ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΥΠΟΛΟΓΙΣΤΙΚΗ ΝΟΗΜΟΣΥΝΗ

Διδάσκοντες: Σ. Λυκοθανάσης, Δ. Κουτσομητρόπουλος

Ακαδημαϊκό Έτος 2023-2024

Εργαστηριακή Άσκηση
Μέρος Β΄

Β. Αποκατάσταση αρχαίων επιγραφών με χρήση ΓΑ.

Πολλές φορές οι αρχαίες επιγραφές είναι μερικώς κατεστραμμένες, παρουσιάζουν κενά ή έχουν δυσανάγνωστες λέξεις.
Θεωρούμε ότι επιγραφές που εντοπίζονται στην ίδια γεωγραφική περιοχή μπορεί να έχουν παραπλήσιο ή συναφές περιεχόμενο και άρα οι λέξεις που λείπουν να ανιχνεύονται ήδη στο σώμα κειμένων των επιγραφών με την ίδια εντοπιότητα. Θεωρούμε επίσης ότι έχουμε μια εικόνα του κενού ή του κατεστραμμένου μέρους που υπάρχει σε μια επιγραφή, ώστε να μπορούμε να εκτιμήσουμε τον αριθμό των λέξεων που λείπουν.
Στην εργασία αυτή σας δίνεται επομένως η παρακάτω φθαρμένη επιγραφή:

[...] αλεξανδρε ουδις [...]

όπου [...] υποδηλώνει την απουσία λέξης. Για την επιγραφή αυτή γνωρίζουμε ότι έχει βρεθεί στην ευρύτερη περιοχή της Συρίας (*Greater Syria and the East,* id = 1683). Σας ζητείται να υλοποιήσετε ΓΑ για να 'μαντέψετε' τις λέξεις που λείπουν και να αποκαταστήσετε την επιγραφή.
Η βασική ιδέα για την αξιοποίηση ΓΑ είναι να αναζητηθούν οι λέξεις που λείπουν, ώστε η συμπληρωμένη περιγραφή να μοιάζει όσο το δυνατόν περισσότερο με τις παρεμφερείς επιγραφές της.

Β1. Σχεδιασμός ΓΑ *[30 μονάδες]*

α) Κωδικοποίηση: Για την αναπαράσταση των δεδομένων σας, μπορείτε να χρησιμοποιήσετε την κωδικοποίηση που ακολουθήσατε στο μέρος Α, δηλαδή BoW με tf-idf. Επίσης μπορείτε να περιορίσετε την αναζήτησή σας στις επιγραφές που έχουν βρεθεί στην ίδια γεωγραφική περιοχή (region_main_id=1683), φτιάχνοντας ένα λεξικό μεγέθους 1678 tokens. Καλό είναι να χρησιμοποιήσετε το σύνολο του λεξικού για την αναπαράσταση των διανυσμάτων, καθώς δεν είναι μεγάλο και αποφεύγεται ο κίνδυνος να αποκοπούν λέξεις που μπορεί να περιλαμβάνουν αυτές που λείπουν.
Όσον αφορά την κωδικοποίηση των ατόμων του πληθυσμού λάβετε υπόψη τα εξής:

i. κάθε άτομο του πληθυσμού αναπαριστά τις 2 λέξεις που πιθανολογείται ότι λείπουν από την επιγραφή και υποθέτουμε ότι υπάρχουν στο λεξικό.
ii. κάθε λέξη που υπάρχει στο λεξικό αντιστοιχεί σε έναν ακέραιο στο διάστημα [1, 1678].
iii. Με αυτές τις 2 λέξεις συμπληρώνεται η επιγραφή (δεν έχει σημασία η σειρά), κατασκευάζεται κατά τα γνωστά το αντίστοιχο διάνυσμα και ελέγχεται η ομοιότητα του.

Εναλλακτικά, μπορείτε να κωδικοποιήσετε κάθε άτομο του πληθυσμού ως ένα πλήρες διάνυσμα 1678 θέσεων που θα αναπαριστά ολόκληρη την ανακτημένη επιγραφή. Σε αυτή την περίπτωση θα πρέπει οι θέσεις που αντιστοιχούν στις δοσμένες λέξεις να παραμένουν σταθερές και να μην ενεργοποιούνται πάνω από 2 θέσεις, κάτι που είναι πιθανό να προκαλέσει αυξημένες μη νόμιμες τιμές και συχνές επιδιορθώσεις.

β) Πλεονάζουσες τιμές: Ανάλογα με την κωδικοποίηση που εφαρμόσατε στο (α) είναι πιθανό να προκύψουν πλεονάζουσες τιμές, για παράδειγμα, τιμές εκτός του εύρους του λεξικού λόγω δυαδικής κωδικοποίησης. Περιγράψτε πώς θα αντιμετωπίσετε το πρόβλημα αυτό. Εξετάστε αν μπορείτε να αποφύγετε τις πλεονάζουσες τιμές, με βάση την κωδικοποίηση που προτείνατε στο (α).

γ) Αρχικός πληθυσμός: Περιγράψτε μια διαδικασία για τη δημιουργία αρχικού πληθυσμού ατόμων. Τα άτομα του πληθυσμού είναι πιθανά ζεύγη λέξεων, όπως αναφέρθηκε στο (α).

δ) Υπολογισμός ομοιότητας: Υπολογίστε την απόσταση της δοσμένης επιγραφής από όλες τις υπόλοιπες στο σύνολο δεδομένων της ίδιας γεωγραφικής περιοχής και βρείτε τις top-5 ή top-10 που είναι πιο κοντά. Για τον υπολογισμό της απόστασης μπορούν να χρησιμοποιηθούν διάφορες μετρικές, όπως *ευκλείδεια απόσταση, απόσταση Manhattan, απόσταση Hamming, συνημίτονο και συσχέτιση Pearson*. Να χρησιμοποιήσετε την ομοιότητα συνημιτόνου και να σχολιάσετε την καταλληλότητά της, σε σχέση και με τις υπόλοιπες, για τη συγκεκριμένη περίπτωση.

ε) Συνάρτηση καταλληλότητας: Ένα άτομο είναι πιο κατάλληλο από άλλα, εφόσον η συμπληρωμένη επιγραφή που προκύπτει είναι πιο κοντά στις top-5 ή top-10. Μπορείτε να χρησιμοποιήσετε είτε άθροισμα είτε Μ.Ο. Ποια είναι η μέγιστη και η ελάχιστη τιμή που μπορεί να έχει η συνάρτηση καταλληλότητας;

στ) Γενετικοί τελεστές: Με βάση την κωδικοποίηση που επιλέξατε, να προτείνετε τους τελεστές επιλογής, διασταύρωσης και μετάλλαξης που θα χρησιμοποιήσετε.

i. Ειδικά για την επιλογή, να αξιολογήσετε τη χρήση *ρουλέτας με βάση το κόστος, με βάση την κατάταξη* και *τουρνουά.*
ii. Ειδικά για τη διασταύρωση, να αξιολογήσετε την καταλληλότητα των ακόλουθων τελεστών: *Διασταύρωση μονού σημείου, διασταύρωση πολλαπλού σημείου, ομοιόμορφη διασταύρωση*.
iii. Ειδικά για τη μετάλλαξη, να αξιολογήσετε τη χρήση *ελιτισμού*.

Β2. Υλοποίηση ΓΑ *[30 μονάδες]*
Να γράψετε ένα πρόγραμμα, σε οποιοδήποτε περιβάλλον ή γλώσσα προγραμματισμού, που να υλοποιεί τον γενετικό αλγόριθμο που σχεδιάσετε.

Β3. Αξιολόγηση και Επίδραση Παραμέτρων *[40 μονάδες]*
α) Να τρέξετε τον αλγόριθμο για τις τιμές των παραμέτρων που φαίνονται στον παρακάτω πίνακα και να τον συμπληρώσετε. Ο αλγόριθμος θα τερματίζει όταν πληρούνται ένα ή περισσότερα από τα *κριτήρια τερματισμού*, δηλαδή όταν:

i. το καλύτερο άτομο της κάθε γενιάς πάψει να βελτιώνεται για ορισμένο αριθμό γενεών ή
ii. βελτιώνεται κάτω από ένα ποσοστό (<1%) ή
iii. έχει ξεπεραστεί ένας προκαθορισμένος αριθμός γενεών (π.χ. 1000)

Α/Α	ΜΕΓΕΘΟΣ ΠΛΗΘΥΣΜΟΥ	ΠΙΘΑΝΟΤΗΤΑ ΔΙΑΣΤΑΥΡΩΣΗΣ	ΠΙΘΑΝΟΤΗΤΑ ΜΕΤΑΛΛΑΞΗΣ	ΜΕΣΗ ΤΙΜΗ ΒΕΛΤΙΣΤΟΥ	ΜΕΣΟΣ ΑΡΙΘΜΟΣ ΓΕΝΕΩΝ
1	20	0.6	0.00		
2	20	0.6	0.01		
3	20	0.6	0.10		
4	20	0.9	0.01		
5	20	0.1	0.01		
6	200	0.6	0.00		
7	200	0.6	0.01		
8	200	0.6	0.10		
9	200	0.9	0.01		
10	200	0.1	0.01		

Προσοχή: *Επειδή οι ΓΑ είναι στοχαστικοί αλγόριθμοι και συνεπώς δεν εξασφαλίζουν την ίδια απόδοση σε κάθε εκτέλεσή τους, θα πρέπει να εκτελέσετε τον αλγόριθμο τουλάχιστον δέκα φορές για κάθε περίπτωση. Στον πίνακα να σημειώσετε το μέσο όρο της απόδοσης της καλύτερης λύσης σε κάθε τρέξιμο.*

β) Για κάθε περίπτωση του παραπάνω πίνακα να σχεδιάστε την καμπύλη εξέλιξης (απόδοση/αριθμό γενιών) της καλύτερης λύσης (της μέσης τιμής αυτής, σε κάθε γενιά, σε κάθε τρέξιμο). Ποια είναι η επιγραφή που προκύπτει;

γ) Με βάση αυτές τις καμπύλες, αλλά και τα αποτελέσματα του παραπάνω πίνακα, να διατυπώσετε αναλυτικά τα συμπεράσματά σας σχετικά με την επίδραση της κάθε παραμέτρου (μέγεθος πληθυσμού, πιθανότητα διασταύρωσης, πιθανότητα μετάλλαξης) στη σύγκλιση του αλγορίθμου.

Παραδοτέα

Η αναφορά που θα παραδώσετε θα πρέπει να περιέχει εκτενή σχολιασμό των πειραμάτων σας καθώς και πλήρη καταγραφή των αποτελεσμάτων και των συμπερασμάτων σας, ανά υπο-ερώτημα. Επίσης, πρέπει να συμπεριλάβετε στην αρχή της αναφοράς σας ένα σύνδεσμο προς τον κώδικα που έχετε χρησιμοποιήσει (σε κάποια file sharing υπηρεσία ή code repo), ώστε να ληφθεί υπόψη.

Μην ξεχάσετε να συμπληρώσετε τα στοιχεία σας, στην αρχή της 1ης σελίδας.

Αξιολόγηση

Η απάντηση των ερωτημάτων Α και Β, έχει βαρύτητα 20% στον τελικό βαθμό του μαθήματος (το σύνολο και των δύο μερών της εργασίας έχει βαρύτητα 40%). Ο βαθμός του Bonus (10%) προστίθεται στο παραπάνω ποσοστό 40%.

Παρατηρήσεις

1. Η αναφορά, σε ηλεκτρονική μορφή, πρέπει να αναρτηθεί στο e-class μέχρι τη Δευτέρα, 10/6/2024, στις 23:59.
2. Για οποιαδήποτε διευκρίνηση / ερώτηση μπορείτε να χρησιμοποιείτε το σχετικό forum στο eclass του μαθήματος.