**Ανάκτηση Πληροφοριών - Μηχανές Αναζήτησης**

***Απαλλακτική Εργασία των φοιτητών:***

***Μπαρμπαγιάννη Ελένη 185239 (***[***lena.mparmpa@gmail.com***](mailto:lena.mparmpa@gmail.com)***)***

***Τσερετέλι Γιώργος 164840 (tsereteligiorgos@gmail.com)***

***ΘΕΜΑ 4ο:***

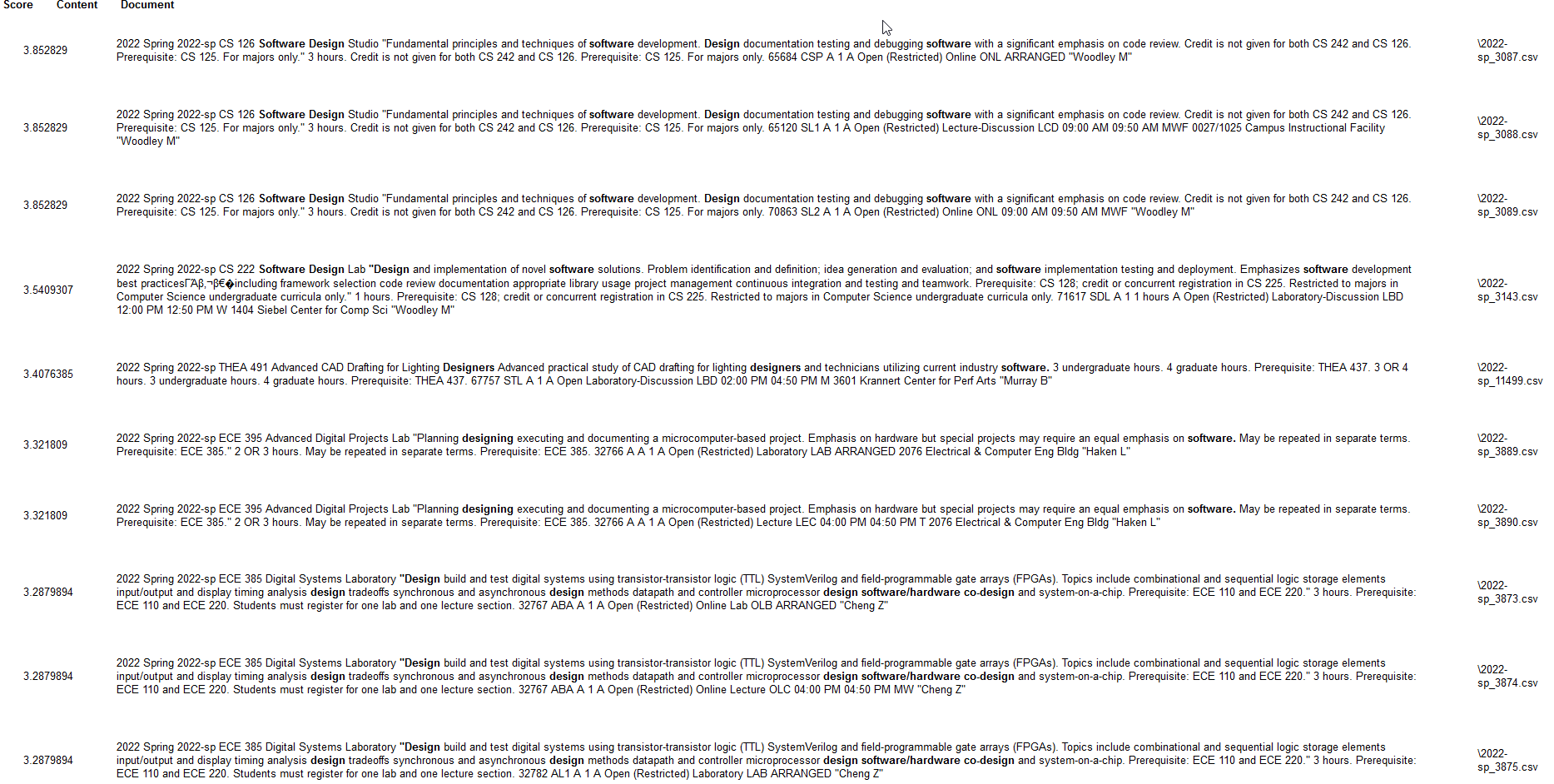
**A.** Αρχικά, λαμβάνοντας υπόψη ότι το dataset που έχουμε επιλέξει (University of Illinois' Course Catalog and Course Schedule) αναφέρεται σε πανεπιστημιακά μαθήματα, είναι λογικό ότι ένας χρήστης θα αναζητούσε σχετικές πληροφορίες.

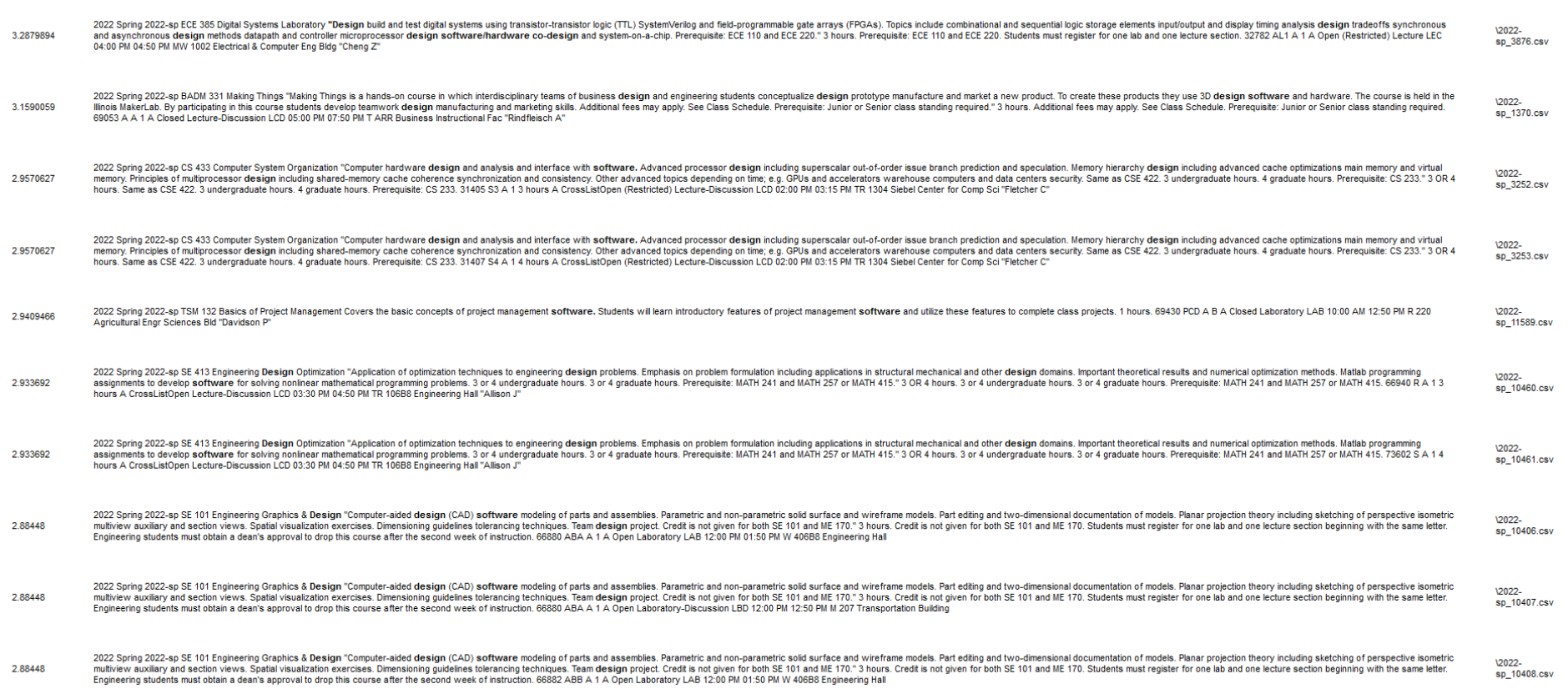
Έτσι, τα ερωτήματα που επιλέχθηκαν είναι σχετικά και με το dataset, άλλα εξετάζουν και κάποια προαναφερόμενα ερωτήματα του 2ου θέματος, όπως θα αναλυθεί στην συνέχεια.

Τα 3 ερωτήματα που επιλέχθηκαν είναι: *software designing, mathematic course, literature of English.* Στα εξής ερωτήματα εξετάζεται αν το σύστημά μας αγνοεί τα stop-words(Υποερώτημα 2 του 2ου θέματος) , αν γίνεται σωστά το stemming (3o υποερώτημα) και πιστεύουμε ότι με τα συγκεκριμένα ερωτήματα, εφόσον υπάρχουν φράσεις και όχι μεμονωμένες λέξεις, θα εξεταστούν καλύτερα τα αποτελέσματα που θα φέρει το σύστημα.

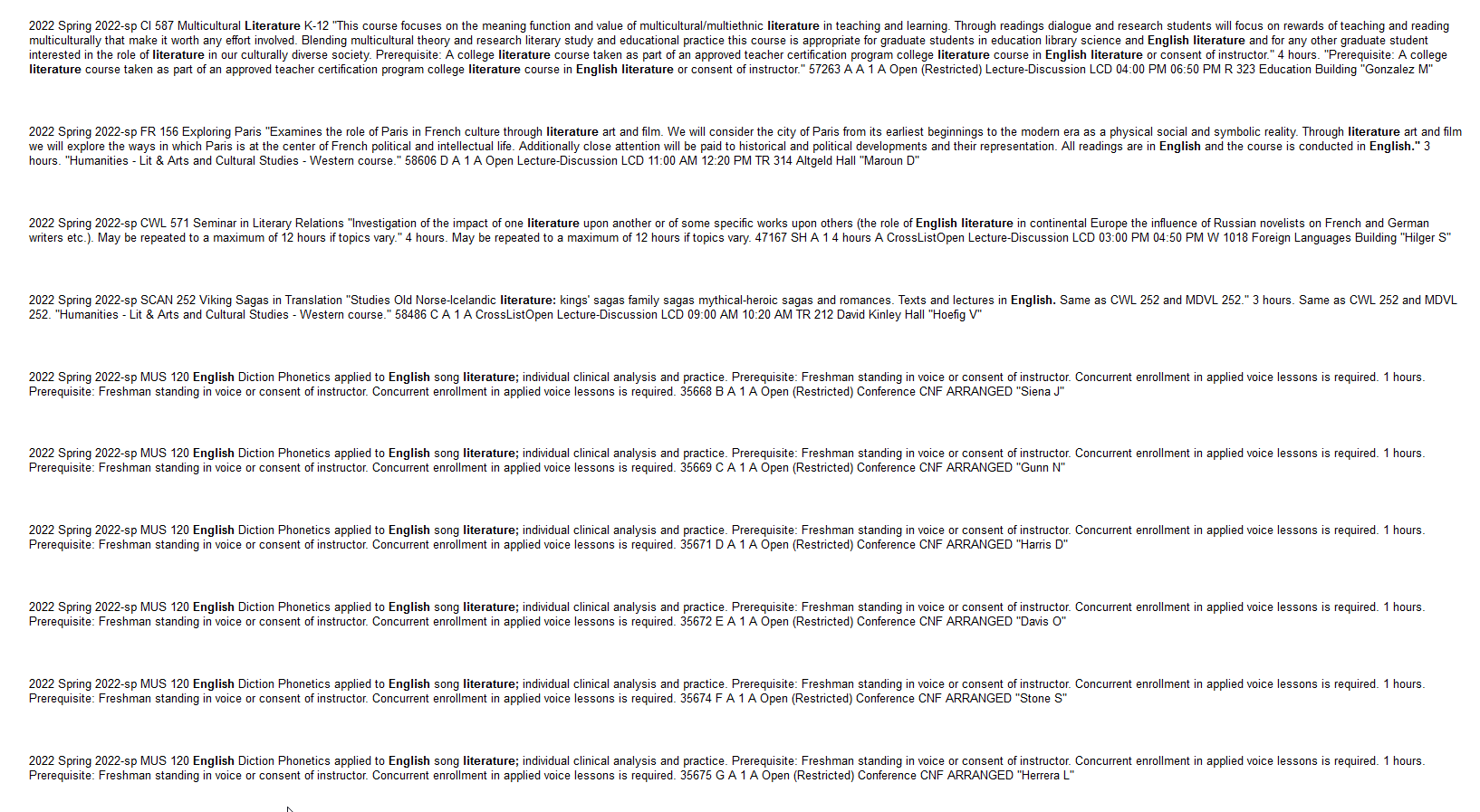
Τα 20 πρώτα αποτελέσματα για κάθε ερώτημα και η σχετικότητα τους σύμφωνα με τα δικά μας κριτήρια ως expert είναι τα εξής:

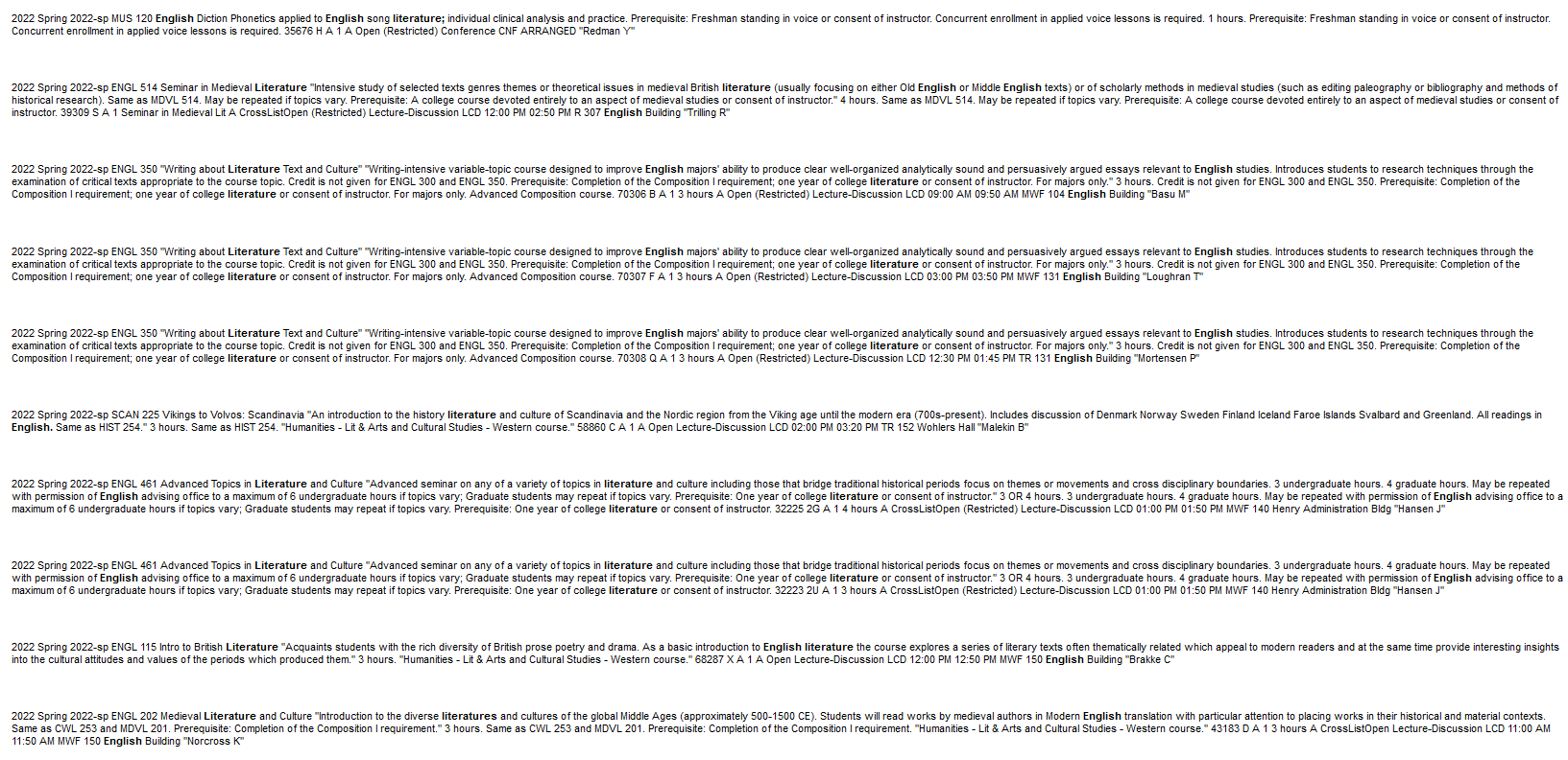
* **Software Designing** (Σχετικά αποτελέσματα 7/20)



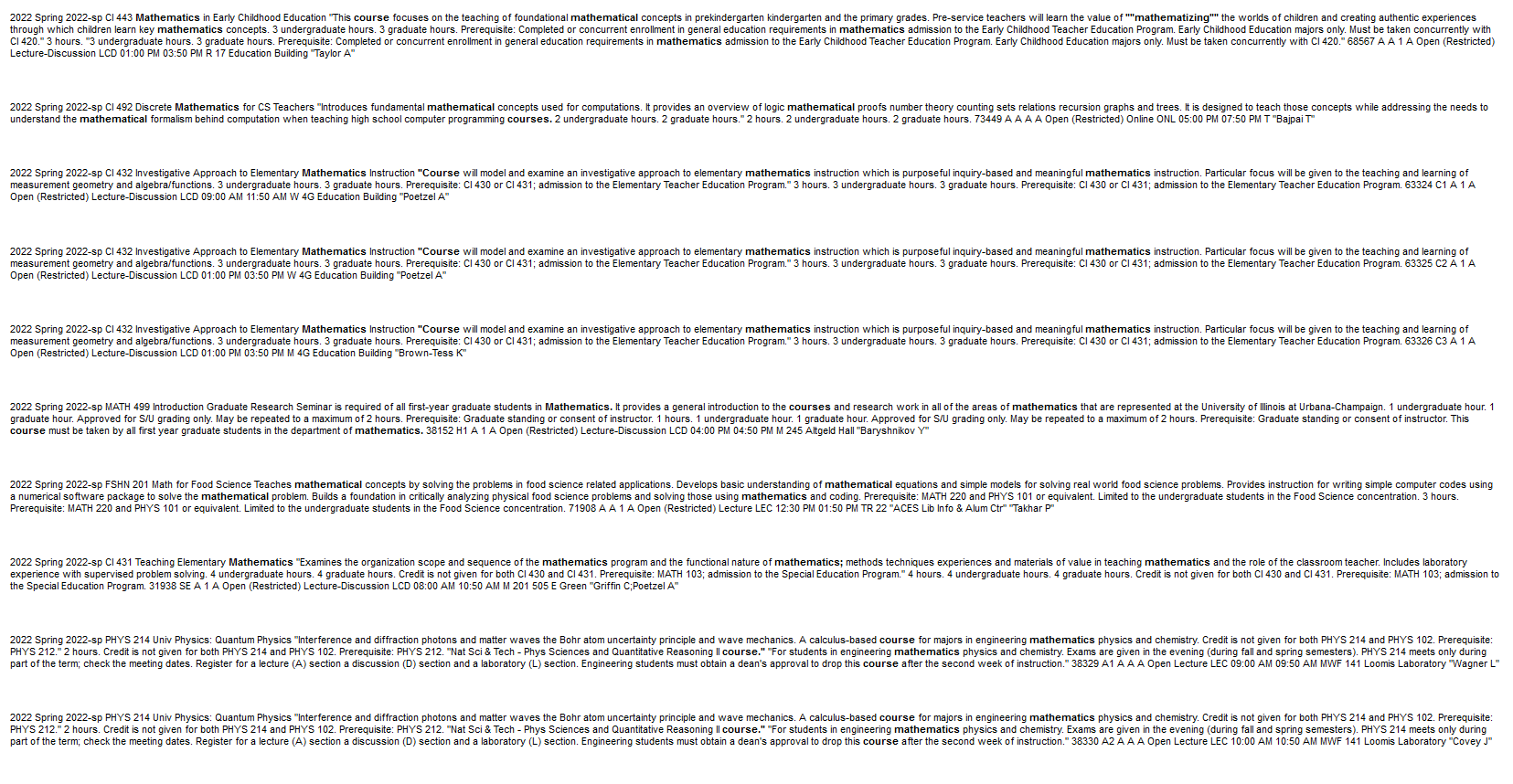


* **Literature of English** (Σχετικά αποτελέσματα 11/20)



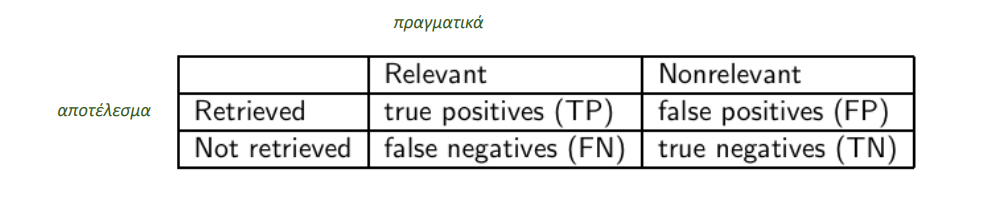


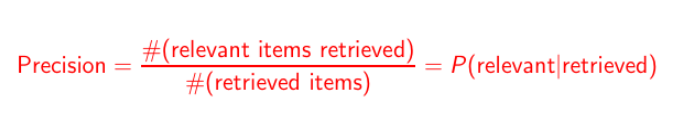
* **Mathematic Course** (Σχετικά αποτελέσματα 7/20)

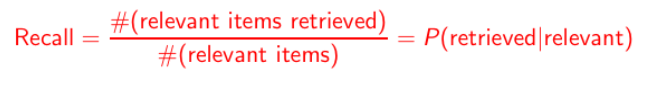




Εφόσον έχουμε τον confusion matrix για τα retrieved και τα not retrieved αποτελέσματα και τους τύπους για το **Recall(R)** και **Precision(P)** :







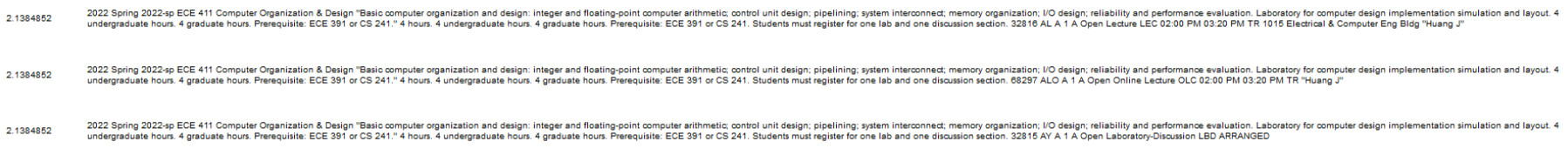
Συμπεραίνουμε ότι το P για κάθε ερώτημα λαμβάνοντας ως σύνολο απαντήσεων τις πρώτες 20 είναι:

* **Software Designing→ P=7/20 =0.35**
* **Literature of English → P=11/20 =0.55**
* **Mathematic Course →P=7/20 =0.35**

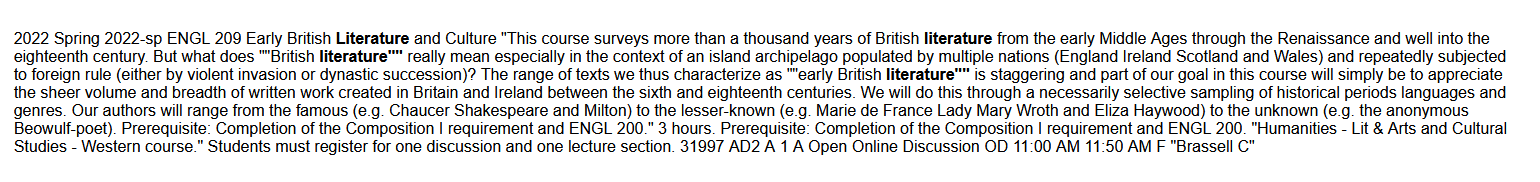
Για να υπολογίσουμε το **Recall** πρέπει αρχικά να βρούμε τα αποτελέσματα που ενδεχομένως να είναι σχετικά με τα ερωτήματά μας ,αλλά δεν εμφανίζονται στην αναζήτηση του χρήστη.

Αναζητώντας στο σύστημα βρήκαμε:

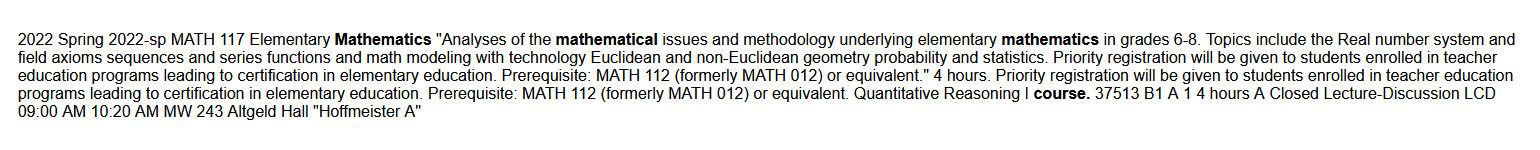
* 3 αποτελέσματα που δεν εμφανίστηκαν για το ερώτημα **Software Designing** (παράδειγμα στην εικόνα)



* 4 αποτελέσματα για το **Literature of English** (παράδειγμα στην εικόνα)

****

* 23 αποτελέσματα για το **Mathematic Course**(παράδειγμα στην εικόνα)

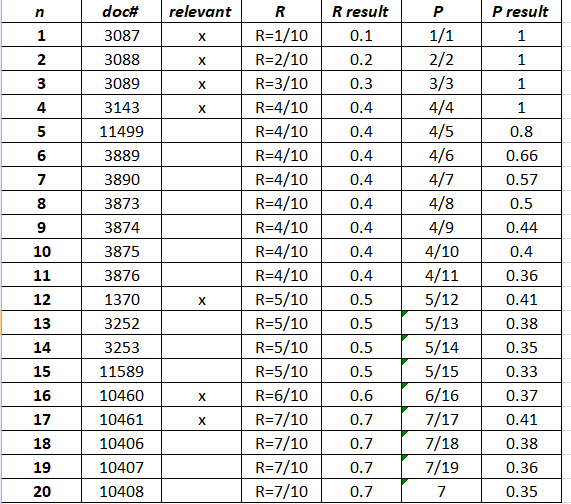


Έτσι, το **Recall** είναι:

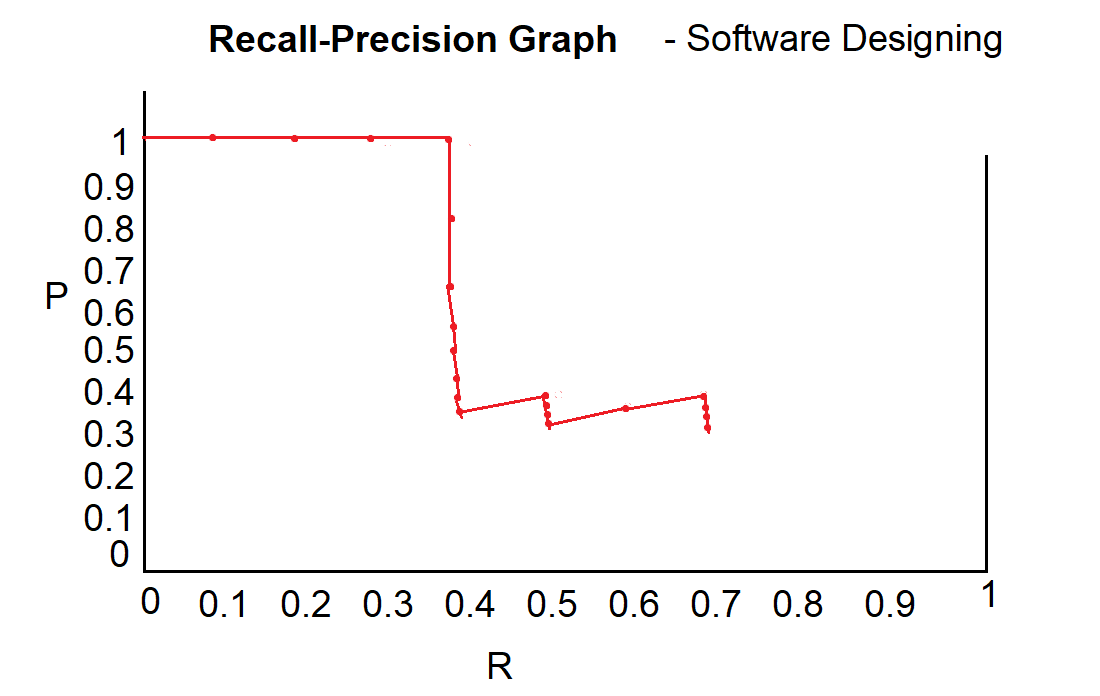
* **Software Designing→ R=7/10 =0.7**
* **Literature of English → R=11/15 =0.733**
* **Mathematic Course →R=7/30 =0.233**

**A.a)** Για την δημιουργία καμπύλης ακρίβειας/ανάκλησης αρχικά υπολογίζουμε τις τιμές precision/recall και στη συνέχεια κάνουμε έλεγχο σε κάθε νέο σημείο recall, όπως φαίνεται στον πίνακα τιμών:

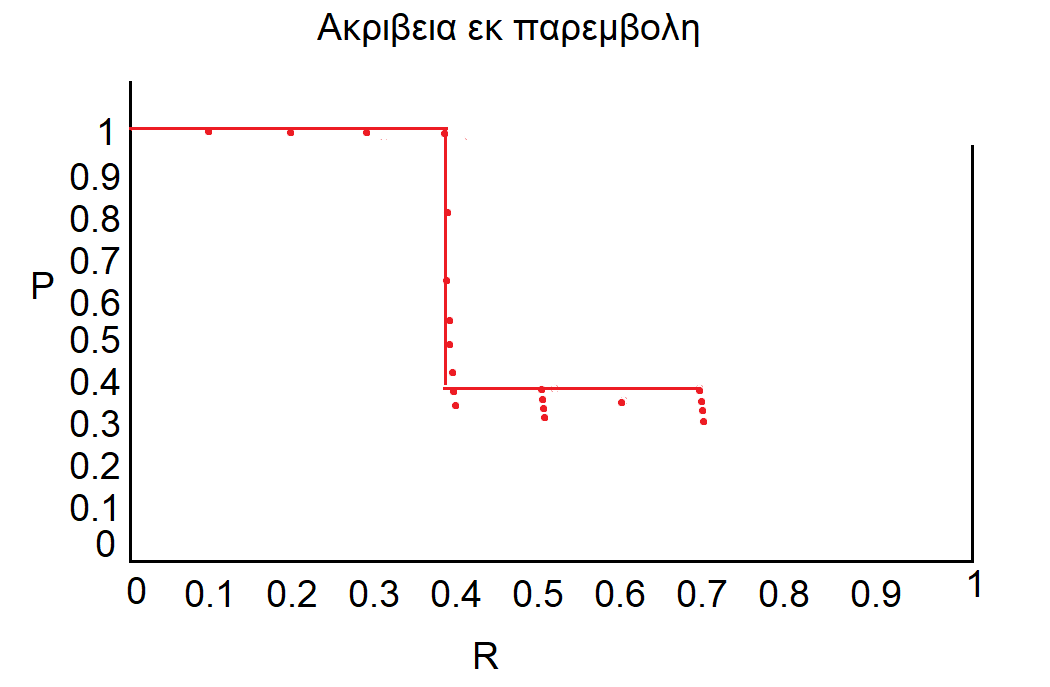
* **Software Designing**

****

Για τον σχεδιασμό καμπύλης 11 σημείων έχουμε:

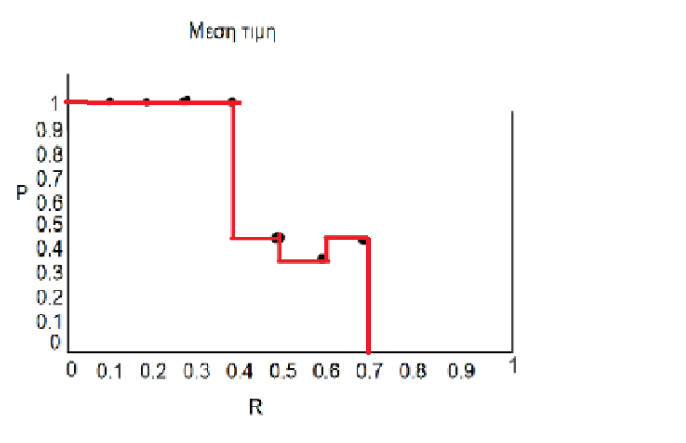


Επομένως, η ακρίβεια εκ παρεμβολής 11 σημείων θα είναι:



Για τον υπολογισμό της μέσης τιμής χρειαζόμαστε μόνο τις τιμές των recall και precision, ανα δεκαδικό ψηφίο (δηλ. 0,1, 0,2 κλπ)

Έτσι, το σχεδιάγραμμα για κάθε δεκαδικό ψηφίο του recall γίνεται:

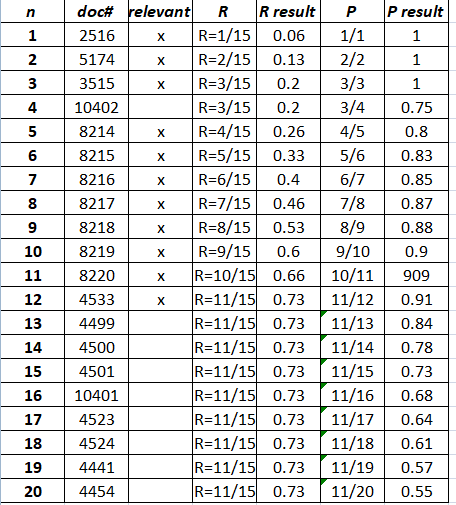


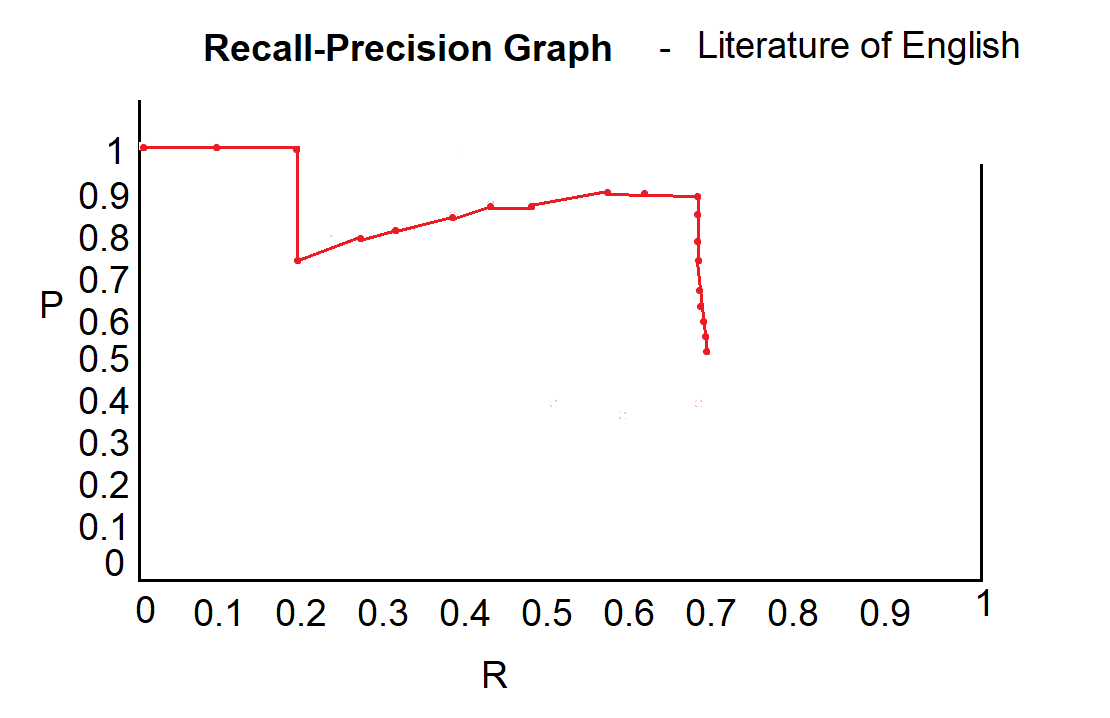
Άρα για τον υπολογισμό της μέσης τιμής έχουμε:

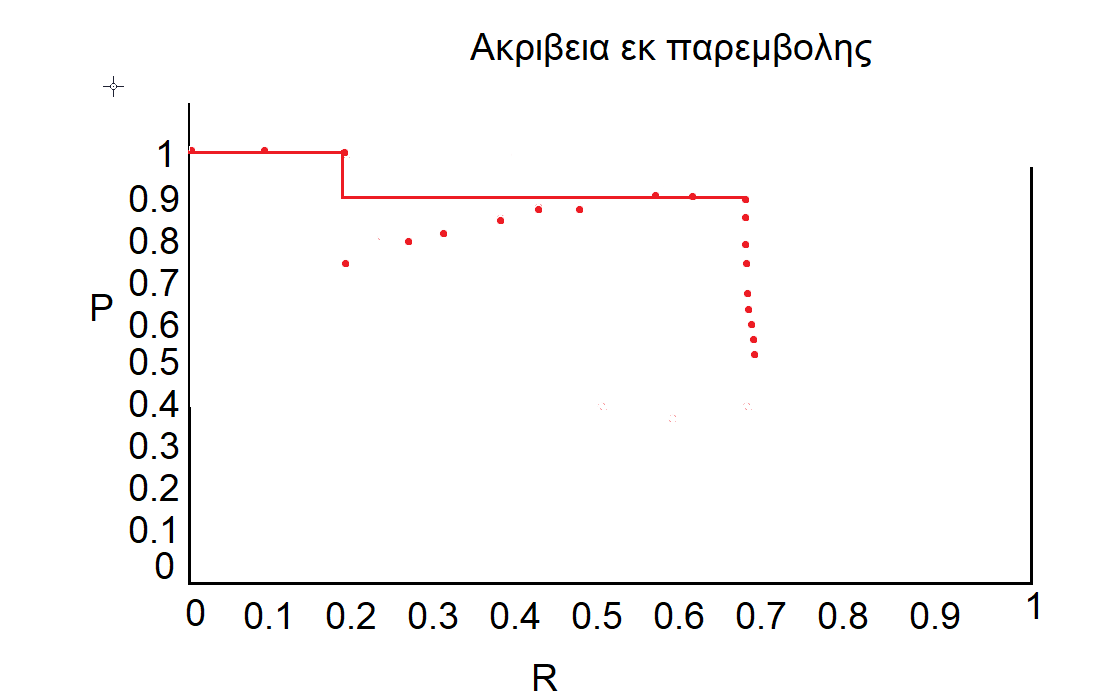
AVG = 1/11 (1+1+1+1+0.41+0.37+0.41 +0+0+0) = 0.380909

Με αντίστοιχο τρόπο κάνουμε:

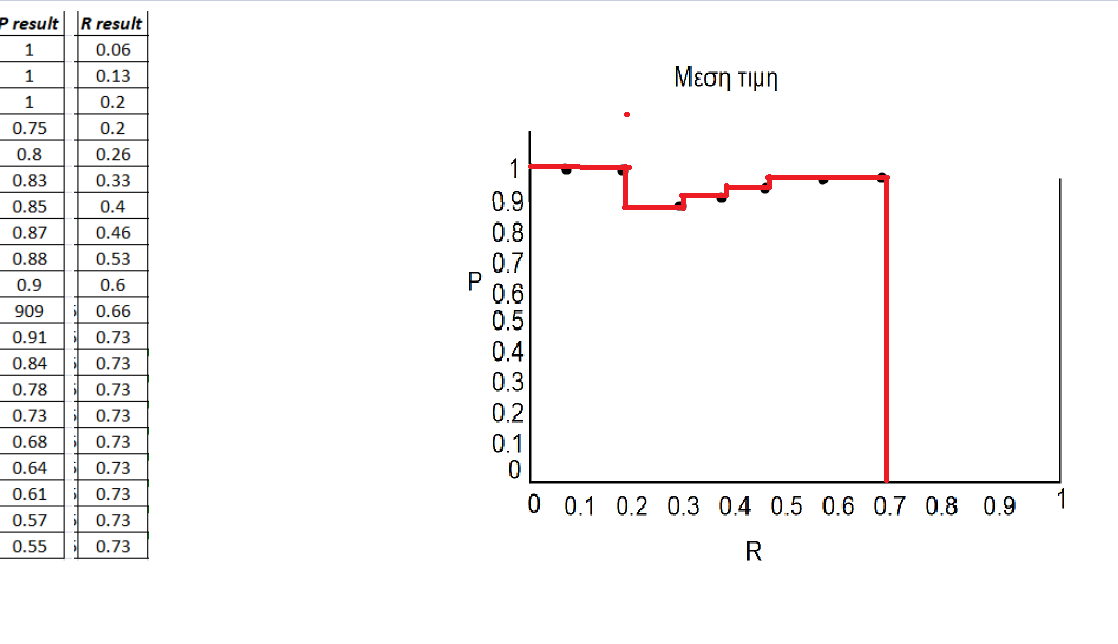
* **Literature of English**

****

****

****

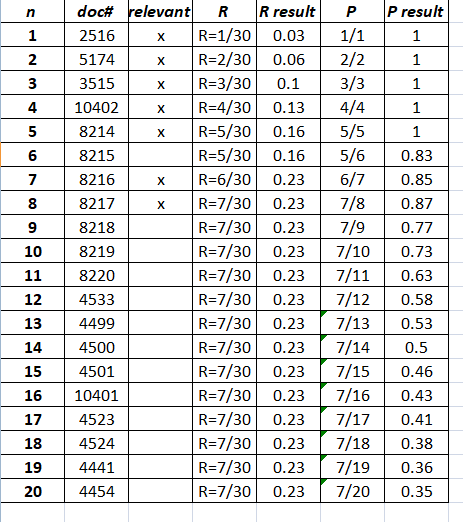
**Μέση τιμή:**

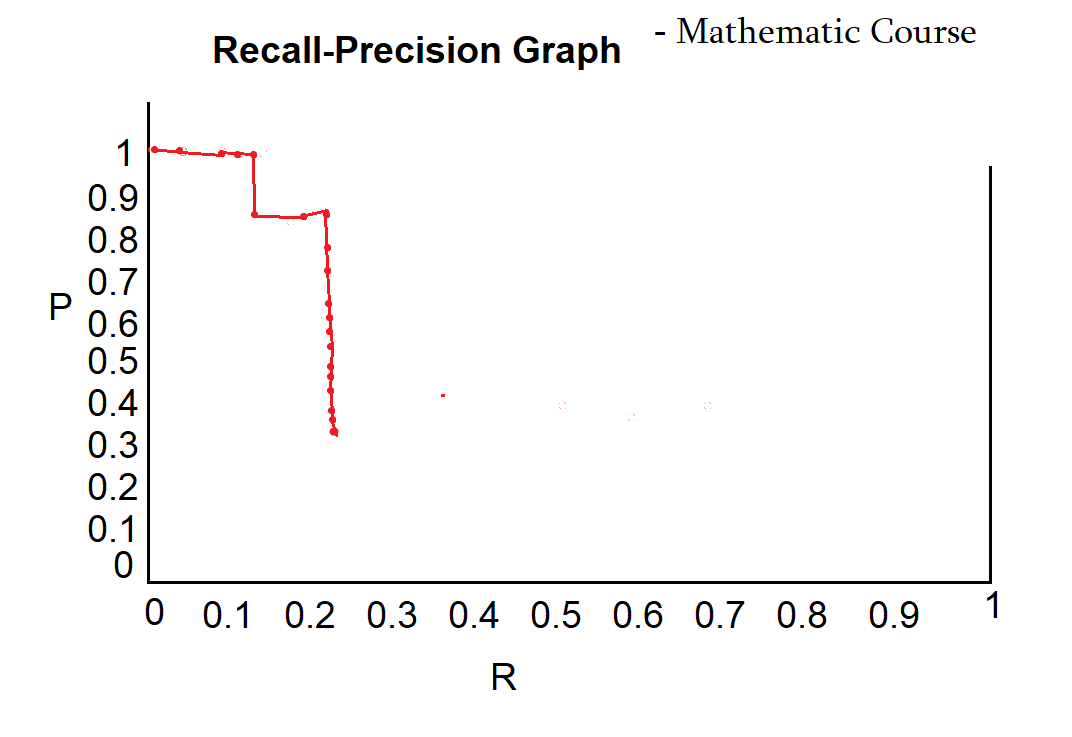
****

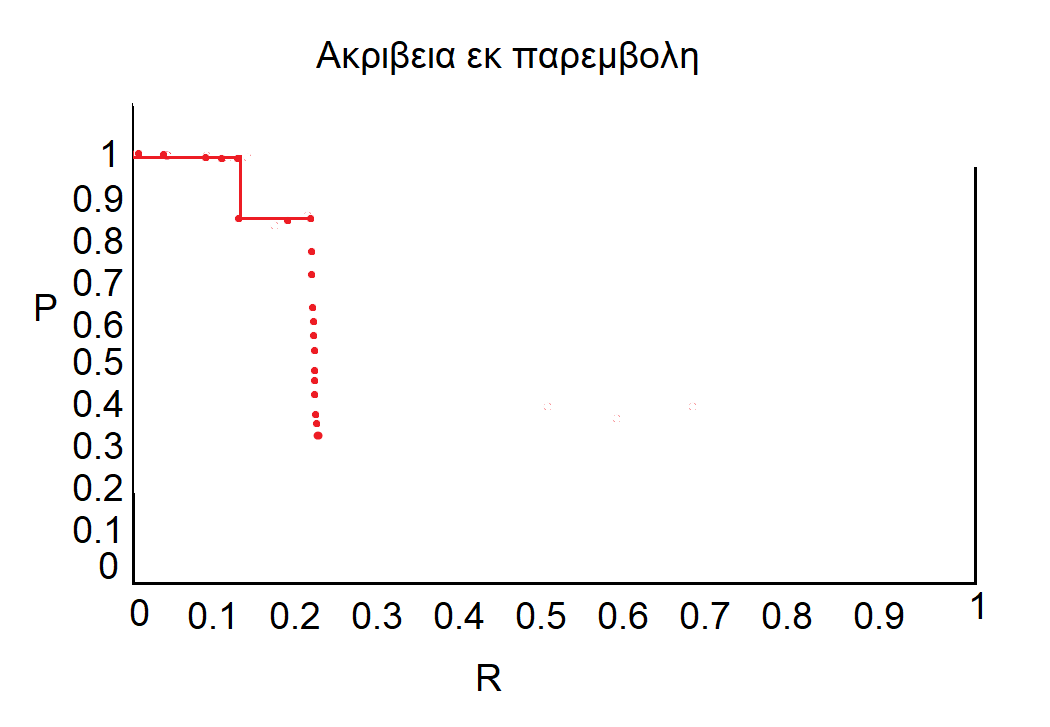
AVG = 1/11(1+1+1+0.83+0.85+0.88+0.9+0.91+0+0+0)=0.67

Και τέλος έχουμε:

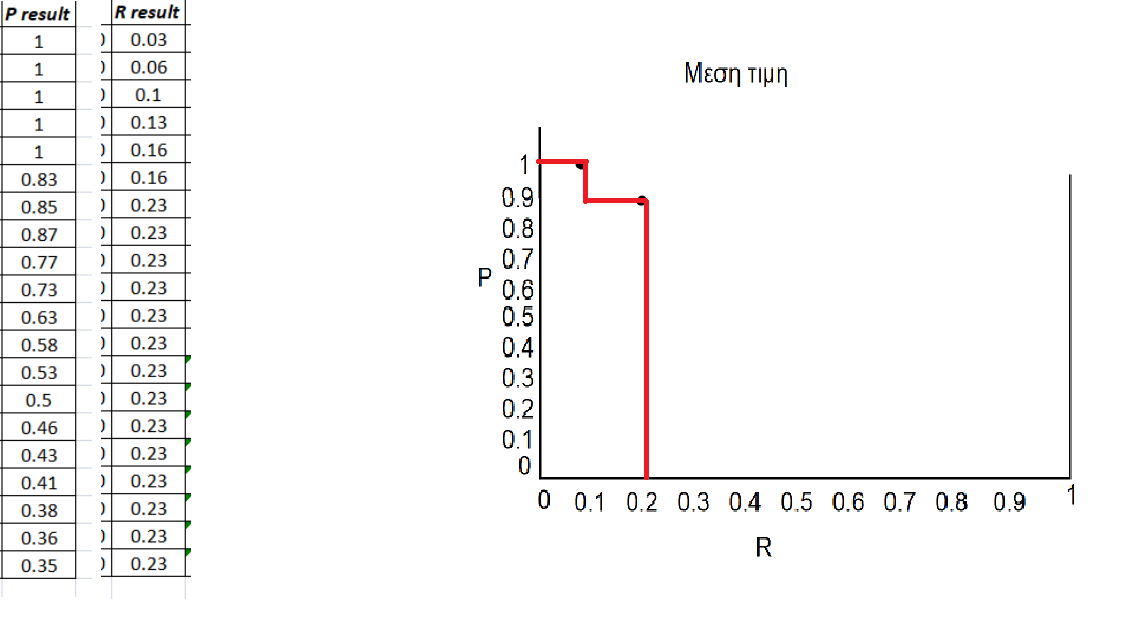
* **Mathematic Course**

****

****

****

Και η μέση τιμή είναι:



AVG= 1/11(1+0.85+0+0…) =0.16818

1. **c) K-Precision - K-recall**

Επειδη εχουμε Κ = 10, παίρνουμε τα Recall και Precision στο Ν = 10

1)Software Designing

K=10

k-Precision = 4/10 = 0.4

k-Recall = 4/10 = 0.4

2) Mathematics Course

k-Precision = 7/10 = 0.7

k-Recall = 7/30 = 0.23

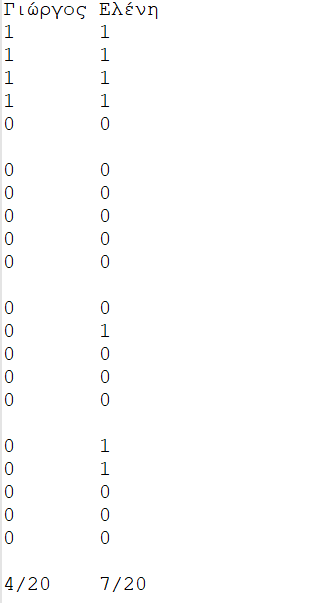
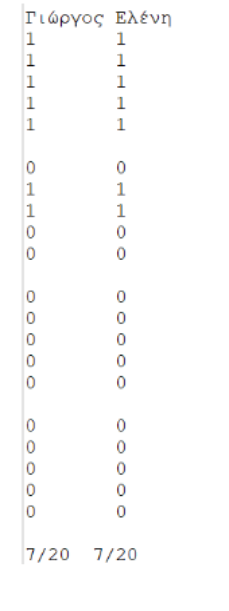
3) Literature of English

k-Precision = 9/10 = 0.9

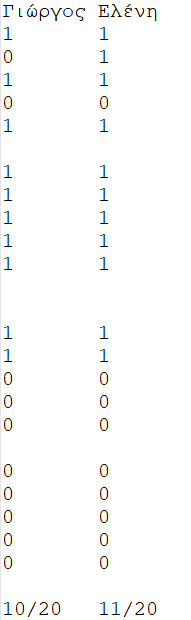
k-Recall = 9/15 = 0.6

**B.1**  Αξιολογώντας ατομικά την συνάφεια των ερωτημάτων ο πίνακας συμφωνιών των δύο κριτών είναι:

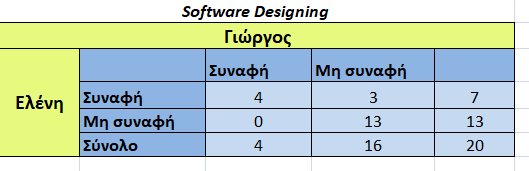
**1)Software Design 3)Mathematic Course**

** **

**2) Literature of English**

****

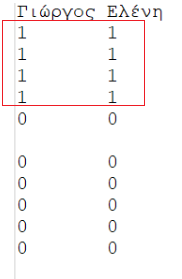
**B.2 1)** Το kappa για το πρώτο ερώτημα είναι :

****

* **P(A)= 4+13/20 = 17/20 = 0.85**
* **p(non relevant) = 13/20 \* 16/20 = 0.65 \* 0.8 = 0.52**
* **P(relevant) = 7/20 \* 4/20 = 0.35 - 0.2=0.15**
* **P(E) = 0.52 + 0.15 = 0.67**

**α)Kappa = (P(A) - P(E) ) / (1 - P(E))= (0.85 - 0.67) /(1-0.67) = 0.18/ 0.33 = 0.545455**

**β)** Η ακρίβεια και η ανάκληση αν υποθέσουμε ότι ένα έγγραφο είναι συναφές μόνο αν **συμφωνούν και οι δύο κριτές:**

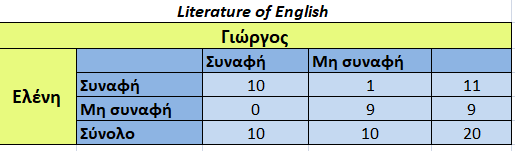
****

* **Precision = 4/20 = 0.2**
* **Recall = 4/10= 0.4**

**γ)** Η ακρίβεια και η ανάκληση αν υποθέσουμε ότι ένα έγγραφο είναι συναφές αρκεί **ένας από τους κριτές** να το θεωρεί συναφές:

* **Precision = 7/20 = 0.35**
* **Recall = 7/10= 0.7**

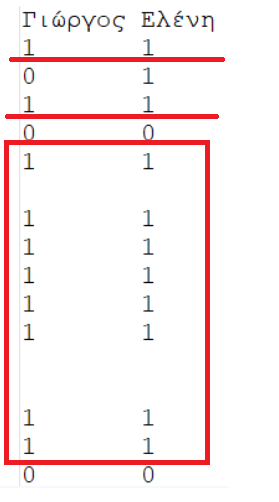
**2)** Το kappa για το δεύτερο ερώτημα είναι :



* **P(A)= 19/20 =0.95**
* **p(non relevant) = 10/20 \* 9/20= 0.5 \* 0.45 = 0.225**
* **P(relevant) = 10/20 \* 11/20 = 0.5 \* 0.55 = 0.275**
* **P(E)= 0.225 + 0.275 = 0.5**

**Kappa = (P(A) - P(E) ) / (1 - P(E))= (0.95-0.5)/(1-0.5)= 0.45/0.5 = 0.9**

**β)** Η ακρίβεια και η ανάκληση αν υποθέσουμε ότι ένα έγγραφο είναι συναφές μόνο αν **συμφωνούν και οι δύο κριτές:**

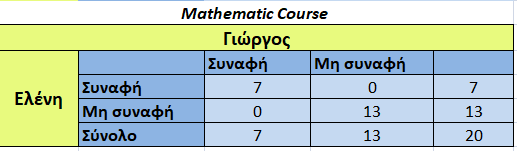
****

* **Precision = 10/20 = 0.5**
* **Recall = 10/15= 0.66**

**γ)** Η ακρίβεια και η ανάκληση αν υποθέσουμε ότι ένα έγγραφο είναι συναφές αρκεί **ένας από τους κριτές** να το θεωρεί συναφές:

* **Precision = 11/20 = 0.55**
* **Recall = 11/15= 0.733**

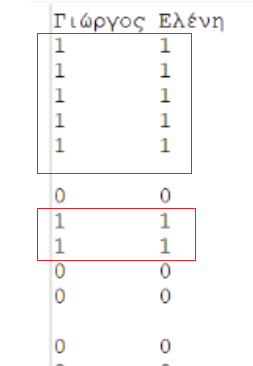
**3)** Το kappa για το τρίτο ερώτημα είναι :

****

* **P(A)= 20/20=1**
* **p(non relevant) = 13/20 \* 13/20= 0.65 \* 0.65 = 0.4225**
* **P(relevant) = 7/20 \* 7/20 = 0.35 \* 0.35 = 0.1225**
* **P(E)= 0.4225 + 0.1225 = 0.545**

**Kappa = (P(A) - P(E) ) / (1 - P(E))= (1-0.545)/(1-0.545)= 1**

**β)** Η ακρίβεια και η ανάκληση αν υποθέσουμε ότι ένα έγγραφο είναι συναφές μόνο αν **συμφωνούν και οι δύο κριτές:**

****

* **Precision = 7/20 = 0.35**
* **Recall = 7/30= 0.233**

**γ)** Η ακρίβεια και η ανάκληση αν υποθέσουμε ότι ένα έγγραφο είναι συναφές αρκεί **ένας από τους κριτές** να το θεωρεί συναφές:

* **Precision = 7/20 = 0.35**
* **Recall = 7/30= 0.233**