

Τρίτη Σειρά Ασκήσεων

Αυτό είναι το δεύτερο μέρος της τρίτης σειράς ασκήσεων. Η ημερομηνία παράδοσης της άσκησης είναι **24 Ιανουαρίου**. Μπορείτε να παραδώσετε είτε ηλεκτρονικά, είτε στο γραφείο. Η ηλεκτρονική παράδοση των ασκήσεων θα γίνει είτε μέσω του turn-in, είτε μέσω email. Λεπτομέρειες για το turn-in στη σελίδα του μαθήματος.

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

Για την ερώτηση αυτή θα υλοποιήσετε τον HITS αλγόριθμο πάνω σε ένα διμερές γράφημα. Στη σελίδα Υλικό του μαθήματος, σας δίνεται ένα αρχείο με ζεύγη (χωρισμένα με tab) από user ids και venue names, όπου ο χρήστης με το συγκεκριμένο id έχει αφήσει ένα tip στο venue στο FourSquare. Θεωρήστε τα venues ως authorities και τους χρήστες ως hubs και χρησιμοποιήστε τον HITS αλγόριθμο για να υπολογίσετε τα authority και hub βάρη. (Προσοχή: Θα υπολογίσετε μόνο το authority βάρος για τα venues, και μόνο το hub βάρος για τους χρήστες). Ταξινομήστε τα venues σε φθίνουσα σειρά με βάση το authority βάρος.

Παραδώσετε τον κώδικα σας, και την ταξινομημένη λίστα με τα venues. Εξετάστε τα πρώτα 20 venues και προσπαθήστε να εξηγήσετε γιατί βγήκαν στην κορυφή.

Ερώτηση 3

Σας δίνεται ένα κείμενο D το οποίο αποτελείται από n προτάσεις. Το κείμενο αναφέρεται σε d διαφορετικές έννοιες $c_1, c_2, \dots, c_d$. Οι έννοιες αυτές αντιστοιχούν σε φράσεις ή λέξεις που εμφανίζονται σαν λήμματα στο Wikipedia. Η αναγνώριση των εννοιών μέσα στο κείμενο έχει γίνει ήδη με κάποιο τρόπο, και ξέρουμε για κάθε πρόταση ποιες έννοιες εμφανίζονται σε αυτή. Μία έννοια μπορεί να εμφανίζεται σε πολλές προτάσεις.

Θέλουμε να δημιουργήσουμε μια περίληψη του κειμένου που να αποτελείται από K προτάσεις του κειμένου, έτσι ώστε στην περίληψη να εμφανίζονται (τουλάχιστον μία φορά) όσο γίνεται περισσότερες από τις d έννοιες του κειμένου.

- Δείξτε ότι υπάρχει ένας greedy αλγόριθμος για το πρόβλημα που έχει σταθερό λόγο προσέγγισης ως προς τον βέλτιστο αλγόριθμο που μεγιστοποιεί τον αριθμό των εννοιών που εμφανίζονται στην περίληψη.
- Τι γίνεται αν θέλουμε να καλύψουμε και τις d έννοιες του κειμένου?