Τμήμα Μηχανικών Η/Υ και Πληροφορικής Πανεπιστημίου Ιωαννίνων
Μεταπτυχιακό Μάθημα: *Αλγόριθμοι Επιστήμης Δεδομένων*
Χειμερινό Εξάμηνο 2019

1η Εργασία (Α' μέρος)

Ημερομηνία παράδοσης: Πέμπτη, 9/1/2020

Άσκηση 1 – Εύρεση όμοιων αντικειμένων

(1α) Έστω ότι μας δίνεται μια συλλογή m συνόλων, τα οποία έχουν το πολύ n στοιχεία το καθένα. Θέλουμε να υπολογίσουμε ακριβώς (χωρίς σφάλμα) τη Jaccard ομοιότητα για κάθε ζεύγος συνόλων. Περιγράψτε την πιο αποδοτική μέθοδο (σε χρόνο εκτέλεσης και απαιτούμενο χώρο) που μπορείτε να σκεφτείτε. Ποιος είναι ο (ασυμπτωτικός) χρόνος εκτέλεσης και ποιος ο απαιτούμενος χώρος της μεθόδου σας (ως συνάρτηση των m και n);

(1β) Σε αυτή την άσκηση θα εφαρμόσουμε τις τεχνικές MinHash και Locality Sensitive Hashing για το γρήγορο έλεγχο ομοιότητας κειμένων. Ως δεδομένα εισόδου θα χρησιμοποιήσουμε τα αρχεία articles_100, articles_1000, articles_2500 και articles_10000, τα οποία περιέχουν 100, 1000, 2500 και 10000 άρθρα αντίστοιχα. Τα μικρότερα αρχεία περιέχουν ένα υποσύνολο των άρθρων του articles_10000 για να μπορέσετε να δοκιμάσετε τις μεθόδους σας σε λιγότερα δεδομένα. Η τελική σας λύση θα πρέπει να είναι σε θέση να επεξεργαστεί όλα τα άρθρα του articles_10000.

Κάθε άρθρο ξεκινά με το χαρακτήρα 't', ακολουθούμενο με τον αριθμό του άρθρου (π.χ., 't980'). Τα αντίστοιχα 'truth' αρχεία περιέχουν ζεύγη όμοιων κειμένων (π.χ., 't1088 t5015').

Στο eCourse θα βρείτε επίσης μερικά αρχεία βοηθητικά αρχεία με κώδικα Python που μπορείτε να χρησιμοποιήσετε.

Για την υλοποίηση του προγράμματός σας, θα πρέπει να ακολουθήσετε τα παρακάτω βήματα:

Βήμα 1: Αναπαράσταση κειμένου ως σύνολο ακέραιων αριθμών

Αρχικά θα πρέπει να γράψετε μια συνάρτηση η οποία διαβάζει ένα άρθρο t και το αποθηκεύει ως σύνολο S_t ακέραιων αριθμών ως εξής. Χωρίζουμε το κείμενο σε μικρές συμβολοσειρές (shingles), και για κάθε συμβολοσειρά σ που παράγουμε, αποθηκεύουμε στο S_t την τιμή $hash(\sigma)$, όπου $hash$ μια συνάρτηση κατακερματισμού η οποία αντιστοιχεί τις συμβολοσειρές σε ακέραιους των 32 bit. Ως $hash$, μπορούμε να χρησιμοποιήσουμε τη συνάρτηση $crc32$ από το άρθρωμα (module) binascii, π.χ. γράφοντας

hashValue = binascii.crc32(sigma.encode('utf-8')) & 0xffffffff

Με τον παραπάνω τρόπο αναπαριστούμε κάθε άρθρο ως ένα σύνολο ακεραίων στο διάστημα $[0, m-1]$, όπου $m = 2^{32}$. Δείτε το αρχείο createSets.py, το οποίο δημιουργεί τις συμβολοσειρές σ ενώνοντας 3 διαδοχικές λέξεις του άρθρου. Στην υλοποίησή σας μπορείτε να χρησιμοποιήσετε αύτη τη μέθοδο (ή κάποια παραλλαγή της), ή τη μέθοδο των k-grams (k διαδοχικοί χαρακτήρες) που είδαμε στο μάθημα. Είναι καλή ιδέα να επεξεργαστείτε πρώτα τα άρθρα, έτσι ώστε να αφαιρέσετε τα σημεία στίξης, να μετατρέψετε τα κεφαλαία γράμματα σε πεζά και να αφαιρέσετε τα κενά διαστήματα.

Δείτε επίσης το αρχείο calculateJaccardSimilarity.py για τον υπολογισμό της Jaccard ομοιότητας δύο συνόλων.

Βήμα 2: Υλοποίηση MinHash

Γράψε μια συνάρτηση η οποία κατασκευάζει τον πίνακα M των MinHash υπογραφών των συνόλων S_t. Για να το πετύχουμε αυτό εύκολα, χωρίς να παράγουμε ολόκληρες τυχαίες μεταθέσεις, μπορούμε να

χρησιμοποιήσουμε μια καθολική οικογένεια $\mathcal{H}$ συναρτήσεων κατακερματισμού $h_{a,b}: \{0, \dots, m-1\} \rightarrow \{0, \dots, m-1\}$. Μια τέτοια οικογένεια υλοποιεί το αρχείο universalHashFunctions.py.

Στη συνέχεια, γράψε μια συνάρτηση η οποία υπολογίζει μια εκτίμηση της Jaccard ομοιότητας δύο συνόλων, χρησιμοποιώντας τον πίνακα M και συγκρίνετε τα αποτελέσματα που λαμβάνετε με τον ακριβή υπολογισμό της Jaccard ομοιότητας δύο συνόλων.

Βήμα 3: Υλοποίηση Locality Sensitive Hashing

Στο τελευταίο μέρος της άσκησης, θα πρέπει να υλοποιήσετε μια συνάρτηση η οποία ομαδοποιεί τα (πιθανά) όμοια άρθρα εφαρμόζοντας την τεχνική του Locality Sensitive Hashing. Η συνάρτηση λαμβάνει στην είσοδο το πίνακα M των MinHash υπογραφών και ένα κατώφλι s για το βαθμό ομοιότητας μεταξύ δύο κειμένων και επιστρέφει ομάδες κειμένων που είναι πιθανό να έχουν βαθμό Jaccard ομοιότητας $\geq s$. Επιλέξτε το πλήθος r των ακέραιων σε κάθε ζώνη (band) του πίνακα M με βάση την τιμή του s, όπως περιγράφεται στις διαφάνειες του μαθήματος.

Χρησιμοποιήστε αυτές τις ομάδες για να επιστρέψετε τα ζεύγη όμοιων κειμένων που έχετε βρει, επιλέγοντας κατάλληλα την τιμή του κατωφλίου s. Συγκρίνετε τα αποτελέσματά σας με τα ζεύγη που αναφέρονται στα 'truth' αρχεία.

Παραδοτέα

Ανεβάστε στο eCourse ένα zip αρχείο το οποίο περιλαμβάνει τα προγράμματά σας σε Python, καθώς και μια σύντομη αναφορά η οποία να περιγράφει την υλοποίησή σας και να αναφέρει τα αποτελέσματα που λάβατε κατά την εκτέλεση των προγραμμάτων.