

MEM704 - Μηχανική Μάθηση

1η Εργαστηριακή Άσκηση

Ο αλγόριθμος PageRank

Παράδοση: 10/03/2022, 18:00,
Εξέταση: 11/3/2022 στο Εργαστήριο

Στη άσκηση αυτή θα υλοποιήσουμε τον αλγόριθμο PageRank για την εκτίμηση της σημαντικότητας ιστοσελίδων στο διαδίκτυο. Στην κέντρο του αλγορίθμου βρίσκεται η μέθοδος των δυνάμεων για την εύρεση της κυρίαρχης ιδιοτιμής ενός κατάλληλου πίνακα.

1 Μέθοδος των δυνάμεων

Έστω $A \in \mathbb{R}^{n,n}$. Η μέθοδος των δυνάμεων για την προσέγγιση της κυρίαρχης ιδιοτιμής λ και του αντίστοιχου κυρίαρχου ιδιοδιανύσματος x , δίνεται από τον παρακάτω ψευδο-αλγόριθμο

Μέθοδος των δυνάμεων για $Ax = \lambda x$

Διαλέγουμε τυχαίο $x_0 \in \mathbb{R}^n$, $x_0 = x_0 / \|x_0\|_1$

Διαλέγουμε k_{max} - μέγιστο αριθμό επαναλήψεων

Διαλέγουμε $\epsilon \sim 10^{-6}$ - σφάλμα

$k = 0$, $d_k = 1$

While $d_k > \epsilon$ and $k < k_{max}$ do

$x_k = Ax_{k-1}$

$x_k = x_k / \|x_k\|_1$

$d_k = \|x_k - x_{k-1}\|_1$

$\lambda = \frac{x_k^T Ax_k}{x_k^T x_k}$

Να γραφτεί ένας κώδικας Python που να υλοποιεί τον παραπάνω αλγόριθμο. Μπορείται να χρησιμοποιηθεί την βιβλιοθήκη Numpy και ειδικότερα το πακέτο linalg. Για την επιλογή του x_0 μπορείται να χρησιμοποιηθεί το πακέτο random της Numpy. Δοκιμάστε τον κώδικά σας με ένα απλό ($n \geq 3$) πίνακα για τον οποίο ξέρετε όλες τις ιδιοτιμές και τα αντίστοιχα ιδιοδιανύσματα του. Επαληθεύστε αριθμητικά την θεωρητική ταχύτητα σύγκλισης της μεθόδου συγκρίνοντας το λόγο $|\lambda_2/\lambda_1|$ με τους διαδοχικούς λόγους d_{k+1}/d_k . Στη συνέχεια, δοκιμάστε τον κώδικά σας για τον τριδιαγώνιο πίνακα $T = triad[-1, 2, -1]$ (2 στη κύρια διαγώνιο και -1 στη 1η υπερδιαγώνιο και 1η υποδιαγώνιο), για διάφορες τιμές του n . Ο T είναι θετικά ορισμένος και οι ιδιοτιμές του είναι $\lambda_k = 2 - 2 \cos\left(\frac{k\pi}{n+1}\right)$ $k = 1, \dots, n$.

2 Ο αλγόριθμος PageRank

Ο αλγόριθμος αυτός αποτελεί τον πυρήνα της μηχανής αναζήτησης Google και δίνει μια εκτίμηση για το πόσο *σημαντική* είναι μια σελίδα του διαδικτύου. Προτάθηκε από τους S. Brin και L. Page το 1998, στην εργασία με τίτλο : *The anatomy of a large-scale hypertextual Web search engine*, Computer Networks and ISDN Systems. 30 (1–7): 107–117. Η μέθοδος βασίζεται στη εύρεση του κυρίαρχου ιδιοδιανύσματος ενός πίνακα τύπου Markov. Οι πίνακες Markov είναι τετραγωνικοί, τα στοιχεία τους εκφράζουν πιθανότητες και το άθροισμα των στοιχείων κάθε στήλης είναι 1. Τα χαρακτηριστικά αυτά έχουν σαν αποτέλεσμα το 1 να είναι η μεγαλύτερη ιδιοτιμή με πολλαπλότητα 1 και τα στοιχεία του αντίστοιχου ιδιοδιανύσματος αθροίζουν σε 1. Οι υπόλοιπες ιδιοτιμές είναι σε απόλυτη τιμή μικρότερες του 1.

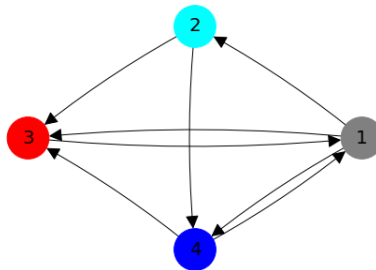
Έστω ότι έχουμε N ιστοσελίδες μεταξύ των οποίων υπάρχουν πολλαπλοί σύνδεσμοι. Ο πίνακας Markov της μηχανής αναζήτησης έχει την μορφή

$$M = d \cdot A + \frac{1-d}{N} B, \quad B = \begin{bmatrix} 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}$$

όπου $d \sim 0.85$, N ο αριθμός ιστοσελίδων και A ο πίνακας με στοιχεία a_{ij}

$$a_{ij} = \begin{cases} \frac{1}{L(j)} & \text{αν υπάρχει σύνδεσμος από την σελίδα } j \text{ στη σελίδα } i \\ 0 & \text{διαφορετικά} \end{cases}$$

όπου $L(j)$ είναι όλοι οι εξερχόμενοι σύνδεσμοι από την σελίδα j . Για παράδειγμα, θεωρούμε τον κατευθυνόμενο γράφο(Directed Graph) της εικόνας, κάθε κόμβος του οποίου αντιπροσωπεύει μια ιστοσελίδα με τους αντίστοιχους συνδέσμους.



και $L(1) = 3$, $L(2) = 2$, $L(3) = 1$, $L(4) = 2$. Ο πίνακας A σε αυτή την περίπτωση είναι

$$A = \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix}$$

Εφαρμόζοντας την μέθοδο των δυνάμεων στον πίνακα M έχουμε την εξής προσέγγιση του κυρίαρχου ιδιοδιανύσματος και της ιδιοτιμής

$$x \approx (0.387, 0.129, 0.29, 0.193), \quad \lambda \approx 0.9999999999656655, \quad \sum_{i=1}^4 x_i = 1$$

Να γραφτεί ένας κώδικας Python ο οποίος θα διαβάσει τα στοιχεία του γράφου, θα φτιάχνει τον πίνακα M και θα υπολογίζει με την μέθοδο των δυνάμεων το κυρίαρχο ιδιοδιάνυσμα και την αντίστοιχη ιδιοτιμή. Το αρχικό διάνυσμα το διαλέγουμε ως $x_{0,i} = \frac{1}{N}$, $i = 1, \dots, N$.

Η κάθε γραμμή του αρχείου(graph0.txt) που περιγράφει τον γράφο είναι της μορφής: $n_i \ n_j$ που σημαίνει ότι υπάρχει ένας σύνδεσμος από τον κόμβο n_i στο κόμβο n_j , π.χ το αρχείο που περιγράφει τον παραπάνω γράφο είναι

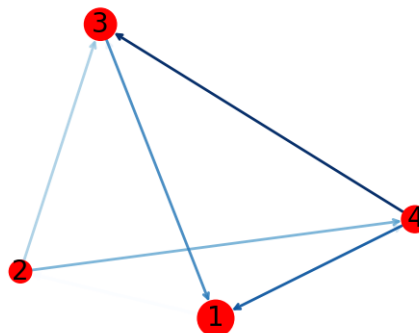
```
1 2
1 3
1 4
2 3
2 4
3 1
4 1
4 3
```

Δοκιμάστε τον κώδικα σας με αυτόν το γράφο και αυτούς των αρχείων graph1.txt, graph2.txt και βεβαιωθείτε ότι δουλεύει σωστά ελέγχοντας ότι για το κυρίαρχο ιδιοζεύγος ισχύουν: $\lambda \approx 1$, $\sum_{i=1}^N |x_i| \approx 1$. Το προγράμμα σας στο τέλος θα τυπώνει τους κόμβους με φθίνουσα σειρά ως προς την σημαντικότητά τους.

3 Η βιβλιοθήκη NetworkX της Python

Στη Python υπάρχει η βιβλιοθήκη **NetworkX** με την οποία μπορούμε με μεγάλη ευκολία να κατασκευάσουμε, χειριστούμε και επεξεργαστούμε γράφους. Στη βιβλιοθήκη, υπάρχουν επίσης και διάφοροι αλγόριθμοι μεταξύ των οποίων και ο PageRank. Εγκαταστήστε πρώτα την βιβλιοθήκη στο περιβάλλον Python που δουλεύετε. Στη συνέχεια φτιάξτε ένα κατευθυνόμενο γράφο(Directed Graph) G , καλέστε τον αλγόριθμο PageRank: `networkx.pagerank(G, d)` και συγκρίνετε με τα δικά σας αποτελέσματα.

Με την βιβλιοθήκη αυτή μπορούμε επίσης να κάνουμε εύκολα γραφική αναπαράσταση του γράφου:



Το μέγεθος των κόμβων του γράφου στη εικόνα είναι ανάλογο της σημαντικότητάς του.

4 Παράδοση - Εξέταση

Για κάθε μέρος της άσκησης να φτιάξετε διαφορετικό κώδικα Python με όνομα π.χ: $\{math, tem, ph\}XXXX_Lab1\{a, b, c\}.py$ όπου $XXXX$ είναι ο αριθμός μητρώου σας. Επίσης στις πρώτες γραμμές του κάθε προγράμματος θα υπάρχουν σαν σχόλιο τα στοιχεία σας: όνομα, επώνυμο και ΑΜ. Θα στείλετε τους κώδικες (ως συνημμένα αρχεία) με email από τον **ιδρυματικό σας λογαριασμό** στη διεύθυνση *mem704labs@gmail.com* το αργότερο μέχρι 18:00, Πέμπτη 10 Μαρτίου. Εκπρόθεσμες ασκήσεις δεν θα βαθμολογούνται. Εργαστείται ατομικά. Κώδικες που είναι προϊόν αντιγραφής θα μηδενίζονται. Πληροφορίες για την άσκηση θα δωθούν στο εργαστήριο της Παρασκευής 4ης Μαρτίου .