Πανεπιστήμιο Ιωαννίνων,
Τμήμα Μηχανικών Πληροφορικής

Εξόρυξη Δεδομένων – ΠΛΕ 059
Χειμερινό Εξάμηνο, 2013

Τελική Εργασία

Στο μάθημα μάθατε μια ποικιλία από τεχνικές για την ανάλυση δεδομένων, όπως την εύρεση συχνών στοιχειοσυνόλων (Frequent Itemset Mining) και κανόνων συσχέτισης (Association Rules), Clustering, κατηγοριοποίηση, κάλυψη, Ranking. Ο στόχος της εργασίας είναι να εφαρμόσετε κάποια από τα εργαλεία που μάθατε στην πράξη.

Για την εργασία θα χρησιμοποιήσετε τα δεδομένα από το Yelp Dataset Challenge. Το Yelp είναι μια σελίδα που συγκεντρώνει κριτικές από χρήστες για διάφορες υπηρεσίες (κυρίως εστιατόρια). Οι χρήστες γράφουν κριτικές, και δίνουν μια αξιολόγηση σε αστέρια, από 1 (κακό) έως 5 (άψογο). Επίσης αξιολογούν τις κριτικές άλλων χρηστών, και μπορούν να κάνουν check-in σε υπηρεσίες. Το dataset περιέχει πληροφορίες για καταστήματα στο Phoenix, Arizona. Ξοδέψτε κάποιο χρόνο στη σελίδα του Yelp για να εξοικειωθείτε με τη λειτουργία του και το είδος του περιεχομένου που έχει. Επίσης διαβάστε προσεκτικά τις οδηγίες στη σελίδα του Yelp Dataset Challenge για να καταλάβετε τι πληροφορία περιέχεται στα δεδομένα.

Για την εργασία θα προτείνετε κάποια ανάλυση που θέλετε να κάνετε στα δεδομένα ώστε να εξάγετε κάποια ενδιαφέρουσα πληροφορία. Το είδος της ανάλυσης που θα κάνετε είναι δικιά σας απόφαση: μπορεί να είναι κάποιας μορφής κανόνες συσχέτισης, κάποιο clustering που πιστεύετε ότι θα δείξει κάτι ενδιαφέρον, κάποιου είδους κατηγοριοποίηση, ή κάποιας μορφής ranking. Ο στόχος είναι να βρείτε κάτι χρήσιμο στα δεδομένα, ή να καταλήξετε ότι κάποια συγκεκριμένη ανάλυση δεν μπορεί να δώσει κάτι ενδιαφέρον. Θα πρέπει να είσαστε ξεκάθαροι στο τι στόχο έχετε και πως θα αξιολογήσετε τα αποτελέσματα. Ως μέρος της εργασίας θα πρέπει να δώσετε και μία αναφορά που θα αναλύετε τα αποτελέσματα που πήρατε.

Η εργασία σας θα έχει τα εξής βήματα:

Βήμα 1: Κατεβάσετε τα δεδομένα από τη σελίδα του Yelp Dataset Challenge, και αποσυμπιέστε τα τοπικά (χρησιμοποιώντας την εντολή tar xvf). Θα πρέπει να πάρετε μια συλλογή από JSON αρχεία. Ξοδέψτε κάποιο χρόνο να καταλάβετε πως δουλεύει το Yelp, και τι πληροφορία περιέχεται στα αρχεία.

Βήμα 2: Τα δεδομένα είναι σε JSON format οπότε θα χρειαστεί να φτιάξετε ένα parser. Φτιάξετε ένα πρόγραμμα που διαβάζει ένα από αυτά τα αρχεία και εξάγει κάποια από τα πεδία. Υπάρχουν πολλές βιβλιοθήκες για την επεξεργασία JSON δεδομένων σε διάφορες γλώσσες (π.χ. κοιτάξτε στο json page). Θα πρέπει να χρησιμοποιήσετε μία από αυτές τις βιβλιοθήκες. Ο στόχος αυτού του βήματος είναι να εξοικειωθείτε με αυτές τις βιβλιοθήκες.

Βήμα 3: Γράψετε μια πρόταση για το είδος της ανάλυσης που θέλετε να κάνετε. Η πρόταση θα πρέπει να έχει τα εξής κομμάτια:

1. Τι θέλετε να τεστάρετε, ή να βρείτε στα δεδομένα και γιατί.
2. Πως σχεδιάζεται να το κάνετε.
3. Πως θα αξιολογήσετε αυτό που κάνατε.

Για την πρόταση η περιγραφή των τριών αυτών σημείων θα πρέπει να είναι επιγραμματική και σε σχετικά υψηλό επίπεδο.

Βήμα 4: Υλοποιήσετε την πρόταση σας και γράψετε μία αναφορά που θα περιγράφει με λεπτομέρεια πλέον τι κάνατε αναφορικά με τα τρία παραπάνω σημεία και πως αντιμετωπίσατε διάφορα προβλήματα. Θα πρέπει επίσης να περιγράψετε τι δείχνουν τα πειράματα.

Θα παραδώσετε τον κώδικα σας και την αναφορά. Θα υπάρξει και εξέταση πάνω στην τελική εργασία.