Εξόρυξη γνώσης από δεδομένα

Εργασία εξαμήνου

AI|LS
ARTIFICIAL INTELLIGENCE
& LEARNING SYSTEMS LABORATORY

Στόχος εργασίας

Τα βασικά

- Μας ενδιαφέρει να εξορύξουμε γνώση από δεδομένα, δηλαδή να μάθουμε πράγματα, ποιοτικά και ποσοτικά, που πριν δεν γνωρίζαμε και που δεν είναι αυτονόητα για κάποιο φαινόμενο, συνήθως πολύπλοκο
- Τα τελικα παραδοτέα θα είναι αναφορές
- Ομάδες των δύο ατόμων (καλύτερα) ή ατομική

Διαφορές με Μηχανική Μάθηση

- Ο πρωτεύων στόχος δεν είναι γνωσιακός (object recognition, sentiment analysis, text mining κλπ) όπως στην Μηχανική Μάθηση αλλά η ανάλυση, η εξερεύνηση, η αιτιολόγηση, η εξαγωγή πιθανοτήτων και συμπερασμάτων “κρυμμένων” μέσα στα δεδομένα.
- Στην εξόρυξη χρησιμοποιούμε τους αλγόριθμους της Μηχανικής Μάθησης ως εργαλεία (πάντα με το σωστό πρωτόκολλο) για ανακάλυψη γνώσης, δεν μας ενδιαφέρει η μελέτη των ιδιοτήτων τους per se.

Στόχος εργασίας

Τί δεν είναι η εξόρυξη γνώσης από δεδομένα

- Η μελέτη ενός μόνο μεμονωμένου dataset σε περίπτωση που αυτό δεν είναι ολόκληρη Βάση Δεδομένων όπως στο Big Query με πίνακες που έχουν δεδομένα που αναφέρονται σε πολλές όψεις του προβλήματος.

Τι δεν θέλουμε στην αναφορά

- Στην αναφορά δεν μας ενδιαφέρουν τετριμμένα συμπεράσματα. Για παράδειγμα, μπορούμε να κάνουμε ανάλυση της συσχέτισης μεταξύ μεταβλητών ως εργαλείο, αλλά στην τελική αναφορά δεν μας χρειάζονται πίνακες συσχέτισης "απλά για να δείξουμε" ότι κάναμε ανάλυση των δεδομένων. Μας ενδιαφέρει όμως εάν σας δείχνει πχ κάτι ιδιαίτερο για δύο μεταβλητές ή κάτι που μπορεί να οδηγήσει σε επιπλέον εξερεύνηση.

Ένα παράδειγμα εργασίας με σχολιασμό των αδυναμιών της.
Ο παραδοσιακός ορισμός του data mining ως KDD (Knowledge Discovery from Databases) είναι πια περιοριστικό με τα τεχνικά μέσα που διαθέτουμε.

Μια δεκτή κατεύθυνση για την εργασία της εξόρυξης κάθε ομάδας

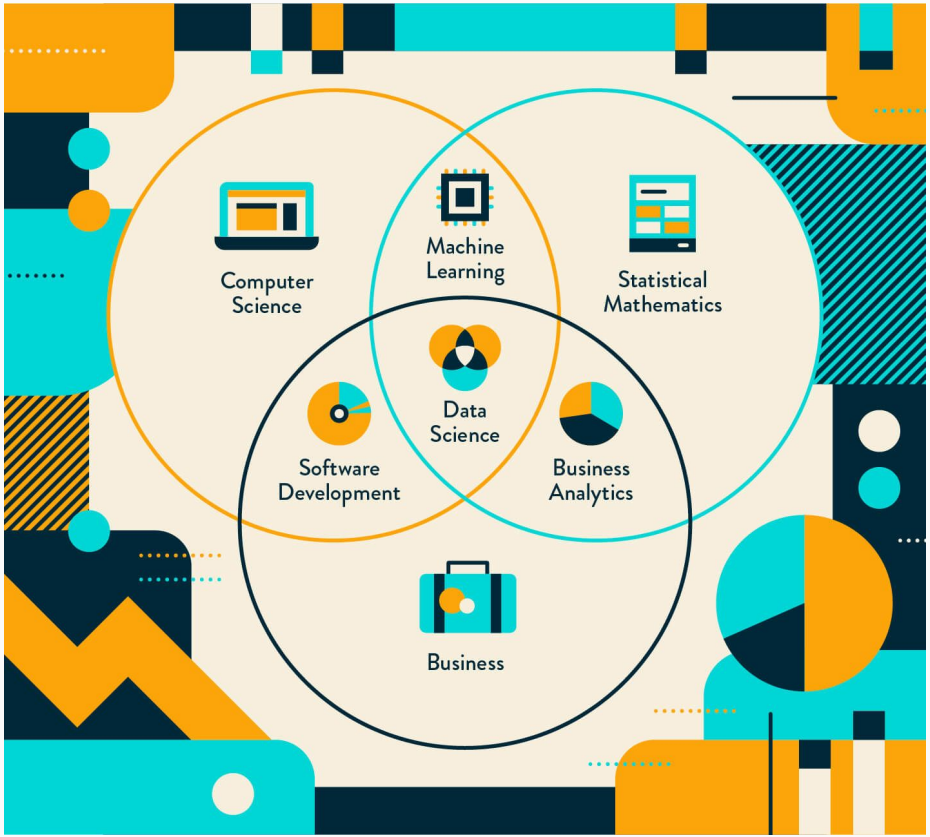
- Το θέμα της εργασίας της κάθε ομάδας δεν πρέπει να ξεκινάει από τα δεδομένα (επιλογή ενός dataset για "ανάλυση") αλλά από την διατύπωση ερωτημάτων για τη μελέτη ενός σύνθετου φαινομένου όπως για παράδειγμα η μελέτη των μετακινήσεων στα αεροδρόμια μιας πόλης συνολικά. Προφανώς θα πρέπει να υπάρχουν δεδομένα για να μπορεί να γίνει η ανάλυση, τα οποίο όμως μπορεί να είναι ετερογενές: datasets βέβαια αλλά και πίνακες, wikipedia, μελέτες (δλδ papers, ακόμα και από τις ανθρωπιστικές επιστήμες)

Δεδομένα και εύρεση θέματος

Α. Κάποια βάση δεδομένων ή μεγάλο dataset είτε από το Azure, το GCP ή το AWS. Η επιλογή ΒΔ ή dataset μπορεί να σας κατευθύνει προς την προσδιορισμό του προβλήματος ανακάλυψης γνώσης που θα είναι το θέμα της εργασίας.

Β. Άλλα dataset και δεδομένα. Μπορείτε να δείτε πολλες συγκεντρωμένες πηγές στο Dataset Sources. Και αυτά μπορούν να σας κατευθύνουν στην περιοχή της μελέτης σας. Για την εξόρυξη γνώσης μας ενδιαφέρει η ΒΔ ή το dataset να είναι σχετικά μεγάλα. Πολλά datasets του Kaggle και του HCI δεν είναι, για αυτό ρωτήστε μας, if in doubt.

Γ. Άλλες πηγές δεδομένων. Πρόκειται για δεδομένα που μπορεί να βρείτε/συλλέξετε αφού έχετε προσδιορίσει το θέμα σας. Μπορούν να είναι από οποιαδήποτε πηγή και κάθε τύπου, δομημένα (πχ άλλη ΒΔ) ή αδόμητα (πχ κείμενα), ενδεχομένως ετερογενή ως προς το βασικό dataset.



Ομάδες και επιλογή θέματος

Δήλωση ομάδας και θέματος: μέχρι Κυριακή 27/12/2021, όχι δεσμευτικά. Δεν υπάρχει λόγος να βιαστείτε, αλλά για να αρχίσετε να ψάχνετε.

- Φόρμα δήλωσης θέματος
- Επιλεγμένα θέματα

Προτεινόμενο template της τελικής αναφοράς

Για ερωτήσεις που αφορούν στην εξαμηνιαία εργαστηριακή άσκηση παρακαλούμε στη σχετική συζήτηση στο forum του μαθήματος.