BEAUTIFUL SOUP

```
import lxml.html as H
import lxml.etree as ET
import BeautifulSoup
```

HTML -> UNICODE

```
def decode_html(filename):
  html_string = file(filename).read()
  converted = BeautifulSoup.UnicodeDammit(html_string, isHTML=True)
  if not converted.unicode:
    return ‘’
  # print converted.originalEncoding
  return converted.unicode
```

GET & PARSE

```
urllib.urlretrieve(url, filename)
…….
root = H.fromstring(decode_html(filename))
res = root.xpath('//*[@id="ctl00_MainContent_divRankList"]/div/table/
tbody/tr')
……
if res:
   for fr in res:
        fieldRank = int(fr.xpath('td[1]//div[@class="left"]')[0].text)
        name = string.join(fr.xpath('td[2]')[0].itertext()).strip()
……
```

UNICODE (2)

```
f = codecs.open('msa.csv','w','utf-8')
print >>f, ‘é’
f.close()
```

Xpath Editor

- PathEnq