Πανεπιστήμιο Ιωαννίνων,
Τμήμα Μηχανικών Η/Υ & Πληροφορικής

Εξόρυξη Δεδομένων – ΜΥΕ012/ΠΛΕ 059
Εαρινό Εξάμηνο, 2017

Σειρά Ασκήσεων Σεπτεμβρίου

Η προθεσμία για την τέταρτη σειρά ασκήσεων είναι στις 24 Σεπτεμβρίου στο τέλος της μέρας. Κάνετε turn-in τον κώδικα σας, με οδηγίες για το πώς τρέχει. Η αναφορά θα πρέπει να έχει λεπτομερείς παρατηρήσεις για τα αποτελέσματα σας. Για καθυστερημένες υποβολές ισχύει η πολιτική στην σελίδα του μαθήματος. Λεπτομέρειες για το turn-in στη σελίδα Ασκήσεις του μαθήματος. Θα γίνει προφορική εξέταση την εβδομάδα μετά την παράδοση.

Ερώτηση 1

Υποθέστε ότι σας δίνεται σαν είσοδος ένας πίνακας με *n* γραμμές και *m* στήλες, με 0/1 τιμές. Θέλετε να βρείτε όλα τα *(r,c)-πλακίδια (tiles)* από 1, δηλαδή συνδυασμούς από *r* γραμμές και *c* στήλες ώστε ο υπο-πίνακας με αυτές τις γραμμές και αυτές τις στήλες να έχει μόνο άσσους. Τα πλακίδια μπορεί να είναι επικαλυπτόμενα. Δώσετε ένα αποτελεσματικό αλγόριθμο για το πρόβλημα χρησιμοποιώντας την ιδέα του APriori.

Ερώτηση 2

Αποδείξτε ότι για ένα μη κατευθυνόμενο γράφο η κατανομή σύγκλισης (stationary distribution) ενός τυχαίου περίπατου είναι ανάλογη του βαθμού του κάθε κόμβου. Δηλαδή αν P είναι ο πίνακας μετάβασης του τυχαίου περίπατου, και π η κατανομή σύγκλισης για την οποία ισχύει ότι $\pi = \pi P$, δείξτε ότι για τον κόμβο i, η πιθανότητα π_i είναι ανάλογη του d_i, όπου d_i είναι ο αριθμός των ακμών με άκρο την κορυφή i.

Ερώτηση 3

Στις Ασκήσεις 1 και 2 σας ζητήθηκε να υλοποιήσετε αλγόριθμους για συχνά στοιχειοσύνολα και για συστάσεις, και να τα εφαρμόσετε σε ένα σύνολο από tweets. Τα δεδομένα περιείχαν μόνο χρήστες που είναι followers του Trump. Στην Άσκηση 3, σας δόθηκε ένα νέο σύνολο δεδομένων που περιείχε και Trump και Clinton followers. Εφαρμόστε τον κώδικα σας και για τα δύο προβλήματα σε αυτά τα δεδομένα και αναφέρετε τα αποτελέσματα ακριβώς όπως και στις δύο πρώτες ασκήσεις. Αναφέρετε επίσης τυχόν διαφορές με τα προηγούμενα αποτελέσματα στα δεδομένα που είχαν μόνο Trump followers. Βλέπετε στοιχειοσύνολα στα αποτελέσματα που να είναι χαρακτηριστικά των δύο διαφορετικών γκρουπ από followers? Βελτιώνεται η ακρίβεια των προβλέψεων?

Μπορείτε να διορθώσετε ή να βελτιώσετε τον κώδικα σας από τις Ασκήσεις 1 και 2 αν το κρίνετε απαραίτητο. Αναφέρετε τις αλλαγές που κάνατε, και παραδώσετε τον ενημερωμένο κώδικα.

Ερώτηση 4

Για την ερώτηση αυτή θα εξετάσετε το πρόβλημα της πρόβλεψης του ποιον υποψήφιο ακολουθεί ένας χρήστης, χρησιμοποιώντας το retweet δίκτυο.

Θα χρησιμοποιήσετε τον retweet γράφο που κατασκευάσατε για την Άσκηση 4. Σπάσετε τους κόμβους σε δύο κομμάτια σε ποσοστό 80/20. Θεωρείστε τους κόμβους στο 80% κομμάτι ως επισημασμένους (labeled) και τους

υπόλοιπους ως μη-επισημασμένους (unlabeled). Στην συνέχεια θα χρησιμοποιήσετε ένα αλγόριθμο διάδοσης επισήμανσης (label propagation) για να βρούμε την κατηγορία των μη-επισημασμένων κόμβων. Για να το κάνουμε αυτό θα δώσετε τιμή -1 στους followers του Trump και +1 στους followers της Clinton. Εφαρμόσετε τον αλγόριθμο που περιγράψαμε στην Διάλεξη 11: οι επισημασμένοι κόμβοι θα λειτουργούν ως απορροφητικοί κόμβοι και η τιμή τους δεν αλλάζει; για τους μη-απορροφητικούς κόμβους θα υπολογίσετε μια τιμή με την επαναληπτική διαδικασία που περιγράψαμε στην Διάλεξη 11. Η διαδικασία συγκλίνει όταν δεν έχουμε ουσιαστική αλλαγή στις τιμές. Δώσετε επισήμανση +1 στους κόμβους με θετική τιμή, και -1 στους κόμβους με αρνητική τιμή. Υπολογίσετε την ακρίβεια της πρόβλεψης και δημιουργήστε τον πίνακα σύγχυσης. Συγκρίνετε τα αποτελέσματα με αυτά του classifier που δημιουργήσατε για την Τέταρτη Άσκηση, τον οποίο θα εκπαιδεύσετε στο 80% των επισημασμένων χρηστών.

Στο δεύτερο κομμάτι της άσκησης, θα δημιουργήσετε ένα νέο retweet γράφο, με το ποιος κάνει retweet ποιον, αλλά αυτή τη φορά θα κρατήσετε και τους χρήστες που δεν εμφανίζονται στο αρχείο με τις κλάσεις. Εφαρμόστε και πάλι την επαναληπτική μέθοδο φιλτραρίσματος και κρατήστε τους χρήστες που έχουν βαθμό μεγαλύτερο του 10, και πάρετε την μεγαλύτερη συνεκτική συνιστώσα (στον τελικό γράφο θα πρέπει να έχετε περίπου 37K χρήστες). Εφαρμόστε τον αλγόριθμο διάδοσης που υλοποιήσατε στο νέο γράφο, χρησιμοποιώντας τις επισημάνσεις των κόμβων που είναι στο αρχείο, και υπολογίσετε μια τιμή για τους κόμβους που δεν είναι στο αρχείο. Χρησιμοποιήσετε την τιμή για να κατατάξετε τους κόμβους ως Trump followers, Clinton followers ή Neither followers (π.χ., μπορείτε να χαρακτηρίσετε ως Trump followers όσους έχουν τιμή κάτω από -0.5, Clinton followers όσους έχουν τιμή πάνω από 0.5 και Neither followers τους υπόλοιπους – διαλέξετε τα κατώφλια κατάλληλα, ανάλογα με τις τιμές που βλέπετε στα δεδομένα). Στη συνέχεια, διαλέξετε με τυχαία δειγματοληψία 20 κόμβους από κάθε κατηγορία και εξετάστε τους χειρωνακτικά για να προσδιορίσετε αν η κατηγοριοποίηση τους είναι σωστή και να πάρετε το ground-truth. Για την αξιολόγηση θα πρέπει να κοιτάξετε το προφίλ και τα tweet των χρηστών και να δικαιολογήσετε την απόφαση σας. Υπολογίσετε την ακρίβεια της κατηγοριοποίησης για κάθε κλάση με βάση τα δείγματα που έχετε πάρει.

Παραδώσετε τον κώδικα σας και την αναφορά σας. Στην αναφορά περιγράψετε σε υψηλό επίπεδο την υλοποίηση και σε λεπτομέρεια τα αποτελέσματα. Θα πρέπει επίσης να δώσετε λεπτομέρειες για το πώς κάνατε την χειρωνακτική αξιολόγηση και πως καταλήξατε στο ground truth labeling.