Αρχικά, χρησιμοποιώ την βιβλιοθήκη re και ως κείμενο εισόδου χρησιμοποιώ το testpage.txt

```
import re
text = open('testpage.txt','r',encoding='utf-8').read()
```

1) Έπειτα κάνω εξαγωγή και εκτύπωση του τίτλου

```
rexp = re.compile('<title>(.*?)</title>')

m = rexp.search(text)
print(m.group(1))
```

2) Κάνω απαλοιφή των σχολίων

```
rexp = re.compile('<!--(.*?)-->', re.DOTALL)

m = rexp.sub('<!-- --> ', text)
```

3) Απαλοιφή των script, style

```
rexp = re.compile('<script>(.*?)</script>', re.DOTALL)

m = rexp.sub(' ', m)

rexp = re.compile('<style>(.*?)</style>', re.DOTALL)

m = rexp.sub(' ', m)
```

4) Την εξαγωγή και την εκτύπωση του συνδέσμου από <a> tags και του κειμένου τους

```
rexp = re.compile('<a(.*?)</a>')

for x in rexp.finditer(m):
    print(x.group(0))
```

5) Απαλοιφή ΟΛΩΝ των tags

```
rexp = re.compile(r'<.*?>')
m = rexp.sub(' ', m)
```

6) Μετατροπή &amp,&gt,&lt, με χρήση συνάρτησης

```
def cp(m):
    if (m.group(0)=='&amp'):
        #print(m.group(0))
        return '&'
    elif (m.group(0)=='&lt'):
        #print(m.group(0))
        return '<'
    elif (m.group(0)=='&gt'):
        return '>'
    elif (m.group(0)==' '):
        return ' '
rexp = re.compile('&amp|&lt|&gt| ')
m = rexp.sub(cp, m)
```

7) Μετατροπή ακολουθιών συνεχόμενων χαρακτήρων whitespace σε ένα ακριβώς κενό

```
def cp2(m):
    return ' '
rexp = re.compile(r'\s+')
m = rexp.sub(cp2, m)
print(m)
```

8) Τύπωση στο output.txt (κώδικας μέσω internet)

```
text_file = open("output.txt", "w", encoding='utf-8')
n = text_file.write(m)
text_file.close()
```

Νικολάου Ευάγγελος ΑΜ:Π2014150 Πηγές:Σημειώσεις μαθήματος, Ιντερνετ, https://pythonexamples.org/python-write-string-to-text-file/