ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
ΗΥ463 – Συστήματα Ανάκτησης Πληροφοριών

Εργασία:
Στελεχωτής Ελληνικής Γλώσσας
(Αξιολόγηση, Επέκταση, Τεκμηρίωση)

Εισαγωγή

Στο μάθημα έχουμε ήδη δει τι είναι και τι εξυπηρετεί ένας στελεχωτής (stemmer). Για την ελληνική γλώσσα υπάρχουν ελάχιστοι στελεχωτές οι οποίοι μάλιστα δεν έχουν ούτε αξιολογηθεί ούτε συγκριθεί μεταξύ τους. Την προηγούμενη χρονιά, και στα πλαίσια του μαθήματος, δημιουργήθηκε ένας στελεχωτής για την ελληνική γλώσσα που έχει ενσωματωθεί στο Μίτο. Εμπίπτει στην κατηγορία των Affix Removal Stemming Algorithms (στην οποία εμπίπτει και ο στελεχωτής του Porter για την Αγγλική γλώσσα), και άρα βασίζεται στην εφαρμογή κανόνων της ελληνικής γραμματικής. Περισσότερες πληροφορίες μπορείτε να βρείτε στο Wiki[1].

Σκοπός της εργασίας αυτής είναι

- η αξιολόγηση του στελεχωτή,
- η βελτίωση του (βάσει και της αξιολόγησης),
- η ενσωμάτωση του στο Μίτο, και
- η υλοποίηση διαφόρων συμπληρωματικών λειτουργιών.

Συλλογή Αξιολόγησης

Η αξιολόγηση πρέπει να γίνει με τυπικό τρόπο και τα αποτελέσματα της να μπορούν να επαναληφθούν. Αρχικά, καλείστε να δημιουργήσετε μια συλλογή αξιολόγησης. Η συλλογή θα αποτελείται από λέξεις οργανωμένες σε δέσμες (blocks). Κάθε δέσμη θα περιέχει εκείνες τις λέξεις που θα θέλαμε να έχουν την ίδια ρίζα, όπως φαίνεται στο παρακάτω παράδειγμα.

αναδιατάσσω
αναδιάταξη
αναδιέταξα

βρίσκω
έβρισκα

επιστήμη
επιστήμονας

Τα παραπάνω πρέπει να είναι αποθηκευμένα σε ένα .txt αρχείο. Επίσης είναι καλό να υπάρχει ένα πρόγραμμα το οποίο να ελέγχει την εγκυρότητα μιας συλλογής και να επιδιορθώνει τυχόν προβλήματα. Συγκεκριμένα, μια λέξη πρέπει να εμφανίζεται σε μία μόνο δέσμη. Αν υπάρχουν δέσμες που έχουν έστω και μία κοινή λέξη πρέπει να συγχωνεύονται σε μια δέσμη. Σημείωση: ένα τέτοιο πρόγραμμα θα μας επέτρεπε να συγχωνεύουμε πολλές συλλογές αξιολόγησης και άρα να φτιάξουμε συνεργατικά μια μεγάλη συλλογή.

[1] http://groogle.csd.uoc.gr/apache2-default/index.php/Stemmer

Εφαρμογή Στελεχωτή στη Συλλογή Αξιολόγησης

Για να αξιολογήσετε ένα στελεχωτή βάσει της συλλογής αξιολόγησης, θα πρέπει να φτιάξετε κατάλληλο πρόγραμμα το οποίο θα τρέχει το στελεχωτή πάνω στις λέξεις της συλλογής και θα δημιουργεί ένα νέο αρχείο, όπως αυτό που περιγράφηκε παραπάνω, στο οποίο οι λέξεις της αρχικής συλλογής θα ομαδοποιούνται με βάση την υπολογιζόμενη από το στελεχωτή ρίζα.

Βαθμολόγηση Στελεχωτή

Μπορούμε να βαθμολογήσουμε το στελεχωτή συγκρίνοντας τα δύο αρχεία: το αρχείο της συλλογής αξιολόγησης με αυτό που παρήγαγε ο στελεχωτής. Παρακάτω περιγράφουμε μια μετρική που θα μπορούσατε να χρησιμοποιήσετε. Βέβαια σας προτείνουμε να σκεφτείτε, να περιγράψετε και να εφαρμόσετε και άλλα μέτρα αξιολόγησης.

Έστω W το σύνολο των λέξεων της συλλογής αξιολόγησης. Επίσης, έστω $A= \{A_1, A_2, ..., A_k\}$ μία διαμέριση (partition) του W και $B= \{B_1, B_2, ..., B_m\}$ μια άλλη διαμέριση του W[2]. Αν μία λέξη $w \in A_i$ και $w \in B_j$ θα γράφουμε $b_A(w) = i$ και $b_B(w) = j$. Για να συγκρίνουμε δύο διαμερίσεις ορίζουμε την απόσταση ενός ζευγαριού λέξεων w και w' ως:

Αν

$b_A(w) = b_A(w')$ *και* $b_B(w) = b_B(w')$

ή $b_A(w) \neq b_A(w')$ *και* $b_B(w) \neq b_B(w')$

τότε

$dist_{A,B}(w, w') = 0$

αλλιώς

$dist_{A,B}(w, w') = 1.$

Η απόσταση δηλαδή είναι ίση με 1 αν τα w και w' ανήκουν στην ίδια δέσμη (block) στο A και όχι στο B, ή το αντίστροφο. Μπορούμε να ορίσουμε τη συνολική απόσταση των διαμερίσεων A και B ως προς το W αθροίζοντας τις αποστάσεις όλων των πιθανών ζευγαριών και κανονικοποιώντας το αποτέλεσμα:

$$Dist(A,B) = \frac{\sum_{i=1}^{|W|-1}\sum_{j=i+1}^{|W|} dist_{A,B}(w_i, w_j)}{\frac{|W|(|W|-1|)}{2}}$$

Όσο πιο μικρή η απόσταση (ιδανικά ίση με 0), τόσο πιο καλή η απόδοση του στελεχωτή στη συγκεκριμένη συλλογή.

Μετρήσεις, Πειράματα, Βελτιώσεις

Έχοντας όλα τα παραπάνω καλείστε να αξιολογήσετε το στελεχωτή (χρησιμοποιώντας μία ή πολλές συλλογές αξιολόγησης). Συνάμα, και έχοντας κατανοήσει τους κανόνες που υποστηρίζει ο υπάρχων στελεχωτής, προσπαθήσετε να τους βελτιώσετε. Παράλληλα βρείτε πιθανούς νέους κανόνες, υλοποιήστε τους και αξιολογείστε τους. Αναδομήστε κατάλληλα τον κώδικα ώστε η διαχείριση των

[2] Μια διαμέριση ενός συνόλου X είναι μια οικογένεια συνόλων X1, .. ,XN, τ.ω. η ένωση τους μας δίνει το X και τα σύνολα αυτά είναι ανά δύο ξένα μεταξύ τους, ήτοι X = X1 ∪…∪XN, και αν i≠j τότε Xi∩ Xj = ∅.

κανόνων να είναι εύκολη και ευέλικτη. Επίσης μπορείτε να συγκρίνετε τον τελικό στελεχωτή με άλλους.

Θα πρέπει να παραδώσετε γραπτή αναφορά που να περιγράφει όλα τα παραπάνω. Συνάμα καλείστε να συντηρείτε τη σελίδα του Wiki που αφορά στο στελεχωτή (καθαρισμός, προσθήκη χρήσιμου υλικού κλπ) η οποία θα είναι κοινή για όλες τις ομάδες

Βελτίωση Κώδικα

Πέραν της βελτίωσης του τρόπου διαχείρισης κανόνων, βελτιώστε τον κώδικα που θα σας δοθεί, τεκμηριώστε τον κατάλληλα και επεκτείνετέ τον. Συστήνεται η χρήση του **JUnit** για τον έλεγχο της ορθότητας του κώδικά σας. Ένα πολύ μικρό και γρήγορο tutorial για JUnit tests, μπορείτε να βρείτε εδώ[3]. Επίσης ενδεικτικά tutorials για JUnits χρησιμοποιώντας Eclipse[4] και NetBeans[5].

Σημείωση: Θα μπορούσατε να χρησιμοποιήσετε JUnit tests ακόμα και για να βοηθηθείτε στην εύρεση αποτελεσματικών κανόνων στελέχωσης.

Συμπληρωματικές Λειτουργίες

Αυτή τη στιγμή στο ευρετήριο του Μίτου αποθηκεύονται μόνο οι ρίζες των λέξεων (όπως προκύπτουν από το στελεχωτή). Εκ τούτου, οι λειτουργίες που προβάλλουν στο χρήστη όρους που επιστρέφονται από τη βάση (π.χ. στο clustering και στο query expansion) αποτελούν τις ρίζες των λέξεων που έχει βρει ο στελεχωτής, οι οποίες δεν είναι κανονικές λέξεις και άρα είναι ακατάλληλες για παρουσίαση (όπως πρέπει να έχετε ήδη διαπιστώσει από τη χρήση του Μίτου).

Για το λόγο αυτό καλείστε κατά τη διάρκεια της διαδικασίας ευρετηρίασης, να αποθηκεύετε εκτός από τη υπολογισμένη από το στελεχωτή ρίζα και μια ενδεικτική κανονική (μη-στελεχωμένη) λέξη. Για την επιλογή της κανονικής λέξης μπορείτε να υποστηρίξετε διάφορες επιλογές όπως: επιλογή της πιο μικρής σε μέγεθος λέξης, επιλογή της πιο συχνά χρησιμοποιούμενης λέξης, κα.

Θα απαιτηθεί και αλλαγή του σχήματος της βάσης, και πιο συγκεκριμένα η προσθήκη ενός νέου πεδίου στον πίνακα word που θα κρατάει τη μη-στελεχωμένη λέξη (δείτε στο Wiki[6]). Επίσης θα πρέπει να υλοποιηθεί η σχετική μέθοδος στον Indexer, η οποία θα μας επιστρέφει τη μη-στελεχωμένη λέξη ενός όρου.

BONUS

Υλοποίηση άλλων προσεγγίσεων στελέχωσης και σύγκριση τους με τον υπάρχοντα στελεχωτή. Για παράδειγμα θα μπορούσατε να πειραματιστείτε με την τεχνική Successor Variety. Σε αυτήν την περίπτωση δημιουργείστε ένα trie για τις λέξεις του λεξιλογίου (αυτό θα διευκολύνει την εφαρμογή της τεχνικής). Και να θυμάστε: ο καλύτερος στελεχωτής θα γίνει ο επίσημος στελεχωτής του Μίτου.

[3] http://www.jaredrichardson.net/articles/junit-tutorial.html
[4] http://www.vogella.de/articles/JUnit/article.html
[5] http://www.fsl.cs.sunysb.edu/~dquigley/cse219/index.php?it=netbeans&tt=junit&pf=y
[6] http://groogle.csd.uoc.gr/apache2-default/index.php/Indexer_DBMS

Λοιπές Οδηγίες και Υλικό

Πηγές:

Διαβάστε το σχετικό υλικό από τις διαλέξεις και τα βιβλία του μαθήματος. Βρείτε και δείτε σχετικά άρθρα. Μερικοί σύνδεσμοι ακολουθούν.

- http://sais.se/mthprize/2007/ntais2007.pdf
- http://iit.demokritos.gr/ bgat/RIAO2000.pdf
- http://clef.isti.cnr.it/2005/working notes/workingnotes2005/tomlinson05.pdf
- http://citeseer.comp.nus.edu.sg/cache/papers/cs/18663/ftp:zSzzSzaiolos.iit.demokritos.grzSzpubzSzskelzSzpaperszSzCOMLEX-2000.pdf/rule-based-named-entity.pdf
- http://hmi.ewi.utwente.nl/dir2006/dir2006.pdf\#page=39
- http://citeseer.ist.psu.edu/fuller98conflationbased.html
- http://acl.ldc.upenn.edu/W/W07/W07-0734.pdf

Επίσης

- Νεοελληνική Γραμματική του Μανόλη Τριανταφυλλίδη
- Το Λεξικό της Νέας Ελληνικής Γλώσσας του Γ. Μπαμπινιώτη
- Μείζον ελληνικό λεξικό των Τεγοπουλου – Φυτράκη

Καλή εργασία!