



ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

Τμήμα Πληροφορικής

ΕΠΛ 660 – Ανάκτηση Πληροφοριών και Μηχανές Αναζήτησης

ΑΣΚΗΣΗ 1 – Ανάπτυξη Συστήματος Αναζήτησης Κριτικών Κρασιών «Dionysos: Wine review search engine»

Διδάσκων: Γιώργος Πάλλης / Υπεύθυνος Εργασίας: Παύλος Αντωνίου

Ημερομηνία Ανάθεσης: Τετάρτη 18/09/19

Ημερομηνία Παράδοσης: Τετάρτη 09/10/19 και ώρα 12:00 (21 μέρες)

(η λύση να υποβληθεί σε zip μέσω του Moodle)

<http://www.cs.ucy.ac.cy/courses/EPL660>

I. Στόχος

Ο Στόχος της παρούσας εργασίας είναι η εξοικειώσή σας με τα αντεστραμμένα ευρετήρια, τις δομές δεδομένων που χρησιμοποιούνται για την υλοποίησή τους και τη χρήση των δενδρικών δομών αναζήτησης.

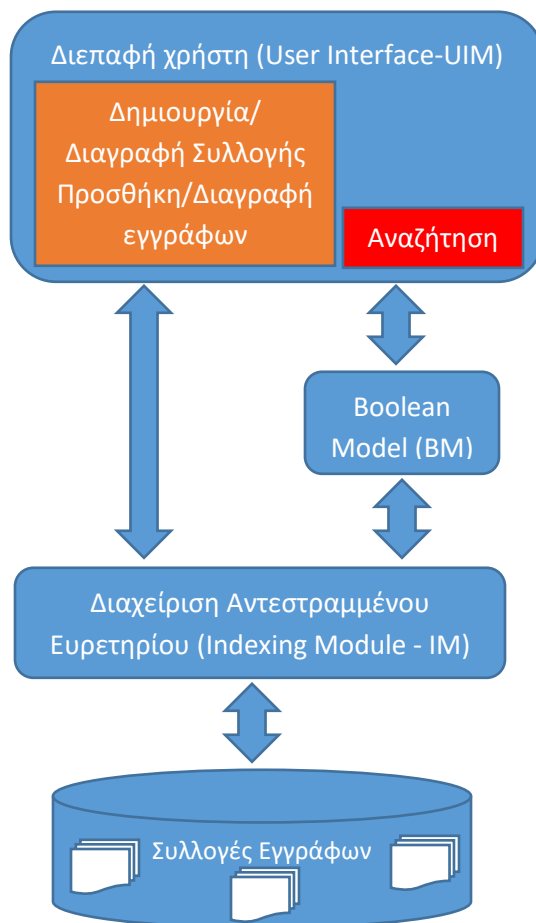
II. Εισαγωγή

Αναπτύξτε ένα σύστημα ανάκτησης κριτικών (reviews) για κρασιά με την ονομασία «Dionysos», το οποίο θα υποστηρίζει βασικές τεχνικές ανάκτησης πληροφοριών. Το σύστημα θα έχει τη δυνατότητα να επεξεργάζεται ερωτήματα λογικών εκφράσεων (boolean queries). Η διαχείριση των αρχείων κειμένου πραγματοποιείται με τη βοήθεια αντεστραμμένου ευρετηρίου (inverted index). Θα χρησιμοποιήσετε το σύστημα για να πραγματοποιείτε επερωτήσεις σε συλλογές αρχείων που περιέχουν κριτικές για κρασιά στα Αγγλικά.

Ένα σχετικό σύστημα φαίνεται εδώ: <https://www.wine-searcher.com/>.

III. Αρχιτεκτονική

Η αρχιτεκτονική του συστήματος δίνεται στο διπλανό σχήμα και εξηγείται από κάτω προς τα πάνω. Η εφαρμογή θα διαχειρίζεται ένα κατάλογο στο σύστημα αρχείων της



δευτερεύουσας μνήμης (σκληρός δίσκος) π.χ. C:/Dionysos (στα windows) ή \$HOME/Dionysos (στο Unix) που αρχικά θα είναι άδειος. Όλα τα έγγραφα που θα εισάγονται στην εφαρμογή θα αποθηκεύονται κάτω από αυτόν τον κατάλογο. Τα έγγραφα της εφαρμογής θα είναι οργανωμένα σε διαφορετικές συλλογές εγγράφων (document collections). Κάθε συλλογή εγγράφων θα πρέπει να αποθηκεύεται σε διαφορετικό κατάλογο μέσα στον κατάλογο της εφαρμογής.

Για κάθε συλλογή θα πρέπει να υπάρχει και ένα διαφορετικό αντεστραμμένο ευρετήριο (inverted index) το οποίο να χρησιμοποιείται για την οργάνωση της συλλογής. Το τμήμα του συστήματος που είναι υπεύθυνο για τη διαχείριση των αντεστραμμένων καταλόγων είναι το τμήμα IM (Indexing Module).

Το ευρετήριο θα ενημερώνεται κάθε φορά που εισάγεται ή διαγράφεται ένα έγγραφο από τη συλλογή. Το ευρετήριο θα είναι φορτωμένο στη πρωτεύουσα μνήμη (RAM) κατά τη διάρκεια που η εφαρμογή τρέχει. **Συνίσταται να αποθηκεύεται στο δίσκο προτού κλείσει (από το χρήστη) η εφαρμογή και να φορτώνεται ξανά μόλις εκκινήσει η εφαρμογή και έτσι θα κρατά τη γνώση για τα έγγραφα της συλλογής.** Εναλλακτικά (δεύτερη λογική), το ευρετήριο μπορεί να χάνεται κατά το κλείσιμο της εφαρμογής και να δημιουργείται ξανά σε κάθε εκκίνηση της εφαρμογής. Σε αυτή την περίπτωση δεν θα κρατά γνώση για τα έγγραφα της συλλογής. Όμως αν ακολουθήσετε η δεύτερη λογική, θα πρέπει να σβήνονται και τα έγγραφα και οι κατάλογοι των συλλογών που ήταν καταχωρημένα πριν κλείσει η εφαρμογή. Θα πρέπει να γράψετε στο αρχείο τεκμηρίωσης (βλέπε ενότητα Παράδοτέα) ποια λογική ακολουθήσατε.

Πάνω στο Indexing Module στηρίζεται το τμήμα που επεξεργάζεται ερωτήματα με βάση το Boolean μοντέλο ανάκτησης (τμήμα Boolean Module - BM).

Στο τελευταίο (υψηλότερο) επίπεδο διακρίνουμε το τμήμα UIM, διεπαφή χρήστη-συστήματος (user interface module), που είναι υπεύθυνο για την αλληλεπίδραση με το χρήστη για δημιουργία/διαγραφή συλλογής, προσθήκη/διαγραφή εγγράφων σε συλλογή και αναζήτηση όρου (term).

IV. Ζητούμενα

Να δημιουργήσετε το λογισμικό σύστημα διαχείρισης συλλογών εγγράφων, που περιλαμβάνει τα ακόλουθα υποσυστήματα:

1. Υποσύστημα Διεπαφής Χρήστη-Συστήματος (από γραμμή εντολών ή γραφική διεπαφή).

Προσφέρει τις ακόλουθες λειτουργίες:

- Δημιουργία νέας συλλογής. Θα πρέπει να δίνεται:
 - όνομα της συλλογής
- Διαγραφή συλλογής. Θα πρέπει να δίνεται:
 - όνομα της συλλογής
- Εισαγωγή αρχείου σε συλλογή. Θα πρέπει να δίνονται:
 - όνομα συλλογής
 - διαδρομή (path) του αρχείου στο δίσκο (αν χρησιμοποιείται η γραμμή εντολών) ή μέσω Browse (αν υπάρχει γραφικό περιβάλλον)

- Διαγραφή αρχείου από τη συλλογή. Θα πρέπει να δίνονται:
 - ο όνομα του αρχείου που θα διαγραφεί
 - ο όνομα συλλογής
- Δυνατότητα μαζικής εισαγωγής εγγράφων που βρίσκονται σε έναν κατάλογο σε συλλογή. Θα πρέπει να δίνονται:
 - ο όνομα συλλογής
 - διαδρομή (path) καταλόγου από τον οποίο θα διαβάζονται τα προς εισαγωγή αρχεία (αν χρησιμοποιείται η γραμμή εντολών) ή μέσω Browse (αν υπάρχει γραφικό περιβάλλον)
- Εκτύπωση περιεχομένων (αρχείων) συλλογής. Θα πρέπει να δίνεται:
 - ο όνομα συλλογής
- Εκτύπωση λεξικού της συλλογής. Θα πρέπει να δίνεται:
 - ο όνομα συλλογής
- Ανάκτηση του ευρετηρίου για μια συλλογή. Θα πρέπει να δίνεται:
 - ο όνομα συλλογής

Σε περίπτωση οποιουδήποτε σφάλματος, θα πρέπει να δίνεται αντίστοιχο μήνυμα προς τον χρήστη. Η χρήση γραφικής διεπαφής μειώνει τα σφάλματα καθώς θα εμφανίζει μόνο υποχρεωτικές συλλογές και καταχωρημένα αρχεία.

2. Υποσύστημα δημιουργίας και διαχείρισης αντεστραμμένου ευρετηρίου.

Κατά τη δημιουργία του αντεστραμμένου ευρετηρίου, το υποσύστημά σας θα πρέπει να πραγματοποιεί στελέχωση κειμένου (stemming) και απαλοιφή των λέξεων αποκλεισμού (stop words).

Για τη στελέχωση του Αγγλικού κειμένου (stemming) χρησιμοποιήστε τον αλγόριθμο Porter Stemmer. Τον κώδικα (αν θα υλοποιήσετε σε γλώσσα προγραμματισμού java) για τον αλγόριθμο Porter Stemmer μπορείτε να βρείτε στην ιστοσελίδα: <http://tartarus.org/martin/PorterStemmer/>.

Μία λίστα με λέξεις αποκλεισμού μπορείτε να την κατεβάσετε από τη διεύθυνση: <http://www.cs.ucy.ac.cy/courses/EPL660/exercises/stoplist.txt>.

Στο αντεστραμμένο ευρετήριο αποθηκεύονται εγγραφές με την ακόλουθη δομή:



όπου το positional posting list περιέχει τις ταυτότητες των αρχείων (π.χ. document id ή document name) που περιλαμβάνουν τον κάθε όρο και τα σημεία στα οποία εντοπίζεται ο κάθε όρος (term), δηλαδή σε ποια θέση (1^η θέση, 2^η θέση κτλ) βρίσκεται ο όρος μέσα στο αρχείο.

Ένας αποδοτικός τρόπος για τη δημιουργία του αντεστραμμένου ευρετηρίου είναι η χρήση δομής κατακερματισμού (hash) ή δομής δέντρου (tree) για την ταχεία αναζήτηση μιας λέξης και του

αντίστοιχου posting list της. Καλείστε να χρησιμοποιήσετε αποδοτικές δομές όπως π.χ. είναι οι κλάσεις HashMap ή TreeMap της γλώσσας Java.

Το παρόν υποσύστημα να προσφέρει τις ακόλουθες λειτουργίες:

- Δημιουργία αντεστραμμένου ευρετηρίου μιας συλλογής, το οποίο να περιλαμβάνει πληροφορίες όρων, τοποθεσίας και συχνότητας
- Δημιουργία δενδρικής δομής αναζήτησης
- Ενημέρωση περιεχομένων ευρετηρίου και δενδρικής δομής αναζήτησης όταν προκύπτουν αλλαγές στη σύνθεση της συλλογής

3. Υποσύστημα αναζήτησης όρων στο αντεστραμμένο ευρετήριο.

Υποστηρίζει τη δυνατότητα εισαγωγής και αναζήτησης ενός ή περισσότερων όρων στο λεξικό. Πιο συγκεκριμένα, το query για την αναζήτηση μπορεί να περιλαμβάνει:

A) Ένα (1) όρο ή

B) Δύο (2) όρους οι οποίοι να συνδέονται με τους λογικούς τελεστές AND, OR, NOT ή

Γ) Περισσότερους από δύο (2) όρους οι οποίοι να συνδέονται με τις λογικές πράξεις AND, OR, NOT (προαιρετικό - bonus)

Σε κάθε αναζήτηση ενός όρου, το πρόγραμμα επιστρέφει, εκτός από τα αρχεία (όνομα και περιεχόμενο) στα οποία περιλαμβάνεται ο/οι αναζητούμενος/οι όρος/οι, και το χρόνο που χρειάστηκε η πραγματοποίηση της αναζήτησης.

4. Αρχεία Συλλογής (dataset).

Για σκοπούς επίδειξης της παρούσας άσκησης θα εντάξετε στο σύστημά σας δεδομένα από τη συλλογή αρχείων του ιστοχώρου [Kaggle](https://www.kaggle.com/datasets/winemag-data-first150k-csv). Πιο συγκεκριμένα, το αρχείο με όνομα [winemag-data_first150k.csv](https://www.kaggle.com/datasets/winemag-data-first150k-csv) που θα χρησιμοποιήσουμε περιέχει πάνω από 150 χιλιάδες κριτικές (reviews) κρασιών με 11 χαρακτηριστικά-πεδία (features) ανά κριτική. Πιο κάτω βλέπετε τις πρώτες 4 γραμμές του αρχείου:

```
id,country,description,designation,points,price,province,region_1,region_2,variety,winery
0,US,"This tremendous 100% varietal wine hails from Oakville and was aged over three years in oak. Juicy red-cherry fruit and a compelling hint of caramel greet the palate, framed by elegant, fine tannins and a subtle minty tone in the background. Balanced and rewarding from start to finish, it has years ahead of it to develop further nuance. Enjoy 2022-2030.",Martha's Vineyard,96,235.0,California,Napa Valley,Napa,Cabernet Sauvignon,Heitz
1,Spain,"Ripe aromas of fig, blackberry and cassis are softened and sweetened by a slathering of oaky chocolate and vanilla. This is full, layered, intense and cushioned on the palate, with rich flavors of chocolaty black fruits and baking spices. A toasty, everlasting finish is heady but ideally balanced. Drink through 2023.",Carodorum Selección Especial Reserva,96,110.0,Northern Spain,Toro,,Tinta de Toro,Bodega Carmen Rodríguez
2,US,"Mac Watson honors the memory of a wine once made by his mother in this tremendously delicious, balanced and complex botrytised white. Dark gold in color, it layers toasted hazelnut, pear compote and orange peel flavors, reveling in the succulence of its 122 g/L of residual sugar.",Special Selected Late Harvest,96,90.0,California,Knights Valley,Sonoma,Sauvignon Blanc,Macauley
```

Η πρώτη γραμμή του αρχείου περιέχει τον τίτλο του κάθε πεδίου, για κάθε πεδίο κριτικής. Κάθε κριτική έχει τη δική της ταυτότητα (id), χώρα προέλευσης κρασιού (country), περιγραφή κριτικής (description), ονομασία κρασιού (designation), βαθμοί αξιολόγησης (points), τιμή κρασιού (price), επαρχία (province), περιοχή 1 (region_1), περιοχή 2 (region_2), ποικιλία (variety), οινοποιείο (winery). Για κάθε γραμμή του αρχείου (εκτός από την πρώτη) δημιουργήσαμε ένα αρχείο που έχει όνομα ίδιο με την ταυτότητα της κάθε κριτικής (π.χ. 0.txt, 1.txt, κτλ.). Το σύνολο των 150935 αρχείων μπορείτε να το κατεβάσετε από το <http://www.cs.ucy.ac.cy/courses/EPL660/exercises/winedata.zip>

Για τους σκοπούς της παρούσας άσκησης, από κάθε κριτική που θα εισάγεται στο σύστημά σας θα χρησιμοποιήσετε μόνο το κείμενο που περιέχεται στο πεδίο description για να δημιουργήσετε το ανάστροφο ευρετήριο.

V. Παραδοτέα

Θα πρέπει να παραδώσετε:

- πηγαίο κώδικα του συστήματός σας με σχόλια (αν υλοποιήσετε την άσκησή σας σε Eclipse, να παραδώσετε ολόκληρο το project),
- τεκμηρίωση (ξεχωριστό αρχείο .docx) όπου να φαίνεται ο τρόπος σκέψης σας για την ανάπτυξη του συστήματος, στο οποίο να αναφέρονται οι διάφορες κλάσεις που αναπτύξατε με τη λειτουργία τους, οι αλγόριθμοι και οι τεχνικές που χρησιμοποιήθηκαν.

Συμπιέστε τον κώδικά (Eclipse project folder) και το αρχείο τεκμηρίωσης σε ένα αρχείο zip. Η υποβολή του αρχείου zip να γίνει μέσω του Moodle. Παρακαλώ όπως υποβάλετε ΜΟΝΟ ΕΝΑ zip file.

ΔΕΝ ΥΠΑΡΧΕΙ ΠΕΡΙΟΡΙΣΜΟΣ ΣΤΗ ΓΛΩΣΣΑ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ ΓΙΑ ΤΗΝ ΥΛΟΠΟΙΗΣΗ.