ΕΠΛ448: ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΣΤΟΝ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ

Διδακτικές Μονάδες: 7,5 ECTS
Εξάμηνο: Εαρινό
Διδάσκοντες: Γιώργος Πάλλης (διαλέξεις, φροντιστήριο) και Παύλος Αντωνίου (εργαστήρια)
Ιστοσελίδα Μαθήματος: https://www.cs.ucy.ac.cy/courses/EPL448/

Προσοχή: Ενδιάμεση Εξέταση 4 Μαρτίου 2025

Στόχοι
Το μάθημα αποσκοπεί να δώσει στους φοιτητές το απαραίτητο υπόβαθρο σε θέματα που αφορούν την εξόρυξη δεδομένων στον Παγκόσμιο Ιστό (Π.Ι). Με δεδομένο ότι ο Π.Ι μπορεί να θεωρηθεί ως μία πολύ μεγάλη και ετερογενής βάση δεδομένων, νέες μεθοδολογίες και τεχνικές μελετώνται με στόχο την εξαγωγή και διαχείριση χρήσιμων προτύπων γνώσης. Το μάθημα προσφέρει μια εισαγωγή στις βασικές έννοιες και τεχνικές του ερευνητικού πεδίου που αφορά στην ανάλυση μεγάλων, σε όγκο και πολυπλοκότητα, συλλογών δεδομένων στον Π.Ι. Αρχικά το μάθημα εισάγει το προγραμματιστικό μοντέλο Map-Reduce. Στη συνέχεια παρουσιάζονται οι βασικές αρχές της εξόρυξης δεδομένων δίνοντας παράλληλα μία γενική εικόνα των βασικών απαιτήσεων και αναγκών για την εφαρμογή νέων αποτελεσματικών μεθόδων και τεχνικών ανάλυσης δεδομένων στον Π.Ι. Στο πλαίσιο αυτό εξετάζονται κανόνες συσχέτισης (association rules) και αλγόριθμοι εποπτευόμενης και μη εποπτευόμενης εξόρυξης δεδομένων, όπως ομαδοποίηση (clustering) και κατηγοριοποίηση (classification). Έμφαση δίνεται στην ανάλυση συνδέσμων ενός Διαδικτυακού τόπου, στα συστήματα προτιμήσεων, στα κοινωνικά δίκτυα και στη Διαφήμιση στον Π.Ι. Στο τέλος παρουσιάζονται κάποιες ερευνητικές μελέτες που σχετίζονται με την εξόρυξη δεδομένων και γνώσης.

Προαπαιτούμενα Μαθήματα
Επιτρέπεται η εγγραφή στο ΕΠΛ451 όσων φοιτητών έχουν ολοκληρώσει επιτυχώς το μάθημα ΕΠΛ231 (Δομές Δεδομένων και Αλγόριθμοι) και το μάθημα ΕΠΛ342 (Βάσεις Δεδομένων).

Τρόπος Διδασκαλίας
Το μάθημα καλύπτει θεωρία και πρακτική και αποτελείται από 4 ώρες διαλέξεων/φροντιστηρίου και εβδομαδιαίου εργαστηρίου. Οι φοιτητές θα εξοικειωθούν με το αντικείμενο του μαθήματος και μέσω εργαστηριακών ασκήσεων και εργασιών.

Αξιολόγηση και Βαθμολογία
Οι φοιτητές θα αξιολογηθούν μέσα από ένα σύνολο τριών εργασιών, με την ολοκλήρωση και παρουσίαση ομαδικής εργασίας εξαμήνου και με γραπτές εξετάσεις (ενδιάμεση και τελική). Ο τελικός βαθμός διαμορφώνεται με βάση τα ποσοστά που δίνονται στον ακόλουθο πίνακα. Σημειώνεται ότι για την επίτευξη προβιβάσιμου βαθμού (5), ο φοιτητής πρέπει να έχει **επιτύχει βαθμό πάνω από 45/100 στον σταθμισμένο μέσο όρο της γραπτής ενδιάμεσης εξέτασης και της τελικής εξέτασης**.

- Ατομικές Εργασίες: 5% [**Κατά την υποβολή τους στο εργαστήριο** – δε θα αξιολογούνται ως προς την ορθότητά τους, π.χ., ένας φοιτητής που υποβάλει όλες τις εργασίες θα πάρει όλες τις μονάδες]• Ομαδική Εργασία Εξαμήνου: 25%
- Γραπτή Ενδιάμεση Εξέταση: 30%
- Γραπτή Τελική Εξέταση: 40%

Αν έχετε πάνω από **3 αδικαιολόγητες απουσίες** (δηλ, απουσιάζεται χωρίς να έχετε ενημερώσει το διδάσκοντα) στα εργαστήρια δε θα προσμετράτε η βαθμολογία σας στις ατομικές εργασίες.

Ομαδική Εργασία Εξαμήνου
- Οι ομάδες είναι 2-3 ατόμων
- Ο κώδικας πρέπει να είναι συνέχεια **up-to-date με ένα Github repository.** Η συνεισφορά του εκάστοτε φοιτητή θα φαίνεται από τα analytics του Github το οποίο θα λαμβάνεται υπόψην στην τελική βαθμολογία για την ομαδική εργασία.
- Θα ζητηθεί να εξηγηθεί η συνεισφορά κάθε φοιτητή στην ομαδική εργασία.

Το report θα πρέπει να περιλαμβάνει τις παρακάτω ενότητες:
- Introduction
- Data cleaning
- Data manipulation (encoding (strings/dates to numerical); data scaling (if needed))
- Features engineering (feature selection / extraction)
- Methodology (training and prediction (cross-validation, multiple predictors (regressors / classifiers), parameter best values selection)
- Exploratory Data Analysis (show important features, feature correlation (with heatmaps & scatterplots), feature distributions, features with outliers)

Σε περίπτωση αξιοποίησης εργαλείων Τεχνητής Νοημοσύνης (ΤΝ) κατά την εκπόνηση εργασιών, ο/η φοιτητής/τήτρια θα πρέπει να αναφέρει **ρητά και επακριβώς το εργαλείο ΤΝ** που χρησιμοποίησε και **με ποιον τρόπο το αξιοποίησε.** Η αυτούσια αντιγραφή από εργαλεία ΤΝ απαγορεύεται.

Βιβλιογραφία
Mining Massive Datasets, by Jure Leskovec, Anand Rajaraman and Jeff Ullman, Cambridge University Press, 2014.

Θα χρησιμοποιηθεί επίσης υλικό και συλλογή άρθρων που θα δοθούν στη διάρκεια του μαθήματος.