Ανάλυση Δεδομένων και Έλεγχος Υποθέσεων με τεχνολογία Python/pandas

Σταύρος Δημητριάδης
sdemetri@csd.auth.gr
Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

Σκοπός

Σκοπός του εργαστηρίου είναι να παρουσιάσει στους συμμετέχοντες τον τρόπο με τον οποίο μπορούν να χρησιμοποιούν το οικοσύστημα ανοικτών τεχνολογιών python/pandas ώστε να:
Α) Εισάγουν δεδομένα από αρχείο (πχ. xls, csv, txt) και τα επεξεργάζονται με ποικίλους τρόπους που υποστηρίζει η βιβλιοθήκη pandas.
Β) Εφαρμόζουν μεθόδους επαγωγικής στατιστικής για τον έλεγχο υποθέσεων (t-test, ANOVA, κλπ.) με χρήση της βιβλιοθήκης Scipy.

Μετά το τέλος του εργαστηρίου οι συμμετέχοντες θα μπορούν να:

1) Χειρίζονται δεδομένα σε ποικίλες δομές επιλέγοντας την κατάλληλη ανάλογα με το είδος του προβλήματος (όπως λίστες, λεξικά, πίνακες, Series & DataFrames).
2) Εισάγουν και επεξεργάζονται δεδομένα με τις τεχνικές που προσφέρει η βιβλιοθήκη pandas.
3) Εφαρμόζουν στατιστικές μεθόδους ελέγχου υποθέσεων με χρήση της βιβλιοθήκης Scipy.

Κοινό

Το σεμινάριο απευθύνεται σε ερευνητές στο χώρο των κοινωνικών/οικονομικών ή άλλων επιστημών οι οποίοι ενδιαφέρονται για προχωρημένου επιπέδου επεξεργασία ποσοτικών δεδομένων και εφαρμογή στατιστικών μεθόδων για τον έλεγχο υποθέσεων.

Μια βασική γνώση της γλώσσας προγραμματισμού Python θα είναι εξαιρετικά βοηθητική για τους συμμετέχοντες καθώς το σεμινάριο δεν στοχεύει στην αναλυτική εκμάθηση της Python αλλά θα προσφέρει μόνο μια συνοπτική επισκόπηση βασικών σημείων που σχετίζονται κυρίως με τις δομές δεδομένων. Απαραίτητο είναι οι συμμετέχοντες να έχουν μια γενική παιδεία σχετικά με τη χρήση υπολογιστή (ικανότητα εγκατάστασης εφαρμογών, δημιουργίας/αποθήκευσης αρχείων, κλπ.).

Ψηφιακό περιβάλλον και προστιθέμενη αξία

Το οικοσύστημα Python (δηλ. η γλώσσα προγραμματισμού Python συνοδευόμενη από μία πλειάδα εξωτερικών πακέτων/βιβλιοθηκών λογισμικού) είναι ένα ισχυρό σύγχρονο τεχνολογικό εργαλείο ανοικτού κώδικα, χρήσιμο όχι μόνον για τη στενά νοούμενη περιοχή της επιστήμης υπολογιστών αλλά και για τους ερευνητές κάθε ειδικότητας που στοχεύουν στην ανάλυση και επεξεργασία δεδομένων.

Σε σχέση με την ανάλυση δεδομένων τα οικοσυστήματα της Python και της R προσφέρουν δύο ισχυρές λύσεις ανοικτών τεχνολογιών σε αντίθεση με κλειστά εμπορικά λογισμικά όπως Matlab, SAS, SPSS. Ένα σημαντικό πλεονέκτημα της Python είναι πως πρόκειται για γενικού σκοπού γλώσσα προγραμματισμού σε αντίθεση με την R η οποία

στοχεύει στενά στη στατιστική ανάλυση και επεξεργασία. Αυτό το στοιχείο καθιστά την Python σημαντικά καλύτερη επένδυση μάθησης καθώς ο χρήστης Python θα είναι σε θέση να γράψει κώδικα και να χρησιμοποιήσει βιβλιοθήκες σχετικές με διάφορες επιστημονικές περιοχές που πιθανώς να τον ενδιαφέρουν μελλοντικά.

Ειδικά τώρα για το αντικείμενο του εργαστηρίου η έμφαση δίνεται σε κάτι που έχει ιδιαίτερη σημασία για τον ερευνητή, δηλ. την εύκολη και αποδοτική επεξεργασία δεδομένων με στόχο την αξιόπιστη εξαγωγή συμπερασμάτων σχετικά με ερευνητικές υποθέσεις. Τα παραδείγματα που θα δοθούν θα αναφέρονται σε σενάρια εκπαιδευτικής έρευνας με ποσοτικά δεδομένα, όπως πχ. επεξεργασία αποτελεσμάτων ερωτηματολογίου μετα-ελέγχου για την αξιολόγηση της μάθησης. Στόχος είναι να αναπτύξουν οι συμμετέχοντες την ικανότητα να χειρίζονται το συγκεκριμένο τεχνολογικό εργαλείο ώστε να το αξιοποιήσουν σε κάθε σχετική ερευνητική τους δραστηριότητα. Επίσης ο εισηγητής του εργαστηρίου θα σχολιάσει σε βάθος το νόημα των στατιστικών ελέγχων (t-test, ANOVA, chi square test, κ.ά) ώστε συνολικά οι συμμετέχοντες να αναπτύξουν κατανόηση των δυνατοτήτων αλλά και περιορισμών των ελέγχων αυτών και όχι απλά να μάθουν την εφαρμογή τους με τη συγκεκριμένη τεχνολογία.

Συνολικά, το σεμινάριο θα δώσει την ευκαιρία στους συμμετέχοντες να γνωρίσουν μια ισχυρή τεχνολογική λύση πολλαπλών δυνατών εφαρμογής μέσα από τη συγκεκριμένη οπτική της ανάλυσης εκπαιδευτικών δεδομένων καθώς και να εμβαθύνουν στον στατιστικό έλεγχο υποθέσεων.

Οργάνωση & Προαπαιτούμενα

Θα χρησιμοποιηθεί το οικοσύστημα Python/pandas που είναι διαθέσιμο μέσω της διανομής (distribution) 'Anaconda' (για την έκδοση 3.x της γλώσσας) και το περιβάλλον εργασίας θα είναι το Jupyter Notebook το οποίο εγκαθίσταται ως τμήμα του Anaconda.

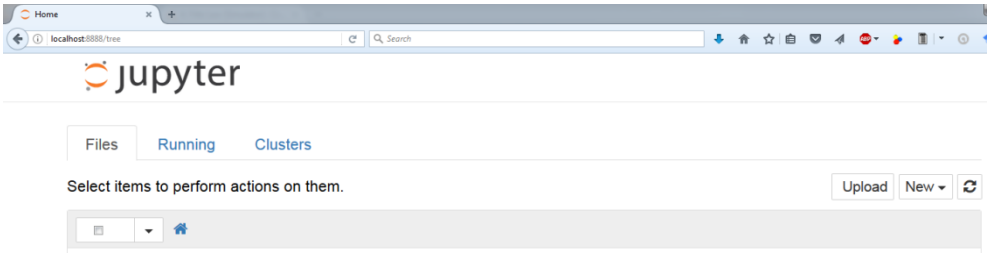
Το εργαστήριο θα ακολουθήσει το μοντέλο 'BYOD' (Bring Your Own Device). Ιδανικά οι συμμετέχοντες θα πρέπει να εγκαταστήσουν σε δικό τους φορητό υπολογιστή το πακέτο Anaconda και να μπορούν να εκκινήσουν το Jupyter Notebook. Στην περίπτωση που αυτό δεν είναι δυνατό οι διοργανωτές του σεμιναρίου θα φροντίσουν να υπάρχει εγκατεστημένο το λογισμικό τοπικά σε υπολογιστές διαθέσιμους στους συμμετέχοντες (θα υπάρξει σχετικά ανακοίνωση).

Η διάρκεια του εργαστηρίου είναι 4ωρη (4 X 50 λεπτά με 10λεπτα διαλείμματα ή 2 X 100 με 30λεπτο ενδιάμεσο διάλειμμα).

Εγκατάσταση λογισμικού: οι συμμετέχοντες θα πρέπει να κατεβάσουν και εγκαταστήσουν στον υπολογιστή τους το:

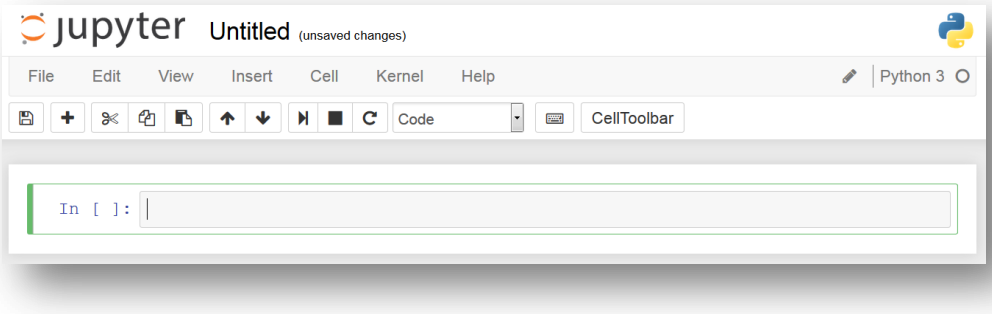
- **Anaconda 4.3.1 (Python 3.6 version)** διαθέσιμο στο: https://www.continuum.io/downloads

Μετά την εγκατάσταση η εκκίνηση του jupyter notebook μπορεί να γίνει εύκολα από τον κατάλογο 'Anaconda' που θα έχει δημιουργηθεί. Η εφαρμογή εκκινεί τοπικά έναν server και το περιβάλλον εμφανίζεται ως σελίδα στον προεπιλεγμένο browser (Εικ. 1)



Εικόνα 1. Αρχική σελίδα Jupyter notebook

Επιλέγοντας 'New / Python 3' ένα νέο notebook εμφανίζεται όπου ο χρήστης μπορεί να γράψει, να εκτελέσει και να αποθηκεύσει κώδικα Python (Εικ. 2).



Εικόνα 2. Μορφή νέου αρχείου Python 3 στο Jupyter notebook

Εκπαιδευτικό υλικό: Το υλικό του εργαστηρίου είναι ήδη ανεπτυγμένο στη διεύθυνση: http://pytolearn.csd.auth.gr/ και εμπλουτίζεται/βελτιώνεται συνεχώς. Οι συμμετέχοντες θα έχουν σύνδεση στο διαδίκτυο και πρόσβαση στο συγκεκριμένο υλικό στη διάρκεια του σεμιναρίου.

Περιεχόμενα: Τα περιεχόμενα του σεμιναρίου είναι:

- Εξοικείωση με το οικοσύστημα και το περιβάλλον Jupyter Notebook. Επισκόπηση της Python με έμφαση σε δομές δεδομένων: Λίστες, Λεξικά, Πίνακες (με χρήση της numpy).
- Εισαγωγή στο πακέτο pandas. Ανάγνωση δεδομένων από εξωτερικό αρχεία και οργάνωση/επεξεργασία τους σε δομές Series και DataFrame. Γραφική αναπαράσταση με χρήση της Matplotlib.
- Επεξεργασία δεδομένων σε δομή DataFrame. Χρήση μεθόδων & συναρτήσεων της βιβλιοθήκης Scipy. Εισαγωγή στις στατιστικές μεθόδους για τον έλεγχο υποθέσεων.
- Παραδείγματα εφαρμογής τεχνικών ελέγχου υποθέσεων σε εκπαιδευτικά δεδομένα.

Με την ολοκλήρωση του εργαστηρίου θα δοθεί στους συμμετέχοντες βεβαίωση παρακολούθησης.

Για κάθε σχετική πληροφορία / διευκρίνιση οι ενδιαφερόμενοι μπορούν να επικοινωνούν με τον διδάσκοντα στη διεύθυνση email: sdemetri@csd.auth.gr

Πηγές στο Διαδίκτυο

Python 3 tutorial (n.d.). Ανακτήθηκε 1 Μαρτίου, 2017, από https://docs.python.org/3/tutorial/index.html

Anaconda documentation (n.d.). Ανακτήθηκε 1 Μαρτίου, 2017, από https://docs.continuum.io/

Jupyter notebook (n.d.). Ανακτήθηκε 1 Μαρτίου, 2017, από http://jupyter.org/

Scipy, (n.d.). Ανακτήθηκε 1 Μαρτίου, 2017, από http://scipy.org/

Matplotlib (n.d.). Ανακτήθηκε 1 Μαρτίου, 2017, από http://matplotlib.org/

Pandas (n.d.). Ανακτήθηκε 1 Μαρτίου, 2017, από http://pandas.pydata.org/pandas-docs/stable/

'pytolearn' web site (n.d.). Ανακτήθηκε 1 Μαρτίου, 2017, από http://pytolearn.csd.auth.gr/

Descriptive statistics (n.d.). Ανακτήθηκε 1 Μαρτίου, 2017, από http://www.socialresearchmethods.net/kb/statdesc.php/

Inferential statistics (n.d.). Ανακτήθηκε 1 Μαρτίου, 2017, από http://www.socialresearchmethods.net/kb/statinf.php/

Concepts & Applications of Inferential Statistics (n.d.). Ανακτήθηκε 1 Μαρτίου, 2017, από http://vassarstats.net/textbook/index.html/