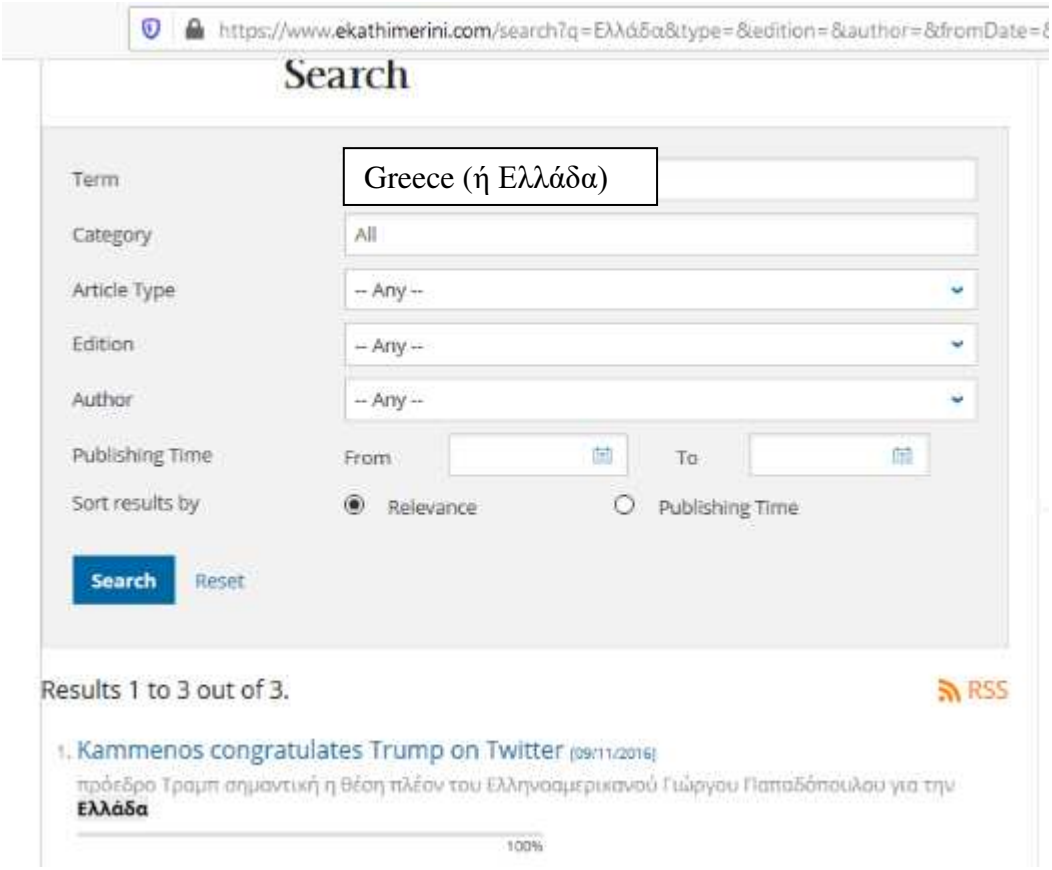
ΠΑΡΑΔΕΙΓΜΑ ΠΡΟΣΩΠΙΚΗΣ ΤΕΚΜΗΡΙΩΣΗΣ / ΕΡΓΑΣΙΑ Python (ΕΚΔΔΑ) / ΑΝΑΛΥΤΙΚΑ #2:

Με ένα πρόγραμμα πλοήγησης, επισκεφθείτε το δικτυακό τόπο της ηλεκτρονικής έκδοσης της εφημερίδας Καθημερινή (https://www.ekathimerini.com). Στο επάνω δεξί μέρος της σελίδας υπάρχει η **λειτουργία της αναζήτησης**, δείτε εικόνα 1 (αν δεν φαίνεται ολόκληρη η εικόνα, τότε στο κάτω μέρος της σελίδας να κάνετε κύλιση/scroll προς τα δεξιά). Εκεί **πληκτρολογούμε (π.χ.) τη λέξη Ελλάδα (ή Greece)** και πατάμε το πλήκτρο της αναζήτησης.



Εικόνα 1. Η σελίδα της εφημερίδας Καθημερινής

Έστω ότι η μηχανή αναζήτησης της εφημερίδας απαντά στην ερώτηση με τα παρακάτω αποτελέσματα, δες εικόνα 2.



Εικόνα 2. Τα αποτελέσματα της αναζήτησης

Όπως βλέπουμε, στο επάνω μέρος των αποτελεσμάτων υπάρχει φόρμα για να προσθέσουμε διάφορα κριτήρια αναζήτησης όπως κατηγορία, τύπος άρθρου κ.λπ. και από κάτω ακριβώς υπάρχουν οι σελίδες 1, 2, 3… με τα αποτελέσματα, όπου η κάθε μια περιέχει συνδέσμους με σχετικά άρθρα. Στην παρούσα άσκηση **δεν μας ενδιαφέρουν τα υπόλοιπα κριτήρια**.

Αν πατήσουμε τη δεύτερη σελίδα των αποτελεσμάτων τότε πάνω στη γραμμή διευθύνσεων θα εμφανιστεί:

https://www.ekathimerini.com/search?q=Greece&t=0&s=r&page=1#searchFull

Στην παραπάνω διεύθυνση μετά το αγγλικό ερωτηματικό **?** υπάρχουν δύο σημαντικές πληροφορίες (π.χ.) **q=Greece** (Ελλάδα) και **p=1** (**που ΔΕΙΧΝΕΙ ΤΗΝ 2η ΣΕΛΙΔΑ ΑΠΟΤΕΛΕΣΜΑΤΩΝ**) που συνδέονται μεταξύ τους με το χαρακτήρα **&**. Αυτά τα δύο δεδομένα είναι παράμετροι που στέλνει ο πελάτης και χρησιμοποιεί ο εξυπηρετητής για να εμφανίζει τις κατάλληλες πληροφορίες.

Για παράδειγμα αν πληκτρολογήσετε στη γραμμή διεύθυνσης:

https://www.ekathimerini.com/search?q=Europe&page=3

θα εμφανιστεί η τέταρτη σελίδα των αποτελεσμάτων σχετικά με την Ευρώπη (η αρίθμηση αρχίζει από 0). Άρα η παράμετρος q=**ΑΛΦΑΡΙΘΜΗΤΙΚΟ** δηλώνει την ερώτηση ενώ το page=**ΑΚΕΡΑΙΟΣ** δηλώνει την σελίδα των αποτελεσμάτων. Το s=r **ή** s=p **σημαίνει ότι τα αποτελέσματα θα ταξινομηθούν είτε ως προς τη σχετικότητα (δες εικόνα 2 Relevance) είτε προς την ημέρα δημοσίευσης (δες εικόνα 2 Publishing)**. Το #searchFull **που υπάρχει στο τέλος δεν χρειάζεται γιατί είναι μια οδηγία προς το πρόγραμμα πλοήγησης**, όταν θα εμφανίσει τα αποτελέσματα τότε να κυλίσει (scroll) την οθόνη ώστε να εμφανίζονται στην αρχή της σελίδας, τα αποτελέσματα της έρευνας.

Ο **σκοπός της εργασίας είναι να αυτοματοποιήσουμε αυτή τη λειτουργία και να δημιουργήσουμε το πρόγραμμα kathimerini.py το οποίο θα ζητά:**

α) Το ερώτημα, έστω ότι πληκτρολογούμε **Ευρώπη**

Προσοχή, αν τυχόν έχετε προβλήματα με την κωδικοποίηση τότε μια λύση είναι να πληκτρολογήσετε λατινικούς χαρακτήρες π.χ. Greece (Ή αν θέλετε, χωρίς να είναι υποχρεωτικό, δείτε εδώ μια λύση: http://stackoverflow.com/questions/11818362/how-to-deal-with-unicode-string-in-url-in-python3)

β) Τη σελίδα έναρξης, έστω ότι πληκτρολογούμε **0**

γ) Την τελική σελίδα, έστω ότι πληκτρολογούμε **2**

Το πρόγραμμα, στο συγκεκριμένο παράδειγμα, θα δημιουργεί 3 αρχεία με ονόματα:

page0.html
page1.html
page2.html

όπου το καθένα θα περιέχει τα αποτελέσματα ανά σελίδα σχετικά με το ερώτημα με αρχή τη σελίδα έναρξης (είναι η β' πληροφορία)

Προσοχή, δεν θα κάνετε καμία επεξεργασία στη σελίδα, όπως τη παραλαμβάνετε, θα την αποθηκεύετε στο δίσκο σας.

Καλή επιτυχία!

ΑΠΑΝΤΗΣΕΙΣ:

Κρίσιμες παραδοχές:

Τα αρχεία αποτελέσματος, αποθηκεύονται πάντα στο τρέχον μονοπάτι εκτέλεσης του προγράμματος, αντικαθιστώντας κάθε προηγούμενο με το ίδιο όνομα.

«Σκέψεις»-Ανάλυση-Προβληματισμοί:

Πρώτα από όλα, ελέγχουμε αρχικά αν ο επιθυμητός αριθμός σελίδας έναρξης που δίνει ο χρήστης είναι ΘΕΤΙΚΟΣ και ΑΝ ΕΙΝΑΙ ΜΙΚΡΟΤΕΡΟΣ/ΙΣΟΣ του αριθμού σελίδας τέλους, διαφορετικά δεν προχωράμε (λογικό σφάλμα) και ενημερώνουμε το χρήστη ανάλογα.

Η υλοποίηση για την υποστήριξη λεκτικού αναζήτησης που ήταν στην Αγγλική γλώσσα, ήταν σχετικά εύκολη. Ωστόσο, τα πράγματα δεν ήταν παρομοίως «εύκολα» με τη χρήση λεκτικού στην Ελληνική γλώσσα: ΑΝΑΛΥΣΗ: Για π.χ. «Ελλάδα», παρατήρησα πως στον φυλλομετρητή μας (browser) το ορθό url θα έπρεπε να είναι: https://www.ekathimerini.com/search?q=%CE%95%CE%BB%CE%BB%CE%AC%CE%B4%CE%B1&page=0

Κάνοντας UTF-8 encoding το λεκτικό «Ελλάδα», παρατήρησα (με print) πως αναγόταν σε \xce\x95\xce\xbb\xce\xbb\xce\xac\xce\xb4\xce\xb1

Οπότε και έφτιαξα δύο (2) βρόχους τελικά προκειμένου να αντικαταστήσω τον χαρακτήρα / με τον % και να ΔΙΑΓΡΑΨΩ τον χαρακτήρα x, για να ΕΧΟΥΜΕ ΕΤΣΙ ΤΗΝ ΤΕΛΙΚΗ ΟΡΘΗ ΑΚΟΛΟΥΘΙΑ (ασχέτως πεζών και κεφαλαίων χαρακτήρων που δεν έπαιζε ρόλο αυτό δοκιμάζοντας) %CE%95%CE%BB%CE%BB%CE%AC%CE%B4%CE%B1

Κώδικας:

```
#(c) Paraskevas G. Papadopoulos / Computer Systems Engineer, H.I. (MSc)
#email: pdesper@gmail.com

#PARADOXES: ta arxeia apotelesmaton pageX.html
#apothikeyontai sto trexon path kai
#den mas apasxolei pou panta antikathistante apo to programma
#se kathe ektelesi tou ek neou!
#TO PROGRAMMA YPOSTHRIZEI KATHE XARAKTHRA LEKTIKOU OPOIASDHPOTE GLOSSAS!!!

#EPIKALOUMASTE ARXIKA TO SXETIKO VOHTHITIKO module tis Python
import urllib.request

#KATHORISMOS EPITHYMHTON ELAXISTON KRITIRION ANAZHTHSHS
print("        *** ΠΡΟΓΡΑΜΜΑ ΕΞΑΓΩΓΗΣ σε αυτόνομες ιστοσελίδες (html) αναλόγως ***")
print("του/των ΕΠΙΘΥΜΗΤΟΥ/ΩΝ ΑΡΙΘΜΟΥ/ΩΝ ΣΕΛΙΔΑΣ/ΩΝ επί των Αποτελεσμάτων Εύρεσης σε επιθυμητό Λεκτικό")
print("      (χωρίς να μας απασχολεί η ταξινόμηση των αποτελεσμάτων με κάποιο κριτήριο),")
print("     κάνοντας χρήση της Μηχανής Αναζήτησης Άρθρων της Εφημερίδας ('Καθημερινή εδώ'):")
print("-------------------------------------------------------------------------------------------")
wordstr=input("Παρακαλούμε, δώστε το ακριβές επιθυμητό (ΣΕ ΟΠΟΙΑΔΗΠΟΤΕ ΓΛΩΣΣΑ) ΛΕΚΤΙΚΟ αναζήτησης (π.χ. Ελλάδα ή Greece) προς αναζήτηση σχετιζόμενων άρθρων: ")

#GIA DOTHEN LEKTIKO (ΣΕΝΑΡΙΟ για ελληνικούς χαρακτήρες που θα είχαμε πρόβλημα!!!)
#TO KODIKOPOIOUME KATA UTF8 KAI TO KANOUME string ( π.χ. TO test GINETAI ETSI b'test' )
wordstr=str(wordstr.encode("utf8"))
#voithitiko keno (arxika) string wordstrGR
wordstrGR=""
#APOMONONOUME TOUS PROTO+DEYTERO+TELEYTAIO XARAKTHRES TOU LEKTIKOU!
for i in range(len(wordstr)):
```

```
    if i==0:
        #PARAKAMPTOUME TON PROTO XARAKTHRA
        pass
    elif i==1:
        #PARAKAMPTOUME TON DEYTERO XARAKTHRA
        pass
    elif i==(len(wordstr)-1):
        #PARAKAMPTOUME TON TELEYTAIO XARAKTHRA
        pass
    else:
        wordstrGR=wordstrGR+wordstr[i]

#ΠΧ: https://www.ekathimerini.com/search?q=%CE%95%CE%BB%CE%BB%CE%AC%CE%B4%CE%B1&page=0
#>>> ΠΑΡΑΤΗΡΗΣΑΜΕ στον BROWSER πως για Ελλάδα ΩΣ ΛΕΞΗ, ΘΕΛΟΥΜΕ όχι ΕΜΦΑΝΙΣΗ ΩΣ
#\xce\x95\xce\xbb\xce\xbb\xce\xac\xce\xb4\xce\xb1 ΣΤΟ ΠΡΟΓΡΑΜΜΑ, ΑΛΛΑ ΩΣ
#%CE%95%CE%BB%CE%BB%CE%AC%CE%B4%CE%B1

#ΟΠΟΤΕ ΚΑΙ ΠΑΜΕ ΝΑ ΑΛΛΑΞΟΥΜΕ ΚΑΘΕ ΧΑΡΑΚΤΗΡΑ \ με τον %
#ΚΑΙ ΝΑ ΔΙΑΓΡΑΨΟΥΜΕ ΚΑΘΕ ΧΑΡΑΚΤΗΡΑ x

#voithitiko keno (arxika) string wordstrGRcode1 ΝΕΑΣ ΑΝΑΓΩΓΗΣ
#pame na antikatastisoume ton xaraktira \ me ton %
wordstrGRcode1=""
for j in range(len(wordstrGR)):
    #xrisi escape character
    if wordstrGR[j]=="\\":
        wordstrGRcode1=wordstrGRcode1+"%"
    else:
        wordstrGRcode1=wordstrGRcode1+wordstrGR[j]

#voithitiko keno (arxika) string wordstrGRcode2 ΝΕΑΣ ΑΝΑΓΩΓΗΣ
#pame na diagrapsoume ton xaraktira x
wordstrGRcode2=""
for z in range(len(wordstrGRcode1)):
    if wordstrGRcode1[z]=="x":
        #DEN TON LAMVANOUME YPOPSH TON XARAKTHRA x
        #wordstrGRcode2=wordstrGRcode2
        pass
    else:
        wordstrGRcode2=wordstrGRcode2+wordstrGRcode1[z]

#ENHMERONOUME TO TELIKO LEKTIKO POU HDH EXEI PROETOIMASTEI
#GIA NA ERMHNEYTEI SOSTA APO URL parakato...
#(ΣΕΝΑΡΙΟ λεκτικού ελληνικών που θα είχαμε πρόβλημα
#ΑΝ ΔΕΝ ΥΛΟΠΟΙΟΥΣΑΜΕ ΤΑ ΠΡΟΑΝΑΦΕΡΟΜΕΝΑ)
wordstr=wordstrGRcode2

#ELEGXOS EGKYROTHTAS ORION ARITHMON SELIDAS
correct=0
while correct==0:
    startstr=input("Παρακαλούμε, δώστε την ΕΠΙΘΥΜΗΤΗ ΣΕΛΙΔΑ ΕΝΑΡΞΗΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ προς εξαγωγή σε αρχείο html (π.χ. 2): ")
```

```
    endstr=input("Παρακαλούμε, δώστε την ΕΠΙΘΥΜΗΤΗ ΣΕΛΙΔΑ ΤΕΡΜΑΤΙΣΜΟΥ ΑΠΟΤΕΛΕΣΜΑΤΩΝ προς εξαγωγή σε αρχείο html (π.χ. 5): ")
    #METATROPES string ORION ARITHMON SE NOUMERA AKERAIA
    start=int(float(startstr))
    end=int(float(endstr))
    #LOGIKOI ELEGXOI ORTHOTHTAS ORION ARITHMON!!!
    if (start>end or start<0):
        #LATHOS ORIA ARITHMON, OPOTE PREPEI NA KSANAZHTHTHOUN!
        #stin print MONO xrisi string!
        print("*** ΚΑΝΑΤΕ ΛΑΘΟΣ: ΔΕΝ ΜΠΟΡΕΙ Η ΣΕΛΙΔΑ ΕΝΑΡΞΗΣ ΝΑ ΕΙΝΑΙ ΜΕΓΑΛΥΤΕΡΗ ΑΠΟ ΤΗ ΣΕΛΙΔΑ ΤΕΡΜΑΤΙΣΜΟΥ ή/και ΝΑ ΕΧΕΤΕ ΔΩΣΕΙ ΑΡΝΗΤΙΚΗ/ΕΣ ΤΙΜΗ/ΕΣ ***")
    else:
        correct=1
print("-----------------------------------------------------------------------------------------------")

#YPOLOGISMOS PLHTHOUS SELIDON
numpages=(end-start)+1
print (">>> ΑΡΙΘΜΟΣ ΕΠΙΘΥΜΗΤΩΝ ΕΞΑΓΩΜΕΝΩΝ html ΣΕΛΙΔΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ = "+str(numpages))
#EPANALHPTIKH DIADIKASIA EKSAGOGHS ARXEION pageX.html (ekkinontas apo to page0.html)
for i in range(numpages):
    url="https://www.ekathimerini.com/search?q="+wordstr+"&page="+str(start)
    print (">>>>> ΤΡΕΧΟΥΣΑ ΔΙΕΥΘΥΝΣΗ URL ΓΙΑ ΕΞΑΓΩΓΗ ΣΕΛΙΔΑΣ (ΣΕ ΑΡΧΕΙΟ ΟΝΟΜΑΤΟΛΟΓΙΑΣ page"+str(start)+".html) = "+url)

    #FTIAXNOUME TO ANTISTOIXO hlp ANTIKEIMENO AITHSHS HTTP
    hlp=urllib.request.urlopen(url)

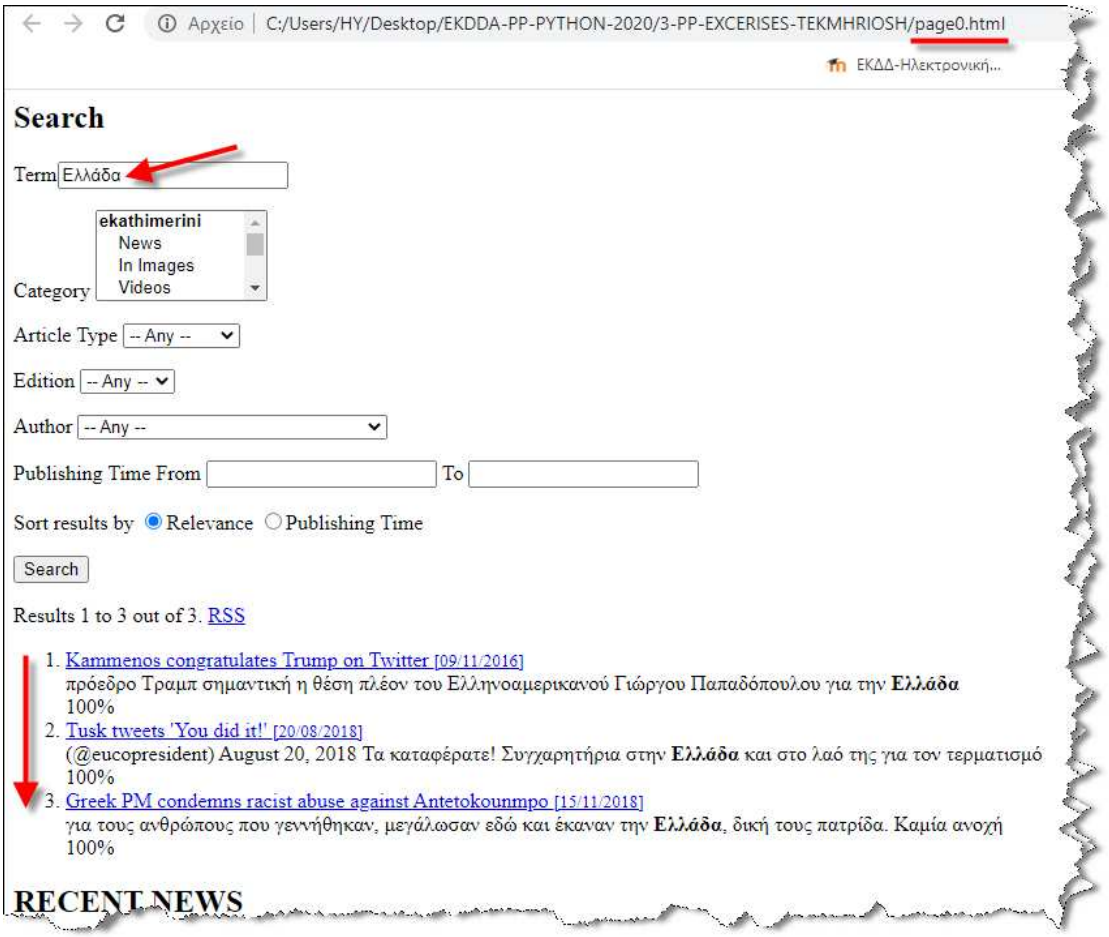
    #DIAVAZOUME TON KODIKA THS TREXOUSAS ISTOSELIDAS APOTELESMATOS
    #ME APOKODIKOPOIHSH UTF-8
    page=hlp.read().decode('utf-8')

    #DHMIOURGIA, ENHMEROSH KAI KLEISIMO TREXONTOS ARXEIOY APOTELESMATOS
    #SOSTO APOTELESMA GIA KATHE GLOSSA ENTOS TOU ARXEIOU,
    #MONO ME PROSTHIKH encoding UTF-8
    #[ALLIOS, XORIS ENCODING, DEN EKSAGETAI EPITYXOS TO ARXEIO KAN!]
    #(ftiaxnoume to onoma tou arxeiou pageX.html pou tha eksagoume)
    fhandle=open("page"+str(start)+".html","w",encoding="utf-8")
    fhandle.write(page)
    fhandle.close()
    print ("> ΑΠΟΘΗΚΕΥΤΗΚΕ ΤΟ ΑΝΤΙΣΤΟΙΧΟ ΑΡΧΕΙΟ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΟΝΟΜΑΤΟΛΟΓΙΑΣ page"+str(start)+".html <")

    #PROETOIMASIA EPOMENHS SELIDAS...
    start=start+1
```

Αποτέλεσμα (π.χ. για τη λέξη «Ελλάδα»…):

```
>>>
= RESTART: C:\Users\HY\Desktop\EKDDA-PP-PYTHON-2020\3-PP-EXCERISES-TEKMHRIOSH\ERGASIA 3-4hs DIDAKTIKHS ENOTHTAS (THEMATIKES 11-17)-kathimerini.py
         *** ΠΡΟΓΡΑΜΜΑ ΕΞΑΓΩΓΗΣ σε αυτόνομες ιστοσελίδες (html) αναλόγως ***
του/των ΕΠΙΘΥΜΗΤΟΥ/ΩΝ ΑΡΙΘΜΟΥ/ΩΝ ΣΕΛΙΔΑΣ/ΩΝ επί των Αποτελεσμάτων Εύρεσης σε επιθυμητό Λεκτικό
      (χωρίς να μας απασχολεί η ταξινόμηση των αποτελεσμάτων με κάποιο κριτήριο),
     κάνοντας χρήση της Μηχανής Αναζήτησης Άρθρων της Εφημερίδας ('Καθημερινή εδώ'):
-------------------------------------------------------------------------------------------
Παρακαλούμε, δώστε το ακριβές επιθυμητό (ΣΕ ΟΠΟΙΑΔΗΠΟΤΕ ΓΛΩΣΣΑ) ΛΕΚΤΙΚΟ αναζήτησης (π.χ. Ελλάδα ή Greece) προς αναζήτηση σχετιζόμενων άρθρων: Ελλάδα
Παρακαλούμε, δώστε την ΕΠΙΘΥΜΗΤΗ ΣΕΛΙΔΑ ΕΝΑΡΞΗΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ προς εξαγωγή σε αρχείο html (π.χ. 2): 0
Παρακαλούμε, δώστε την ΕΠΙΘΥΜΗΤΗ ΣΕΛΙΔΑ ΤΕΡΜΑΤΙΣΜΟΥ ΑΠΟΤΕΛΕΣΜΑΤΩΝ προς εξαγωγή σε αρχείο html (π.χ. 5): 1
-------------------------------------------------------------------------------------------
>>> ΑΡΙΘΜΟΣ ΕΠΙΘΥΜΗΤΩΝ ΕΞΑΓΟΜΕΝΩΝ html ΣΕΛΙΔΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ = 2
>>>>> ΤΡΕΧΟΥΣΑ ΔΙΕΥΘΥΝΣΗ URL ΓΙΑ ΕΞΑΓΩΓΗ ΣΕΛΙΔΑΣ (ΣΕ ΑΡΧΕΙΟ ΟΝΟΜΑΤΟΛΟΓΙΑΣ page0.html) = https://www.ekathimerini.com/search?q=%ce%95%ce%bb%ce%bb%ce%ac%ce%b4
%ce%bl&page=0
> ΑΠΟΘΗΚΕΥΤΗΚΕ ΤΟ ΑΝΤΙΣΤΟΙΧΟ ΑΡΧΕΙΟ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΟΝΟΜΑΤΟΛΟΓΙΑΣ page0.html <
>>>>> ΤΡΕΧΟΥΣΑ ΔΙΕΥΘΥΝΣΗ URL ΓΙΑ ΕΞΑΓΩΓΗ ΣΕΛΙΔΑΣ (ΣΕ ΑΡΧΕΙΟ ΟΝΟΜΑΤΟΛΟΓΙΑΣ page1.html) = https://www.ekathimerini.com/search?q=%ce%95%ce%bb%ce%bb%ce%ac%ce%b4
%ce%bl&page=1
> ΑΠΟΘΗΚΕΥΤΗΚΕ ΤΟ ΑΝΤΙΣΤΟΙΧΟ ΑΡΧΕΙΟ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΟΝΟΜΑΤΟΛΟΓΙΑΣ page1.html <
>>>
```



ΤΕΛΟΣ ΑΝΑΠΤΥΞΗΣ

Τεκμηρίωση (Νοέμβριος 2020):
© Παρασκευάς Γ. Παπαδόπουλος
Μηχανικός Η/Υ Συστημάτων, Η.Ι. (MSc)
Email : pdesper@gmail.com