" id="time" name="sortImg" src="/images/sort/time/on_desc.gif" title="시간 내림차순"/>
 </a>
 <a href="#crp1" id="crp1" onclick="setOrder(crp);  return false;">
  <img alt="회사별 오름차순" id="crp" name="sortImg" src="/images/sort/crp/off_asc.gif" title="회사별 오름차순"/>
 </a>
 <a href="#rpt1" id="rpt1" onclick="setOrder(rpt);  return false;">
  <img alt="보고서별 오름차순" id="rpt" name="sortImg" src="/images/sort/rpt/off_asc.gif" title="보고서별 오름차순"/>
 </a>
</div>
<!--목록 -->
```

기본 정보들

Q) 그런데 전체 정보가 364건 인데, 지금 페이지에는 다 안나타나고, 2페이지로 넘어 가야 하는데??

17:13	기	한국투자증권	일괄신고추가서류(파생결합증권-주식워런트증권)	한국투자증권	2016.07.13
17:13	기	원주제이차	감사보고서 (2016.03)	효림회계법인	2016.07.13
17:12	기	와이케이플랜	감사보고서 (2016.03)	효림회계법인	2016.07.13
17:11	유	한국자산신탁	최대주주등소유주식변동신고서	한국자산신탁	2016.07.13 유
17:11	기	와이케이에이티에스	감사보고서 (2016.03)	효림회계법인	2016.07.13
17:10	기	예스플러스제일차	감사보고서 (2016.03)	효림회계법인	2016.07.13
17:09	유	미래에셋증권	투자설명서(일괄신고)	미래에셋증권	2016.07.13
17:09	기	국민은행	증권발행실적보고서	국민은행	2016.07.13
17:08	기	엠스퀘어평촌제일차	감사보고서 (2016.03)	효림회계법인	2016.07.13
17:06	코	드래곤플라이	교환사채(해외교환사채포함)발행후만기전사채취득	드래곤플라이	2016.07.13 고

« ‹ 1 2 3 4 › » [1/4] [총 364 건]

→ 뒤에서 이 부분에 대한 것을 해결할 것임!

```
<td>
 <span class="nobr1" style="max-width:150px;">
  <img alt="코스닥시장" src="/images/ico_kosdaq.gif" title="코스닥시장"/>
  <a href="/dsae001/selectPopup.ax?selectKey=00206686" onclick="openCorpInfo('00206686'); return false;" title="케이피엠테크 기업개황 새창">
   케이피엠테크
  </a>
 </span>
</td>
<td>
 <a href="/dsaf001/main.do?rcpNo=20160713000492" id="r_20160713000492" onclick="openReportViewer('20160713000492'); return false;"
itle="주요사항보고서(유상증자결정) 공시뷰어 새창">
  주요사항보고서(유상증자결정)
 </a>
</td>
<td title="케이피엠테크">
 <div class="nobr" style="width:95px">
  케이피엠테크
 </div>
</td>
<td class="cen_txt">
 2016.07.14
</td>
<td class="cen_txt end">
 <img alt="본 보고서 제출 후 정정신고가 있으니 관련 보고서를 참조하시기 바람" hspace="1" src="/images/sub/remark01.gif" title="본 보고서 제출 후 정정신고가 있
으니 관련 보고서를 참조하시기 바람"/>
</td>
</tr>
<tr>
<td class="cen_txt">
 18:32
</td>
<td>
 <span class="nobr1" style="max-width:150px;">
  <img alt="유가증권시장" src="/images/ico_kospi.gif" title="유가증권시장"/>
```

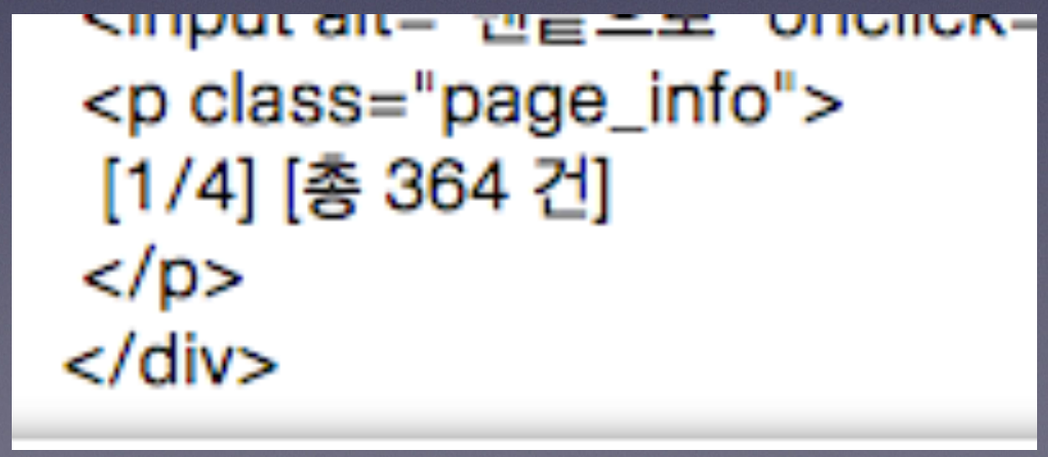
1개 회사에 대한 정보의 범위!
뭔가 테그가 td/tr에서 묶이고 있구나!

잠시) html의 구조에 대해서 간략하게…

- 요소(elements) : 시작 태그와 종료 태그로 이루어진 모든 명령어들을 의미함. (<body></body>
- 테그(tag) : 요소(elements)의 일부로, 시작 태그(<body>)와 종료태그(</body>)로 구성되어 있음.
- 속성(attributes) : 요소의 시작태그 안에서 사용되는 것으로 점 더 구체화된 명령어 체계임(<body align=~~>)
- 변수(arguments) : 속성에 대한 구체적인 값(<body align=“center”>)

Step4)
이제 필요한 정보 하나씩 가져오는 규칙 찾기!

- step4-1) 전체 건수 정보 : 전체 건수를 바탕으로 page롤링을 해야하니 필요함
  - 정보 위치 : 맨 위에도 있고, 맨 아래에도 있음.
  - 여기서는 맨 아래의 정보를 가지고 추출할 것임.(중간에 정규식에 대한 부분을 진행하기 위함)
  - p 테그 이용
  - 정규식을 이용하여 파싱



- Step4-2) 빈 데이터 fataframe 만들기

-

	Cat	ComName	Content	colD	pubDate	pubTime	rcpNo	reqDate
0								

- Step4-3) 전체 페이지 롤링 : 아래의 currentPage=1~이렇게 처리하면 전체 데이터 모두 가져올 수 있을 것!(기본적으로 한 페이지에 100개 정보만 주기 때문에)

```
from bs4 import BeautifulSoup as bs
import requests

date = "2016.07.13"
url_part1 = """http://dart.fss.or.kr/dsac001/search.ax?selectDate="""
url_part2 = """&sort=&series=&mdayCnt=0&currentPage="""
url = url_part1 + date + url_part2

res = requests.get(url)
soup=bs(res.text, 'html.parser')
```

- 구조) 전체 페이지를 계산해서 이에 대한 가변적인 url을 생성하여 요청하기, 그 안에서 각기 페이지에 대한 정보들을 가지고 추출하여 dataframe에 완성하기
    - for 전체 페이지 관련 반복:
        - 이 안에서 가변적으로 페이지별 정보 요청
        - for 각 페이지에서 전체 데이터 수만큼 반복:
            - 각 데이터에서 원하는 정보 추출하고 dataframe에 대입

- 주의!!)
  - 전체 데이터를 관리하는 인덱스와 페이지를 관리하는 인덱스 생각!!
  - : 페이지 관리 인덱스는 앞에서 선정
  - : 그럼 전체 데이터 dataframe.ix[i, 컬럼]에서 i를 어떻게 처리해야 하나??
  - for 문에서 반복되는 인덱스와 같은가? 다른가? 다르면 어떻게 해야하나???

- Step4-4) 각 페이지에서 받아온 공시정보의 숫자는 몇 개인가?
    - 각기 공시에 대한 정보가 tr이라는 테그로 구분이 되어 있음.
    - find_all 이 해당 테그에 대한 전체를 “리스트”로 가져다 주니 이를 활용!

```
# 각 페이지에서 얻은 정보의 숫자 확인
tempNumCont = len(soup.find_all("tr"))
print tempNumCont
```

-

- Step4-5) 1개 회사에 대한 pubTime확인
  - find_all에서 단순 테그 이외에 “속성-변수”를 지정해서 구체적인 것을 찾을 수 있음
  - find_all(“td”, class_=“cen_txt”)
  - 그리고 문자열에서 빈공란의 값을 제거하는 .strip()메소드 이용 “ abc “ -> “abc”
  -

```
<tr>
<td class="cen_txt">
                                    18:32
                                </td>
<td>
<span class="nobr1" style="max-width:150px;">
```

- Strp4-6) 회사 이름 찾기

```
<tr>
<td class="cen_txt">
                                    18:32
                                </td>
<td>
<span class="nobr1" style="max-width:150px;">
<img alt="유가증권시장" src="/images/ico_kospi.gif" title="유가증권시장"/>
                                    <a href="/dsae001/selectPopup.ax?selectKey=00164645" onclick="openCorpInfo('00164645'); return false;" title="현대상선 기업개황 새창">
                                        현대상선
                                    </a>
</span>
```

- 테그가 “a”이고, 그 중에서 가장 처음에 나오니 find사용
- 문자열 인코딩을 위해서 확실하게 .encode(“utf-8”)로 설정함
- 빈 공란 제거는 앞에서 .strip()사용

- Step4-7) 회사 종류 정보 찾기

```
<td class="cen_txt">
                                18:32
                            </td>
<td>
<span class="nobr1" style="max-width:150px;">
<img alt="유가증권시장" src="/images/ico_kospi.gif" title="유가증권시장"/>
                                <a href="/dsae001/selectPopup.ax?selectKey=00164645" onclick="openCorpInfo('00164645'); return false;" title="현
대상선 기업개황 새창">
                                    현대상선
                                </a>
</span>
</td>
```

- 위의정보는 “td”테그 중에서 다시 테그가 “img”이면 되나, 원하는 회사종류는 테그 사이에 있는 값이 아니라, 테그에 있는 “변수”임!
- 이를 위해서는 find가 아닌 get이라는 메소드를 사용함. get(“title”)
- 그리고 인코딩 문제 혹시 발생할 수 있으니 적용

- Step4-8)회사 id 찾기

```
<img alt="유가증권시장" src="/images/ico_kospi.gif" title="유가증권시장" />
                                <a href="/dsae001/selectPopup.ax?selectKey=00164645" onclick="openCorpInfo('00164645'); return false;" title="현
대상선 기업개황 새창">
                                    현대상선
                                </a>
```

- 앞에서 사용한 것으로 각기 속성에 대한 값으로 접근(get)

- But)그러나 원하는 것은 변수 중에서도 다시 일부분임!!! —> 문자열로 간주 —> 정규식 사용

- import re 를 선언해야 함!

- re.findall(r"정규식", 찾을 문자열)

- Step4-9) 앞의 내용들을 이용해서 남은 “공시내용”, “요청날짜”,”공시날짜”, “rcpNo”를 추출해서 완성하기!

-

Try!!!!

```
# 데이터 앞/뒤 확인
```

```
resultData.head(3)
```

	Cat	ComName	Content	colD	pubDate	pubTime	rcpNo	reqDate
0	코스닥시장	케이피엠테크	주요사항보고서(유상증자결정)	00206686	2016.07.13	18:40	20160713000492	2016.07.14
1	유가증권시장	현대상선	유상증자신주발행가액(안내공시)	00164645	2016.07.13	18:32	20160713800475	2016.07.13
2	유가증권시장	미원상사	주요사항보고서(자기주식취득결정)	00121932	2016.07.13	18:18	20160713000472	2016.07.14

```
resultData.tail(3)
```

	Cat	ComName	Content	colD	pubDate	pubTime	rcpNo	reqDate
361	기타법인	에이플러스라이프	주요사항보고서(타법인주식및출자증권양도결정)	00838652	2016.07.13	07:00	20160712000351	2016.07.13
362	기타법인	징크옥사이드코퍼레이션	[기재정정]특수관계인의유상증자참여	00892696	2016.07.13	07:00	20160712000352	2016.07.13
363	기타법인	징크옥사이드코퍼레이션	특수관계인의유상증자참여	00892696	2016.07.13	07:00	20160712000348	2016.07.13

```
len(resultData)
```

```
364
```