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

HMY01K06
Επιστημονικός Προγραμματισμός με Python

6 +1 ερωτήσεις επανάληψης για τη Διάλεξη 11

Λήψη και επεξεργασία πληροφορίας από το διαδίκτυο

Φθινόπωρο 2025

Ερώτηση 1/7

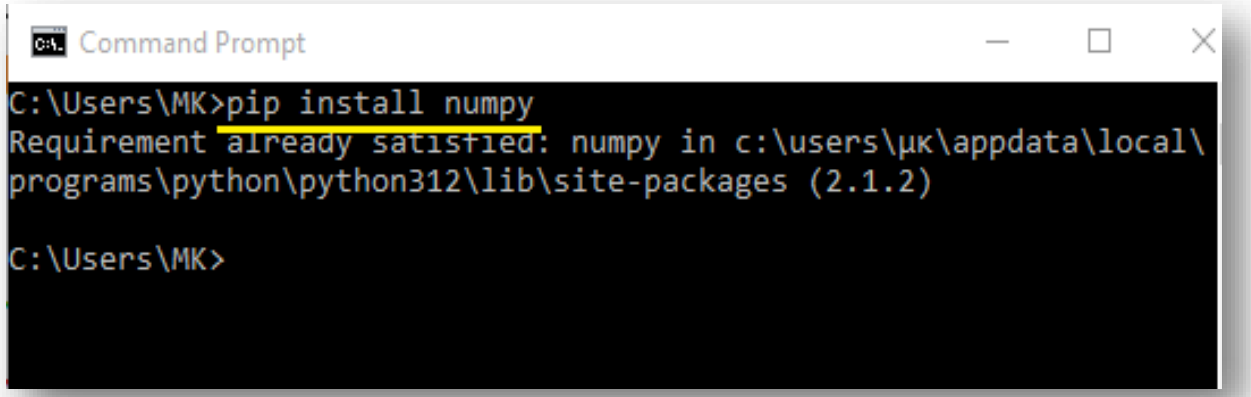
Πως μπορούμε να εγκαταστήσουμε ένα νέο πακέτο (βιβλιοθήκη) στην Python;

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
```



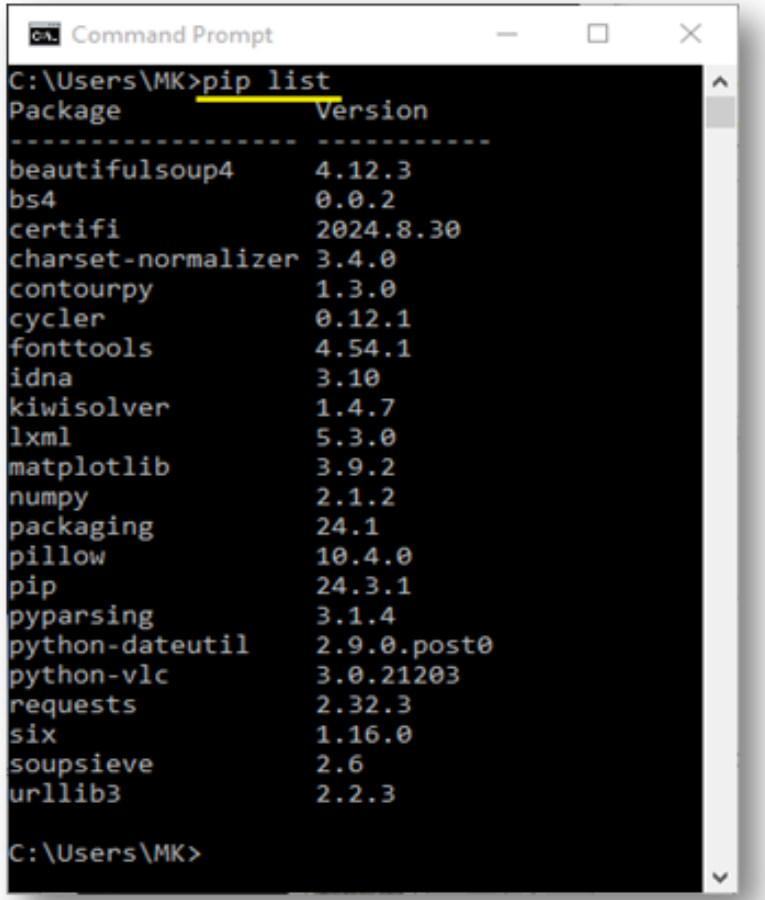
Ερώτηση 2/7

Πως μπορούμε να δούμε ποια πακέτα (βιβλιοθήκες) είναι εγκατεστημένα στον υπολογιστή μας

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)*



Ερώτηση 3/7

Τι είναι το *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

Ερώτηση 4/7

Ποιοι είναι οι δυο βασικές κατηγορίες εργαλείων 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*

Ερώτηση 5/7

Ποιος είναι απλούστερος κώδικας για να πάρουμε το περιεχόμενο μιας ιστοσελίδας σε μορφή κειμένου html με τη χρήση της βιβλιοθήκης Requests;

```
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')
```

Ερώτηση 6/7

Τι ακριβώς είναι η βιβλιοθήκη 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.)*

Bonus Ερώτηση 7/7

Πώς μπορούμε να βρούμε και να εξάγουμε σε μια λίστα όλους τους συνδέσμους που περιέχει μια ιστοσελίδα *(πχ. meteo.gr)*;

```
import requests
from bs4 import BeautifulSoup

response = requests.get("https://meteo.gr/")
html_content = response.content

soup = BeautifulSoup(html_content, "html.parser")

all_links = [link.get("href") for link in soup.find_all("a")]

for link in all_links:
    print(link)
```

```
https://www.facebook.com/Meteo.Gr
https://twitter.com/meteogr
https://www.youtube.com/channel/UCtLRyc7qTR242G1ufaNquUw
https://www.instagram.com/meteogr/
. . .
```