ΓΛΩΣΣΙΚΗ ΤΕΧΝΟΛΟΓΙΑ

PROJECT 2020-2021

Σεπτέμβριος 2021

Γενικές Πληροφορίες

Η εργασία είνα ατομική. Η προθεσμία παράδοσης είναι μέχρι 11/09/2021 και η παράδοση θα γίνει μέσω της πλατφόρμας eclass (σύνδεσμος εργασίες). Θα βγει σχετική ανακοίνωση για την προφορική εξέταση στην οποία θα γίνει αναλυτική επίδειξη και αξιολόγηση των συστημάτων που αναπτύχθηκαν και των λειτουργιών τους.

ΜΕΡΟΣ Α

Σκοπός είναι η υλοποίηση ενός ολοκληρωμένου συστήματος συγκομιδής και δεικτοδότησης ιστοσελίδων. Η αρχιτεκτονική του συστήματος θα αποτελείται από έναν προσκομιστή ιστοσελίδων ο οποίος θα παρακολουθεί ειδησεογραφικές ιστοσελίδες και θα κατεβάζει νέες σελίδες ειδήσεων καθώς και από ένα υποσύστημα επεξεργασίας του περιεχομένου των ιστοσελίδων το οποίο στόχο έχει την δεικτοδότηση σε ανεστραμμένο ευρετήριο που θα σχεδιάσετε.

Αρχικά θα χρειαστεί να σχεδιάσετε και να χρησιμοποιηθεί ένας προσκομιστής ιστοσελίδων (crawler) ο οποίος θα παρακολουθεί ειδησεογραφικές ιστοσελίδες της επιλογής σας, θα κατεβάζει νέα άρθρα και θα τα αποθηκεύει στην βάση του συστήματος για ανάλυση. Στην συνέχεια θα αναπτύξετε ευρετήριο με βάση τα άρθρα που περιέχει η συλλογή.

Για να γίνει η αξιολόγηση της 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 αξιολογηθεί με βάση το κατά πόσο ανταποκρίνεται στις ανάγκες πραγματικών συστημάτων δεικτοδότησης. Η απόδοση των σχεδιαστικών επιλογών θα πρέπει να πραγματοποιηθεί με τέτοιο τρόπο ώστε να επιτρέπεται όσο το δυνατόν ταχύτερη συγκομιδή και δεικτοδότηση των ιστοσελίδων. Επίσης θα υλοποιήσετε μηχανισμό αποθήκευσης και επαναφόρτωσης του ευρετηρίου έτσι ώστε να μην υπάρχει ανάγκη ανακατασκευής του κάθε φορά που υποβάλλονται ερωτήματα σε αυτό.

Συνιστάται η χρήση PYTHON, καθώς αρκετά από τα εργαλεία που θα χρειαστείτε είναι ήδη υλοποιημένα στο NLTK.

Σχεδιασμός Ανεστραμμένου Ευρετηρίου

Η απόδοση του συστήματος που θα υλοποιήσετε εξαρτάται σε μεγάλο βαθμό από την επιλογή της κατάλληλης δομής δεδομένων για την φόρτωση του ανεστραμμένου ευρετηρίου στη

μνήμη. Η δομή δεδομένων που θα χρησιμοποιήσετε πρέπει να ελαχιστοποιεί το χρόνο αναζήτησης στο ευρετήριο, αλλά και να εξασφαλίζει ικανοποιητικό χρόνο κατασκευής του.

Το ανεστραμμένο ευρετήριο είναι μια συλλογή εγγραφών της μορφής:

<λήμμα, {<_σελίδας1, βάρος1>, <_σελίδας2, βάρος2>,…}>

Για κάθε μοναδικό λήμμα που συναντάται στη συλλογή ιστοσελίδων δημιουργείται μια εγγραφή στο ανεστραμμένο ευρετήριο. Η εγγραφή περιέχει το λήμμα καθώς και το σύνολο των ιστοσελίδων στις οποίες εμφανίζεται. Για κάθε ιστοσελίδα όπου εμφανίζεται, αποθηκεύεται επίσης το βάρος του λήμματος σε αυτή.

Σε κάθε ιστοσελίδα που ανήκει στη συλλογή θα πρέπει να δίνεται ένα id. Το id μιας ιστοσελίδας είναι ένα μοναδικό αναγνωριστικό της. Μπορείτε να δώσετε δικό σας id σε κάθε σελίδα για να το χρησιμοποιείτε εσωτερικά, να χρησιμοποιήσετε το path που έχει αποθηκευτεί τοπικά η ιστοσελίδα ή το url της. Σε κάθε περίπτωση θα πρέπει με δεδομένο το id να μπορείτε να ανακτήσετε το path και το url.

Βασικά υποσυστήματα που θα υλοποιήσετε

~~Προσκομιστής Ιστοσελίδων.~~

~~Ο προσκομιστής ιστοσελίδων (crawler) που θα υλοποιήσετε θα παρακολουθεί μια σειρά απο ειδησεογραφικές ιστοσελίδες (στην αγγλική γλώσσα) θα κατεβάζει νέα άρθρα στην βάση του συστήματος, τα οποία έχουν μορφοποίηση html. Για τον σχεδιασμό του crawler προτείνεται να βασιστείτε στον προσκομιστή Scrapy (https://scrapy.org/). Θα πρέπει να κατεβάζετε σελίδες μορφής html και όχι δεδομένα άλλου τύπου όπως doc, pdf, xls, php, js κλπ, καθώς απαιτείται παραπάνω επεξεργασία για την εξαγωγή καθαρού κειμένου από αυτές.~~

Προεπεξεργασία δεδομένων

Το συγκεκριμένο υποσύστημα εξάγει το καθαρό κείμενο από τις ιστοσελίδες που συγκεντρώσατε. Ο καθαρισμός αφορά την απομόνωση του κειμενικού περιεχομένου των html σελίδων σύμφωνα με το τρόπο που δομούνται οι ειδήσεις στο κάθε ειδησιογραφικό site που παρακολουθεί ο προσκομιστής.

Μορφοσυντακτική Ανάλυση

Σχολιάστε μορφοσυντακτικά τις λέξεις του κάθε tokenized κειμένου. Για το μορφοσυντακτικό σχολιασμό χρησιμοποιήστε κάποιον PoSTagger ανάλογα με το περιβάλλον υλοποίησης που έχετε επιλέξει. Στο τέλος της μορφοσυντακτικής ανάλυσης, κάθε κείμενο της συλλογής ιστοσελίδων θα διαθέτει μορφοσυντακτικό σχολιασμό (PoStags) για κάθε λέξη που περιέχει. Τα μορφοσυντακτικά σχολιασμένα κείμενα της συλλογής θα πρέπει να αποθηκευτούν σε βοηθητικό ενδιάμεσο αρχείο για μελλοντική χρήση.

Αναπαράσταση ιστοσελίδων στο Μοντέλο Διανυσματικού Χώρου.

Για να αναπαραστήσετε το περιεχόμενο κάθε κειμένου ως διάνυσμα θα χρησιμοποιήσετε τα μορφοσυντακτικά σχολιασμένα κείμενα (που θα προκύψουν από την έξοδο της μορφοσυντακτικής ανάλυσης) και αρχικά θα αφαιρέσετε τους τερματικούς όρους (stopwords) 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.

Στο παρακάτω link: http://www.infogistics.com/tagset.html θα βρείτε δύο πίνακες, έναν με τα PoStags για open class categories και έναν με τα PoStags για closed class categories. Τα openclasscategories είναι γραμματικές κατηγορίες των λέξεων που έχουν σημασιολογικό περιεχόμενο και άρα τις χρειαζόμαστε. Αντίθετα, τα closedclasscategories είναι γραμματικές κατηγορίες για λέξεις άνευ σημασιολογικού περιεχομένου, δηλ., stopwords. 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 closedclasscategorytag.

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).

Δημιουργία του ευρετηρίου

Στην φάση αυτή της υλοποίησης θα ολοκληρώνεται η κατασκευή του ανεστραμμένου ευρετηρίου. Για ολόκληρη τη συλλογή ιστοσελίδων που έχετε συγκεντρώσει, θα εντοπίζονται τα μοναδικά λήμματα, καθώς και τα κείμενα στα οποία εμφανίζεται το κάθε λήμμα. Θα δημιουργείται η αντίστοιχη εγγραφή στο ανεστραμμένο ευρετήριο και θα υπολογίζονται τα αντίστοιχα βάρη. Το βάρος του κάθε λήμματος για ένα κείμενο αντιπροσωπεύει το βαθμό σπουδαιότητας του λήμματος για το συγκεκριμένο κείμενο και θα το υπολογίσετε χρησιμοποιώντας τη μετρική TF-IDF .

Αποθήκευση και επαναφόρτωση ευρετηρίου.

Θα υλοποιήσετε τις απαραίτητες συναρτήσεις έτσι ώστε να είναι δυνατή η αποθήκευση και η επαναφόρτωση του ανεστραμμένου ευρετηρίου.

Το ευρετήριό σας θα αποθηκεύεται στην ακόλουθη μορφή:

<inverted_index>
<lemma name="orange">
<document id="…" weight="0.4"/>
<document id="…" weight="0.34"/>
</lemma>
<lemma name="apple">
<document id="…" weight="0.65"/>
document ="…" weight="0.87"/>
document ="…" weight="0.45"/>
</>
</>

Επίσης θα υπάρχει δυνατότητα ανάγνωσης ενός αρχείου με τη συγκεκριμένη μορφή και φόρτωσης των περιεχομένων στη δομή του ανεστραμμένου ευρετηρίου. Στόχος είναι να μην επαναλαμβάνεται κάθε φορά η κατασκευή του ευρετηρίου, αλλά να είναι δυνατή η απευθείας φόρτωσή του κατά τη φάση της αξιολόγησης.

Αξιολόγηση ευρετηρίου.

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 id - url?? των ιστοσελίδων τα οποία περιέχουν το λήμμα ή τα λήμματα του ερωτήματος.

Η λίστα των ιστοσελίδων 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.

Θα μετρήσετε το μέσο χρόνο απόκρισης του ευρετηρίου σας, υποβάλλοντας 20 ερωτήματα της μιας λέξης, 20 ερωτήματα των δύο λέξεων,30 των τριών λέξεων σε αυτό και 30 των τεσσάρων λέξεων. Θα μετρήσετε το συνολικό χρόνο και θα διαιρέσετε με τον αριθμό των ερωτημάτων για να υπολογίσετε το μέσο χρόνο απόκρισης. (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).

ΜΕΡΟΣ Β

Σκοπός της άσκησης Α είναι η υλοποίηση ενός συστήματος κατηγοριοποίησης κειμένων σε προκαθορισμένες θεματικές κατηγορίες. Η είσοδος του συστήματος αποτελείται από δύο σύνολα από έγγραφα (συλλογή Ε και συλλογή Α), και από ένα σύνολο θεματικών κατηγοριών ΘΚ. Η συλλογή Ε αποτελείται από έγγραφα ήδη κατηγοριοποιημένα στις προκαθορισμένες θεματικές κατηγορίες ΘΚ. Η συλλογή Α αποτελείται από έγγραφα τα οποία δεν είναι κατηγοριοποιημένα στις κατηγορίες του συνόλου ΘΚ, και τα οποία πρέπει να κατηγοριοποιηθούν στις θεματικές κατηγορίες ΘΚ.
Η κατηγοριοποίηση ενός εγγράφου Χ σε μια θεματική κατηγορία Κ γίνεται συγκρίνοντας το έγγραφο Χ με όλα τα μοντέλα των κατηγοριών ΘΚ, και επιλέγοντας την κατηγορία Κ από το ΘΚ που ταιριάζει περισσότερο με το έγγραφο Χ. Κάθε έγγραφο Χ θα αναπαρασταθεί από ένα διάνυσμα χαρακτηριστικών σταθερού μήκους. Κάθε κατηγορία Κ θα αναπαρασταθεί από ένα σύνολο διανυσμάτων χαρακτηριστικών σταθερού μήκους, το οποίο περιέχει όλα τα διανύσματα των εγγράφων της συλλογής Ε που ανήκουν στην κατηγορία Κ. Συνεπώς, η άσκηση αποτελείται από τρεις βασικές υπο-εργασίες:

- Ορισμός ενός χώρου χαρακτηριστικών S, πεπερασμένου μεγέθους (π.χ. 8.000 χαρακτηριστικά). Τα χαρακτηριστικά αυτά θα είναι θέματα (stems) από λέξεις που θα επιλεγούν από τα κείμενα της συλλογής Ε.

- Κατασκευή διανυσμάτων χαρακτηριστικών για όλα τα έγγραφα της συλλογής Ε. Το διάνυσμα του εγγράφου Χ θα περιέχει τα βάρη για όλα τα χαρακτηριστικά του χώρου S. Κάθε βάρος θα είναι το κανονικοποιημένο TF-IDF του χαρακτηριστικού στο κείμενο Χ.
- Κατηγοριοποίηση εγγράφου Χ από την συλλογή Α: δημιουργείται ένα διάνυσμα χαρακτηριστικών όπως στην περίπτωση των εγγράφων της συλλογής Ε, το οποίο συγκρίνεται με όλα τα διανύσματα των εγγράφων της συλλογής Ε, βάση συναρτήσεων σχετικότητας (similarityfunctions). Το έγγραφο Χ κατηγοριοποιείται στην κατηγορία του εγγράφου με το οποίο είναι πιο σχετικό.

Σαν συναρτήσεις σχετικότητας μπορούν να χρησιμοποιηθούν μετρικές όπως οι cosine similarity, Tanimoto, Jaccard. Σαν συλλογή κειμένων θα χρησιμοποιηθεί το σώμα κειμένων "20 news groups corpus", το οποίο είναι διαθέσιμο από την διεύθυνση http://qwone.com/~jason/20Newsgroups/.

Βασικά υποσυστήματα

Προ-επεξεργασία των συλλογών Ε και Α

Η άσκηση απαιτεί την αναγνώριση λέξεων και την θεματοποίησή τους. Μπορούν να χρησιμοποιηθούν έτοιμα εργαλεία γλωσσικής τεχνολογίας για αυτές τις εργασίες για την Αγγλική γλώσσα.

Δημιουργία χώρου χαρακτηριστικών

Θα πρέπει να υπολογιστεί το TF-IDF για όλα τα θέματα όλων των λέξεων όλων των εγγράφων της συλλογής Ε, και να επιλεγούν τα Ν (ανάλογα με το επιθυμητό μέγεθος του χώρου χαρακτηριστικών) με την μεγαλύτερη τιμή σύμφωνα με το TF-IDF.

Δημιουργία διανυσμάτων χαρακτηριστικών

Θα πρέπει να είναι δυνατή η δημιουργία διανυσμάτων τόσο για έγγραφα της συλλογής Ε όσο και της συλλογής Α. Το IDF στην περίπτωση των εγγράφων της συλλογής Α, θα ταυτίζεται με το IDF του χαρακτηριστικού στην συλλογή Ε. Είναι επιθυμητό η αναπαράσταση των διανυσμάτων να γίνει με αραιό τρόπο (sparse vectors).

Σύγκριση διανυσμάτων χαρακτηριστικών

Θα πρέπει να υλοποιηθεί υποσύστημα σύγκρισης δύο διανυσμάτων σταθερού μήκους, χρησιμοποιώντας τουλάχιστον 2 μετρικές σχετικότητας.

Παραδοτέα

1. Ο πηγαίος κώδικας όλων των παραπάνω συστημάτων. 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 εκτεταμένη αναφορά παρουσίασης του συστήματος που θα περιέχει τα αποτελέσματα των μετρήσεων όπου απαιτούνται από την άσκηση. Επίσης για κάθε ένα από τα υποσυστήματα που περιγράφονται, θα συμπεριλάβετε για τη βασική λειτουργία ένα σχόλιο για το κατά πόσον χρησιμοποιήσατε έτοιμη υλοποίηση (π.χ. από το NLTK) ή υλοποιήσατε τη μέθοδο εσείς.
3. Το αρχείο xml που θα περιέχει το ανεστραμμένο ευρετήριο.