Πανεπιστήμιο Ιωαννίνων,
Τμήμα Μηχανικών Η/Υ & Πληροφορικής

Εξόρυξη Δεδομένων – ΜΥΕ012/ΠΛΕ 059
Εαρινό Εξάμηνο, 2017

Τέταρτη Σειρά Ασκήσεων

Η προθεσμία για την τέταρτη σειρά ασκήσεων είναι στις 21 Ιουνίου 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

Στην τάξη αποδείξαμε ότι για το MAX-COVERAGE πρόβλημα ο Greedy αλγόριθμος που επαναληπτικά διαλέγει το σύνολο που καλύπτει τα περισσότερα στοιχεία που δεν έχουμε ήδη καλύψει έχει προσεγγιστικό λόγο $\left(1-\frac{1}{e}\right)$. Θεωρείστε μια παραλλαγή του MAX-COVERAGE προβλήματος όπου η συλλογή των συνόλων είναι διαμερισμένη σε K κατηγορίες, και κάθε σύνολο ανήκει σε μία από αυτές τις κατηγορίες. Θέλουμε να επιλέξουμε K σύνολα, όπου θα έχουμε ακριβώς ένα σύνολο από κάθε κατηγορία, ώστε να μεγιστοποιήσουμε τον αριθμό των στοιχείων που καλύπτουμε.

Δείξετε ότι σε αυτή την περίπτωση ο προσεγγιστικός λόγος του Greedy αλγόριθμου είναι $\frac{1}{2}$. Δώστε ένα παράδειγμα εισόδου όπου ο Greedy αλγόριθμος επιτυγχάνει αυτόν τον προσεγγιστικό λόγο.

Ερώτηση 2 (Κατηγοριοποίηση)

Για την ερώτηση αυτή θα εξασκηθείτε στην εφαρμογή αλγορίθμων κατηγοριοποίησης (classification).

Θα χρησιμοποιήσετε το αρχείο “clinton_trump_tweets.txt” που χρησιμοποιήσατε και στην Άσκηση 3. Ο στόχος είναι να δημιουργήσετε ένα μοντέλο κατηγοριοποίησης που προβλέπει αν κάποιος χρήστης ακολουθεί τον Trump ή την Clinton.

Η ερώτηση αυτή έχει δύο κομμάτια. Στο πρώτο κομμάτι, θα χρησιμοποιήσετε τα δεδομένα (μαζί με τις κλάσεις στο αρχείο “clinton_trump_user_classes.txt”) και για εκμάθηση και τεστάρισμα. Επεξεργαστείτε τα δεδομένα και αφαιρέσετε τους χρήστες που έχουν λιγότερα από 10 tweets (οποιασδήποτε μορφής). Για τους χρήστες που έμειναν, χρησιμοποιήσετε την πληροφορία που έχετε για να εξάγετε χαρακτηριστικά τα οποία θεωρείται κατάλληλα για την κατηγοριοποίηση. Χρησιμοποιήσετε την φαντασία σας για την εξαγωγή των χαρακτηριστικών από την πληροφορία που έχετε συνολικά για τους χρήστες (μπορείτε να χρησιμοποιήσετε τα εργαλεία που είδαμε στα φροντιστήρια). Δοκιμάσετε τέσσερεις κατηγοριοποιητές που μάθαμε στο μάθημα (Decision Trees, SVM, Logistic Regression, k-NN). Χρησιμοποιήσετε 5-fold cross validation για να αξιολογήσετε τους κατηγοριοποιητές. Στην αναφορά σας περιγράψετε τα χαρακτηριστικά που χρησιμοποιήσατε, και αναφέρετε την ακρίβεια των classifiers στο 5-fold cross validation.

Στο δεύτερο κομμάτι της άσκησης θα χρησιμοποιήσετε τον κατηγοριοποιητή που φτιάξατε στο πρώτο κομμάτι για να κάνετε την ίδια πρόβλεψη για ένα νέο σύνολο από χρήστες που δεν είναι στα δεδομένα που σας δόθηκαν. Για το στόχο αυτό δημιουργήθηκε ένας διαγωνισμός στο Kaggle για το μάθημα (εδώ είναι ο σύνδεσμος για τον διαγωνισμό). Δημιουργήστε ένα account με το email του πανεπιστημίου. Θα σας δοθεί πρόσβαση στον διαγωνισμό του μαθήματος και θα μπορέστε να καταθέσετε μια λύση για τον διαγωνισμό. Υπάρχει μία κατάταξη στο οποίο μπορείτε να δείτε την θέση σας σε σχέση με άλλες λύσεις. Μια καλή θέση θα ενισχύσει τον βαθμό σας. Το σημαντικό είναι να πετύχετε ένα καλό σκορ στην ακρίβεια του classifier σε σχέση με τις υπόλοιπες λύσεις. Μπορείτε να χρησιμοποιήσετε όποιο αλγόριθμο θέλετε. Προσθέσετε στην αναφορά σας τα αποτελέσματα που είχατε στο Kaggle.

Στην αναφορά αναφέρετε επίσης και πειράματα που κάνατε και δεν δούλεψαν, ή πως βελτιώσατε μια λύση που δεν απέδιδε καλά (περιλάβετε και αποτελέσματα από τα πειράματα).

Ερώτηση 3 (Ανάλυση Δικτύων)

Για την ερώτηση αυτή θα εξασκηθείτε στην εφαρμογή αλγορίθμων ανάλυσης δικτύων.

Θα χρησιμοποιήσετε και πάλι το αρχείο “clinton_trump_tweets.txt” που χρησιμοποιήσατε στην προηγούμενη ερώτηση. Χρησιμοποιώντας αυτό το αρχείο θα δημιουργήσετε ένα δίκτυο με το ποιος κάνει retweet ποιον. Δημιουργήστε τον γράφο προσθέτοντας μια ακμή (xxx,yyy) αν ο χρήστης με screen name xxx έχει κάνει retweet κάποιο μήνυμα από τον χρήστη με handle @yyy. Για παράδειγμα, η γραμμή

```
Saint   Saint2205       1537088066      70      133                     Fri Oct 28 20:27:12
EEST 2016   792055126408048640      en      null    27      0       RT @greeneyes0084:
Wikileaks Email: Hillary Campaign Struggles to Reach F**king Dumb Young People
https://t.co/S6QZhY9rBH via @realalexjo
```

στο αρχείο θα έχει ως αποτέλεσμα την δημιουργία της ακμής (saint2205,greeneyes0084) στον γράφο.

Δημιουργήστε αυτό τον γράφο και στη συνέχεια αφαιρέστε τους κόμβους που δεν είναι στο αρχείο (δηλαδή, τους κόμβους που δεν έχουν κάνει αυτοί κάποιο tweet). Επίσης αφαιρέσετε επαναληπτικά όλους τους κόμβους που έχουν βαθμό μικρότερο από 10. Πάρετε την μεγαλύτερη συνεκτική συνιστώσα αυτού του γράφου. Αυτός είναι ο γράφος με τον οποίο θα δουλέψετε.

Στόχος μας είναι να μελετήσουμε αυτό το γράφο. Πρώτα υπολογίσετε το ποσοστό των Trump και Clinton followers στον γράφο. Μετά, τρέξετε τους αλγορίθμους PageRank και HITS και αναφέρετε τους πρώτους 10 κόμβους για κάθε αλγόριθμο. Μελετήστε τα προφίλ των κόμβων, και προσπαθήσετε να εξηγήσετε γιατί αυτοί οι κόμβοι μπορεί να είναι σημαντικοί στο δίκτυο.

Στη συνέχεια θα προσπαθήσουμε να ξεχωρίσουμε τους followers του Trump και της Clinton χρησιμοποιώντας αλγορίθμους για community detection. Θα υλοποιήσετε τον αλγόριθμο των Girvan-Newman όπου επαναληπτικά αφαιρούμε την ακμή με το μεγαλύτερο betweenness centrality για να βρείτε δύο κοινότητες. Τρέξετε και υπολογίσετε τον χρόνο που παίρνει μια επανάληψη του αλγορίθμου. Πόσο χρόνο θα πάρει αν χρειαστεί να αφαιρέσουμε 100, 500, ή 1000 ακμές? Τροποποιήσετε τον αλγόριθμο ώστε να αφαιρεί ακμές σε batches μεγέθους K (τις K με το μεγαλύτερο betweenness), και δοκιμάσετε τον αλγόριθμο για διαφορετικά K (20, 50, 100, 250, 500). Αναφέρετε τον χρόνο εκτέλεσης, το confusion matrix και την επιτυχία με την οποία ο αλγόριθμος ξεχωρίζει τους followers του Trump και Clinton.

Τέλος τρέξετε τους αλγορίθμους PageRank και HITS και για τους υπογράφους που δημιουργήσατε και αναφέρετε και πάλι τους πρώτους 10 κόμβους για κάθε υπογράφο. Σχολιάσετε τους κόμβους και συγκρίνετε με τα προηγούμενα αποτελέσματα.

Για την άσκηση θα σας βοηθήσει να χρησιμοποιήσετε τη βιβλιοθήκη networkx για να υπολογίσετε τους υπογράφους των συνεκτικών συνιστωσών, το betweeness centrality, και τα PageRank και HITS.

Ερώτηση 4

Ο στόχος αυτής τη ερώτησης είναι να ερευνήσετε αν υπάρχει πολιτισμικό χάσμα μεταξύ των followers του Trump και της Clinton και να προσπαθήσετε να το χαρακτηρίσετε. Ξέρουμε ότι τα δύο group διαφέρουν μεταξύ τους ως προς τις πολιτικές τους πεποιθήσεις και υπάρχουν διάφορα hashtags τα οποία είναι πολύ χαρακτηριστικά για να ξεχωρίσουν τους μεν και τους δε (π.χ., το #maga για τους followers του Trump και το #imwithher για τους followers της Clinton). Πέραν όμως της πολιτικής μπορούμε από τα tweets να καταλάβουμε το πολιτιστικό προφίλ του μέσου οπαδού? Π.χ., είναι οι οπαδοί του Trump πιο πιθανό να κάνουν tweet για Nascar ή baseball, και οι οπαδοί της Clinton για θέατρο και τη ζωή στην πόλη?

Αυτή είναι μια ανοιχτή ερώτηση στην οποία καλείστε να αυτοσχεδιάσετε. Μπορείτε να χρησιμοποιήσετε οποιαδήποτε από τις τεχνικές που μάθατε στο μάθημα, και να επεξεργαστείτε και να τροποποιήσετε τα δεδομένα όπως πιστεύετε ότι είναι καλύτερο. Προτείνετε και ορίσετε την προσέγγιση σας για το πρόβλημα, υλοποιήσετε την, και αναφέρετε τα αποτελέσματα. Στη βαθμολόγηση της άσκησης θα έχει μεγάλο βάρος ο τρόπος που θα προσεγγίσετε και θα διατυπώσετε το πρόβλημα.