

Ανάπτυξη συστήματος για τη δημιουργία βάσεων δεδομένων από το web με Python

Στα πλαίσια του μαθήματος Επιχειρηματική Ευφυΐα και Διαχείριση Γνώσης



Εισαγωγή

Σκοπός της εργασίας ήταν ανάπτυξη ενός συστήματος δημιουργίας βάσης δεδομένων με εξόρυξη δεδομένων από το web σε γλώσσα προγραμματισμού Python. Μετά από έρευνα διάφορων ιστοτόπων που προσφέρουν σημαντικά δεδομένα στον τελικό χρήστη, η online υπηρεσία εύρεσης tracking code των Ελληνικών Ταχυδρομείων παρουσίασε μια σημαντική ευκαιρία για την εξόρυξη μεγάλης ποσότητας δεδομένων, με ανοιχτά στο κοινό δεδομένα αλλά όχι εύκολα προσβάσιμα. Το πρόβλημα αυτό αντιμετωπίζεται με την παραγωγή πλήθους τυχαίων tracking codes που στέλνονται με τη μορφή request στον server των ΕΛΤΑ. Τα δεδομένα τελικά αποθηκεύονται σε μορφή JSON και είναι έτοιμα για εισαγωγή σε συστήματα διαχείρισης βάσεων δεδομένων τύπου NoSQL. Μετά από μετατροπή σε CSV τα δεδομένα μπορούν να εισαχθούν και σε σχεσιακά συστήματα διαχείρισης βάσεων δεδομένων.

Πηγή δεδομένων

Η πηγή των δεδομένων είναι προσβάσιμη στο κοινό στην παρακάτω διεύθυνση:

<https://www.elta-courier.gr/search>

ΕΛΤΑ Courier

Εταιρεία Υπηρεσίες Αποστολή Επικοινωνία Ερωτήσεις Εργαλεία

Αναζήτηση... **11120** ΕΞΥΠΗΡΕΤΗΣΗ ΠΕΛΑΤΩΝ
ΑΠΟ ΣΤΑΘΕΡΟ Η ΚΙΝΗΤΟ ΜΕ ΑΣΤΙΚΗ ΧΡΕΩΣΗ

Εταιρεία

- Αναζήτηση Αποστολής
- Υπολογισμός Κόστους
- Ογκομετρητής
- Οδηγίες Συσκευασίας
- Δίκτυο Καταστημάτων

Αναζήτηση Αποστολής

Αναζητήστε έως 5 αποστολές ταυτόχρονα. Πληκτρολογήστε μέχρι 5 αριθμούς αποστολής χωρισμένους με enter, που βρίσκονται στο πάνω δεξιά μέρος του Συνοδευτικού Δελτίου Ταχυμεταφοράς (Voucher) και κάντε κλικ στο "Αναζήτηση".

πληκτρολογήστε τον αριθμό...

ANAZHTHSH

NZ990003254GR

NEA Διαβάστε περισσότερα...

29-12-2015 ΠΑΡΑΔΟΣΗ ΑΝΘΡΩΠΙΣΤΙΚΗΣ ΒΟΗΘΕΙΑΣ ΓΙΑ ΠΡΟΣΦΥΓΕΣ ΚΑΙ ΜΕΤΑΝΑΣΤΕΣ	28-12-2015 ΕΛΤΑ Courier 2015	24-12-2015 Ε.Ε. κατά ταχυμεταφορών για τα υψηλά έξοδα αποστολής	18-12-2015 Κατατέθηκε στη Ρ.Α.Ε. η αίτηση των ΕΛΤΑ για Άδεια Προμήθειας και Εμπορίας Ηλεκτρικής Ενέργειας
--	---------------------------------	---	--

Εικόνα 1 - Αναζήτηση αποστολής χρήστη με την εισαγωγή του tracking code

Τα δεδομένα εμφανίζονται στην σελίδα που βλέπει ο χρήστης μέσω ενός άλλου script που λαμβάνει ως παράμετρο "number" στο url, τον tracking code που έχει πληκτρολογηθεί:

<https://www.elta-courier.gr/track.php>

Δομή δεδομένων

Τα δεδομένα ανταλλάσσονται με την μέθοδο POST και επιστρέφονται σε μορφή JSON. Παρακάτω φαίνεται ένα παράδειγμα του response που λαμβάνεται για έναν τυχαίο κωδικό (CP329354847GR) με τα αντίστοιχα πεδία:

```
{
  "status": 1,
  "result": {
    "CP329354847GR": {
      "status": 1,
      "result": [
        {
          "date": "20-02-2015",
          "time": "14:01",
          "place": "ΒΑΡΚΙΖΑΣ",
          "status": "ΠΑΡΑΛΑΒΗ ΑΠΟ RECEPTION ΣΤΑΘΜΟΥ"
        },
        {
          "date": "24-02-2015",
          "time": "07:39",
          "place": "ΒΑΡΚΙΖΑΣ",
          "status": "ΚΛΕΙΣΙΜΟ ΑΠΟ ΒΑΡΚΙΖΑΣ"
        },
        {
          "date": "24-02-2015",
          "time": "19:20",
          "place": "Κ.Δ. ΔΕΜΑΤΩΝ ΑΘΗΝΑΣ",
          "status": "ΚΛΕΙΣΙΜΟ ΑΠΟ Κ.Δ. ΔΕΜΑΤΩΝ ΑΘΗΝΑΣ"
        },
        {
          "date": "25-02-2015",
          "time": "11:55",
          "place": "ΑΧΑΡΝΩΝ (Κ.Κ.)",
          "status": "ΑΦΙΞΗ ΣΕ ΑΧΑΡΝΩΝ (Κ.Κ.)"
        },
        {
          "date": "06-03-2015",
          "time": "10:37",
          "place": "ΑΧΑΡΝΩΝ (Κ.Κ.)",
          "status": "ΑΠΟΣΤΟΛΗ ΠΑΡΑΔΟΘΗΚΕ-ΧΩΡΙΣ ΣΤΟΙΧΕΙΟ ΠΑΡΑΔΟΣΗΣ"
        }
      ]
    }
  }
}
```

Σε αυτά τα δεδομένα παρατηρούμε τα πεδία με τα status, τον κωδικό που στείλαμε, ημερομηνίες, ώρα, τοποθεσία (κατάστημα ΕΛΤΑ), και κατάσταση για την συγκεκριμένη ώρα που μετακινήθηκε η αποστολή με τον εκάστοτε κωδικό.

Το πρόγραμμα Python

Προαπαιτούμενα

Για την σωστή λειτουργία του κώδικα η Python χρειάζεται να έχει εγκατεστημένες τις παρακάτω βιβλιοθήκες:

- requests
- random
- time
- pause
- json
- tqdm
- codecs
- fileinput
- os
- math

Το πρόγραμμα έχει γραφεί και δοκιμαστεί σε Python 2.7.

Δημιουργία tracking codes

Η δημιουργία των κωδικών γίνεται με βάση το πρότυπο [UPU S10](#).

Για αρχή επιλέγονται τυχαία 60000 αριθμοί ανάμεσα στο 0 και στο 99999999.

```
το πρόθεμα CP (δέματα)
+ "3"
+ τους αριθμούς που επιλέχθηκαν (με συμπληρωματικά μηδενικά στην αρχή)
+ το υπολογιζόμενο ψηφίου ελέγχου (όπως ορίζει το πρότυπο S10)
+ "GR" (δέματα που αποστέλλονται απο Ελλάδα)
```

δίνει έναν έγκυρο αριθμό αποστολής που γίνεται να σταλεί στο σύστημα των ΕΛΤΑ.

π.χ.

CP+3+2935484+7+GR

Requests

Στη συνέχεια του προγράμματος αποστέλλονται 60000 requests, ένα για κάθε tracking code που υπολογίστηκε. Τα requests γίνονται σειριακά και με καθυστέρηση 10ms για αποφυγή εμποδισμού του server των ΕΛΤΑ απο υπερβολικά requests. Η βιβλιοθήκη requests και pause βοηθούν σε αυτό το κομμάτι της υλοποίησης.

Αποθήκευση δεδομένων

Κάθε ανταπόκριση απο τον server των ΕΛΤΑ εφόσον δεν είναι κενή αποθηκεύεται απευθείας στο ανοιγμένο αρχείο RESULTS.json και για να είναι έγκυρο το τελικό αρχείο, το πρόγραμμα πρέπει να τελειώσει τα requests και για τους 60000 κωδικούς. Οι ελληνικοί χαρακτήρες αποθηκεύονται σε μορφή unicode, που όμως αποκωδικοποιείται εύκολα απο προγράμματα που “διαβάζουν” json. Οι κωδικοί που υπολογίστηκαν σε προηγούμενο βήμα αποθηκεύονται και αυτοί στο αρχείο CODES.txt. Το τελικό αρχείο θα έχει μεγάλο μέγεθος, γύρω στα 17mb.

Λειτουργία

Το πρόγραμμα ολοκληρώνει την λειτουργία του μέσα σε 4-5 ώρες. Απο το command line, είναι ορατή η πρόοδος, ο αριθμός των έγκυρων κωδικών και αυτών με δεδομένα, για την διευκόλυνση του χρήστη ενώ γίνεται σαφής οποιαδήποτε προσπάθεια διαγραφής αρχείου. Είναι σημαντικό να σημειωθεί πως κάθε φορά που τρέχει το πρόγραμμα, παράγονται διαφορετικοί κωδικοί και έτσι τα τελικά δεδομένα θα διαφέρουν απο προηγούμενες εκτελέσεις.

Χρησιμοποιήθηκαν ακόμη τα πακέτα WinPython και PyInstaller για την υλοποίηση του κώδικα και του εκτελέσιμου για Windows. Για περισσότερες πληροφορίες για το πρόγραμμα μπορείτε να δείτε σχόλια μέσα στον κώδικα (codes.py)