Ελληνικό Μεσογειακό Πανεπιστήμιο
Σχολή Μηχανικών / Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

HMY01K06
Επιστημονικός Προγραμματισμός με Python

Διάλεξη Ενδέκατη
Λήψη και επεξεργασία πληροφορίας από το διαδίκτυο

Εγκατάσταση νέων πακέτων στη Python - Πρωτόκολλο http(s) και γλώσσα html
Εργαλεία λήψης, ανάλυσης και αποκωδικοποίησης περιεχομένου ιστοσελίδων

Φθινόπωρο 2025

Εγκατάσταση πακέτων στην Python με το "pip"

- Η Python συνοδεύεται από περισσότερα από *200 προεγκατεστημένα πακέτα*
- Καθένα από αυτά τα πακέτα περιέχει ένα ή περισσότερα *δομοστοιχεία (modules)* που μπορούμε να *εισάγουμε* στο πρόγραμμά μας με την εντολή `import` και στη συνέχεια να *χρησιμοποιούμε* τις συναρτήσεις που περιέχει στο πρόγραμμά μας
- Ωστόσο, μπορούμε να εγκαταστήσουμε και *μόνοι μας* άλλα πακέτα με ένα ειδικό εργαλείο που ονομάζεται `pip` *(Package Installer for Python)*
- Το `pip` είναι ένα εργαλείο *γραμμής εντολών* που εγκαθίσταται μαζί με την Python και επιτρέπει την *εγκατάσταση*, *αναβάθμιση* και *αφαίρεση* πακέτων λογισμικού (βιβλιοθηκών) της γλώσσας

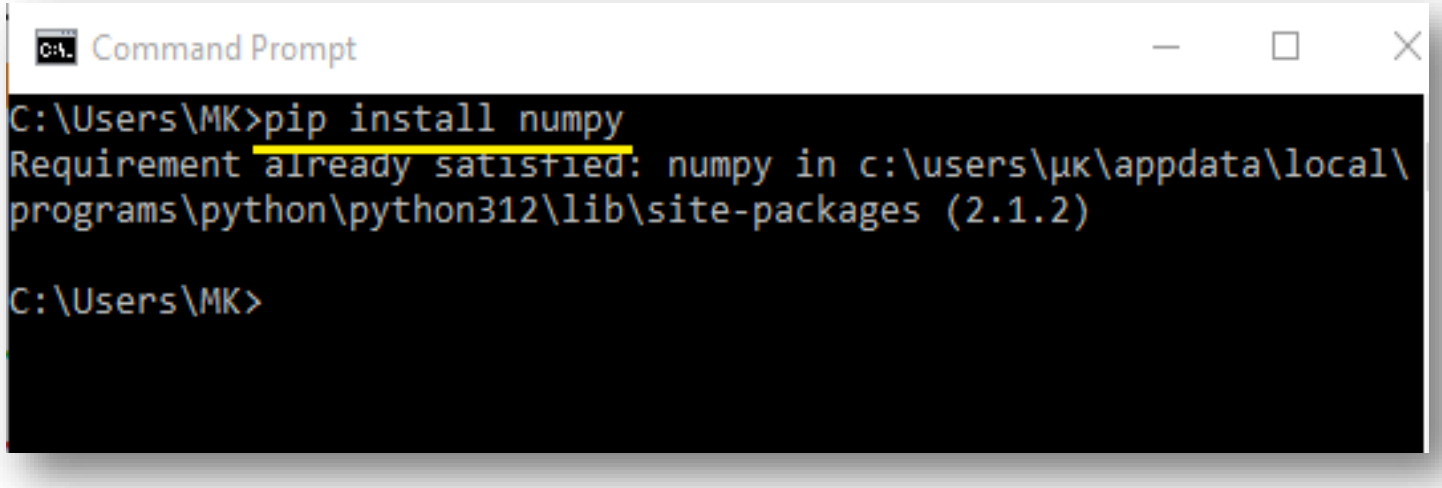
Εγκατάσταση πακέτων στην Python με το "pip"

Για να εγκαταστήσουμε ένα πακέτο, ανοίγουμε ένα παράθυρο γραμμής εντολών *(Command Prompt)* και δίνουμε την ακόλουθη εντολή:

```
pip install <όνομα πακέτου>
```

πχ. για να εγκαταστήσουμε την βιβλιοθήκη συναρτήσεων αριθμητικής ανάλυσης numpy δίνουμε στο παράθυρο γραμμής εντολών την εντολή

```
pip install numpy
```



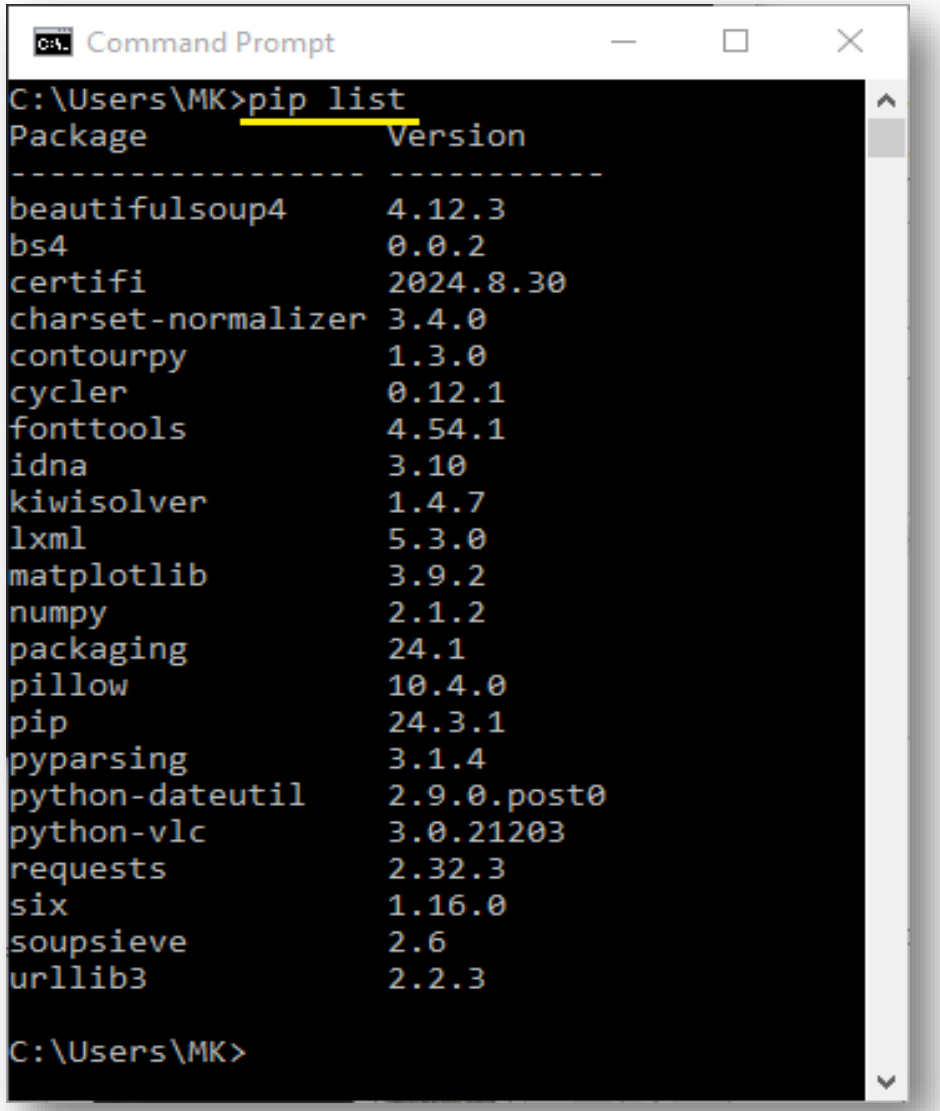
Εδώ βγάζει το μήνυμα "Requirement already satisfied" γιατί το πακέτο numpy είναι ήδη εγκατεστημένο στην Python σε αυτόν τον υπολογιστή

Εγκατάσταση πακέτων στην Python με το "pip"

Για να δούμε ποια πακέτα (βιβλιοθήκες) είναι εγκατεστημένα στον υπολογιστή μας, δίνουμε την εντολή `pip list`

Η λίστα αυτή περιέχει *μόνο τα πακέτα που έχουν εγκατασταθεί με το pip*, και όχι όσα έχουν ήδη προεγκατασταθεί μαζί με την Python

Στο σύνδεσμο *https://docs.python.org/3/py-modindex.html* θα βρείτε έναν κατάλογο με όλα τα modules που περιέχονται σε μια τυπική εγκατάσταση της Python *(περισσότερα από 210 στην έκδοση 3.14)*



Μορφές διακίνησης πληροφορίας στο διαδίκτυο

Ποικίλλουν ανάλογα με

- τον *τύπο* των δεδομένων *(κείμενο, εικόνα, ήχος, πολυμέσα δομημένα δεδομένα)*
- την *ύπαρξη ή μη αλληλεπίδρασης* με τον χρήστη *(μονόδρομη / αμφίδρομη επικοινωνία)*

URL (*Uniform Resource Locator*): Η *μοναδική* διεύθυνση μιας *πηγής πληροφοριών* στο διαδίκτυο

`https://eclass.hmu.gr/courses/ECE102/`

Οι *αρχικοί χαρακτήρες* κάθε διεύθυνσης URL προσδιορίζουν *το πρωτόκολλο μεταφοράς δεδομένων*

- *http, https:* Web browsers *(Chrome, Edge, Safari, Mozilla Firefox, Opera, Tor…)*
  hyper text transfer protocol (secure)
- *ftp:* FTP clients *(Telnet, FileZilla, WS-ftp…)*
  file transfer protocol
- *smtp:* Mail clients *(Outlook, Thunderbird, Pegasus, Foxmail…)*
  simpler mail transfer protocol

Ανάλογα με το πρωτόκολλο μεταφοράς δεδομένων χρησιμοποιούνται διαφορετικές εφαρμογές (προγράμματα) διεπικοινωνίας με τους διακομιστές *(servers)* που περιέχουν την πληροφορία.

Πρωτόκολλο http(s) και γλώσσα html

- Το http(s) είναι το πιο διαδεδομένο *πρωτόκολλο μεταφοράς δεδομένων* στο διαδίκτυο

- Βασίζεται στη γλώσσα html (Hypertext Markup Language / Γλώσσα Σήμανσης Υπερκειμένου) η οποία αποτελεί τη *βασική δομική γλώσσα* για την κατασκευή ιστοσελίδων στον παγκόσμιο ιστό

- Οι ιστοσελίδες html αποτελούνται από απλό κείμενο και ανάμεσά του *σύνολα ετικετών* (tags) που καθορίζουν τον τρόπο εμφάνισης της ιστοσελίδας στην οθόνη

- Το περιεχόμενο των ιστοσελίδων html μπορεί να κυμαίνεται από *απλό κείμενο* και μερικούς *συνδέσμους*, έως και *πολύπλοκες διαδραστικές συνθέσεις πολυμέσων*

- Η γλώσσα html συνήθως συνδυάζεται με την γλώσσα *CSS* (Cascading Style Sheets) για την καθορισμό του *στυλ* και της *εμφάνισης*, καθώς και με την γλώσσα *JavaScript* για τη *δυναμική συμπεριφορά* των ιστοσελίδων

Το βασικό στοιχείο της γλώσσας html

Οι ετικέτες (tags) είναι τα βασικά στοιχεία σήμανσης σε ένα αρχείο html

- Οι ετικέτες html περιέχουν *εντολές / οδηγίες μορφοποίησης* του περιεχομένου ενός εγγράφου html που αναγνωρίζονται και ερμηνεύονται ειδικά προγράμματα *(πχ web browsers που απεικονίζουν ιστοσελίδες)*
- Κάθε ετικέτα html περικλείεται ανάμεσα τους χαρακτήρες `'<'` και `'>'`.
- Συνήθως οι ετικέτες εμφανίζονται σε ζεύγη της μορφής

  `<ετικέτα_Χ>` απλό κείμενο `</ετικέτα_Χ>`

  ή `< ετικέτα_Χ παράμετροι_ετικέτας>` απλό κείμενο `</ετικέτα_Χ>`

  υποδεικνύοντας την αρχή και το τέλος της μορφοποίησης ενός στοιχείου κειμένου

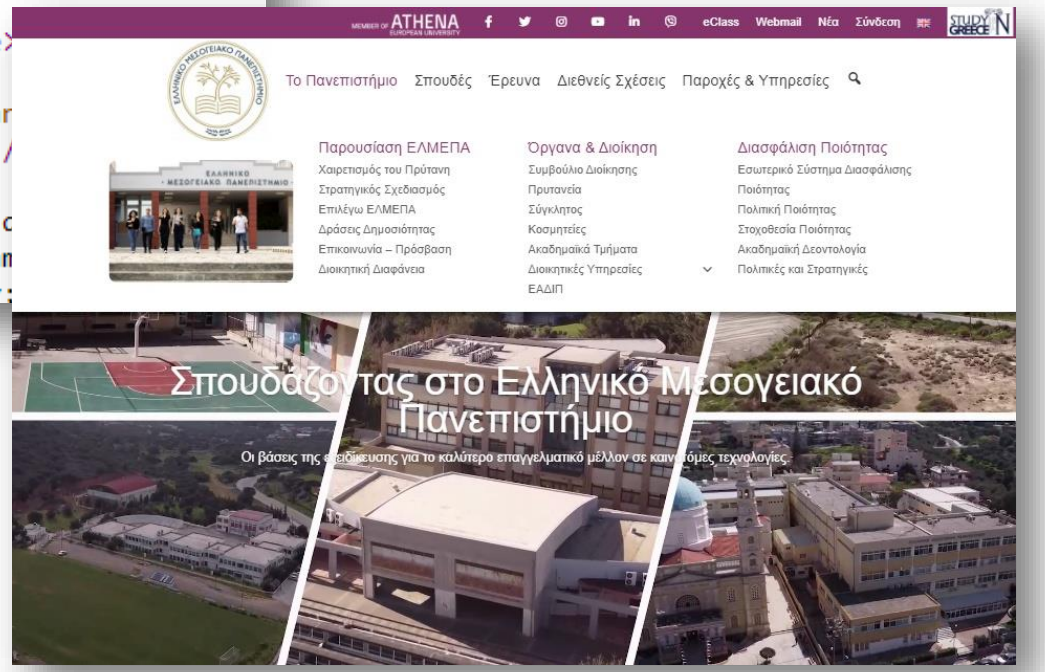
  Παράδειγμα: `<th>` ΣΥΓΚΕΝΤΡΩΤΙΚΑ ΣΤΟΙΧΕΙΑ `</th>` *(επικεφαλίδα πίνακα)*

- Εκτός από ζεύγη ετικετών υπάρχουν και μόνες ετικέτες (self closing tabs)

  Παραδείγματα: `<br>` *(line break / νέα γραμμή)* `<img>` *(ενσωμάτωση εικόνας)*

Από html σε Ιστοσελίδα

```
<html lang="el">
<head>
    <meta charset="UTF-8" />
<meta http-equiv="X-UA-Compatible" content="IE=edge">
    <link rel="pingback" href="https://hmu.gr/xmlrpc.php" />

    <script type="text/javascript">
        document.documentElement.className = 'js';
    </script>

    <title>ΕΛΜΕΠΑ | Ελληνικό Μεσογειακό Πανεπιστήμιο</title>
<meta name='robots' content='max-image-preview:large' />
<link rel="alternate" href="https://hmu.gr/en/home/" hreflan
<link rel="alternate" href="https://hmu.gr/" hreflang="el" /
<script type="text/javascript">
            let jqueryParams=[],jQuery=function(r){return jq
[...jqueryParams,r],jQuery},$=function(r){return jqueryParam
[...jqueryParams,r],$};window.jQuery=jQuery,window.$=jQuery;
```



Οι ετικέτες html περιέχουν εντολές μορφοποίησης του περιεχομένου ενός εγγράφου html ώστε τα προγράμματα περιήγησης ιστού να το απεικονίζουν ως ιστοσελίδα

Ενδεικτικές ετικέτες της γλώσσας html

- *Σύνδεσμοι* (Links): Οι ετικέτες `<a>` χρησιμοποιούνται για τη δημιουργία συνδέσμων προς άλλες σελίδες ή πόρους

  Παράδειγμα: Μετάβαση στο site του ΕΛΜΕΠΑ

- *Λίστες* (Lists): Οι ετικέτες `<ul>` *(απλή λίστα)*, `<ol>` *(διατεταγμένη λίστα)* και `<li>` *(στοιχείο λίστας)* χρησιμοποιούνται για τη δημιουργία λιστών

- *Πίνακες* (Tables): Οι ετικέτες `<table>`, `<tr>` *(σειρά πίνακα)*, `<th>` *(επικεφαλίδα πίνακα)*, και `<td>` *(κελί πίνακα)* χρησιμοποιούνται για τη δημιουργία πινάκων

- *Εικόνες* (Images): Οι ετικέτες `<img>` χρησιμοποιούνται για την ενσωμάτωση εικόνων στη σελίδα

Εξαγωγή πληροφορίας από τον κώδικα html μιας ιστοσελίδας *(web scraping)*

Το *web scraping* αναφέρεται στη διαδικασία εξαγωγής δεδομένων από ιστοσελίδες

Γίνεται με τη χρήση κώδικα προγραμματισμού *(εδώ με Python)* , και *εξειδικευμένων εργαλείων* που επιτρέπουν την ανάκτηση πληροφοριών *αναλύοντας τον κώδικα html* των συγκεκριμένων ιστοσελίδων

Πραγματοποιείται για διάφορους λόγους:

- *Συγκέντρωση δεδομένων* και πληροφοριών από διάφορες ιστοσελίδες, όπως τιμές μετοχών, κριτικές προϊόντων, κλπ.
- *Ανάλυση και σύνθεση πληροφοριών* για δημιουργία πλήρων συνόλων δεδομένων ή ανάλυση τάσεων
- *Παρακολούθηση τιμών προϊόντων* ή υπηρεσιών σε διάφορες ιστοσελίδες και λήψη ειδοποιήσεων όταν υπάρχουν αλλαγές
- *Σύγκριση πληροφοριών* από διάφορες πηγές για κατανόηση διαφορών και ομοιοτήτων

Εργαλεία για web scraping

- Εργαλεία *λήψης* περιεχομένου ιστοσελίδας

  Ο χρήστης δίνει *την διεύθυνση url μιας ιστοσελίδας*, και λαμβάνει το *περιεχόμενο* αυτής της ιστοσελίδας ως "*ακατέργαστο κείμενο*" *(raw text)* που συνήθως είναι σε μορφή *html* *(μπορεί να έχει και άλλη μορφή πχ xml)*

  Στην Python γίνεται από εξειδικευμένες βιβλιοθήκες (*urllib*, *Requests*, …)

- Εργαλεία *ανάλυσης* και *αποκωδικοποίησης* του περιεχομένου ιστοσελίδας *(που ήδη υπάρχει ως ακατέργαστο κείμενο)* για *εξαγωγή πληροφορίας*

  Στην Python γίνεται τόσο από εξειδικευμένες βιβλιοθήκες (*bs4*, *lxml*) όσο και με την βοήθεια κανονικών εκφράσεων *(regex)* για μεγαλύτερη ισχύ και ευελιξία

Προτού εφαρμόσουμε web scraping σε έναν ιστότοπο, πρέπει να ελέγχουμε τους *όρους χρήσης του* καθώς κάποιοι ιστότοποι είτε *απαγορεύουν εντελώς* το scraping, είτε το επιτρέπουν *μόνο υπό συγκεκριμένους όρους*

Η ενσωματωμένη βιβλιοθήκη της Python για πρόσβαση στο διαδίκτυο

Το πακέτο `urllib` (και τα συμπληρωματικά του `urllib2` και `urllib3`) αποτελούν ενσωματωμένες βιβλιοθήκες της Python που μας δίνουν τη δυνατότητα να *επικοινωνούμε απ' ευθείας με τους διακομιστές* στο διαδίκτυο *χωρίς την βοήθεια ειδικών προγραμμάτων όπως browsers, ftp clients και mail clients* αποκτώντας έτσι απ' ευθείας πρόσβαση στην *'ακατέργαστη'* πληροφορία

Το πακέτο `urllib` περιλαμβάνει μεταξύ άλλων και τα ακόλουθα *δομοστοιχεία*

- `urllib.request:` μεταφορά, άνοιγμα και ανάγνωση περιεχομένου διευθύνσεων URL
- `urllib.error:` διαχείριση των εξαιρέσεων που προκύπτουν από το urllib.request
- `urllib.parse:` ανάλυση της δομής των αρχείων που λαμβάνονται από το urllib.request
- `urllib.robotparser:` ανάλυση της δομής των ειδικών αρχείων *robots.txt* που χρησιμοποιούνται από τους διακομιστές για να ελέγχουν την πρόσβαση του περιεχομένου τους από *αυτοματοποιημένα προγράμματα μαζικής ανάκτησης πληροφορίας* (*internet bots*)

Από τα παραπάνω δομοστοιχεία, το βασικότερο όλων είναι το πρώτο (`urllib.request`) επειδή περιέχει όλα τα απαραίτητα εργαλεία να *κατεβάσουμε* το περιεχόμενο μιας δοσμένης διεύθυνσης URL σε μορφή *ακατέργαστου κειμένου (html, xml..)*

Υπάρχει καλύτερη εναλλακτική

Requests: Μια πολύ καλύτερη εναλλακτική

Βιβλιοθήκη συναρτήσεων για *αποστολή αιτήσεων* και *λήψη περιεχομένου http* σε διακομιστές, με εξαιρετικά απλό και φιλικό τρόπο.

Χρησιμοποιείται στην θέση των ενσωματωμένων βιβλιοθηκών της Python *urllib, urllib2* και *urllib3*, οι οποίες παρέχουν τις βασικές συναρτήσεις πρόσβασης σε σελίδες διακομιστών.

Χρειάζεται ξεχωριστή εγκατάσταση `(pip install requests)`

Εξαιρετικά απλή στην χρήση

Παράδειγμα:

```
import requests
url = 'https://www.python.org'
x = requests.get(url).text
print(x)
```

Λήψη περιεχομένου μιας ιστοσελίδας σε 3 γραμμές κώδικα

```
b'<!doctype html>\n<!--[if lt IE 7]>
<html class="no-js ie6 lt-ie7 lt-ie8 lt-ie9">
. . . (ακολουθούν άλλες 1070 γραμμές 'κειμένου')
```

Εξαγωγή πληροφορίας από τον κώδικα html μιας ιστοσελίδας με χρήση Κανονικών Εκφράσεων

Παράδειγμα:

Έστω ότι η μεταβλητή `txt` περιέχει το παρακάτω *απόσπασμα κώδικα html* από μια ιστοσελίδα

```
'<p>Λήξη μαθημάτων </p></td><td style="width:49%;"><p>16 Ιανουαρίου 2026</p>'
```

Μπορούμε να *ταιριάσουμε (απομονώσουμε) το 'καθαρό' κείμενο* που περιέχεται ανάμεσα σε `<p>` και `</p>` χρησιμοποιώντας την κανονική έκφραση `'<p>(.*?)</p>'`

ταίριασμα non-greedy

```
a = re.findall('<p>(.*?)</p>', txt)
print(a)
```

```
['Λήξη μαθημάτων ', '16 Ιανουαρίου 2026']
```

Θυμόμαστε ότι η **findall** επιστρέφει μια λίστα που περιέχει όλα τα ταιριάσματα

Θυμόμαστε επίσης ότι η Python κάνει εξ' ορισμού ταίριασμα greedy:

Αν είχαμε χρησιμοποιήσει την Κ.Ε. `'<p>(.*)</p>'`, η λίστα θα περιείχε ενα μόνο στοιχείο *(γιατί;)*

```
['Λήξη μαθημάτων </p></td><td style="width:49%;"><p>16 Ιανουαρίου 2026']
```

Παράδειγμα 1

Εξαγωγή ημερήσιου δελτίου καιρού από rss news feed του meteo.gr *(με χρήση regex)*

www.meteo.gr/rss/news.cfm

```
import requests
import re
url='https://www.meteo.gr/rss/news.cfm'
try:
    rss= requests.get(url).text
except:
    print ('Πρόβλημα σύνδεσης!')
    raise SystemExit
raw_forecast = re.search('<item>.*?</item>', rss, re.S).group()
title = re.search('<title>(.*)</title>', raw_forecast).group(1)
descr = re.search('<description>(.*)</description>', raw_forecast, re.DOTALL).group(1).strip()
descr= re.sub('\n', '\n\n', descr)
print(title, '\n', '-'*len(title), '\n')
print(descr)
```

Αναλυτική παρουσίαση του παραδείγματος στην Άσκηση Πράξης

Το κείμενο που κατεβάζουμε από την ιστοσελίδα

(εκχωρείται στην συμβολοσειρά raw_forecast)

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?><?xml-stylesheet type="text/css"
href="rss.css"?><rss version="2.0">
<channel>
<title>Meteo.gr: Weather forecasts for Greece</title>
<link>
#XmlFormat("http://www.meteo.gr")#
</link>
<language>en</language>
<generator>http://www.meteo.gr</generator>
<item>
<title>ΓΕΝΙΚΗ ΠΡΟΓΝΩΣΗ ΓΙΑ: 09/12/2025</title>
<link>
http://www.meteo.gr
</link>
<category>weather forecasts for Greece</category>
<description> Την Τρίτη 9 Δεκεμβρίου 2025 αναμένονται πρόσκαιρες τοπικές βροχές σε σημαντικό
τμήμα της χώρας. Φυσιολογικές για την εποχή θερμοκρασίες. Άνεμοι έως 6 μποφόρ στο Αιγαίο.
Πιο αναλυτικά, στα βορειοανατολικά, ανατολικά και νοτιοανατολικά ηπειρωτικά, στη Χαλκιδική,
στην Εύβοια, στην Κρήτη και στο Αιγαίο αναμένονται νεφώσεις κατά διαστήματα. Πρόσκαιρες
τοπικές βροχές θα εκδηλωθούν κυρίως στη Χαλκιδική, στις Σποράδες, στην Εύβοια...
```

Το τμήμα του πηγαίου κώδικα της ιστοσελίδας που μας ενδιαφέρει εντοπίζεται από την Κ.Ε. '<item>.*?</item>'

```
. . .
<item>
<title>ΓΕΝΙΚΗ ΠΡΟΓΝΩΣΗ ΓΙΑ: 09/12/2025</title>
<link>
http://www.meteo.gr
</link>
<category>weather forecasts for Greece</category>
<description> Την Τρίτη 9 Δεκεμβρίου 2025 αναμένονται πρόσκαιρες τοπικές βροχές σε σημαντικό
τμήμα της χώρας. Φυσιολογικές για την εποχή θερμοκρασίες έως 6 μποφόρ στο Αιγαίο.
Πιο αναλυτικά, στα βορειοανατολικά, ανατολικά κ ηπειρωτικά, στη Χαλκιδική,
στην Εύβοια...

...Οι άνεμοι θα πνέουν έως 3 μποφόρ. Η θερμοκρασία στο κέντρο της
Θεσσαλονίκης θα κυμανθ βαθμούς Κελσίου.</description>
<pubDate>10 Dec 2023 14 /pubDate>
</item>
. . .
```

Εντοπίζεται *(απομονώνεται)* στη συνέχεια από την Κ.Ε. '<title>(.*)</title>'

Εντοπίζεται *(απομονώνεται)* στη συνέχεια από την Κ.Ε. '<description>(.*)</description>'

Αποτέλεσμα εκτέλεσης του προγράμματος

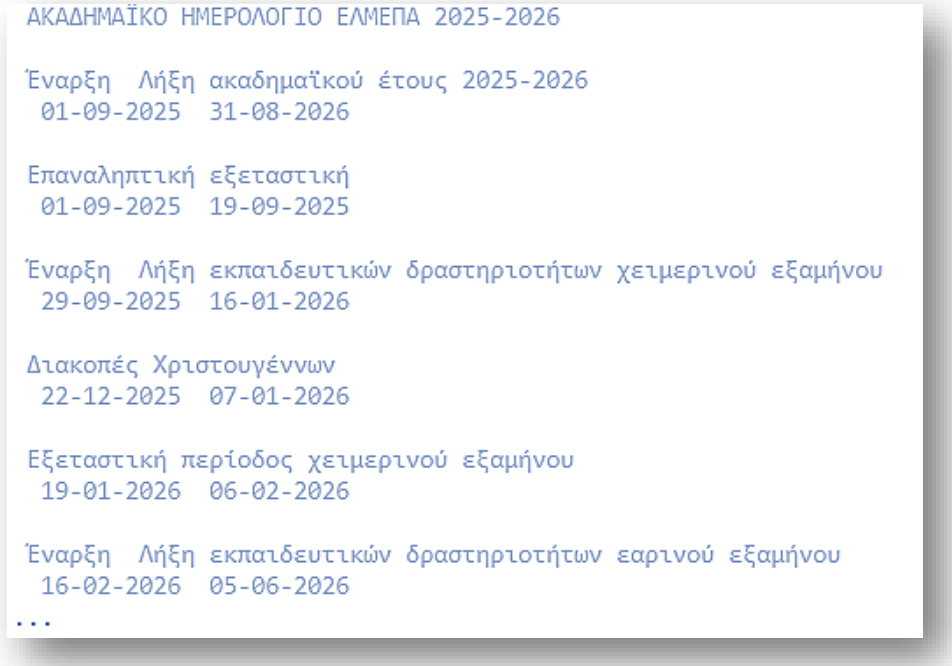
```
raw_forecast = re.search('<item>.*?</item>', rss, re.S).group()
title = re.search('<title>(.*)</title>', raw_forecast).group(1)
descr = re.search('<description>(.*)</description>', raw_forecast, re.DOTALL).group(1).strip()
descr= re.sub('\n', '\n\n', descr)
print(title, '\n', '-'*len(title), '\n')
print(descr)
```

```
ΓΕΝΙΚΗ ΠΡΟΓΝΩΣΗ ΓΙΑ: 09/12/2025
 ------------------------------

Την Τρίτη 9 Δεκεμβρίου 2025 αναμένον
την εποχή θερμοκρασίες. Άνεμοι έως 6

Πιο αναλυτικά, στα βορειοανατολικά,
Κρήτη και στο Αιγαίο αναμένονται νεφ
Χαλκιδική, στις Σποράδες, στην Εύβοια, στην Ανατολική Στερεά και Πελοπόννησο και στην Κρήτη. Στην υπόλοιπη χώρα θα
υπάρχει ηλιοφάνεια με πρόσκαιρες μόνο νεφώσεις κατά τόπους.

Η θερμοκρασία στη Δυτική Μακεδονία θα κυμανθεί από -1 έως 10 βαθμούς Κελσίου στην υπόλοιπη Μακεδονία και στη Θράκη
από 3 έως 13 βαθμούς Κελσίου, στη Θεσσαλία από 5 έως 14, στην Ήπειρο από 3 έως 14, στη Στερεά από 5 έως 15, στην
Πελοπόννησο από 6 έως 16, στα νησιά του Ιονίου από 9 έως 16, στα νησιά του Βορείου και Ανατολικού Αιγαίου από 7
έως 12, στις Κυκλάδες από 10 έως 14, στα Δωδεκάνησα από 10 έως 17 και στην Κρήτη από 11 έως 17 βαθμούς Κελσίου.

Οι άνεμοι στο Αιγαίο οι άνεμοι θα πνέουν από βόρειες διευθύνσεις 4 έως 6 μποφόρ. Στο Ιόνιο οι άνεμοι θα πνέουν από
μεταβαλλόμενες διευθύνσεις έως 3 μποφόρ.

Στην Αττική αναμένονται νεφώσεις κατά διαστήματα ενώ πρόσκαιρες τοπικές βροχές θα εκδηλωθούν κυρίως στα ορεινά του
νομού. Οι άνεμοι θα πνέουν από βόρειες διευθύνσεις 3 έως 5 μποφόρ. Η θερμοκρασία στο κέντρο των Αθηνών θα κυμανθεί
από 11 έως 15 βαθμούς Κελσίου.

Στον νομό Θεσσαλονίκης αναμένεται ηλιοφάνεια με πρόσκαιρες μόνο νεφώσεις κατά τόπους. Οι άνεμοι θα πνέουν από
βόρειες διευθύνσεις έως 3 μποφόρ. Η θερμοκρασία στο κέντρο της Θεσσαλονίκης θα κυμανθεί από 7 έως 13 βαθμούς
Κελσίου.
```

Παράδειγμα 2

Εξαγωγή του ακαδημαϊκού ημερολόγιου του ΕΛ.ΜΕ.ΠΑ *(με χρήση regex)*
https://hmu.gr/akadimaiko-imerologio-2025-2026

```
import requests
import re
url='https://hmu.gr/akadimaiko-imerologio-2025-2026/'
html_txt= requests.get(url).text
mobj = re.search('<tbody>(.*?)</tbody>', html_txt, re.S)
L = re.findall('<td.*?>(.*?)</td>',mobj.group(0), re.S)
for i in range(len(L)):
    L[i] = re.sub('<.*?>', '', L[i])
    L[i] = re.sub('&.*?;', '', L[i])
for x in L:
    print(x)
```



```
ΑΚΑΔΗΜΑΪΚΟ ΗΜΕΡΟΛΟΓΙΟ ΕΛΜΕΠΑ 2025-2026

Έναρξη  Λήξη ακαδημαϊκού έτους 2025-2026
 01-09-2025  31-08-2026

Επαναληπτική εξεταστική
 01-09-2025  19-09-2025

Έναρξη  Λήξη εκπαιδευτικών δραστηριοτήτων χειμερινού εξαμήνου
 29-09-2025  16-01-2026

Διακοπές Χριστουγέννων
 22-12-2025  07-01-2026

Εξεταστική περίοδος χειμερινού εξαμήνου
 19-01-2026  06-02-2026

Έναρξη  Λήξη εκπαιδευτικών δραστηριοτήτων εαρινού εξαμήνου
 16-02-2026  05-06-2026
...
```

Αναλυτική παρουσίαση του παραδείγματος στην Άσκηση Πράξης

Το τμήμα του πηγαίου κώδικα της ιστοσελίδας που μας ενδιαφέρει

```
... <table style="height: 741px; border-style: solid; width: 798px;" border="2"
width="798" cellspacing="1" cellpadding="1">
<tbody>
<tr>
<td style="width: 649px;" colspan="3"><strong>ΑΚΑΔΗΜΑΪΚΟ ΗΜΕΡΟΛΟΓΙΟ ΕΛΜΕΠΑ 2025-
2026</strong></td>
</tr>
<tr>
<td style="width: 418.062px;">Έναρξη – Λήξη ακαδημαϊκού έτους 2025-2025</td>
<td style="width: 111.359px;">01/09/2025</td>
<td style="width: 107.578px;">31/08/2026</td>
</tr>
<tr>
<td style="width: 418.062px;">Επαναληπτική εξεταστική</td>
<td style="width: 111.359px;">01/09/2025</td>
<td style="width: 107.578px;">19/09/2025</td>
</tr>
<tr>
..............
..............
</tbody>
</table>
<p> </p>
```

Εντοπίζεται από την Κ.Ε.
`'<tbody>(.*?)</tbody>'`

Τα σημεία που εντοπίζει η Κ.Ε. '<td.*?>(.*?)</td>'

```
... <table style="height: 741px; border-style: solid; width: 798px;" border="2" width="798"
cellspacing="1" cellpadding="1">
<tbody>
<tr style="height: 49px;">
<td style="width: 563.672px; height: 49px;"><strong>ΑΚΑΔΗΜΑΪΚΟ ΗΜΕΡΟΛΟΓΙΟ ΕΛΜΕΠΑ 2025-
2026</strong></td>
<td style="width: 104.016px; height: 49px;"> </td>
<td style="width: 110.312px; height: 49px;"> </td>
</tr>
<tr style="height: 26px;">
<td style="width: 563.672px; height: 26px;">Έναρξη &#8211; Λήξη ακαδημαϊκού έτους 2025-
2026</td>
<td style="width: 104.016px; height: 26px; text-align: right;">01-09-2025</td>
<td style="width: 110.312px; height: 26px; text-align: right;">31-08-2026</td>
</tr>
<tr style="height: 30px;">
<td style="width: 563.672px; height: 30px;">Επαναληπτική εξεταστική</td>
<td style="width: 104.016px; height: 30px; text-align: right;">01-09-2025</td>
<td style="width: 110.312px; height: 30px; text-align: right;">19-09-2026</td>
</tr>
<tr style="height: 28px;">
<td style="width: 563.672px; height: 28px;">Έναρξη &#8211; Λήξη εκπαιδευτικών δραστηριοτήτων
χειμερινού εξαμήνου</td>...
```

tr: table row td: table data

Τα σημεία που εντοπίζουν οι Κ.Ε. '<.*?>' και '&.*?;'
μέσα στο προηγούμενο -ήδη εντοπισμένο- κείμενο

```
... <table style="height: 741px; border-style: solid; width: 798px;" border="2" width="798"
cellspacing="1" cellpadding="1">
<tbody>
<tr style="height: 49px;">
<td style="width: 563.672px; height: 49px;"><strong>ΑΚΑΔΗΜΑΪΚΟ ΗΜΕΡΟΛΟΓΙΟ ΕΛΜΕΠΑ 2025-
2026</strong></td>
<td style="width: 104.016px; height: 49px;"> </td>
<td style="width: 110.312px; height: 49px;"> </td>
</tr>
<tr style="height: 26px;">
<td style="width: 563.672px; height: 26px;">Έναρξη &#8211; Λήξη ακαδημαϊκού έτους 2025-
2026</td>
<td style="width: 104.016px; height: 26px; text-align: right;">01-09-2025</td>
<td style="width: 110.312px; height: 26px; text-align: right;">31-08-2026</td>
</tr>
<tr style="height: 30px;">
<td style="width: 563.672px; height: 30px;">Επαναληπτική εξεταστική</td>
<td style="width: 104.016px; height: 30px; text-align: right;">01-09-2025</td>
<td style="width: 110.312px; height: 30px; text-align: right;">19-09-2026</td>
</tr>
<tr style="height: 28px;">
<td style="width: 563.672px; height: 28px;">Έναρξη &#8211; Λήξη εκπαιδευτικών δραστηριοτήτων
χειμερινού εξαμήνου</td>...
```

Παράδειγμα 3

Εξαγωγή των νέων του Τμήματος ΗΜΜΥ *(με χρήση regex)*

ece.hmu.gr/news_gr/

```
import requests
import re

url= 'https://ece.hmu.gr/news_gr/'
try:
    raw_text= requests.get(url).text
except:
    print ('Connection Error!')
    raise SystemExit

NewsList = re.findall('<article.*?</article>', raw_text, re.S)
```

Ο κώδικας συνεχίζεται στην μεθεπόμενη διαφάνεια

Αναλυτική παρουσίαση του παραδείγματος στην Άσκηση Πράξης

Παράδειγμα 3 - συνέχεια

```
NewsList = re.findall('<article.*?</article>', raw_text, re.S)
```

Τυπικό στοιχείο της λίστας NewsList

```
<article id="post-7863" class="et_pb_post clearfix et_pb_no_thumb et_pb_blog_item_0_1 post-7863
post type-post status-publish format-standard hentry category-news category-events category-
epikaira"><h2 class="entry-title">
<a href="https://ece.hmu.gr/katataktitries-exetaseis-toy-tmimatos-immy-tis-scholis-
michanikon/">Κατατακτήριες εξετάσεις του τμήματος ΗΜΜΥ της Σχολής Μηχανικών</a>
</h2>
<p class="post-meta"><span class="published">Νοέ 29, 2022</span> |
<a href="https://ece.hmu.gr/category/news/" rel="tag">Ανακοινώσεις</a>,
<a href="https://ece.hmu.gr/category/events/" rel="tag">Εκδηλώσεις</a>,
<a href="https://ece.hmu.gr/category/epikaira/" rel="tag">επικαιρα</a>
</p>
<div class="post-content"><div class="post-content-inner">
<p>Οι κατατακτήριες εξετάσεις του τμήματος ΗΜΜΥ της Σχολής Μηχανικών θα πραγματοποιηθούν σύμφωνα με
το παρακάτω πρόγραμμα εξετάσεων  Δευτέρα 12/12: 10.00 -- 13.00  ΑΙΘΟΥΣΑ Β1 ΗΜΜΥ:
ΔΟΜΗΜΕΝΟΣ  ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΣ Τρίτη 13/12: 10.00 -- 13.00  ΑΙΘΟΥΣΑ Β1 ΗΜΜΥ: ΗΛΕΚΤΡΙΚΑ...</p>
</div></div></article>
```

Παράδειγμα 3 - συνέχεια

(συνέχεια κώδικα)

```
for news_item in NewsList:
    news_title_mobj= re.search('<a href=.*?>(.*?)</a>', news_item, re.S)
    news_title = news_title_mobj.group(1) # λόγω των παρενθέσεων στην Κ.Ε. που ορίζουν ενα capturing group
    news_text_mobj= re.search('<p>(.*?)</p>', news_item, re.S)
    news_text = news_text_mobj.group(1) # λόγω των παρενθέσεων στην Κ.Ε. που ορίζουν ενα capturing group
    news_url_mobj= re.search('<a href="(.*?)"', news_item, re.S)
    news_url = news_url_mobj.group(1) # λόγω των παρενθέσεων στην Κ.Ε. που ορίζουν ενα capturing group
    print(news_title, '\n','-'*len(news_title) )
    print(news_text)
    print('Περισσότερα στο: ', news_url, '\n')
```

re.S: διότι μπορεί να εκτείνεται σε πάνω από μια γραμμές

Αν δεν υπήρχε capturing group, τότε το ταίριασμα θα βρισκόταν στο mobj.group(0)

```
<a href="https://ece.hmu.gr/katataktitries-exetaseis-toy-tmimatos-immy-tis-scholis-michanikon/">Κατατακτήριες εξετάσεις του τμήματος ΗΜΜΥ της Σχολής Μηχανικών</a>
. . . . . .
<p>Οι κατατακτήριες εξετάσεις του τμήματος ΗΜΜΥ της Σχολής Μηχανικών θα πραγματοποιηθούν σύμφωνα με το παρακάτω πρόγραμμα εξετάσεων  Δευτέρα 12/12: 10.00 -- 13.00  ΑΙΘΟΥΣΑ Β1 ΗΜΜΥ: ΔΟΜΗΜΕΝΟΣ  ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΣ Τρίτη 13/12: 10.00 -- 13.00  ΑΙΘΟΥΣΑ Β1 ΗΜΜΥ: ΗΛΕΚΤΡΙΚΑ... </p>
```

Παράδειγμα 3 - αποτελέσματα

Προθεσμίες δήλωσης και διανομής συγγραμμάτων

Η καταληκτική ημερομηνία για τη δήλωση των συγγραμμάτων του χειμερινού εξαμήνου από τους φοιτητές είναι η Παρασκευή 22 Δεκεμβρίου 2023. Η καταληκτική ημερομηνία για τη διανομή των συγγραμμάτων στους φοιτητές είναι η Παρασκευή 5 Ιανουαρίου...

Περισσότερα στο: https://ece.hmu.gr/prothesmies-dilosis-kai-dianomis-syggrammaton/

Ομάδα Προσαρμογής και Ενδυνάμωσης

Συνάντηση Ομάδας Προσαρμογής & Ενδυνάμωσης ΕΛΜΕΠΑ Αγαπητοί /-ές φοιτητές /-τριες του Ελληνικού Μεσογειακού Πανεπιστημίου, Την Τρίτη 5 Δεκεμβρίου 2023, στο Κτίριο Κ24 (Πολυόροφο), στο Γραφείο Συμβουλευτικής και Ψυχοκοινωνικής Στήριξης (ισόγειο), στις 15:00, θα...

Περισσότερα στο: https://ece.hmu.gr/omada-prosarmogis-kai-endynamosis/

Απόφαση της συνέλευσης των φοιτητών

Από την Γενική Συνέλευση των φοιτητών της σχολής Μηχανικών Την Πέμπτη 7/12 στις 12:00 καλούνται όλοι οι φοιτητές της σχολής Μηχανικών στο κεντρικό προαύλιο (μπροστά από την λέσχη) για να συζητήσουμε με τον Πρύτανη για το πολύ σοβαρό ζήτημα των επαγγελματικών...

Περισσότερα στο: https://ece.hmu.gr/apofasi-tis-syneleysis-ton-foititon/

Web scrapping Με BeautifulSoup

- Η BeautifulSoup είναι μια βιβλιοθήκη της Python φτιαγμένη αποκλειστικά για web scraping
- Μπορεί να εξάγει δεδομένα από κώδικες html είτε XML
- Εγκαθίσταται ως bs4 (`pip install bs4`)
- Εισάγεται ως εξής: `from bs4 import BeautifulSoup`

Παράδειγμα χρήσης

```
import requests
from bs4 import BeautifulSoup
html_page = requests.get('https://ece.hmu.gr/news_gr/').text
soup = BeautifulSoup(html_page, 'html.parser')
```

Το soup είναι ένα αντικείμενο τύπου BeautifulSoup το οποίο περιέχει *με δομημένο τρόπο* όλο το *περιεχόμενο της ιστοσελίδας* για εξαγωγή κάθε είδους πληροφορίας *(κείμενο, τμήματα κώδικα, σύνδεσμοι κλπ.)*

Βασικότερες μέθοδοι του αντικειμένου BeautifulSoup

Έστω `soup` ένα αντικείμενο τύπου `BeautifulSoup` που περιέχει τον κώδικα μιας ιστοσελίδας

`soup.get_text()` : επιστρέφει το *"καθαρό" κείμενο* της ιστοσελίδας

`soup.find(<html_tag>)` : επιστρέφει όλα τα συγκεκριμένα *html tags* της ιστοσελίδας

πχ. το `soup.find('title')` επιστρέφει ο,τιδήποτε έχει τη μορφή `<title>...</title>` *(titles)*

πχ. το `soup.find('td')` επιστρέφει ο,τιδήποτε έχει τη μορφή `<td>...</td>` *(table data)*

`soup.find(id = 'x')` : επιστρέφει το/τα *τμήμα(τα) της ιστοσελίδας* με αυτό το id

πχ. το `soup.find(id ='header')` επιστρέφει το τμήμα της επικεφαλίδας της ιστοσελίδας

`soup.find_all('a', href=True)` : επιστρέφει όλους τους *συνδέσμους* της ιστοσελίδας

`soup.prettify()` : επιστρέφει τον κώδικα της ιστοσελίδας κατάλληλα διαμορφωμένο *για φιλική στην όψη εκτύπωση (ξεχωριστές ενότητες, κατάλληλες εσοχές και κενά…)*

Παράδειγμα 4

Εξαγωγή των νέων του Τμήματος ΗΜΜΥ *(με BeautifulSoup)*

ece.hmu.gr/news_gr/

```
import requests
from bs4 import BeautifulSoup

url= 'https://ece.hmu.gr/news_gr/'

try:
    html_txt= requests.get(url).text
except:
    print ('Connection Error!')
    raise SystemExit

soup = BeautifulSoup(html_txt, 'html.parser')

L = soup.find_all('article')
```

κατέβασε ως κείμενο τον κώδικα html από την ιστοσελίδα

δημιούργησε το αντικείμενο soup από το κείμενο του html_txt

δημιούργησε μια λίστα L με όλα τα άρθρα των νέων που υπάρχουν στο soup (δηλαδή τα τμήματα που αρχίζουν από <article> και τελειώνουν σε </article>)

συνέχεια στην επόμενη διαφάνεια →

Παράδειγμα 4 - συνέχεια

```
for article in L:
    link = article.find('a') # βρες το πρώτο link στο article (πιθανόν να υπάρχουν κι' άλλα)
    news_title = link.text   # κείμενο του link
    news_url = link.get('href') # url του link

    # το κείμενο της ανακοίνωσης βρίσκεται στη δεύτερη "παράγραφο" <p...> ... </p>
    news_text = article.find_all('p')[1].text
```

```
<article id="post-7863" class="et_pb_post clearfix et_pb_no_thumb et_pb_blog_item_0_1 post-7863
post type-post status-publish format-standard hentry category-news category-events category-
epikaira"><h2 class="entry-title">
<a href="https://ece.hmu.gr/katataktitries-exetaseis-toy-tmimatos-immy-tis-scholis-
michanikon/">Κατατακτήριες εξετάσεις του τμήματος ΗΜΜΥ της Σχολής Μηχανικών</a>
</h2>
[0] <p class="post-meta"><span class="published">Νοέ 29, 2022</span> |
<a href="https://ece.hmu.gr/category/news/" rel="tag">Ανακοινώσεις</a>,
<a href="https://ece.hmu.gr/category/events/" rel="tag">Εκδηλώσεις</a>,
<a href="https://ece.hmu.gr/category/epikaira/" rel="tag">επικαιρα</a>
</p>
[1] <div class="post-content"><div class="post-content-inner">
<p>Οι κατατακτήριες εξετάσεις του τμήματος ΗΜΜΥ της Σχολής Μηχανικών θα πραγματοποιηθούν σύμφωνα με
το παρακάτω πρόγραμμα εξετάσεων  Δευτέρα 12/12: 10.00 -- 13.00  ΑΙΘΟΥΣΑ Β1 ΗΜΜΥ:
ΔΟΜΗΜΕΝΟΣ  ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΣ Τρίτη 13/12: 10.00 -- 13.00  ΑΙΘΟΥΣΑ Β1 ΗΜΜΥ: ΗΛΕΚΤΡΙΚΑ...</p>
</div></div></article>
```

Παράδειγμα 4 - συνέχεια

```
for article in L:
    link = article.find('a') # βρες το πρώτο link στο article (πιθανόν να υπάρχουν κι' άλλα)
    news_title = link.text   # κείμενο του link
    news_url = link.get('href') # url του link

    # το κείμενο της ανακοίνωσης βρίσκεται στη δεύτερη "παράγραφο" <p...> ... </p>
    news_text = article.find_all('p')[1].text

    # είμαστε έτοιμοι, και τώρα μπορούμε να εκτυπώσουμε
    # τα news_title, news_text και news_url
    print(news_title, '\n','-'*len(news_title) )
    print(news_text)
    print('Περισσότερα στο: ', news_url, '\n')
```

Παράδειγμα 4 - αποτελέσματα

```
Προθεσμίες δήλωσης και διανομής συγγραμμάτων

-----------------------------------------

Η καταληκτική ημερομηνία για τη δήλωση των συγγραμμάτων του χειμερινού εξαμήνου από τους φοιτητές είναι η
Παρασκευή 22 Δεκεμβρίου 2023. Η καταληκτική ημερομηνία για τη διανομή των συγγραμμάτων στους φοιτητές είναι
η Παρασκευή 5 Ιανουαρίου...

Περισσότερα στο:  https://ece.hmu.gr/prothesmies-dilosis-kai-dianomis-syggrammaton/


Ομάδα Προσαρμογής και Ενδυνάμωσης

--------------------------------

Συνάντηση Ομάδας Προσαρμογής &amp;amp; Ενδυνάμωσης ΕΛΜΕΠΑ Αγαπητοί /-ές φοιτητές /-τριες του Ελληνικού
Μεσογειακού Πανεπιστημίου, Την Τρίτη 5 Δεκεμβρίου 2023, στο Κτίριο Κ24 (Πολυόροφο), στο Γραφείο
Συμβουλευτικής και Ψυχοκοινωνικής Στήριξης (ισόγειο), στις 15:00, θα...

Περισσότερα στο:  https://ece.hmu.gr/omada-prosarmogis-kai-endynamosis/


Απόφαση της συνέλευσης των φοιτητών

---------------------------------

Από την Γενική Συνέλευση των φοιτητών της σχολής Μηχανικών Την Πέμπτη 7/12 στις 12:00  καλούνται όλοι οι
φοιτητές της σχολής Μηχανικών στο κεντρικό προαύλιο (μπροστά από την λέσχη) για να συζητήσουμε με τον
Πρύτανη για το πολύ σοβαρό ζήτημα των επαγγελματικών...

Περισσότερα στο:  https://ece.hmu.gr/apofasi-tis-syneleysis-ton-foititon/
```

Ίδια ακριβώς αποτελέσματα με εκείνα του Παραδείγματος 3

Τέλος Διάλεξης

Ερωτήσεις;

Η συνέχεια στην Ασκηση Πράξης >>>

Τμήματα αυτής της διάλεξης περιέχουν στοιχεία από πηγές που είναι ελεύθερες στο διαδίκτυο όπως η Βικιπαιδεια και ανοιχτές σημειώσεις παρεμφερών διαλέξεων