|  |
| --- |
|  |
| page1image29306176  Τμήμα Μηχανικών Η/Υ και Πληροφορικής  Πανεπιστήμιο Πατρών  Πολυτεχνική Σχολή  Τομέας Λογικού των Υπολογιστών  Διδάσκων: Χρήστος Μακρής  Ακαδημαϊκό Έτος: 2023 – 2024  Ημ / νία Παράδοσης: 28 / 01 / 2024 |

Ανάκτηση Πληροφορίας

Επιλεγόμενο Μάθημα – CEID\_NE5597

Εργαστηριακή Άσκηση

Χειμερινό Εξάμηνο 2023

Χρυσαυγή Πατέλη | 1084513 | up1084513@ac.upatras.gr

Μηλτιάδης Μαντές | 1084661 | up1084661@ac.upatras.gr

ΠΕΡΙΕΧΟΜΕΝΑ

**1 ΕΙΣΑΓΩΓΗ 1.1** Ερώτημα 1 …………………………………………………………………………………….. 3 **1.2** Ερώτημα 2  **1.3** Ερώτημα 3 **1.4** Ερώτημα 4

**2 ΥΛΟΠΟΙΗΣΗ 2.1** Ερώτημα 1 **2.2** Ερώτημα 2  **2.3** Ερώτημα 3 **2.4** Ερώτημα 4

**3 ΑΠΟΤΕΛΕΣΜΑΤΑ – ΠΑΡΑΤΗΡΗΣΕΙΣ 3.1** Ερώτημα 1 **3.2** Ερώτημα 2  **3.3** Ερώτημα 3 **3.4** Ερώτημα 4

**4 ΑΝΑΦΟΡΕΣ**

**5 ΠΑΡΑΡΤΗΜΑ**

**1 ΕΙΣΑΓΩΓΗ**

**1.2 Ερώτημα 2**

Στο συγκεκριμένο ερώτημα μας ζητείται να υλοποιήσουμε το **Μοντέλο Διανυσματικού Χώρου** (Vector Space Model) πάνω στο ανεστραμμένο αρχείο που εξάγαμε στο προηγούμενο ερώτημα.

Το **Μοντέλο Διανυσματικού Χώρου** χρησιμοποιεί την έννοια των διανυσμάτων στον πολυδιάστατο χώρο προκειμένου να αναπαραστήσει αλγεβρικά την φυσική γλώσσα των κειμένων της συλλογής εγγράφων. Πιο συγκεκριμένα, το σκεπτικό πίσω από αυτό το σύστημα ανάκτησης πληροφορίας έγκειται στην ευκολία εύρεσης κοινών όρων (terms) ανάμεσα σε κείμενα και ερωτήματα χρήστη (queries) αν τα ζυγισμένα διανύσματα που τα αναπαριστούν τοποθετηθούν σε διανυσματικό χώρο *V*.

Σε μια πιο γενική περιγραφή του μοντέλου ισχύουν τα εξής:

1. Έστω {t1, t2, …, tn} οι όροι ενός κειμένου di, i = 1, 2, …, m από την συλλογή μας. Υποθέτοντας χωρίς βλάβη γενικότητας ότι για κάθε διακριτό όρο ti υπάρχει στο χώρο ένα μοναδιαίο ορθοκανονικό διάνυσμα **ti** στο οποίο αντιστοιχίζεται, τότε ο διανυσματικός χώρος *V* θα είναι το span{ti} και έτσι το διάνυσμα **di** του κειμένου μπορεί να δοθεί ως γραμμικός συνδυασμός των **ti**.

**di =**

1. Έστω {T1, T2, …, Tn} οι όροι ενός ερωτήματος, οι οποίοι σχετίζονται με το σύνολο {t1, t2, …, tn}. Για κάθε ερώτημα q όμοια προκύπτει ότι το αντίστοιχο του διάνυσμα **q** του διανυσματικού χώρου *V* θα είναι πάλι γραμμικός συνδυασμός των διανυσμάτων **Τi**.

**q =**

Εικόνα που περιέχει γραμμή

Περιγραφή που δημιουργήθηκε αυτόματαΚάθε διακριτός όρος σε κάθε περίπτωση αντιπροσωπεύει μια διάσταση του χώρου μας. Συνεπώς, στόχος μας είναι στο χώρο *V* για κάθε διάνυσμα **q** να υπολογίσουμε την απόστασή του από κάθε διάνυσμα **di**. Για να γίνει αυτό επιδιώκουμε να προσεγγίσουμε τη γωνία που σχηματίζεται ανάμεσά τους και πιο συγκεκριμένα το συνημίτονό της, cos(θ), το οποίο δίνεται από το εσωτερικό γινόμενο:

cos(θ) =

Στο τέλος, χρησιμοποιούμε ως κριτήριο ταξινόμησης τη τιμή του συνημίτονου που προκύπτει προκειμένου να διατάξουμε τα κείμενα μας με βάση τη σχετικότητα τους ως προς κάθε ερώτημα χρήστη.

Ειδικότερα, στη δικιά μας υλοποίηση επιλέγουμε για απλοποίηση τα διανύσματα **di**, **qi** να έχουν ως διάσταση το πλήθος των διακριτών όρων του κειμένου και του ερωτήματος αντίστοιχα και κάθε γραμμή του διανύσματος **qi** να περιέχει μια boolean τιμή, με το 0 να συμβολίζει την μη – ύπαρξη του συγκεκριμένου όρου στο διάνυσμα κειμένου **di** και το 1 να συμβολίζει προφανώς την ύπαρξή του.

Εικόνα που περιέχει κείμενο, γραμμή, γραμματοσειρά, αριθμός

Περιγραφή που δημιουργήθηκε αυτόματαΤώρα, για την καλύτερη εφαρμογή του **Vector Space Model** είναι απαραίτητη η αξιοποίηση της τεχνικής του *term – weighting*, καθώς οι διαφορετικοί όροι μπορεί να έχουν διαφορετική σημασία και άρα και επίδραση στα συμφραζόμενα του κειμένου. Αρχικά, όσον αφορά το local term – weighting, επιλέγουμε μια από τις ακόλουθες τεχνικές υπολογισμού του term frequency (TF), καθώς μας αφορά μόνο η συχνότητα με την οποία εμφανίζονται οι όροι σε κάθε κείμενο ανεξάρτητα, την οποία και έχουμε αποθηκευμένη στο ανεστραμμένο αρχείο.

**Εικόνα 1**:

Η πρώτη μας σκέψη είναι να επιλέξουμε την *Απλή Συχνότητα Εμφάνισης*, η οποία υπολογίζει πόσες φορές εμφανίζεται ένας όρος στο κείμενο. Άλλωστε, όσο περισσότερες φορές εμφανίζεται ένας όρος στο κείμενο, τόσο πιο πιθανό είναι το κείμενο αυτό να είναι σχετικό με το ερώτημά μας. Ωστόσο, το μειονέκτημα αυτού του σκεπτικού είναι κυρίως ότι ευνοεί υπερβολικά τους όρους που εμφανίζονται με αυξημένη συχνότητα, καθώς η μεγάλη συχνότητα εμφάνισης δεν μας εξασφαλίζει υποχρεωτικά ότι οι όροι αυτοί θα είναι και πιο σημαντικοί από κάποιους όρους που εμφανίζονται πιο σπάνια. Επίσης, η χρήση δυαδικών βαρών είναικαι αυτή αρκετά περιοριστική. Γι’ αυτό καταλήγουμε στην *Απλή Λογαριθμική Κανονικοποίηση*, η οποία προσπαθεί να αντισταθμίσει την επίδραση της *Απλής Συχνότητας Εμφάνισης* μειώνοντας μέσω του λογαρίθμου την επίδραση πολύ μεγάλων διαφορών στις συχνότητες εμφάνισης, έτσι ώστε να έχουμε δικαιότερη απόδοση βαρών στο τελικό διάνυσμα **di**.

Έπειτα, όσον αφορά το global term – weighting, θέλουμε να εξετάσουμε πώς η κατανομή και η συχνότητα εμφάνισης ενός όρου σε όλα τα κείμενα της συλλογής επηρεάζει το πόσο σημαντικός είναι. Για το λόγο αυτό επιλέγουμε την αποδοτικότερη από τις πιο κάτω τεχνικές για τον υπολογισμό του Inverted Document Frequency (IDF):

**Εικόνα 2**:

Εικόνα που περιέχει κείμενο, γραμμή, γραμματοσειρά, αριθμός

Περιγραφή που δημιουργήθηκε αυτόματαΕδώ επιλέγουμε την τεχνική της Απλής Ανάστροφης Συχνότητας Εμφάνισης, η οποία «λογαριθμίζει» την αναλογία των κειμένων μέσα στη συλλογή μας που περιέχουν ένα συγκεκριμένο όρο. Η τιμή της αυξάνεται όσο το πλήθος των κειμένων στο οποίο εντοπίζεται αυτός ο όρος μειώνεται και έτσι ψηλή τιμή σημαίνει συνήθως ότι ένας όρος εμφανίζεται στο συγκεκριμένο έγγραφο με μεγαλύτερη συχνότητα από το συνηθισμένο.

Το τελικό σημείο ενδιαφέροντος στον καθορισμό των βαρών είναι η κανονικοποίησή τους, την οποία εμείς επιλέγουμε να μην την εφαρμόσουμε. Η κύρια χρησιμότητά της είναι

Έχοντας ολοκληρώσει αυτή τη διαδικασία τα τελικά μας βάρη θα πρέπει να είναι