ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

Διαδικαστικά

Συστάσεις Ι

- Ποιός είμαι εγώ:
  - Email: tsap@cs.uoi.gr
  - Γραφείο: Β.3
  - Προτιμώμενες ώρες γραφείου: 11:00-18:00

  - Ενδιαφέροντα
    - Web mining, Social networks, User Generated Content
    - Bioinformatics
    - Συνδυασμός θεωρίας και πράξης

Συστάσεις II

- Ποιοί είσαστε εσείς:
  - Συμπληρώστε τη φόρμα με τα στοιχεία σας για την email λίστα του μαθήματος.

Γενικές πληροφορίες για το μάθημα

- Διαλεξεις: Τρίτη 12:00 – 3:00 μ.μ., αίθουσα Ι2
  - Οι διαφάνειες θα είναι στα αγγλικά, αλλά η διάλεξη θα γίνεται στα ελληνικά.
  - Θα έχουμε και κάποια επιπλέον μαθήματα/αναπληρώσεις κάποιες εβδομάδες.
    - Πότε σας βολεύει?

- Web: http://www.cs.uoi.gr/~tsap/teaching/cse012/
  - Ανακοινώσεις, ασκήσεις, υλικό για διάβασμα διαφάνειες από τις διαλέξεις
  - Γραφτείτε και στη σελίδα του μαθήματος στο ecourse.

- Βαθμολογία:
  - Το μάθημα θα έχει απαλλακτικές εργασίες. Δεν υπάρχει τελική εξέταση ούτε τον Ιανουάριο ούτε τον Σεπτέμβριο. Θα υπάρχει προσωπική εξέταση για τις εργασίες.
  - Φέτος μπορεί να έχουμε κάποια εξέταση
  - Πολιτική για καθυστερημένες εργασίες:
    - Μία μέρα καθυστέρηση -10%, δύο μέρες -20%, τρεις μέρες -40%, τέσσερεις μέρες -70%, πέντε μέρες -100%.
    - Free pass policy: Έχετε 4 free passes τα οποία μπορείτε να χρησιμοποιήσετε όποτε θέλετε για να καθυστερήσετε την παράδοση μιας εργασίας. Το κάθε pass σας δίνει μία μέρα επιπλέον.

Ασκήσεις

- Οι ασκήσεις θα έχουν (συνήθως) δύο τύπους ερωτήσεων: θεωρητικές και προγραμματιστικές.
  - Θεωρητικές: Θα σας ζητηθεί να σχεδιάσετε ένα αλγόριθμο, ή να αποδείξετε κάποια ιδιότητα
  - Αλγοριθμικές: Θα σας ζητηθεί να σχεδιάσετε ένα μια λύση για ένα πρόβλημα.
  - Προγραμματιστικές: Θα σας ζητηθεί να υλοποιήσετε ένα αλγόριθμο, ή να χρησιμοποιήσετε κάποιο έτοιμο εργαλείο σε κάποια δεδομένα.
- Αναφορά: Στις κάποιες ερωτήσεις θα πρέπει να παραδώσετε μία αναφορά. Η αναφορά αυτή μετράει ένα σημαντικό ποσοστό του βαθμού της ερώτησης και πρέπει να γίνεται προσεκτικά. Τις περισσότερες φορές σας ζητείται να εξηγήσετε τα αποτελέσματα κάποιου πειράματος.
- Προγραμματισμός: Η επεξεργασία μεγάλων ποσοτήτων δεδομένων απαιτεί έξυπνο και αποτελεσματικό προγραμματισμό.
  - Πρέπει να αποφεύγετε δαπανηρές λειτουργίες.
  - Πρέπει να χρησιμοποιείτε τις κατάλληλες δομές.
  - Πρέπει να προσπαθείτε να χρησιμοποιείτε λίγη μνήμη.
  - Κάποιες φορές το πρόγραμμα σας μπορεί να πάρει μερικές ώρες να τελειώσει.

«Προαπαιτούμενα»

- Δεν υπάρχουν προαπαιτούμενα αλλά καλό θα είναι να έχετε κάποια άνεση με:
  - Προγραμματισμός + Python: Ευκολία στην εκμάθηση νέων εργαλείων. Θα χρησιμοποιήσουμε κάποιες νέες βιβλιοθήκες python.
  - Βασεις δεδομένων: Χρήση βασικών SQL λειτουργειών
  - Αλγορίθμους: γνώση βασικών αλγορίθμων (π.χ., sorting), και σχεδίασης αλγορίθμων (greedy algorithms, dynamic programming).
  - Δομές δεδομένων: χρήση βασικών δομών δεδομένων.
  - Πιθανότητες: Άνεση με βασικές γνώσεις πιθανοτήτων.
  - Γραμμική άλγεβρα: πίνακες, διανύσματα, ιδιοδιανύσματα.
  - Γραφήματα: βασικές έννοιες γραφημάτων

Στόχοι του μαθήματος

- Να μάθετε βασικές έννοιες του data mining, που καλύπτουν και το θεωρητικό υπόβαθρο, και την εφαρμογή στην πράξη.
- Να καταλάβετε το είδος των προβλημάτων που μπορείτε να λύσετε χρησιμοποιώντας τεχνικές data mining.
- Να καταλάβετε τη θεωρία και τα μαθηματικά πίσω από τους αλγόριθμους και τις τεχνικές
- Να αποκτήσετε ένα σύνολο από εργαλεία (toolbox) για εξόρυξη δεδομένων.
- Να παίξετε με πραγματικά δεδομένα και να δείτε κάποια ενδιαφέροντα πραγματικά προβλήματα (ελπίζω).
- Να μάθετε κάτι ενδιαφέρον.

Μάθημα

- Η παρακολούθηση και συμμετοχή είναι απαραίτητες
  - Κάνετε ερωτήσεις. Κάποια πράγματα δεν θα είναι ξεκάθαρα και θα πρέπει να τα επαναλάβω.
  - Αν κάτι στηρίζεται σε παλαιότερη γνώση που δεν θυμάστε ζητήστε να κάνουμε μια (σύντομη) επισκόπηση.
  - Αν υπάρχει πρόβλημα με αγγλική ορολογία και τις διαφάνειες μπορούμε να κάνουμε κάποιες ρυθμίσεις.
- Για τα εργαλεία που θα χρησιμοποιήσουμε θα προσπαθήσω να κάνουμε ένα ξεχωριστό φροντιστήριο.

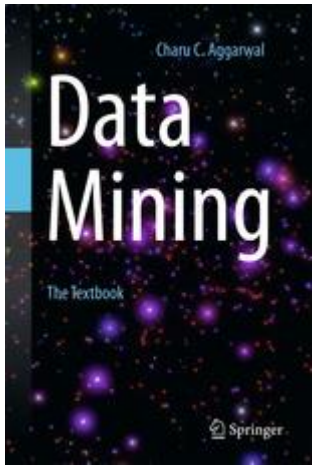
Θέματα που θα καλύψουμε

- Κάποιο υποσύνολο από τα παρακάτω
  - Frequent itemsets and association rules (συσχετισμοί)
  - Definitions and Computation of Similarity
  - Clustering (συσταδιοποίηση), co-clustering, compression
  - Regression και Classification (κατηγοριοποίηση)
  - Dimensionality Reduction
  - Ranking (ιεραρχηση/ταξινόμηση)
  - Recommendation systems
  - Graph Analysis
  - Covering problems
  - Map-Reduce tools
  - Time-series analysis
  - Aggregation
  - Privacy preserving data mining

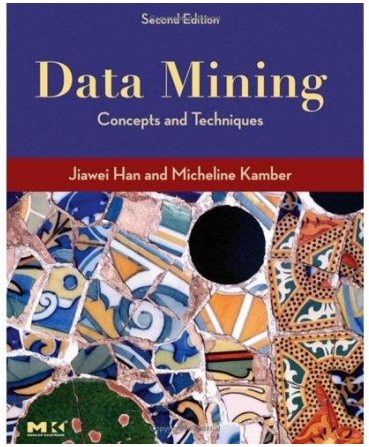
Βιβλιογραφία (ελληνικά)

- A. Rajaraman, J. D. Ullman. Εξόρυξη από Μεγάλα Σύνολα Δεδομένων, ΕΚΔΟΣΕΙΣ ΝΕΩΝ ΤΕΧΝΟΛΟΓΙΩΝ, 2014
- P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining Addison Wesley, 2006, Β. Βερύκιος και Σ. Σουραβλάς, Εκδόσεις Τζιόλα (2010).
- ΕΞΟΡΥΞΗ ΚΑΙ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ: ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ ΚΑΙ ΑΛΓΟΡΙΘΜΟΙ, MOHAMMED J. ZAKI, WAGNER MEIRA JR

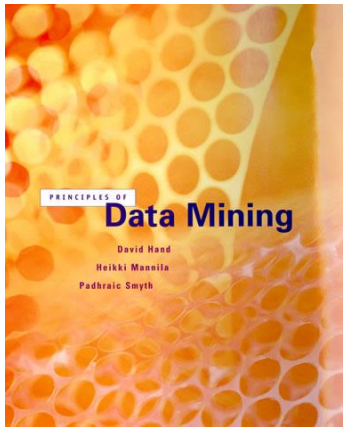
Επιπλέον Βιβλιογραφία (αγγλικά)



Charu Aggarwal, Data Mining, The Textbook, Springer, 2015

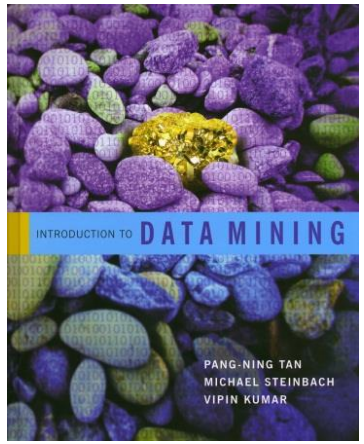


J. Han and M. Kamber. Data Mining: Concepts and Techniques, Morgan Kaufmann, 2006

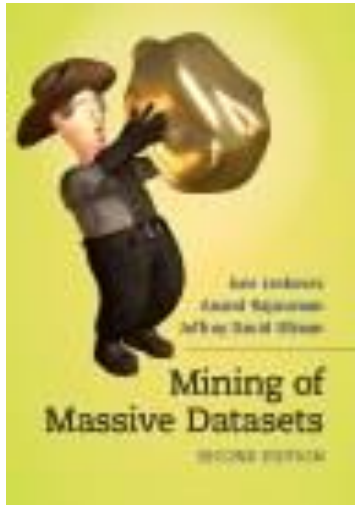


Hand, Mannila, Smyth. Principles of Data Mining

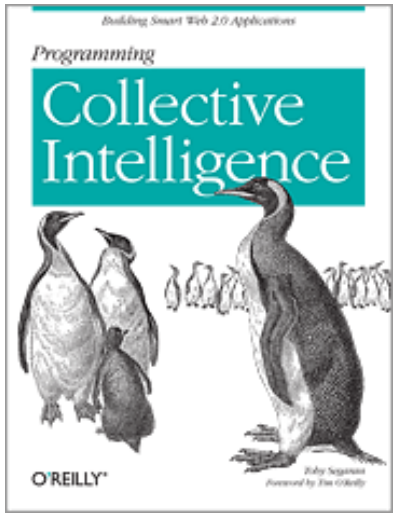
Online βιβλία (αγγλικά)



P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Addison Wesley, 2006
Διαφάνειες και υλικό online



Anand Rajaraman, Jeff Ullman and Jure Leskovec, Mining Massive Datasets.
Δεύτερη έκδοση. Διατίθεται δωρεάν online.



Toby Segaran, Programming Collective Intelligence. Building Smart Web 2.0 Applications

Υλικό

- Εκτός από βιβλία θα χρησιμοποιήσουμε υλικό και από δημοσιευμένα άρθρα

- Για τις διαφάνειες θα δανειστούμε από πολλές πηγές
  - Εξόρυξη δεδομένων, Ε. Πιτουρά
  - Data Mining, E. Terzi
  - Data Mining, Aris Anagnostopoulos
  - P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Addison Wesley, 2006
  - J. Han and M. Kamber. Data Mining: Concepts and Techniques, Morgan Kaufmann, 2006
  - Anand Rajaraman and Jeff Ullman Mining Massive Datasets.

Ερωτηματολόγιο

- Σύντομο ερωτηματολόγιο για να δω τι ξέρετε
  - Χρησιμεύει για να πάρω μια ιδέα του τι κενά μπορεί να χρειαστεί να καλύψουμε.
  - Δεν επηρεάζει βαθμό ή κάτι άλλο.