ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΑΣ
ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ
ΕΞΕΤΑΣΕΙΣ ΦΕΒΡΟΥΑΡΙΟΥ 2021
Διδάσκων: Μακρής Χρήστος
ΔΙΑΡΚΕΙΑ 1:30
(ΟΜΑΔΑ 8)

ΘΕΜΑ 1 (*Μονάδες 3.5*)
α.Λόγω του μεγάλου όγκου δεδομένων που έχουν να χειριστούν οι μηχανές ψαξίματος στον παγκόσμιο ιστό, συχνά διατηρούν δύο ανεστραμμένα αρχεία, ένα που περιέχει πληροφορία για όλα τα κείμενα και αποθηκεύεται στο δίσκο και ένα μικρότερου μεγέθους που περιέχει ένα υποσύνολο της συνολικής πληροφορίας και αποθηκεύεται στην κύρια μνήμη. Τα δύο αυτά ανεστραμμένα αρχεία λειτουργούν με τη λογική ότι σε δοθέν ερώτημα, πρώτα θα ερωτηθεί η δομή στην κύρια μνήμη και αν η απάντηση δεν είναι ικανοποιητική θα ερωτηθεί και η δομή στον δίσκο. Προτείνετε πολιτικές δημιουργίας και ενημέρωσης της δομής ανεστραμμένων αρχείων της κύριας μνήμης, υποθέτοντας ότι η μηχανή ακολουθεί ένα συνδυασμό μοντέλου διανυσματικού χώρου, και μίας link μετρικής (ανεξάρτητης περιεχομένου), στη βαθμολόγηση των ιστοσελίδων.
β. Καταγράψτε τους βασικούς αλγορίθμους συμπίεσης ανεστραμμένων αρχείων, και συγκρίνετέ τους μεταξύ τους.

ΘΕΜΑ 2 (*Μονάδες 3.5*)
α. Θεωρώντας ότι ο αριθμός των επιστρεφόμενων κειμένων είναι 30, θεωρήστε τις ακόλουθες δύο μηχανές που απαντάνε στο ίδιο ερώτημα (δίνονται ο αριθμός των σχετικών κειμένων και η θέση τους στο αποτέλεσμα) :
Μηχανή1 Αριθμός: 10, Θέση: 1, 3, 6, 9, 13, 18, 24, 25, 27, 30
Μηχανή2 Αριθμός: 10, Θέση: 1, 2, 3, 7, 10, 19, 22, 26, 27, 30.
Θεωρήστε ότι οι εμπειρογνώμονες εκτός από το αν ένα κείμενο είναι σχετικό ή όχι δίνουν και βαρύτητα σχετικότητας : 1 (αρκετά σχετικό), και 2 (πολύ σχετικό), και υποθέστε ότι στις δύο ανωτέρω μηχανές συμβαίνει κατά την ανάκτηση τα πέντε πρώτα σχετικά κείμενα να έχουν βαρύτητα σχετικότητας 2, και τα επόμενα έχουν βαρύτητα σχετικότητας 1.
ΣΥΓΚΡΙΝΕΤΕ ΤΙΣ ΔΥΟ ΜΗΧΑΝΕΣ χρησιμοποιώντας τη μετρική NDCG
(ξεκινήστε να υπολογίζετε τον τύπο, και φθάστε το μέχρι το σημείο που μπορεί να γίνει η σύγκριση).
β. Εξηγήστε πως ο αλγόριθμος βαθμολόγησης σελίδων Pagerank του Google και HITS του CLEVER μπορεί να επεκταθεί έτσι ώστε να ενσωματώνει πληροφορίες σχετικά με τη δεδομένα κίνησης πληροφορίας στο διαδίκτυο και προσπελασιμότητας των ιστοσελίδων από τους χρήστες.

ΘΕΜΑ 3 **(**Μονάδες 3)
Υποθέστε ότι το γράφημα του Web είναι αποθηκευμένο στον δίσκο ενός υπολογιστικού συστήματος, με τη μορφή ενός σύνολου adjacency lists (λίστες γειτνιάσεως) που είναι βασικά λίστες που περιέχουν για κάθε κορυφή **v** του

γραφήματος τη λίστα των άμεσα προσβάσιμων κορυφών από τον κόμβο **v**. Με τον τρόπο αυτό μπορούμε να ρωτήσουμε για τους γείτονες που έχει ένας συγκεκριμένος κόμβος και να τους προσπελάσουμε σαν μία λίστα ΑΜΕΣΑ από τον δίσκο.

α. Περιγράψτε τα βασικά βήματα του αλγόριθμο που υπολογίζει το PageRank στο σενάριο αυτό.

β. Υποθέστε ότι σε κάθε σελίδα απεικονίζουμε ένα ακέραιο αριθμό από **1** έως **N** (**N** το πλήθος των σελίδων). Πως μπορεί να συμπιεστεί η αναπαράστασή μας του γραφήματος στο σχήμα αυτό;

γ. Εξηγήστε γιατί η λεξικογραφική διάταξη των αλφαριθμητικών των URLs των σελίδων και η ανάθεση αύξοντων αριθμών σε αυτές οδηγεί σε καλή συμπίεση του γραφήματος.