ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΙΟΙΚΗΣΗ ΨΗΦΙΑΚΗΣ ΕΠΙΧΕΙΡΗΣΗΣ

ΕΡΓΑΣΙΑ 4: ΑΝΑΖΗΤΗΣΗ & RECOMMENDER SYSTEMS

Ειρήνη Δόντη

ΑΜ: 03119839

8ο εξάμηνο

Αθήνα 2023

1.ΑΝΑΖΗΤΗΣΗ – TFIDF

Υποθέστε ότι πραγματοποιείτε μια αναζήτηση σε μια μηχανή αναζήτησης με τους όρους “earth, mission, astro”. Θεωρείστε σαν μοναδικά δεδομένα της μηχανής τις παρακάτω περιγραφές ταινιών.

a). Χρησιμοποιείστε τη μετρική TF-IDF (χωρίς TF normalization) για να βρείτε ποια θα είναι η κατάταξη των αποτελεσμάτων. Να χρησιμοποιηθεί Cosine Similarity και να ληφθούν υπόψη το κείμενο των τίτλων και οι παράγωγες/σύνθετες λέξεις (Stemming). Σχολιάστε τα αποτελέσματα.

Για την αναζήτηση με χρήση TFIDF ακολουθείται η εξής διαδικασία:

Για κάθε λέξη σε ένα query [$term_A$, $term_B$, $term_C$] (με $term_A$ = earth , $term_B$ = mission , $term_C$ = astro) αναζητούμε και υπολογίζουμε τη συχνότητα εμφάνισης της (TF: Term Frequency) σε κάθε έγγραφο που δίνεται. Παραθέτουμε τον παρακάτω πίνακα:

	Europa Report	**The Hitchhiker’s Guide to the Galaxy**	**The Martian**	**Interstellar**	**Elysium**
earth	0	2	1	3	3
mission	1	0	2	1	1
astro	1	0	1	0	0

Λαμβάνονται υπόψη οι σύνθετες λέξεις που περιέχουν τα $term_A$, $term_B$ και $term_C$.

Στη συνέχεια, υπολογίζεται το IDF από τον τύπο: IDF = $\log(\frac{|D|}{|d: ti \epsilon d|})$ =

$\log(\frac{\text{total number of documents}}{\text{\# documents with word x in it}})$ με D το πλήθος των κειμένων:

IDF(earth) = $\log(\frac{5}{4})$ = 0.0969

IDF(mission) = $\log(\frac{5}{4})$ = 0.0969

IDF(astro) = $\log(\frac{5}{2})$ = 0.3979

Στη συνέχεια, πολλαπλασιάζουμε το TF με το IDF:

	Europa Report	**The Hitchhiker's Guide to the Galaxy**	**The Martian**	**Interstellar**	**Elysium**
earth	0	1.938	0.0969	2.907	2.907
mission	0.0969	0	1.938	0.0969	0.0969
astro	0.3979	0	0.3979	0	0

Με τη βοήθεια του Cosine Similarity, έχουμε:

Cosine Similarity(qi, doci) = $\frac{qi*doci}{||qi||*||doci||}$

Όπου ||doci|| = $\sqrt{0.0969^2 + 0.0969^2 + 0.3979^2}$ = 0.4208

<u>Europa Report</u>

q1*doc1 = (0*0.0969) + (0.0969*0.0969) + (0.3979*0.3979) = 0.1677

||q1|| = $\sqrt{0^2 + 0.0969^2 + 0.3979^2}$ = 0.4095

||doc1|| = $\sqrt{0.0969^2 + 0.0969^2 + 0.3979^2}$ = 0.4208

Cosine Similarity(q1, doc1) = $\frac{q1 * doc1}{||q1|| * ||doc1||}$ = 0.9732

The Hitchhiker's Guide to the Galaxy

q2*doc2 = (1.938*0.0969) + (0 *0.0969) + (0 *0.3979) = 0.1878

||q2|| =$\sqrt{1.938^2 + 0^2 + 0^2}$ = 1.938

||doc2|| = $\sqrt{0.0969^2 + 0.0969^2 + 0.3979^2}$ = 0.4208

Cosine Similarity(q2, doc2) = $\frac{q2 * doc2}{||q2|| * ||doc2||}$ = 0.2303

The Martian

q3*doc3 = (0.0969*0.0969) + (1.938*0.0969) + (0.3979*0.3979) = 0.3555

||q3|| = $\sqrt{0.0969^2 + 1.938^2 + 0.3979^2}$ = 1.9808

||doc3|| = $\sqrt{0.0969^2 + 0.0969^2 + 0.3979^2}$ = 0.4208

Cosine Similarity(q3, doc3) = $\frac{q3 * doc3}{||q3|| * ||doc3||}$ = 0.4265

Interstellar

q4*doc4 = (2.907*0.0969) + (0.0969*0.0969) + (0*0.3979) = 0.2911

||q4|| =$\sqrt{2.907^2 + 0.0969^2 + 0^2}$ = 2.909

||doc4|| = $\sqrt{0.0969^2 + 0.0969^2 + 0.3979^2}$ = 0.4208

Cosine Similarity(q4, doc4) = $\frac{q4 * doc4}{||q4|| * ||doc4||}$ = 0.2378

Elysium

q5*doc5 = (2.907*0.0969) + (0.0969*0.0969) + (0*0.3979) = 0.2911

||q5|| =$\sqrt{2.907^2 + 0.0969^2 + 0^2}$ = 2.909

||doc5|| = $\sqrt{0.0969^2 + 0.0969^2 + 0.3979^2}$ = 0.4208

Cosine Similarity(q5, doc5) = $\frac{q5 * doc5}{||q5|| * ||doc5||}$ = 0.2378

Βάσει των αποτελεσμάτων, η μηχανή αναζήτησης θα εμφανίσει τα έγγραφα με τον εξής τρόπο: Europa Report, The Martian, Elysium, Interstellar, The Hitchhiker's Guide to the Galaxy.

Η σειρά των αποτελεσμάτων είναι λογική, αν παρατηρηθεί πόσες κοινές λέξεις περιέχουν από τους όρους αναζήτησης.

(Στην περίπτωση ισοβαθμίας του cosine similarity, θεωρούμε ότι επιλέγεται ο τίτλος ανάλογα με το πρώτο γράμμα κατά αύξουσα σειρά)

2.ΑΝΑΖΗΤΗΣΗ - PRECISION / RECALL

a. Ποια αποτελέσματα αναφέρονται στην ταινία και ποια όχι; Εξηγήστε με συντομία. Ποια από τα αποτελέσματα είναι σωστά(true positive) και ποια λάθος (false positive);

b. Αν γνωρίζουμε ότι υπάρχουν ακόμη 450 αποτελέσματα που σχετίζονται με την ταινία και δεν βρέθηκαν – false negative – υπολογίστε τα παρακάτω

i. Precision

ii. Recall

iii. F-Measure

Σχολιάστε τα αποτελέσματα.

a. Η μηχανή επιστρέφει 26 αποτελέσματα εκ των οποίων τα 6 (2, 6, 10, 13, 20, 21 στη σειρά) αφορούν την ταινία που αναζητείται. Οπότε, τα 6 αποτελέσματα είναι true positive και τα υπόλοιπα 20 αποτελέσματα είναι false positive. Είναι λογικό να προκύπτουν και αποτελέσματα που δεν αφορούν αυτό που αναζητούμε, καθώς δεν προσδιορίστηκε ότι αναζητείται η ταινία. Οπότε, ένας διαφορετικός χρήστης μπορεί να αναζητά βοήθεια για contact στο WhatsApp ή την ετυμολογία της λέξης contact στα αγγλικά.

b. Από την εκφώνηση και από τα παραπάνω: TP = 6, FP = 20 & FN = 450

   (i) Precision: Precision = $\frac{\text{TP}}{\text{TP + FP}}$ = 0.2308. Το Precision αφορά την ακρίβεια ή πιστότητα των αποτελεσμάτων, αφού υπολογίζεται ο αριθμός των σωστών αποτελεσμάτων προς το σύνολο του αθροίσματος των σωστών και λανθασμένων αποτελεσμάτων.

(ii) Recall: Recall = $\frac{TP}{TP + FN}$ = 0.0132. Το Recall αφορά την πληρότητα των αποτελεσμάτων που προέκυψαν. Το αποτέλεσμα έχει χαμηλή τιμή, καθώς υπήρξαν 5 σωστά αποτελέσματα, ενώ υπάρχουν άλλα 450 αποτελέσματα που δεν εμφανίστηκαν.

(iii) F-Measure: F-Measure = $\frac{2*Precision*Recall}{Precision + Recall}$ = 0.025. Το F-Measure είναι ένας σταθμισμένος αρμονικός μέσος όρος μεταξύ Precision & Recall.

3.RECOMMENDER SYSTEMS

Έξι άνθρωποι αξιολόγησαν τις ταινίες του ερωτήματος 1 με βαθμολογία από 1 (καθόλου καλή) έως 10 (εξαιρετική) και τα αποτελέσματα φαίνονται στον παρακάτω πίνακα. Αντικαταστήστε όπου X τον τελευταίο ψηφίο του Αριθμού Μητρώου σας, αφού προσθέσετε 1.

	Sunshine	The Martian	Moon	Contact	Gattaca
Χρήστης 1	X	3		7	7
Χρήστης 2	7	1		2	5
Χρήστης 3	8	2	4	2	X
Χρήστης 4	3	3	2	7	3
Χρήστης 5		3	X	7	6
Χρήστης 6	9	4	5	4	9

a. Υπολογίστε την ομοιότητα (similarity) μεταξύ των 6 χρηστών χρησιμοποιώντας δυο μεθόδους: Ευκλείδεια απόσταση και Pearson Correlation.

b. Χρησιμοποιώντας K-Nearest Neighbors με k=2 και weighted average και με τις δύο μετρικές του ερωτήματος i. υπολογίστε πως περιμένουμε να αξιολογήσει την ταινία Moon ο χρήστης ii . Σχολιάστε τα αποτελέσματα.

c. Αν υποθέσουμε ότι χρησιμοποιούμε τις προτιμήσεις των χρηστών στις ταινίες για να προτείνουμε φίλους, τότε ποιες σχέσεις σας φαίνονται πιο πιθανές; Εξηγήστε.

Ο πίνακας στον οποίο θα εργαστούμε, είναι ο παρακάτω:

	Sunshine	**The Martian**	**Moon**	**Contact**	**Gattaca**
Χρήστης 1	10	3		7	7
Χρήστης 2	7	1		2	5
Χρήστης 3	8	2	4	2	10
Χρήστης 4	3	3	2	7	3
Χρήστης 5		3	10	7	6
Χρήστης 6	9	4	5	4	9

a.

<u>Ευκλείδεια Απόσταση</u>:

Στην περίπτωση που υπάρχει κενό σε τουλάχιστον ένα κελί, το ζευγάρι δεν υπολογίζεται.

Ισχύει ότι sum = sum+ (rating (User u, Item j) – rating (User z, Item j))^2,

dist((x, y), (a, b)) = (x-a)^2 + (y-b)^2 και similarity(0,1) = $\frac{1}{1+\sqrt{sum}}$

Πίνακας ευκλείδειας απόστασης :

	Χρήστης 1	**Χρήστης 2**	**Χρήστης 3**	**Χρήστης 4**	**Χρήστης 5**	**Χρήστης 6**
Χρήστης 1	-	6.481	6.245	8.062	1	3.873
Χρήστης 2		-	5.196	7	5.477	5.745
Χρήστης 3			-	10.198	8.832	3.317
Χρήστης 4				-	8.544	9.539
Χρήστης 5					-	6.633
Χρήστης 6						-

Πίνακας Ομοιότητας:

	Χρήστης 1	**Χρήστης 2**	**Χρήστης 3**	**Χρήστης 4**	**Χρήστης 5**	**Χρήστης 6**
Χρήστης 1	-	0.134	0.138	0.110	0.5	0.205
Χρήστης 2		-	0.161	0.125	0.154	0.148
Χρήστης 3			-	0.089	0.102	0.232
Χρήστης 4				-	0.105	0.095
Χρήστης 5					-	0.131
Χρήστης 6						-

Pearson Correlation:

Ισχύει ότι: Sum1=sum(ratings user1), Sum2=sum(ratings user2) Sum1Sq=sum[(ratings user1)^2], Sum2Sq=sum[(ratings user2)^2], pSum = sum((rating user 1, item k) x (rating user 2, item k))

num = pSum - $\frac{sum1sum2}{n}$, den = $\sqrt{(sum1Sq - \frac{sum1^2}{n})(sum2Sq - \frac{sum2^2}{n})}$

Correlation = $\frac{num}{den}$, similarity = $\frac{1+correlation}{2}$

Πίνακας Συσχέτισης:

	Χρήστης 1	**Χρήστης 2**	**Χρήστης 3**	**Χρήστης 4**	**Χρήστης 5**	**Χρήστης 6**
Χρήστης 1	-					
Χρήστης 2	0.899	-				
Χρήστης 3	0.521	0.802	-			
Χρήστης 4	0.387	-0.044	-0.409	-		
Χρήστης 5	-0.673	-0.693	-0.273	0.080	-	
Χρήστης 6	0.632	0.894	0.978	-0.377	-0.433	-

Πίνακας Ομοιότητας:

	Χρήστης 1	**Χρήστης 2**	**Χρήστης 3**	**Χρήστης 4**	**Χρήστης 5**	**Χρήστης 6**
Χρήστης 1	-					
Χρήστης 2	0.949	-				
Χρήστης 3	0.760	0.901	-			
Χρήστης 4	0.693	0.478	0.295	-		
Χρήστης 5	0.163	0.153	0.364	0.540	-	
Χρήστης 6	0.816	0.947	0.989	0.312	0.283	-

b.

(i) Χρησιμοποιούμε k-Nearest Neighbors με k = 2 και weighted average:

Από τον πίνακα ομοιότητας που υπολογίσαμε μέσω της *Ευκλείδειας Απόστασης*, προκύπτει ο πίνακας γειτνίασης για τον Χρήστη 2.

Πίνακας Ομοιότητας:

	Χρήστης 1	**Χρήστης 2**	**Χρήστης 3**	**Χρήστης 4**	**Χρήστης 5**	**Χρήστης 6**
Χρήστης 1	-	0.134	0.138	0.110	0.5	0.205
Χρήστης 2		-	0.161	0.125	0.154	0.148
Χρήστης 3			-	0.089	0.102	0.232
Χρήστης 4				-	0.105	0.095
Χρήστης 5					-	0.131
Χρήστης 6						-

	Χρήστης 1	**Χρήστης 2**	**Χρήστης 3**	**Χρήστης 4**	**Χρήστης 5**	**Χρήστης 6**
Χρήστης 2	0.134	-	0.161	0.125	0.154	0.148

Οπότε, οι πιο κοντινοί γείτονες του Χρήστη 2 είναι ο Χρήστης 3 και ο Χρήστης 5. Και οι δύο χρήστες έχουν βαθμολογήσει την ταινία Moon, οπότε:

$$\text{PredictedRatingU2} = \frac{\text{sim(u2,u3)} * \text{rating(u3)} + \text{sim(u2,u5)} * \text{rating(u5)}}{\text{sim(u2,u3)} + \text{sim(u2,u5)}} = 6{,}933$$

Οπότε, η αναμενόμενη βαθμολογία του Χρήστη 2 για την ταινία Moon είναι 6,933.

Από τον πίνακα ομοιότητας που υπολογίσαμε μέσω της *Pearson Correlation* , προκύπτει ο πίνακας γειτνίασης για τον Χρήστη 2.

Πίνακας Ομοιότητας:

	Χρήστης 1	**Χρήστης 2**	**Χρήστης 3**	**Χρήστης 4**	**Χρήστης 5**	**Χρήστης 6**
Χρήστης 1	-					
Χρήστης 2	0.949	-				
Χρήστης 3	0.760	0.901	-			
Χρήστης 4	0.693	0.478	0.295	-		
Χρήστης 5	0.163	0.153	0.364	0.540	-	
Χρήστης 6	0.816	0.947	0.989	0.312	0.283	-

	Χρήστης 1	**Χρήστης 2**	**Χρήστης 3**	**Χρήστης 4**	**Χρήστης 5**	**Χρήστης 6**
Χρήστης 2	0.949	-	0.901	0.478	0.153	0.947

Οπότε, οι πιο κοντινοί γείτονες του Χρήστη 2 είναι ο Χρήστης 1 και ο Χρήστης 6.

Όμως, ο Χρήστης 1 δεν έχει βαθμολογήσει την ταινία Moon και συνεπώς εξετάζουμε τον επόμενο πιο κοντινό Χρήστη, τον Χρήστη 3 που την έχει βαθμολογήσει.

$$\text{PredictedRatingU2} = \frac{\text{sim(u2,u3)}*\text{rating(u3)} + \text{sim(u2,u6)}*\text{rating(u6)}}{\text{sim(u2,u3)} + \text{sim(u2,u6)}} = 4.512$$

Οπότε, η αναμενόμενη βαθμολογία του Χρήστη 2 για την ταινία Moon είναι 4.512.

(ii) Παρατηρείται ότι η μέθοδος της Ευκλείδειας Απόστασης και η μέθοδος Pearson Correlation διαφέρουν κατά δύο μονάδες περίπου, γεγονός που δικαιολογείται από τον τρόπο που υπολογίζεται η κάθε μέθοδος. Στην πρώτη περίπτωση, υπολογίζεται το διάνυσμα της σχετικής διαφοράς, ενώ στη δεύτερη περίπτωση υπολογίζεται η weighted average των βαθμολογιών.

c.

Για την *Ευκλείδεια Απόσταση*:

Πίνακας Ομοιότητας:

	Χρήστης 1	**Χρήστης 2**	**Χρήστης 3**	**Χρήστης 4**	**Χρήστης 5**	**Χρήστης 6**
Χρήστης 1	-	0.134	0.138	0.110	0.5	0.205
Χρήστης 2		-	0.161	0.125	0.154	0.148
Χρήστης 3			-	0.089	0.102	0.232
Χρήστης 4				-	0.105	0.095
Χρήστης 5					-	0.131
Χρήστης 6						-

Προτείνονται οι σχέσεις για τους Χρήστες (1, 5), (3, 6), (1, 6), αφού έχουν μεγαλύτερη ομοιότητα (similarity) 0.5, 0.232 και 0.205.

Για την *Pearson Correlation*:

Πίνακας Ομοιότητας:

	Χρήστης 1	**Χρήστης 2**	**Χρήστης 3**	**Χρήστης 4**	**Χρήστης 5**	**Χρήστης 6**
Χρήστης 1	-					
Χρήστης 2	0.949	-				
Χρήστης 3	0.760	0.901	-			
Χρήστης 4	0.693	0.478	0.295	-		
Χρήστης 5	0.163	0.153	0.364	0.540	-	
Χρήστης 6	0.816	0.947	0.989	0.312	0.283	-

Προτείνονται οι σχέσεις για τους Χρήστες (6, 3), (2, 1), (6, 2), αφού έχουν μεγαλύτερη ομοιότητα (similarity) 0.989, 0.949 και 0.947.