Τμήμα Μηχανικών Η/Υ & Πληροφορικής
Πανεπιστήμιο Πατρών

Εξόρυξη Δεδομένων και Αλγόριθμοι Μάθησης

Εργαστηριακή Άσκηση
Εαρινό Εξάμηνο 2021-22

Διδάσκοντες:
Καθηγητής Β. Μεγαλοοικονόμου ,
Αναπληρωτής Καθηγητής Χ. Μακρής

Γλώσσα Υλοποίησης

Ως γλώσσα υλοποίησης της άσκησης ορίζεται η python. Είστε ελεύθεροι να χρησιμοποιήσετε όποια βιβλιοθήκη επιθυμείτε αρκεί να την συμπεριλάβετε στην αναφορά σας.

Ερώτημα 1

Στους φακέλους *demand* και *sources* θα βρείτε αρχεία *csv* που περιέχουν τις ηλεκτρικές ενεργειακές ανάγκες της Πολιτείας της Καλιφόρνια καθώς και τις πηγές από τις οποίες αυτές καλύπτονται για κάθε ημέρα τους έτους από 1/1/2019 μέχρι 31/12/2021 σε χρονική ανάλυση πέντε λεπτών. Από εσάς ζητείται:

Α. Να ενοποιήσετε τα αρχεία και να πραγματοποιήσετε ανάλυση του dataset και γραφική αναπαράσταση αυτής. Πιο συγκεκριμένα, καλείστε να υπολογίστε τα βασικά στατιστικά μεγέθη για τις δοθέντες τιμές, να ανακαλύψετε αν η μορφή των γραφικών παραστάσεων ακολουθεί συγκεκριμένο μοτίβο, να προσπαθήσετε να εντοπίσετε συσχετίσεις που είναι εμφανείς με γυμνό μάτι κοκ.

Β. Να επιχειρήσετε να ομαδοποιήσετε τις τιμές της ζήτησης αλλά και των διαθέσιμων πηγών κάθε ημέρας. Σκοπός σας είναι να εντοπίσετε ημέρες-outliers κατά τις οποίες η ζήτηση ή η παραγωγή δεν είχαν τις αναμενόμενες τιμές. Επιλέξτε κάποιον αλγόριθμο ομαδοποίησης που είναι κατάλληλος για αυτό τον σκοπό.

Γ. Το σημαντικότερο πρόβλημα των ανανεώσιμων πηγών ενέργειας είναι πως δεν είναι διαθέσιμες κατά το δοκούν. Για να καλυφθούν οι επιπλέον ανάγκες οι πάροχοι ηλεκτρικής ενέργειας χρησιμοποιούν κατά κύριο λόγο ορυκτά καύσιμα. Προσπαθήστε να εκπαιδεύσετε έναν παλινδρομητή βασισμένο σε LSTM νευρωνικά δίκτυα ο οποίος να μαντεύει για κάθε στιγμή της ημέρας πόση ενέργεια απαιτείται να

παραχθεί από μη ανανεώσιμες πηγές. Αξιολογήστε το μοντέλο σας με βάση τις γνωστές μετρικές για την παλινδρόμηση αλλά και τις ιδιαιτερότητες του προβλήματος που επιλύετε.

Ερώτημα 2

Σε αυτό το ερώτημα σας δίνεται το αρχείο *amazon.csv* το οποίο περιέχει δύο στήλες. Η πρώτη στήλη περιλαμβάνει το κείμενο από διάφορες κριτικές χρηστών για προϊόντα ενώ η δεύτερη στήλη είναι η βαθμολογία του προϊόντος. Σκοπός σας είναι να προσπαθήσετε να μαντέψετε την πληροφορία της δεύτερης στήλης χρησιμοποιώντας έναν RandomForest. Για να μετασχηματίσετε το σύνολο δεδομένων που σας δόθηκε έτσι ώστε να μπορέσετε να το εισάγετε στο νευρωνικό σας δίκτυο θα πρέπει να μετατρέψετε τα κείμενα σε διανύσματα εκπαιδεύοντας ένα μοντέλο με την τεχνική των Word Embeddings. Μετά τη δημιουργία του τελικού μητρώου, καλείστε να το χωρίσετε σε training-test dataset με αναλογία 80%-20%. Αξιολογήστε την απόδοσή του συστήματός σας με τις μετρικές f1 score, precision και recall.

Παραδοτέα

1. Τα αρχεία κώδικα που υλοποιούν τα ζητούμενα των ασκήσεων.
2. Μια αναφορά σε μορφή pdf η οποία θα πρέπει να περιέχει τα ακόλουθα:
    - Αναλυτική καταγραφή του περιβάλλοντος υλοποίησης (βιβλιοθήκες λογισμικού κτλ.) καθώς και τα βήματα που απαιτούνται για την εγκατάστασή του.
    - Σύντομη περιγραφή της διαδικασία υλοποίησης.
    - Σχολιασμό των τελικών αποτελεσμάτων.

Διαδικαστικά

1. Η άσκηση μπορεί να υλοποιηθεί είτε **ατομικά** είτε σε **ομάδες των δύο**.
2. Η άσκηση μπορεί να υποβληθεί έως και **τρεις ημέρες πριν την ημερομηνία της γραπτής εξέτασης** του μαθήματος στις **23:59**.
3. Η άσκηση θα εξεταστεί προφορικά σε ημερομηνία που θα ανακοινωθεί στο τέλος του εξαμήνου.
4. Η υποβολή της άσκησης πρέπει να γίνει μέσω του eclass του μαθήματος.
5. Η άσκηση μπορεί να αποσταλεί πολλές φορές αλλά θα βαθμολογηθεί μόνο η τελευταία της υποβολή.