Διαδικτυακές βιβλιογρφικές πηγές (Μοριακής) Βιολογίας και Βιοπληροφορικής

BIO 230 - Εισαγωγή στην Υπολογιστική Βιολογία

Βασίλης Προμπονάς

Λευκωσία 2015-2019

Στόχοι

- Εξοικείωση των φοιτητών με έννοιες από το πεδίο της Ανάκτησης Πληροφορίας (π.χ. Boolean αναζήτηση, ακρίβεια–precision, ανάκληση–recall, αστοχία–fall-out/false positive rate).

- Εξοικείωση των φοιτήτών με τις βασικές λειτουργίες διαθέσιμες στη βάση δεδομένων PubMed®.

- Αυτόνομη διερεύνηση από τους φοιτητές άλλων σχετικών διαδικτυακών πηγών (π.χ. arXiv.org, bioRxiv, CiteSeerX, Google Scholar, Scopus), των κύριων μοντέλων δημοσίευσης (π.χ. ανοικτής πρόσβασης) και του ηλεκτρονικού καταλόγου της βιβλιοθήκης του Πανεπιστημίου Κύπρου.

Ανάκτηση Πληροφορίας – βασικά στοιχεία

Ορισμός:

Με τον όρο Ανάκτηση Πληροφορίας (Information Retrieval – IR) αναφερόμαστε στην “εύρεση υλικού (συνήθως εγγράφων) το οποίο είναι εν γένει μη δομημένο (συνήθως σε μορφή κειμένου) που ικανοποιεί μια ανάγκη πληροφόρησης μέσα από μεγάλες συλλογές (συνήθως σε ψηφιακή μορφή)” (Manning et al., 2008).

Παραδείγματα

1. Αναζήτηση ιστοσελίδων του διαδικτύου που αφορούν ένα συγκεκριμένο αντικείμενο (π.χ. Υπολογιστική Βιολογία).

2. Αναζήτηση στον υπολογιστή μας για αρχεία που περιέχουν πληροφορίες για ένα συγκεκριμένο μάθημα (π.χ. BIO230).

3. Αναζήτηση στα μηνύματα του ηλεκτρονικού μας ταχυδρομείου για να εντοπίσουμε ανακοινώσεις του μαθήματος BIO230.

Συλλογή κειμένων - corpus (μη-δομημένα έγγραφα)

K1: Η Βιοπληροφορική είναι ένα γνωστικό πεδίο που σχετίζεται με τη Βιολογία και την Πληροφορική.

K2: Τα πεδία της Γονιδιωματικής, και Επιγονιδιωματικής βασίζονται στη χρήση Βιοπληροφορικών τεχνικών για την ανάλυση μεγάλων όγκων δεδομένων.

K3: Στο μάθημα BIO001 δεν διδάσκονται αντικείμενα σχετικά με τη Βιοπληροφορική.

K4: Ο Γιαννής και ο Κωστής είναι αδέλφια.

K5: Αυτό το εξάμηνο παρακολουθώ το μάθημα BIO003 γιατί με ενδιαφέρει η Βιοπληροφορική, κλάδος που καταπιάνεται με την διαχείριση και ανάλυση βιολογικών δεδομένων.

Keyword search

Kw: “Βιοπληροφορική”

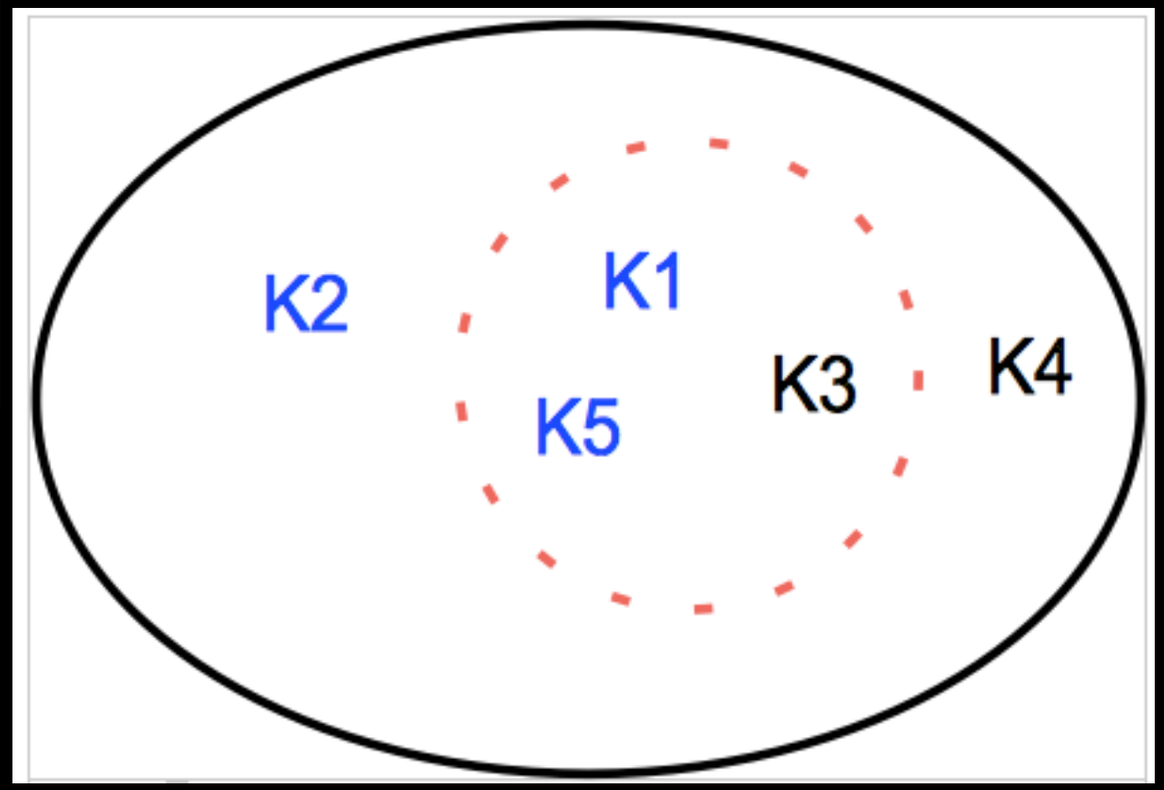
K1: Η Βιοπληροφορική είναι ένα γνωστικό πεδίο που σχετίζεται με τη Βιολογία και την Πληροφορική.

K2: Τα πεδία της Γονιδιωματικής, και Επιγονιδιωματικής βασίζονται στη χρήση Βιοπληροφορικών τεχνικών για την ανάλυση μεγάλων όγκων δεδομένων.

K3: Στο μάθημα BIO001 δεν διδάσκονται αντικείμενα σχετικά με τη Βιοπληροφορική.

K4: Ο Γιαννής και ο Κωστής είναι αδέλφια.

K5: Αυτό το εξάμηνο παρακολουθώ το μάθημα BIO003 γιατί με ενδιαφέρει η Βιοπληροφορική, κλάδος που καταπιάνεται με την διαχείριση και ανάλυση βιολογικών δεδομένων.



Πόσο καλά πήγαμε;

Αληθώς θετικά (True Positives): είναι τα κείμενα τα οποία **καλώς** ανακτήθηκαν από το ερώτημα Q που πραγματοποιήσαμε (K1 και K5).

Ψευδώς θετικά (False Positives): είναι τα κείμενα τα οποία **κακώς** ανακτήθηκαν από το Q (K3).

Αληθώς Αρνητικά (True Negatives): είναι τα κείμενα τα οποία **καλώς** δεν ανακτήθηκαν από το Q (K4).

Ψευδώς Αρνητικά (False Negatives): είναι τα κείμενα τα οποία **κακώς** δεν ανακτήθηκαν από το Q (K2).[3]

Πόσο καλά πήγαμε;

Αληθώς θετικά (True Positives): είναι τα κείμενα τα οποία **καλώς** ανακτήθηκαν από το ερώτημα Q που πραγματοποιήσαμε (K1 και K5).

Ψευδώς θετικά (False Positives): είναι τα κείμενα τα οποία **κακώς** ανακτήθηκαν από το Q (K3).

Αληθώς Αρνητικά (True Negatives): είναι τα κείμενα τα οποία **καλώς** δεν ανακτήθηκαν από το Q (K4).

Ψευδώς Αρνητικά (False Negatives): είναι τα κείμενα τα οποία **κακώς** δεν ανακτήθηκαν από το Q (K2).

ακρίβεια–precision:

$$precision = \frac{TP}{(TP + FP)} \quad (1)$$

ανάκληση–recall:

$$recall = \frac{TP}{(TP + FN)} \quad (2)$$

αστοχία–fall-out/false positive rate:

$$FPR = \frac{FP}{(FP + TN)} \quad (3)$$