Πανεπιστήμιο Ιωαννίνων,
Τμήμα Μηχανικών Η/Υ & Πληροφορικής

Εξόρυξη Δεδομένων – ΜΥΕ012/ΠΛΕ 059
Χειμερινό Εξάμηνο, 2022

Τρίτη Σειρά Ασκήσεων

Η προθεσμία για την τρίτη σειρά ασκήσεων είναι την Κυριακή 19 Φεβρουαρίου 11:55 μ.μ. Παραδώστε Notebooks με τον κώδικα και τις αναφορές. Για την Ερώτηση 1 μπορείτε να παραδώσετε και pdf με την απόδειξη, ή φωτογραφίες από χειρόγραφα. Για καθυστερημένες υποβολές ισχύει η πολιτική στην σελίδα του μαθήματος. Η παράδοση θα γίνει μέσω του ecourse. Λεπτομέρειες στη σελίδα Ασκήσεις του μαθήματος. Η άσκηση είναι **ατομική**.

Ερώτηση 1

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 $\pi = \pi \cdot P$, 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 2

Στην άσκηση αυτή θα εξασκηθείτε στην εφαρμογή αλγορίθμων κατηγοριοποίησης. Θα χρησιμοποιήσετε τα δεδομένα που χρησιμοποιήσατε και στην Τρίτη Ερώτηση της Δεύτερης Άσκησης. Θα πάρετε πάλι 951 επιχειρήσεις-εστιατόρια, που ανήκουν σε τρεις κατηγορίες: Japanese, Italian, Burgers. Η λίστα με αυτά τα business ids μαζί με την κατηγορία του εστιατορίου σας δίνεται στο αρχείο philly_restaurants_categories.csv στη σελίδα Ασκήσεων του μαθήματος. Θα τραβήξετε και πάλι τα κείμενα από τα reviews γι αυτά τα εστιατόρια και θα τα συνενώσετε σε ένα μεγάλο String. Ο στόχος είναι να χρησιμοποιήσετε το κείμενο για να κατηγοριοποιήσετε τα εστιατόρια στις τρεις κατηγορίες.

Η Ερώτηση έχει δύο βήματα:

1. Στο πρώτο βήμα θα πάρετε την tf-idf αναπαράσταση των reviews (όπως στην Δεύτερη Άσκηση), και θα πειραματιστείτε με τρεις classifiers: Logistic Regression, SVM, και K-NN. Για την αξιολόγηση θα χρησιμοποιήσετε 5-fold cross validation. Θα κάνετε shuffle τα δεδομένα και θα χρησιμοποιήσετε την μέθοδο KFold για να πάρετε τα 5 train-test υποσύνολα (ή μπορείτε να κάνετε μόνοι σας το σπάσιμο). Σε κάθε fold, θα δημιουργείτε ένα διαφορετικό tf-idf vectorizer με τα train δεδομένα, θα κάνετε train τους classifiers, και θα τους τεστάρετε στα test δεδομένα.
   Αναφέρετε το μέσο confusion matrix από τα 5 folds, και τις μέσες τιμές για τις μετρικές accuracy, precision, recall και F1-measure (για τις τρεις τελευταίες ανά κλάση). Μπορείτε να πειραματιστείτε με διάφορες παραλλαγές του tf-idf vectorizer (π.χ., συγκεκριμένο αριθμό από features, κλπ).
   Για τον Logistic Regression classifier στο τελευταίο fold, βρείτε τις 20 λέξεις που ο classifier δίνει το μεγαλύτερο θετικό βάρος και τις 20 λέξεις με το μικρότερο αρνητικό βάρος. Σχολιάστε τις λέξεις που είναι σημαντικές για την κατηγοριοποίηση.

2. Στο δεύτερο βήμα θα χρησιμοποιήσετε τα ίδια δεδομένα όπως και στο Βήμα 1, αλλά θα εξάγετε τα features χρησιμοποιώντας embeddings. Θα εκπαιδεύσετε τρία διαφορετικά embeddings: Ένα CBOW model, ένα Skipgram model και ένα Doc2Vec model. Για τα word embeddings η αναπαράσταση του κειμένου θα είναι η μέση τιμή των embeddings των λέξεων, όπως δείξαμε στο φροντιστήριο. Κάνετε την ίδια αξιολόγηση όπως στο Βήμα 1. Δηλαδή θα κάνετε k-fold cross validation με k=5, και θα χρησιμοποιήσετε ακριβώς τα ίδια folds όπως και πριν, και θα μετρήσετε τις ίδιες μετρικές. Για καθένα από τα training sets θα εκπαιδεύσετε ένα διαφορετικό embedding. Συγκρίνετε τα αποτελέσματα σε αυτή την προσέγγιση με αυτά με την tf-idf αναπαράσταση.
   Bonus: Χρησιμοποιήσετε τα word embeddings του Google για να εκπαιδεύσετε ένα classifier. Κάνετε την ίδια αξιολόγηση όπως και στα προηγούμενα βήματα και συγκρίνετε με τις άλλες προσεγγίσεις.

Παράδοση: Παραδώστε ένα notebook με τους υπολογισμούς σας, τα αποτελέσματα, και το κείμενο του σχολιασμού. Στο notebook θα πρέπει να είναι σαφή τα διαφορετικά βήματα της άσκησης.

Ερώτηση 3

Σε αυτή την ερώτηση θα χρησιμοποιήσετε το κοινωνικό δίκτυο μεταξύ των χρηστών του Yelp για να προβλέψετε τα ratings τους για νέες επιχειρήσεις. Για την πρόβλεψη θα υλοποιήσετε τον αλγόριθμο για value propagation τον οποίο περιγράψαμε στην τάξη.

Για την υλοποίηση σας θα ακολουθήσετε τα εξής βήματα.

Βήμα 1: Θα χρησιμοποιήσετε παρόμοια δεδομένα με αυτά που δημιουργήσατε για την Δεύτερη Σειρά Ασκήσεων, για τα συστήματα συστάσεων. Η διαφορά είναι ότι θα κάνετε πιο επιθετικό pruning. Συγκεκριμένα, στα τελικά σας δεδομένα θα έχετε ένα σύνολο από χρήστες U, και ένα σύνολο από επιχειρήσεις B, όπου ο κάθε χρήστης στο U θα έχει **τουλάχιστον 30 reviews** σε επιχειρήσεις στο B, και η κάθε επιχείρηση στο B θα έχει **τουλάχιστον 50 reviews** από χρήστες στο U. Για διευκόλυνση σας δίνονται τα δεδομένα χρήστης-επιχείρηση-rating για την πόλη της Φιλαδέλφειας τα οποία θα πρέπει να κλαδέψετε στο αρχείο philly_users_businesses_stars.csv στη σελίδα των Ασκήσεων.

Δημιουργείστε ένα γράφημα με κορυφές τους χρήστες στο σύνολο U και ακμές τις φιλίες μεταξύ των χρηστών, τις οποίες θα πάρετε από το αρχείο yelp_academic_dataset_user.json . Από αυτό το γράφημα κρατήστε τη μεγαλύτερη συνεκτική συνιστώσα (θα πρέπει να είναι ίδια με όλο το γράφημα). Αυτή θα ορίσει το γράφημα G με το οποίο θα δουλέψετε. Μετά από αυτή την διαδικασία θα πρέπει να μείνετε με 1504 χρήστες και 887 επιχειρήσεις.

Βήμα 2: Δημιουργείστε τα training και test δεδομένα. Από τους χρήστες διαλέξτε τυχαία 100 χρήστες. Από την ένωση όλων των επιχειρήσεων επιλέξτε τυχαία 100 επιχειρήσεις. Οι τριάδες χρήστης-επιχείρηση-rating με τους χρήστες και τις επιχειρήσεις που επιλέξατε θα είναι τα test δεδομένα, D_{test}. Οι υπόλοιπες τριάδες θα είναι τα train δεδομένα D_{train}. Υλοποιήστε την επιλογή των δεδομένων αλλά για τα επόμενα βήματα χρησιμοποιήστε τα δεδομένα από τα αρχεία train_data.csv και test_data.csv που σας δίνονται στη σελίδα των Ασκήσεων.

Βήμα 3: Ο στόχος είναι να προβλέψουμε τα ratings στο D_{test} πραγματοποιώντας διάχυση τιμών (value propagation) στο γράφημα G. Έστω B_{test} το σύνολο των 100 επιχειρήσεων που επιλέξατε στο Βήμα 2, και εμφανίζονται στο D_{test}. Θα πραγματοποιήσετε την διαδικασία του value propagation για κάθε επιχείρηση $b \in B_{test}$. Για κάθε τριάδα χρήστη-επιχείρηση-rating (u, b, r) που εμφανίζεται στο D_{train} ο κόμβος u στο γράφημα θα έχει σταθερή τιμή r (γίνεται απορροφητικός), και για κάθε άλλο χρήστη-κόμβο v (μη απορροφητικός), θα υπολογίσετε την τιμή $R(v, b)$ χρησιμοποιώντας τη διαδικασία του value propagation που περιγράψαμε στην τάξη. Για κάθε (μη απορροφητικό) κόμβο v, θέτουμε

$$R(v, b) = \frac{1}{|N_v|} \sum_{u \in N_v} R(u, b)$$

όπου N_v είναι οι γείτονες του κόμβου v στο γράφημα G. Ο υπολογισμός αυτός γίνεται επαναληπτικά μέχρι να συγκλίνει (δηλαδή η μέγιστη διαφορά της νέας τιμής με την παλιά να είναι κάτω από ένα threshold, π.χ. 10^{-6}). Για ένα ζευγάρι χρήστη-επιχείρηση (x, b) η πρόβλεψη σας θα είναι η τιμή $R(x, b)$. Υπολογίσετε το Root Mean Square Error (RMSE) για αυτή τη μέθοδο.

Βήμα 4: Τρέξτε τους αλγορίθμους UCF, ICF, UA, IA που υλοποιήσατε στην Δεύτερη Άσκηση για αυτό το dataset και συγκρίνετε το Root Mean Square Error (RMSE) με τη μέθοδο Propagation. Παρουσιάστε τα αποτελέσματα σας και γράψετε τις παρατηρήσεις σας.

Ερώτηση 4

Ο στόχος αυτής της ερώτησης είναι να κατασκευάσετε ένα λεξικό με θετικές και αρνητικές λέξεις (ή εκφράσεις που αποτελούνται από πάνω από μία λέξη) που σχολιάζουν εστιατόρια. Για παράδειγμα το θετικό λεξικό σας μπορεί να περιέχει τη λέξη excellent και την έκφραση low prices, και το αρνητικό τη λέξη poor και την έκφραση long waiting. Θα χρησιμοποιήσετε το κείμενο των κριτικών για εστιατόρια γι αυτό το σκοπό. Θα πρέπει να σχεδιάσετε ένα αλγόριθμο και να τον υλοποιήσετε. Παραδώστε τον κώδικα σας και τα λεξικά που δημιουργήσατε. Θα πρέπει να αξιολογήσετε τα λεξικά που φτιάξατε και να αναφέρετε κάποιες μετρικές για την αξιολόγηση. Η ερώτηση είναι «ανοιχτή» οπότε έχετε μεγάλο περιθώριο για το πως θα σχεδιάσετε τον αλγόριθμο σας και πως θα τον αξιολογήσετε (στην αξιολόγηση μπορεί να χρειαστεί να κάνετε και χειρωνακτική αξιολόγηση).