ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΥΠΟΛΟΓΙΣΤΙΚΗ ΝΟΗΜΟΣΥΝΗ

Διδάσκων: Δ. Κουτσομητρόπουλος

Ακαδημαϊκό Έτος 2021-2022

Εργαστηριακή Άσκηση
Μέρος Β΄

Β. Εντοπισμός σχετικών λέξεων από ένα σώμα κειμένου με χρήση ΓΑ

Ένα κείμενο μπορεί να διαμορφώνεται κάνοντας χρήση μεμονωμένων στοιχείων, των λέξεων. Οι λέξεις αυτές αντλούνται από μεγάλα ευρετήρια-λεξικά, με κάποιες από αυτές να μην συνεισφέρουν σημαντικά στη σημασία του κειμένου, πχ. stop words, σε αντίθεση με άλλες που συνδέονται σημασιολογικά με την βασική έννοια έχοντας, έτσι, κυρίαρχο ρόλο στο κείμενο.

Στην εργασία αυτή σας ζητείται να προτείνετε και να υλοποιήσετε **Γενετικό Αλγόριθμο** που θα χρησιμοποιηθεί για τον εντοπισμό των πιο σημαντικών λέξεων σε ένα σώμα κειμένου. Για τις ανάγκες της άσκησης, θα χρησιμοποιήσετε το dataset που σας δόθηκε στο μέρος Α[1].

Σκοπός του αλγορίθμου είναι να εντοπίσει ποιες από τις 8.520 λέξεις που διαθέτει συνολικά το λεξικό είναι οι πιο σημαντικές για το συγκεκριμένο dataset. Για να υπάρχει ικανοποιητική δειγματοληψία, θεωρήστε ως ελάχιστο πλήθος λέξεων 1000 (περίπου 10% του λεξικού). Ως μετρική της σημαντικότητας μιας λέξης μπορεί να χρησιμοποιηθεί ο Μ.Ο. του tf-idf για όλα τα κείμενα του dataset.

Β1. Σχεδιασμός ΓΑ *[30 μονάδες]*

α) Κωδικοποίηση: Να προτείνετε μια κωδικοποίηση για τα άτομα του πληθυσμού. Λάβετε υπόψη τα παρακάτω:

i. Ένα άτομο αναπαριστά το ίδιο το λεξικό, αλλά με διαφορετική επιλογή λέξεων, των πιο σημαντικών κάθε φορά.

ii. Η επιλογή ή μη μιας λέξης είναι δυαδική (0 ή 1) και υπάρχουν συνολικά 8.520 λέξεις.

β) Αρχικός πληθυσμός: Περιγράψτε μια διαδικασία για τη δημιουργία αρχικού πληθυσμού ατόμων. Τα άτομα του πληθυσμού αναπαριστούν το λεξικό με τις πιο σημαντικές λέξεις.

γ) Διαδικασία επιδιόρθωσης: Σε κάθε άτομο μπορεί να είναι επιλεγμένες από καμία έως και το σύνολο των λέξεων του λεξικού. Παρόλα αυτά, έχουν τεθεί περιορισμοί στο κάτω όριο, τουλάχιστον 1000 επιλεγμένες λέξεις, ώστε να υπάρχει ικανοποιητική δειγματοληψία. Η μη τήρηση της συνθήκης αυτής είναι πιθανό να δημιουργεί μη νόμιμες λύσεις. Προδιαγράψτε μια διαδικασία χειρισμού των μη νόμιμων λύσεων, σχολιάζοντας και αξιολογώντας τις παρακάτω εναλλακτικές:

i. Απόρριψη της μη νόμιμης λύσης από τον πληθυσμό και αντικατάστασής της από κάποιο άλλο άτομο (τυχαία ή με ελιτισμό).

[1] https://archive.ics.uci.edu/ml/datasets/DeliciousMIL%3A+A+Data+Set+for+Multi-Label+Multi-Instance+Learning+with+Instance+Labels

ii. Επιδιόρθωση: Διαδικασία επιδιόρθωσης (repair procedure) η οποία αντιστοιχίζει τη μη νόμιμη λύση σε μια νόμιμη, π.χ. επιλογή επιπλέον λέξεων όταν το πλήθος των επιλεγμένων λέξεων είναι κάτω από το όριο που έχει τεθεί.

iii. Εφαρμογή ποινής: Μια μη νόμιμη λύση γίνεται αποδεκτή, αλλά της εφαρμόζεται ανάλογη ποινή από την συνάρτηση καταλληλότητας. Να περιγράψετε μια διαδικασία εφαρμογής ποινής για τέτοιες λύσεις.

Αξιολογήστε τις παραπάνω μεθόδους και προτείνετε την καταλληλότερη για το πρόβλημά σας.

δ) Υπολογισμός tf-idf: Χρησιμοποιήστε τη στατιστική μετρική tf-idf (term frequency–inverse document frequency[2]), βάση της οποίας η σημαντικότητα μιας λέξης σε ένα κείμενο σχετίζεται με το σύνολο των κειμένων του dataset (σώμα κειμένου). Υπολογίστε για κάθε λέξη τη συχνότητα εμφάνισής της στο κείμενο αλλά και στο σώμα κειμένου με βάση τον ακόλουθο τύπο $tf.idf(t,d,D) = tf(t,d) * idf(t,D)$ όπου t ο όρος-λέξη, d το κείμενο και D το σώμα κειμένου. Αναλυτικά οι όροι του γινομένου έχουν ως εξής:

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \qquad [1]$$

όπου $f_{t,d}$ το πλήθος εμφάνισης της λέξης t στο κείμενο d και $\sum_{t' \in d} f_{t',d}$ το πλήθος όλων των λέξεων του κείμενου

$$idf(t,D) = log\ \frac{N}{|\{d \in D : t \in d\}|} \qquad [2]$$

όπου *N* το πλήθος των κειμένων στο dataset και $|\{d \in D : t \in d\}|$ το πλήθος των κειμένων *d* που εμφανίζεται η λέξη *t*.

Παραδείγματα:

- λέξη με μεγάλο πλήθος εμφανίσεων σε ένα συγκεκριμένο κείμενο και τουλάχιστον μία εμφάνιση σε όλα τα άλλα κείμενα, θα έχει τιμή μηδέν. Από τύπο [2] : $log\frac{N}{N} = log1 = 0.$
- λέξη με μεγάλο πλήθος εμφανίσεων σε ένα συγκεκριμένο κείμενο και καμία εμφάνιση στα άλλα κείμενα, θα έχει αυξημένη τιμή. Από τύπο [2] : $log\frac{N}{1} = logN > 1.$

ε) Συνάρτηση καταλληλότητας: Ένα άτομο είναι πιο κατάλληλο από άλλα, εφόσον:

i. Οι λέξεις (γονίδια) που το σχηματίζουν είναι οι πιο σημαντικές.

ii. Έχει μικρότερο αριθμό επιλεγμένων λέξεων, με κάτω όριο τις 1000.

Επομένως η συνάρτηση καταλληλότητας θα πρέπει να συνδυάζει αυτά τα δυο ανταγωνιστικά κριτήρια. Για το i. μπορείτε να χρησιμοποιήσετε τον Μ.Ο. της τιμής αξιολόγησης από όλα τα γονίδια-λέξεις του ατόμου που έχουν επιλεγεί. Το κάθε γονίδιο-λέξη θα υπολογίζεται κάνοντας χρήση του Μ.Ο. της μετρικής tf-idf για τη συγκεκριμένη λέξη πάνω σε όλα τα κείμενα του συνόλου εκπαίδευσης. Για το ii. θα πρέπει να εφαρμόζετε μια ποινή στα άτομα που έχουν υψηλό αριθμό λέξεων. Η ποινή αυτή θα πρέπει να είναι σε κλίμακα ανάλογη με την τιμή στο i., έτσι ώστε αφενός να μην κυριαρχεί κατά την αξιολόγηση ενός ατόμου, αφετέρου να παίζει ικανό ρολό κατά την επιλογή. Να αιτιολογήσετε επαρκώς τη συνάρτηση καταλληλότητας στην οποία καταλήξατε. Ποια θα είναι η μέγιστη τιμή που μπορεί να έχει;

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.*

[2] https://en.wikipedia.org/wiki/Tf–idf

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 *[30 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
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 τρέξιμο).

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.

Β4. Επιλογή χαρακτηριστικών ΤΝΔ *[10 μονάδες]*
Η επιλογή χαρακτηριστικών (feature selection) είναι η διαδικασία μείωσης του αριθμού των εισόδων κατά τον σχεδιασμό και την εφαρμογή ενός αλγορίθμου μηχανικής μάθησης. Ένας αυξημένος αριθμός χαρακτηριστικών οδηγεί σε αύξηση του χρόνου εκπαίδευσης, ενώ παράλληλα δυσχεραίνει τη μάθηση και μπορεί να οδηγήσει σε υπερεκπαίδευση (κατάρα της διαστατικότητας). Γνωστές μέθοδοι για επιλογή χαρακτηριστικών αποτελούν τα ίδια τα ΤΝΔ, η

συσχέτιση Pearson, η Ανάλυση Κύριων Συνιστωσών (PCA), η ομαλοποίηση, το dropout κ.α.

Για να εφαρμόσετε επιλογή χαρακτηριστικών, μπορείτε να επανεκπαιδεύσετε το βέλτιστο ΤΝΔ του μέρους Α (μοντέλο BoW) χρησιμοποιώντας αυτή τη φορά ως εισόδους μόνο τις πιο σημαντικές λέξεις, όπως προέκυψαν από τα καλύτερο αποτέλεσμα του ΓΑ παραπάνω. Να συγκρίνετε την απόδοσή του με αυτή που είχατε βρει στο μέρος Α και να διατυπώσετε τα συμπεράσματά σας ως προς:

i. την γενικευτική ικανότητα των δύο δικτύων.
ii. την επίδραση της επιλογής (μείωσης) των χαρακτηριστικών στην απόδοση του δικτύου.

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 υποερώτημα. Επίσης, πρέπει να συμπεριλάβετε στην αρχή της αναφοράς σας ένα link προς τον κώδικα που έχετε χρησιμοποιήσει (σε κάποια file sharing υπηρεσία ή code repo).

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, 13/6/2022 στις 23:59.
2. Για οποιαδήποτε διευκρίνηση / ερώτηση μπορείτε να χρησιμοποιείτε το σχετικό forum στο eclass του μαθήματος.