# ΑΣΚΗΣΗ

# ΜΕΡΟΣ Α′

# Σεπτέμβριος 2010

## Δεικτοδότηση Συλλογής Κειμένων σε Ανεστραμμένο Ευρετήριο

Τα ανεστραμμένα αρχεία αποτελούν μια βασική μορφή ευρετηρίου και μας επιτρέπουν να εντοπίσουμε ποια κείμενα της συλλογής που δεικτοδοτείται περιέχουν συγκεκριμένους όρους. Σκοπός της άσκησης είναι να επεξεργαστείτε μία συλλογή κειμένων η οποία είναι διαθέσιμη στο site του μαθήματος (http://www.dblab.upatras.gr/gr/GlwssikiTexnologia.html) και τη δεικτοδοτήσετε σε ένα ανεστραμμένο ευρετήριο που θα υλοποιήσετε. Προκειμένου να αξιολογήσετε την απόδοση του ευρετηρίου σας θα υλοποιήσετε έναν μηχανισμό υποβολής ερωτημάτων στα δεικτοδοτημένα κείμενα της συλλογής, μέσω του οποίου θα υποβάλετε ερωτήματα προς το ευρετήριο και θα ανακτάτε τα κείμενα της συλλογής που σχετίζονται με αυτά (δηλ. περιέχουν τους όρους των ερωτημάτων).

Η υλοποίησή σας θα πρέπει να ανταποκρίνεται στις ανάγκες πραγματικών συστημάτων δεικτοδότησης, δηλ. το ευρετήριο να φορτώνεται στη μνήμη, ώστε να μην υπάρχει ανάγκη ανακατασκευής ευρετηρίου κάθε φορά που υποβάλλονται ερωτήματα σε αυτό.

Διαβάσετε προσεχτικά την εκφώνηση και απαντήσετε με σαφήνεια στα ερωτήματα, χρησιμοποιώντας μία από τις ακόλουθες γλώσσες προγραμματισμού: C# ή Python.

## Βήματα Υλοποίησης

**1. Tokenization (5%).** Κατεβάστε τη συλλογή κειμένων Wikipedia.zip (συλλογή άρθρων από την αγγλική Wikipedia). Εφαρμόστε σε κάθε κείμενο τεχνικές tokenization προκειμένου να το μορφοποιήσετε έτσι ώστε στο tokenized κείμενο που θα προκύψει να εμφανίζεται μία λέξη ανά γραμμή ή ένα σημείο στίξης ανά γραμμή. Ως σημεία στίξης θεωρήστε όσα αναγράφονται στον παρακάτω πίνακα.

Αποθηκεύστε την tokenized μορφή κάθε κειμένου σε αντίστοιχο αρχείο.

Προσοχή, όταν χρησιμοποιείται απόστροφος για να δηλώσει γενική κτητική π.χ. John’s το tokenization θα πρέπει να έχει τη μορφή:

*John*
*’*
*s*

Επίσης θα πρέπει να προσέξετε την ορθή αναγνώριση αριθμών (ύπαρξη τελείας, κόμματος σαν υποδιαστολή).

***Πίνακας σημείων στίξης***

| | | | | |
|---|---|---|---|---|
| . | ‘ | ’ | < | > |
| { | } | [ | ] | ( |
| ) | \ | / | ? | ; |
| : | « | » | ~ | ` |
| “ | ” | ! | @ | # |
| $ | % | ^ | & | * |
| - | + | … | \| | |

**2. Μορφοσυντακτική Ανάλυση (10%)**. Σχολιάστε μορφοσυντακτικά τις λέξεις του κάθε tokenized κειμένου. Για το μορφοσυντακτικό σχολιασμό χρησιμοποιήστε κάποιον από τους προτεινόμενους PoS-Taggers (ανάλογα με το περιβάλλον υλοποίησης που έχετε επιλέξει) που θα κατεβάσετε από το site του μαθήματος. Στο τέλος του βήματος 2 κάθε κείμενο της συλλογής θα διαθέτει μορφοσυντακτικό σχολιασμό (PoS-tags) για κάθε λέξη που περιέχει. Τα μορφοσυντακτικά σχολιασμένα κείμενα της συλλογής θα πρέπει να αποθηκευτούν σε βοηθητικό-ενδιάμεσο αρχείο για μελλοντική χρήση.

*Σημείωση 2.1*: για κάθε λέξη του κειμένου που αναγνωρίζει ο tagger επιστρέφει εκτός από το PoS-tag που περιγράφει τη γραμματική κατηγορία της λέξης και το λήμμα (δηλ., τον πρώτο κλιτικό τύπο) της λέξης. Η διαδικασία αυτή ονομάζεται λημματοποίηση (lemmatization), την εκτελεί by default ο Tagger και ουσιαστικά συγχωνεύει σε έναν τύπο (λήμμα) όλες τις κλιτικές μορφές μιας λέξης. Για παράδειγμα, το λήμμα *παιδί* συγχωνεύει τους κλιτικούς τύπους *παιδιού*, *παιδιά*, *παιδιών*.

**3. Αναπαράσταση κειμένων στο Μοντέλο Διανυσματικού Χώρου (15%).** Για να αναπαραστήσετε το περιεχόμενο κάθε κειμένου ως διάνυσμα θα χρησιμοποιήσετε τα μορφοσυντακτικά σχολιασμένα κείμενα (που θα προκύψουν από το βήμα 2) και αρχικά θα αφαιρέσετε τους τερματικούς όρους (stop-words) από κάθε κείμενο. Οι τερματικοί όροι είναι λέξεις που δεν έχουν σημασιολογικό περιεχόμενο και εμφανίζονται σε όλα τα κείμενα, με αποτέλεσμα να μην αποτελούν χρήσιμους όρους δεικτοδότησης.

*Σημείωση 3.1:* στο παρακάτω link http://www.infogistics.com/tagset.html θα βρείτε δύο πίνακες, έναν με τα PoS tags για open class categories και έναν με τα PoS tags για closed class categories. Τα open class categories είναι γραμματικές κατηγορίες των λέξεων που έχουν σημασιολογικό περιεχόμενο και άρα τις χρειαζόμαστε. Αντίθετα, τα closed class categories είναι γραμματικές κατηγορίες για λέξεις άνευ σημασιολογικού περιεχομένου, δηλ., stop-words. Συνεπώς, για να εξαλείψετε τους τερματικούς όρους από κάθε μορφοσυντακτικά σχολιασμένο κείμενο της συλλογής θα πρέπει να αφαιρέσετε τις λέξεις στις οποίες έχει ανατεθεί ένα closed class category tag.

Αφού αφαιρέσετε τους τερματικούς όρους από κάθε κείμενο της συλλογής, στη συνέχεια για κάθε **μοναδικό λήμμα** του κειμένου θα μετρήσετε τη συχνότητα εμφάνισής του στο κείμενο (πόσες φορές εμφανίζεται το λήμμα και όχι η λέξη) και θα δημιουργήσετε ένα νέο αρχείο για κάθε κείμενο, κάθε γραμμή του οποίου θα είναι της μορφής:

***Λήμμα*** ***PoS_tag*** ***συχνότητα_εμφάνισης (freq)***

Για παράδειγμα ένα μέρος του αρχείου θα μπορούσε να είναι

*apple* *N* *4*

*sell* *V* *6*

το οποίο σημαίνει πως στο κείμενο βρέθηκε το λήμμα apple 4 φορές και το λήμμα sell 6 φορές.

*Σημείωση 3.2:*. Για τις λέξεις που ο tagger επιστρέφει λήμμα ***unknown*** (ανεξάρτητα από το αν τους έχει δώσει κάποιο PoS tag ή όχι) **δεν** θα πρέπει να αποθηκεύεται στο αρχείο το λήμμα unknown.

*Σημείωση 3.3:* Φροντίστε να μετατρέπετε όλους τους χαρακτήρες των λημμάτων σε πεζούς, έτσι ώστε να αποφύγετε να ομαδοποιήσετε λάθος τα λήμματα.

**4. Δημιουργία ανεστραμμένου ευρετηρίου (20%).** Στο βήμα αυτό θα δημιουργήσετε το ανεστραμμένο ευρετήριο για τη συλλογή κειμένων. Συγκεκριμένα, για όλα τα κείμενα της συλλογής θα επιλέξετε τα μοναδικά λήμματα που εμφανίζονται στη συλλογή. Για κάθε ένα από τα λήμματα της συλλογής θα εντοπίσετε τα κείμενα στα οποία εμφανίζεται από την έξοδο του προηγούμενου βήματος καθώς και τη συχνότητα εμφάνισης του λήμματος στο κάθε κείμενο. Η δομή αποθήκευσης που θα επιλέξετε για το ανεστραμμένο ευρετήριο θα έχει συνεπώς εγγραφές της μορφής:

***<λήμμα, {<docid1, freq1>, <docid2, freq2>,…}>***

*Σημείωση 4.1*: δεν είναι αναγκαίο να δημιουργήσετε ξεχωριστό αρχείο εξόδου για το βήμα αυτό καθώς το ευρετήριο δεν θα είναι ολοκληρωμένο. Η ολοκλήρωσή του θα γίνει στο επόμενο βήμα οπότε και θα είναι έτοιμο να φορτωθεί στη μνήμη. Αν όμως, κρίνετε αναγκαίο βάσει του σχεδιασμού σας να δημιουργήσετε ενδιάμεσο αρχείο εξόδου για το βήμα 4 μπορείτε να το κάνετε αφού αιτιολογήσετε την επιλογή σας.

*Σημείωση 4.2*: Προσοχή! Ο τρόπος με τον οποίο θα σχεδιάσετε την υλοποίηση του βήματος 4 παίζει πολύ σημαντικό ρόλο στην αποδοτικότητα της αναζήτησης κειμένων στο ευρετήριο.

**5. Ανάθεση βαρών τους όρους του ευρετηρίου (20%).** Χρησιμοποιήστε τη μετρική TF-IDF και υπολογίστε για κάθε λήμμα του ευρετηρίου το βαθμό σπουδαιότητάς του (βάρος) για κάθε κείμενο της συλλογής στο οποίο περιέχεται. Στο τέλος του βήματος αυτού το ευρετήριό σας θα πρέπει να έχει τη μορφή:

***<λήμμα, {<docid1, weight1>, <docid2, weight2>,…}>***

όπου το weight θα είναι μια τιμή που θα υπολογίζει η μετρική TF-IDF.

Σημείωση 5.1: Για τον υπολογισμό του IDF χρησιμοποιούμε log10 και το βάρος κάθε λήμματος θα υπολογίζετε για κάθε κείμενο στο οποίο εμφανίζεται.

Καλείστε να αποθηκεύσετε το ευρετήριο σας σε μορφή **XML** η οποία θα είναι ως εξής:

```
<inverted_index>
    <lemma name="orange">
        <document id="1" weight="0.4"/>
        <document id="2" weight="0.34"/>
    </lemma>
    <lemma name="apple">
        <document id=1 weight="0.65"/>
        <document id=2 weight="0.87"/>
        <document id=3 weight="0.45"/>
    </lemma>
</inverted_index>
```

Επίσης καλείστε να υλοποιήσετε τη λειτουργία φόρτωσης του ευρετηρίου απευθείας από xml αρχείο της προηγούμενης μορφής, έτσι ώστε αν έχει υπολογιστεί το ευρετήριο να μην απαιτείται η επανάληψη όλων των βημάτων για τη δημιουργία του.

**6. Αξιολόγηση ευρετηρίου (10%).** Υλοποιήστε έναν απλό μηχανισμό υποβολής ερωτημάτων στο ευρετήριο σας. Ο μηχανισμός θα δέχεται input από τον χρήστη ένα ερώτημα (που θα αποτελείται από ένα ή περισσότερα λήμματα), το οποίο θα ταυτοποιεί (με χρήση

string matching) στα λήμματα του ευρετηρίου και θα επιστρέφει στο χρήστη τα *ids* των κειμένων τα οποία περιέχουν το λήμμα ή τα λήμματα του ερωτήματος.

Προσοχή! Η λίστα των κειμένων που θα επιστρέφεται θα πρέπει να είναι ταξινομημένη σε φθίνουσα σειρά με βάση το TF-IDF βάρος που έχει το λήμμα του ερωτήματος για το κάθε κείμενο. Αν το ερώτημα έχει περισσότερα από ένα λήμματα, η ταξινόμηση θα γίνεται με βάση το άθροισμα των βαρών των λημμάτων που εντοπίστηκαν στο κείμενο.

Για να ελέγξετε την ορθότητα του μηχανισμού ερωτημάτων, κατεβάστε το αρχείο *query.txt* το οποίο περιέχει μια λίστα ερωτημάτων (ένα ερώτημα ανά γραμμή). Υποβάλλετε στο ευρετήριό σας τα ερωτήματα και υπολογίστε τον **χρόνο απόκρισης του ευρετηρίου για όλα τα ερωτήματα συνολικά.** Υπολογίστε το μέσο χρόνο απόκρισης του ευρετηρίου διαιρώντας το συνολικό χρόνο που μετρήσατε με τον αριθμό των ερωτημάτων. Αν οι χρόνοι είναι πολύ μικροί για να μετρηθούν, επαναλάβατε πολλές φορές πριν υπολογίσετε το μέσο χρόνο και διαιρέστε το συνολικό χρόνο με τις φορές επανάληψης του πειράματος επί τον αριθμό των ερωτημάτων.

Καταγράψτε επίσης για κάθε ένα από τα ερωτήματα που περιέχονται στην λίστα τα πρώτα 10 αποτελέσματα και το βάρος του κάθε αποτελέσματος.

*Σημείωση 6.1:* Προσοχή! Μην συνυπολογίσετε στη μέτρηση των χρόνων την είσοδο ή την έξοδο από αρχείο. Πριν αρχίσετε να μετράτε χρόνο θα πρέπει να έχετε φορτώσει στη μνήμη όλα τα ερωτήματα και μπορείτε να γράψετε τα αποτελέσματα σε αρχείο μόνο αφού έχετε ολοκληρώσει τη μέτρηση του χρόνου. Το I/O σε αρχείο είναι δυσανάλογα μεγαλύτερο από τους υπολογισμούς στη μνήμη.

## Παραδοτέα Άσκησης Α′

1. Ο πηγαίος κώδικας για την εκτέλεση όλων των παραπάνω βημάτων. Ο σχολιασμός του κώδικα να γίνει απαραίτητα σε επίπεδο συναρτήσεων (λειτουργία, ορίσματα, έξοδος) αλλά και εσωτερικά των συναρτήσεων όπου κρίνεται αναγκαίο για να γίνει κατανοητός.
2. Μια σύντομη αναφορά που να περιγράφει πώς ακριβώς εργαστήκατε σε κάθε βήμα.
3. Δώστε το συνολικό αριθμό λημμάτων της συλλογής και τον αριθμό των μοναδικών λημμάτων της συλλογής αφού απαλειφθούν τα stop-words
4. Τα tokenized κείμενα (το αρχείο που θα προκύψει στο τέλος του βήματος 1)
5. Τα PoS tagged κείμενα (το αρχείο που θα προκύψει στο τέλος του βήματος 2)
6. Τη διανυσματική αναπαράσταση των κειμένων (το αρχείο που θα προκύψει στο τέλος του βήματος 3)
7. Το ανεστραμμένο ευρετήριο της συλλογής κειμένων σε XML (το αρχείο που θα προκύψει στο τέλος του βήματος 5)
8. Δώστε το μέσο χρόνο απόκρισης του ευρετηρίου σας για όλα τα ερωτήματα που περιέχονται στο αρχείο *query.txt*.
9. Καταγράψτε για κάθε ερώτημα του αρχείου *query.txt* τα 10 πρώτα αποτελέσματα (*docids*) και το βάρος που έχουν οι όροι του ερωτήματος σε κάθε αποτέλεσμα

# ΜΕΡΟΣ Β′

## Εφαρμογή Τεχνικών Σύγκρισης Ομοιότητας Κειμένων

Πολλές φορές διαφορετικά κείμενα περιέχουν πανομοιότυπο περιεχόμενο· με άλλα λόγια, είτε το ένα είναι ακριβές αντίγραφο του άλλου, είτε μέρος τους ενός κειμένου περιέχεται αυτούσιο σε άλλο κείμενο. Για να υπολογίσουμε την επανάληψη περιεχομένου μεταξύ δύο κειμένων χρησιμοποιούμε τη μετρική *containment*, η οποία υπολογίζει το ποσοστό ενός κειμένου που περιέχεται αυτούσιο σε κάποιο άλλο κείμενο και παίρνει τιμές μεταξύ 0 και 1, όπου η τιμή 0 δηλώνει πως τα δύο εξεταζόμενα κείμενα είναι εντελώς διαφορετικά και η τιμή 1 δηλώνει πως τα δύο εξεταζόμενα κείμενα είναι ακριβώς ίδια!

Για να εφαρμοστεί η μετρική *containment* πρέπει να αναπαραστήσουμε το περιεχόμενο κάθε κειμένου σε *w-shingles*, όπου ένα *w-shingle* είναι ένα μοναδικό σύνολο *w* διαδοχικών στοιχείων. Για παράδειγμα, έστω το κείμενο:

***This is a beautiful day.***

Αν θέλουμε να αναπαραστήσουμε το περιεχόμενό του σε *w-shingles* όπου *w* =4 θα είχαμε την ακόλουθη αναπαράσταση:

| | | | |
|---|---|---|---|
| ***this*** | ***is*** | ***a*** | ***beautiful*** |
| ***is*** | ***a*** | ***beautiful*** | ***day*** |
| ***a*** | ***beautiful*** | ***day*** | **.** |
| ***beautiful*** | ***day*** | **.** | |
| ***day*** | **.** | | |
| **.** | | | |

Σκοπός της άσκησης είναι να υπολογίσετε το ποσοστό επανάληψης πληροφορίας στα κείμενα που χρησιμοποιήσατε για το πρώτο μέρος της άσκησης.

## Βήματα Υλοποίησης

**1. Αναπαράσταση κειμένων σε *w-shingles* (10%).** Αναπαραστήστε το περιεχόμενο κάθε tokenized κειμένου που αποθηκεύσατε στο αρχείο εξόδου του βήματος 1 του προηγούμενου μέρους της άσκησης σε *w-shingles*, μεγέθους 2 (*w* = 2). Κάθε κείμενο το περιεχόμενο του οποίου θα αναπαραστήσετε με 2-*shingles* θα το αποθηκεύσετε σε ξεχωριστό αρχείο εξόδου, το οποίο θα περιέχει ένα shingle ανά γραμμή (όπως φαίνεται και στο παράδειγμα παραπάνω). Επίσης μετατρέψτε όλα τα κεφαλαία γράμματα σε πεζά σε όλες τις λέξεις που διαβάζετε.

**2. Υπολογισμός επανάληψης περιεχομένου σε κείμενα - *containment* (10%)** Επιλέξτε τα πέντε μεγαλύτερα κείμενα (σε αριθμό λέξεων) και υπολογίστε το ποσοστό του περιεχομένου τους που επαναλαμβάνεται (περιέχεται) στα υπόλοιπα κείμενα της συλλογής.

*Σημείωση 2.1:* Προσοχή! Στο Β′ μέρος της άσκησης χρησιμοποιούμε λέξεις και όχι λήμματα, συνεπώς δουλεύουμε με τα tokenized texts, δηλ., το αρχείο εξόδου του βήματος 1 του πρώτου μέρους.

Για τον υπολογισμό της επανάληψης περιεχομένου χρησιμοποιήστε τη μετρική *containment* η οποία δίνεται από τον τύπο:

$$Containment = \frac{\left|V(A) \cap V(Collection - A)\right|}{\left|V(A)\right|}$$

όπου $V(A)$ είναι το σύνολο των 2-*shingles* του κειμένου Α, $V(Collection - A)$ είναι το σύνολο των 2-*shingles* της υπόλοιπης συλλογής, δηλαδή όλα τα κείμενα εκτός από το Α, $V(A) \cap V(Collection - A)$ είναι τα *2-shingles* του κειμένου Α τα οποία επαναλαμβάνονται (περιέχονται) στην υπόλοιπη συλλογή.

*Σημείωση 2.2*:Για τον υπολογισμό του *Containment* ενός κειμένου (έστω A) θα πρέπει να διαβάσετε όλα τα αρχεία με τα 2-*shingles* των υπολοίπων κειμένων (όλων εκτός από το A) να τα συμπεριλάβετε σε ένα σύνολο και να σχηματίσετε το $V\left(Collection - A\right)$

Σύμφωνα με τα παραπάνω, υπολογίστε το ποσοστό που το περιεχόμενο σε καθένα από τα 5 μεγαλύτερα κείμενα επαναλαμβάνεται (*Containment*) στα υπόλοιπα κείμενα της συλλογής και αποθηκεύστε τα αποτελέσματα σας σε ένα αρχείο.

## Παραδοτέα Άσκησης Β′

10. Ο πηγαίος κώδικας για την εκτέλεση όλων των παραπάνω βημάτων. Ο σχολιασμός του κώδικα να γίνει απαραίτητα σε επίπεδο συναρτήσεων (λειτουργία, ορίσματα, έξοδος) αλλά και εσωτερικά των συναρτήσεων όπου κρίνεται αναγκαίο για να γίνει κατανοητός.
11. Τα 2-shingles για κάθε κείμενο της συλλογής (το αρχείο που θα προκύψει στο τέλος του βήματος 1 του Β′ μέρους της άσκησης)
12. το ποσοστό Containment των 5 μεγαλύτερων κειμένων στα υπόλοιπα κείμενα της συλλογής (αποτέλεσμα βήματος 2 του Β′ μέρους της άσκησης)

# Διαδικαστικά θέματα

**Ημερομηνία Παράδοσης:** 30 Σεπτεμβρίου 2010

Η άσκηση είναι των **2 ατόμων** και **υποχρεωτική**. Σε περίπτωση που κάποιος επιθυμεί να δουλέψει μόνος του ή σε συνεργασία με άλλα 2 άτομα θα πρέπει να το δηλώσει.

**Προσοχή,** στην εργασία που θα παραδώσετε να αναγράφεται το ονομά σας, το επίθετό σας, το ΑΜ σας και το έτος σπουδών σας. Κάθε απάντηση να είναι **αιτιολογημένη**.

Τυχόν απορίες και ερωτήσεις, παρακαλώ να κοινοποιούνται μέσω forum προκειμένου να γνωστοποιούνται σε όλους οι απαντήσεις τους.

Η τελική βαθμολογία στο μάθημα θα προκύψει κατά 80% από την επίδοσή σας στην άσκηση (υλοποίηση και προφορική εξέταση στα ερωτήματα) και κατά 20% από την επίδοσή σας στην προφορική εξέταση στην ύλη που καλύπτουν οι διαλέξεις του μαθήματος. Η βαθμολογική συμμετοχή των επιμέρους ερωτημάτων της άσκησης (μέρος Α′ και Β′) φαίνεται στην εκφώνηση.

**ΠΡΟΣΟΧΗ!!** Όσοι υλοποιήσετε την άσκηση σε Python **ΔΕΝ** πρέπει να χρησιμοποιήσετε το NLTK

ΚΑΛΗ ΕΠΙΤΥΧΙΑ!!