**k-Εγγύτεροι Γείτονες & Παραγωγή Συστάσεων**

(Δ. Δέρβος, Τμ. Μηχ. Πληροφορικής & Ηλεκτρονικών Συστημάτων, ΔΙ.ΠΑ.Ε.)

Από: http://www.billqualls.com/pa/Qualls_ECT584_RecommenderSystems.pdf

Ο πίνακας που ακολουθεί καταχωρεί δεδομένα αξιολόγησης βιβλίων από αναγνώστες-πελάτες ηλεκτρονικού βιβλιοπωλείου. Είκοσι (20) πελάτες (U1-U20) αξιολογούν το κάθε ένα από οκτώ (8) βιβλία βαθμολογώντας τα σε μία κλίμακα από το 1=κακό έως το 5=εξαιρετικό. Δύο νέοι πελάτες (NU1 κα;ι NU2) έχουν διαβάσει και αξιολογήσει μερικά από τα βιβλία. Στον Πίνακα 1 που ακολουθεί, οι κενές κυψέλες δηλώνουν την απουσία αξιολόγησης:

| | TRUE BELIEVER | THE DA VINCI CODE | THE WORLD IS FLAT | MY LIFE SO FAR | THE TAKING | THE KITE RUNNER | RUNNY BABBIT | HARRY POTTER |
|---|---|---|---|---|---|---|---|---|
| **U1** | 1 | 5 | | 3 | | | 3 | 5 |
| **U2** | 5 | 4 | | | 3 | 2 | 1 | |
| **U3** | 3 | | 1 | 2 | 2 | | | 5 |
| **U4** | | 3 | | | 4 | 1 | | 3 |
| **U5** | 2 | 4 | 3 | | | 2 | 2 | |
| **U6** | 5 | | | 3 | 1 | | 3 | 1 |
| **U7** | 1 | 4 | 5 | 5 | 2 | | | 4 |
| **U8** | 2 | 1 | | | 4 | 5 | 1 | |
| **U9** | | | 3 | 2 | 2 | | | 5 |
| **U10** | 3 | 5 | 1 | | | | 4 | 4 |
| **U11** | | | 2 | 1 | | 2 | | 3 |
| **U12** | 4 | 4 | | 2 | | 1 | 1 | 4 |
| **U13** | | | 2 | | 4 | | 4 | 5 |
| **U14** | | 5 | 3 | 3 | 2 | | 1 | 1 |
| **U15** | | 2 | | | 3 | 3 | | 2 |
| **U16** | | 3 | 2 | 1 | 1 | | 4 | 4 |
| **U17** | 1 | 5 | 1 | 2 | | 4 | | 4 |
| **U18** | 5 | | 4 | | 3 | 3 | 4 | 5 |
| **U19** | | 4 | | 2 | | 5 | 1 | 5 |
| **U20** | 2 | 5 | 1 | 1 | 5 | 3 | | 4 |
| | | | | | | | | |
| **NU1** | 3 | | 5 | 4 | 2 | 3 | | 5 |
| **NU2** | | 5 | 2 | 2 | 4 | | 1 | 3 |

**Πίνακας 1**. Είκοσι (U1-U20) συν δύο (NU1-NU2) πελάτες-αναγνώστες αξιολογούν οκτώ (8) βιβλία

Να χρησιμοποιηθεί ο αλγόριθμος των k-Εγγύτερων Γειτόνων για να προβλεφθούν οι τιμές με τις οποίες θα αξιολογήσουν οι πελάτες NU1 και NU2 τα δύο βιβλία τα οποία ο κάθε ένας τους δεν έχει αξιολογήσει, ακόμη. Ως μέτρο ομοιότητας (απόστασης) να χρησιμοποιηθεί η Pearson correlation coefficient. Πιο συγκεκριμένα:

1. Να χρησιμοποιηθεί το συνοδεύον αρχείο MS-Excel για τον υπολογισμό των τιμών συσχέτισης (correlation) μεταξύ των δύο νέων πελατών (NU1 και NU2) και όλους τους υπόλοιπους είκοσι (20) πελάτες (U1-U20). Στη συνέχεια και για κάθε έναν από τους NU1 και NU2 να εντοπιστούν οι τρεις (3) εγγύτεροι γείτονές του. Η μεθοδολογία παρέκτασης/extrapolation που χρησιμοποιείται για τον υπολογισμό του βαθμού αξιολόγησης που καταχωρείται στα κενά κελιά του πίνακα οφέλους (utility matrix) βασίζεται στην εύρεση του σταθμισμένου μέσου (weighted average function):

$$r(NU, I_t) = \frac{\sum_{i=1}^{K} r(U_i, I_t) \times sim(NU, U_i)}{\sum_{l=1}^{K} sim(NU, U_l)}$$

… όπου $r(U_i, I_t)$ αναπαριστά το βαθμό που δίνει ο πελάτης $U_i$ στο βιβλίο $I_t$ και $sim(NU, U_i)$ είναι η ομοιότητα των πελατών $NU$ και $U_i$.
**Ισοδύναμα**: οι τιμές με τις οποίες οι k-εγγύτεροι γείτονες αξιολογούν το βιβλίο $I_t$ σταθμίζονται βάσει της ομοιότητάς τους προς τον πελάτη NU και το άθροισμα των σταθμισμένων αυτών τιμών διαιρείται δια του αθροίσματος των ομοιοτήτων των K-γειτόνων προς τον πελάτη NU.
**Σημείωση-1**: Να γίνει χρήση της συνάρτησης CORREL() του MS-Excel
**Σημείωση-2**: Για τον εντοπισμό των k=3 εγγύτερων γειτόνων να αγνοηθούν οι γείτονες που πρεσβεύουν μέτρο ομοιότητας με τιμή μικρότερη του μηδέων (0)
**Σημείωση-3**: Στην περίπτωση όπου συμβεί ένας η περισσότεροι από τους k=3 εγγύτερους γείτονες να μην έχει(-ουν) αξιολογήσει το επίμαχο βιβλίο, αυτός(-οί) να αγνοηθεί(-ούν) και οι υπολογισμοί να γίνουν με τους υπόλοιπους (k<3) από τους εγγύτερους γείτονες (σ.σ. ΚΑΙ στον αριθμητή, ΚΑΙ στον παρανομαστή της παραπάνω αλγεβρικής παράστασης).

2. Για την αξιολόγηση της ποιότητας στην πρόβλεψη μέσω των του παραπάνω (1), να υπολογιστεί το Μέσο Απόλυτο Σφάλμα (MAE, Mean Absolute Error) των προβλέψεων που υπολογίζονται για τους πελάτες NU1 και NU2. Αυτό μπορεί να γίνει με εφαρμογή της διαδικασίας του παραπάνω (1) για την πρόβλεψη τιμών αξιολόγησης τα οποία έχουν ήδη αξιολογήσει οι πελάτες NU1 και NU2. Πιο συγκεκριμένα, για τον υπολογισμό προβλεπόμενων τιμών με τα οποία αξιολογεί ο πελάτης NU1 όλα τα βιβλία εκτός των “The DaVinci Code” και “Runny Babbit”. Επίσης, για τον υπολογισμό προβλεπόμενων τιμών με τα οποία αξιολογεί ο πελάτης NU2 όλα τα βιβλία εκτός των “True Believer” και “The Kite Runner”. Στη συνέχεια υπολογίζεται η απόλυτη τιμή της διαφοράς της προβλεπόμενης από την πραγματική τιμή αξιολόγησης και υπολογίζονται οι μέσες τιμές του σφάλματος στην πρόβλεψη: μία τιμή MAE για τον πελάτη NU1 και μία τιμή MAE για τον πελάτη NU2.
**Σημείωση**: μπορεί να γίνει χρήση των συναρτήσεων IF() και AVERAGE() στο περιβάλλον του MS-Excel.

~~~~~~~~~

Με ακρίβεια τριών (3) δεκαδικών ψηφίων, οι σωστές απαντήσεις για τα παραπάνω (1) και (2) έχουν ως εξής:

1. r(NU1, “The DaVinci Code”) = 4.360
   r(NU1, “Runny Babbit”) = 2.500
   r(NU2, “True Believer”) = 3.040
   r(Nu2, “The Kite Runner”) = 2.319
2. MAE(NU1) = 0.721
   MAE(NU2) = 0.539

~~~~~~~~~~