Αλέξανδρος Ξιάρχος 1059619

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ · ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

ΕΡΓΑΣΤΗΡΙΑΚΗ ΑΣΚΗΣΗ ·2023-2024

ΠΕΡΙΕΧΟΜΕΝΑ

1 ΕΙΣΑΓΩΓΗ

1.1 VECTOR SPACE MODEL

Το μοντέλο διανυσματικού χώρου (Vector Space Model) αποτελεί μία από τις πιο διαδεδομένες μεθόδους για την αναπαράσταση και επεξεργασία κειμένων σε συστήματα ανάκτησης πληροφορίας. Σύμφωνα με αυτό, κάθε κείμενο αναπαρίσταται ως ένα διάνυσμα σε ένα πολυδιάστατο χώρο, όπου κάθε διάσταση αντιστοιχεί σε ένα όρο του λεξιλογίου. Το μήκος του διανύσματος αντιστοιχεί στη σημασία του κειμένου, ενώ η κατεύθυνση του διανύσματος αντιστοιχεί στο περιεχόμενο του κειμένου. Η ομοιότητα μεταξύ δύο κειμένων μπορεί να υπολογιστεί ως η συνημιτονική γωνία μεταξύ των αντίστοιχων διανυσμάτων. Με αυτόν τον τρόπο, μπορούμε να ταξινομήσουμε τα έγγραφα μιας συλλογής ως προς την ομοιότητά τους με ένα δεδομένο ερώτημα.

1.2 ΥΠΟΛΟΓΙΣΜΟΣ ΒΑΡΩΝ TF & IDF

Καταρχάς πρέπει να επιλέξουμε την παραλλαγή των βαρών TF και IDF για έγγραφα και ερωτήματα που είναι καταλληλότερη για τη συλλογή μας.

Όσον αφορά τα **έγγραφα**: Μιας και η συλλογή αφορά βάση δεδομένων για την Κυστική Ίνωση, δηλαδή πρόκειται για συλλογή με τεχνικές –ιατρικές συγκεκριμένα– ορολογίες (technical vocabulary and meaningful terms [MED collections]) [1], θα χρησιμοποιήσουμε τη **διπλή 0,5 κανονικοποίηση** (augmented normalized TF):

$$0.5 + 0.5\frac{F_{ij}}{max_k F_{kj}}$$

για το βάρος που αφορά τα έγγραφα, όπου $\boldsymbol{F_{ij}}$ οι φορές που ο όρος εμφανίζεται σε ένα έγγραφο και $\mathbf{max_k}\ \boldsymbol{F_{kj}}$ το μεγαλύτερο πλήθος εμφανίσεων κάποιου όρου σε ένα έγγραφο.

Όσον αφορά τα **queries**: κάθε λήμμα από τα ερωτήματα είναι σημαντικό (σχεδόν κάθε λέξη είναι ιατρική ορολογία), άρα θα χρησιμοποιήσουμε πάλι τη **διπλή 0,5 κανονικοποίηση** για το TF βάρος.

Για το IDF βάρος και σε έγγραφα και σε ερωτήματα, χρησιμοποιούμε την **απλή ανάστροφη συχνότητα εμφάνισης**:

$$\log \frac{\boldsymbol{N}}{\boldsymbol{n_i}}$$

όπου $\boldsymbol{N}$ ο συνολικός αριθμός των εγγράφων και $\boldsymbol{n_i}$ ο αριθμός των εγγράφων στα οποία εμπεριέχεται ο όρος.

white	Βάρος εγγράφων	Βάρος ερωτημάτων
TF	$\mathbf{0.5 + 0.5\frac{F_{ij}}{max_k F_{kj}}}$	$\mathbf{0.5 + 0.5\frac{F_{ij}}{max_k F_{kj}}}$
IDF	$\mathbf{\log \frac{N}{n_i}}$	$\mathbf{\log \frac{N}{n_i}}$

[1] Gerard Salton, Christopher Buckley, Term-weighting approaches in automatic text retrieval, Information Processing Management, Volume 24, Issue 5, 1988, Pages 513-523, ISSN 0306-4573

2 ΥΛΟΠΟΙΗΣΗ