Διαχείριση Δεδομένων Μεγάλης Κλίμακας
Περικλής Ανδρίτσος

Καταληκτική ημερομηνία υποβολής 09.06.2023 @ 11:59μμ

Εισαγωγή
Η εργασία έχει ως στόχο την ανάλυση μεγάλων δεδομένων και την παρουσίαση των αποτελεσμάτων σας. Η εργασία θα περιλαμβάνει 2 μέρη, το πρώτο μέρος θα είναι συγγραφή μιας αναφοράς 6 σελίδων και ενώ το δεύτερο η δημιουργία κώδικα για την εκτέλεση των πειραμάτων σας. Η ημερομηνία υποβολής της θα είναι η 9η Ιουνίου 2023, η τελευταία εβδομάδα των μαθημάτων.

Γραπτή αναφορά

Η αναφορά θα πρέπει να ακολουθεί τα εξής:

- Να είναι έως 6 σελίδες, χρησιμοποιώντας γραμματοσειρά μεγέθους 11 pt, μονό διάστημα (single space).
- Να συμπεριλαμβάνει τις ακόλουθες υποενότητες:
    - Εισαγωγή
    - Ορισμός προβλήματος και κίνητρο (καλό είναι να συμπεριλάβουμε ένα παράδειγμα χρήσης των αποτελεσμάτων της, π.χ. για ποιόν είναι χρήσιμα ?)
    - Σύντομη περιγραφή του συνόλου δεδομένων που χρησιμοποιήσατε
    - Περιγραφή της μεθόδου ανάλυσης των δεδομένων (←ιδιαίτερη έμφαση σε αυτό)
    - Πειραματικά Αποτελέσματα (←ιδιαίτερη έμφαση σε αυτό)
    - Συζήτηση/Κριτική αποτίμηση αποτελεσμάτων
    - Συμπεράσματα

Η υποενότητα της ανάλυσης των δεδομένων θα πρέπει να περιγράφει τις τεχνικές που χρησιμοποιήσατε και μια εξήγηση γιατί! Πολύ σημαντικό είναι να προσπαθήσετε να πείσετε τον αναγνώστη ότι μια συγκεκριμένη τεχνική που χρησιμοποιείται είναι αυτή που ταιριάζει στο πρόβλημα. Να είστε σαφείς και περιεκτικοί.

Η ενότητα των πειραματικών αποτελεσμάτων θα πρέπει να περιλαμβάνει όλα τα πειράματα που χρησιμοποιήσατε. Συζητήστε τις παραμέτρους και ιδιαίτερα τους χρόνους εκτέλεσης, καθώς και τυχόν μέτρα αξιολόγησης που χρησιμοποιήθηκαν. Συμπεριλάβετε πίνακες/σχήματα όπως κρίνετε απαραίτητο (τα περισσότερα έγγραφα ανάλυσης δεδομένων τα διαθέτουν). Σημειώστε ότι δεν βαθμολογείστε για την «ομορφιά» των γραφημάτων σας, αλλά για το μήνυμα που μεταφέρουν και πόσο ξεκάθαρα περιγράφεται.

Κώδικας

Τα πειράματά σας θα πρέπει να γίνουν στη γλώσσα Python. Μπορείτε να χρησιμοποιήσετε οποιαδήποτε πλατφόρμα υλοποίησης, π.χ. Jupyter-lab, Google Colab, IDLE κλπ. Το βασικό είναι να μπορούν να αναπαραχθούν. Δώστε μεγάλη προσοχή στη χρήση σχολίων στον κώδικά σας.

Σημείωση: αν χρησιμοποιήσετε εξωτερικές πηγές, θα πρέπει να τις αναφέρετε σε σχόλια μέσα στον κώδικα.

Πηγές δεδομένων

Για την εκπόνηση της εργασίας μπορείτε να χρησιμοποιήσετε δεδομένα και από τις εξής ενδεικτικές πηγές:

Google Data set search:
https://datasetsearch.research.google.com/

KDnuggets Datasets:
https://www.kdnuggets.com/datasets/index.html

kaggle Datasets:
https://www.kaggle.com/datasets

144 libraries of datasets:
https://data.world/datasets/library

Τι θα υποβάλλετε

1. Το έγγραφο της τελικής αναφοράς της εργασία σε μορφή pdf
2. Τον κώδικα Python είτε σε αρχείο . ipynb είτε σε αρχείο .py (είτε και στα δύο)

Groups εργασίας

Η εργασία μπορεί να γίνει σε groups των 2 ή 3 φοιτητών. Τα groups θα πρέπει να δηλωθούν στο eClass

Τελική βαθμολογία

Υπενθύμιση πως η τελική βαθμολογία θα γίνει ως εξής

- 60% από την τελική εξέταση
- 40% από την εργασία