

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΩΝ

ΘΕΜΑ 1.

Θεωρείστε την ακόλουθη συλλογή πέντε κειμένων. Για τη δική σας ευκολία παρατίθεται και η μετάφραση στα ελληνικά

Κείμενα	Μετάφραση
K1: banking <u>on</u> banks <u>to</u> raise <u>the</u> interest rate	Βασιζόμαστε στις τράπεζες να αυξήσουν τα επιτόκια
K2: jogging <u>along</u> <u>the</u> river bank <u>to</u> look <u>at</u> <u>the</u> sailboats	Τρέχοντας κατά μήκος της όχθης του ποταμού για να δω τα ιστιοπλοϊκά
K3: jogging <u>to</u> the bank <u>to</u> look <u>at</u> <u>the</u> interest rate	Τρέχοντας στη τράπεζα για να δω τα επιτόκια
K4: buzzer-beating shot banked <u>in</u> !	Έβαλε καλάθι τη στιγμή που σήμανε η λήξη του παιχνιδιού!
K5: scenic outlooks <u>on</u> <u>the</u> banks <u>of</u> <u>the</u> river	Είναι πολύ γραφικά στις όχθες του ποταμού

Κατά την προ-επεξεργασία των κειμένων, έχουν απομονωθεί οι κοινές γραμματικές ρίζες των λέξεων (stemming) καθώς επίσης έχουν αφαιρεθεί όλες οι λέξεις που έχουν υπογραμμιστεί στο κείμενο και οι οποίες δεν έχουν κάποιο σημασιολογικό περιεχόμενο (stopwords).

α) Με τη βοήθεια του λογικού μοντέλου απαντήστε τις παρακάτω ερωτήσεις (1,5 μονάδα)

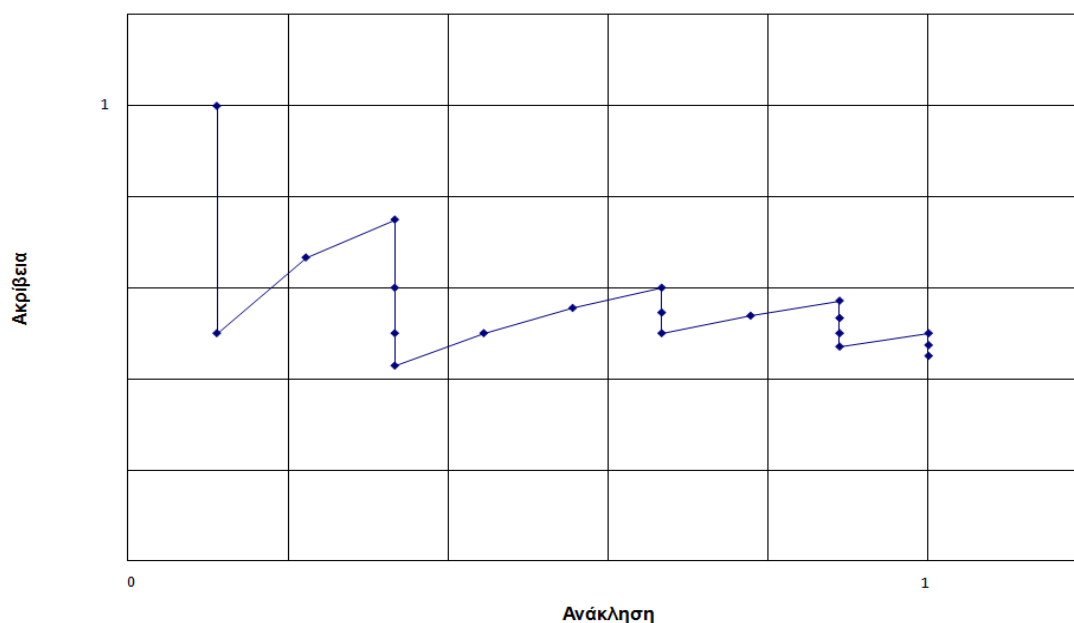
1) interest AND rate and NOT jog

2) bank AND (scenic OR jog)

γ) Με τη βοήθεια του διανυσματικού μοντέλου απαντήστε την ερώτηση “bank interest” (2 μονάδες)

Σχολιάστε τα αποτελέσματα του διανυσματικού μοντέλου.

ΘΕΜΑ 2.



Στο προηγούμενο σχήμα φαίνεται ένα διάγραμμα ακρίβειας /ανάκλησης για μία ερώτηση στο οποίο δεν έχει εφαρμοσθεί αναγωγή στα 11 καθορισμένα επίπεδα ανάκλησης (0%, 10%,...,100%).

α) «Αποκωδικοποιήστε» το διάγραμμα δηλ. καθορίστε ποια από τα κείμενα που επιστράφηκαν ήταν σχετικά και ποια όχι. **(0,75 μονάδες)**

β) Υπολογίστε το μέτρο απόδοσης «Average Precision at Seen Relevant Documents». **(0,5 μονάδες)**

γ) Υπολογίστε το μέτρο απόδοσης R-precision **(0,5 μονάδες)**

δ) Σχεδιάστε το διάγραμμα ακρίβειας /ανάκλησης μετά την αναγωγή στα 11 καθορισμένα επίπεδα ανάκλησης (0%, 10%,...,100%). **(0,25 μονάδες)**

ΘΕΜΑ 3.

α) Δώστε το δέντρο των καταλήξεων (suffix trie) για τη συμβολοσειρά «αβααβαβααβα» **(1 μονάδα)**

β) Γενικά, αν για μία συμβολοσειρά S έχει δημιουργηθεί το δέντρο των καταλήξεων, περιγράψτε συνοπτικά πως μπορούμε να απαντήσουμε τις παρακάτω ερωτήσεις με την βοήθεια της δομής αυτής:

i) Πόσες φορές εμφανίζεται η συμβολοσειρά $S1$ εντός της συμβολοσειράς S ; **(0,75 μονάδες)**

ii) Ποια είναι η μεγαλύτερου μήκους επαναλαμβανόμενη υπό-συμβολοσειρά στη συμβολοσειρά S ; **(0,75 μονάδες)**

ΘΕΜΑ 4.

Κάποιος προσπαθεί να ανεβάσει τη βαθμολογία της σελίδας του όπως αυτή προκύπτει από την PageRank τεχνική (PageRank score). Για το σκοπό αυτό, εφαρμόζει την τεχνική της φάρμας των συνδέσμων (link spam farm).

Συμβολίζουμε με p τη σελίδα της οποίας τη βαθμολογία προσπαθεί το συγκεκριμένο άτομο να ανεβάσει. Θεωρείστε επίσης ότι με κάποιο τρόπο το άτομο αυτό έχει καταφέρει m «καλές» ιστοσελίδες να δείχνουν στην σελίδα p . Έστω $g_1, g_2, \dots, g_m$ οι ιστοσελίδες αυτές και θεωρείστε επίσης κάθε μία από αυτές τις σελίδες έχει βαθμολογία γ σύμφωνα με την PageRank τεχνική καθώς επίσης ότι όλες αυτές οι σελίδες έχουν ίδιο πλήθος συνδέσμων, o , που δείχνουν σε άλλες σελίδες ($\text{outdegree}=o$). Επιπλέον, το συγκεκριμένο άτομο έχει δημιουργήσει k πλασματικές σελίδες $b_1, b_2, \dots, b_k$ που δείχνουν στη p . Συγκεκριμένα, κάθε μία από αυτές τις σελίδες έχει σχεδιαστεί ώστε να δείχνει την p και η σελίδα p έχει τροποποιηθεί ώστε να δείχνει σε κάθε μία από τις k πλασματικές σελίδες. Πρέπει να σημειωθεί ότι η σελίδα p δεν έχει άλλους συνδέσμους που να δείχνουν σε σελίδες άλλες πέραν των k παραπάνω σελίδων. Επιπλέον, κάθε πλασματική σελίδα δείχνει στις $k-1$ υπόλοιπες σελίδες. Έτσι τελικά κάθε σελίδα b_i ($i=1 \dots k$) έχει k συνδέσμους συνολικά που δείχνουν σε άλλες σελίδες: ένα προς τη σελίδα p και $k-1$ στις υπόλοιπες πλασματικές σελίδες.

Υπολογίστε τη βαθμολογία PageRank της σελίδας p .

Σημείωση: Θα πρέπει να καταλήξετε σε «κλειστό» τύπο και όχι αναδρομικό. Θεωρείστε επίσης ότι N είναι το πλήθος των ιστοσελίδων του παγκόσμιου ιστού και ότι d είναι ο συντελεστής dumping της τεχνικής PageRank. **(2 μονάδες)**