ΥΠΟΛΟΓΙΣΤΙΚΗ ΝΟΗΜΟΣΥΝΗ

Διδάσκων: Δ. Κουτσομητρόπουλος
Ακαδημαϊκό Έτος 2021-2022

Εργαστηριακή Άσκηση
Μέρος Β΄

Φοιτητής: Τσιαμήτρος Κωνσταντίνος ΑΜ 235913

Εντοπισμός σχετικών λέξεων από ένα σώμα κειμένου με χρήση ΓΑ

Β1. Σχεδιασμός Γενετικού Αλγορίθμου

Α) Κωδικοποίηση

Δεδομένου ότι ένα άτομο αναπαριστά το ίδιο το λεξικό, ότι το λεξικό αποτελείται από 8520 λέξεις, και η επιλογή μιας λέξης είναι δυαδική, αρκεί να αντιστοιχίσω ένα Bit σε κάθε μια λέξη. Το bit αυτό θα είναι 0 αν η λέξη δεν περιέχεται στις σημαντικότερες λέξεις και 1 αλλιώς.

Άρα για την αναπαράσταση ενός ατόμου – ή αλλιώς χρωμοσώματος, (επειδή έχουμε 1 χρωμόσωμα ανά άτομο) του πληθυσμού χρειάζονται:

$$8520\ bit\ (γονίδια)$$

Β) Αρχικός πληθυσμός

Δεδομένης της παραπάνω αναπαράστασης, αρκεί να «τραβήξουμε» 8520 αριθμούς από μια «σακούλα» με τυχαίους αριθμούς (γεννήτρια ψευδοτυχαίων αριθμών) οι οποίοι θα έχουν μέγιστη τιμή την συμβολοσειρά 11....11 (8520 bit - δηλαδή αν επιλεχθούν όλες οι λέξεις του λεξικού ως σημαντικές) και ελάχιστη τιμή, κάθε συμβολοσειρά 8520 (bit) που να αριθμεί 1000 άσσους (δηλαδή, αποδεκτό πλήθος άσσων: > 1000 άσσοι). Επίσης, απορρίπτονται συμβολοσειρές που έχουν επιλεγεί ήδη (δηλαδή, ο αρχικός πληθυσμός περιέχει μοναδικά άτομα).

- **αρχεία** random_generator.py, normal_init_pop.py, Initial_pop_testbench.py

εκτελώντας το αρχείο Initial_pop_testbench.py είναι ορατές οι δύο κατανομές των αρχικών πληθυσμών που λήφθηκαν υπόψιν:

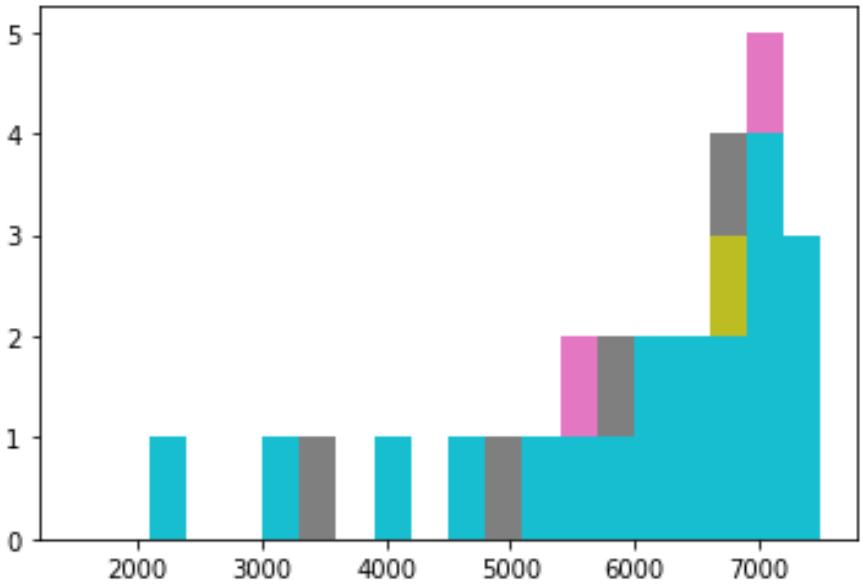


Figure 1 - Exponential Distribution

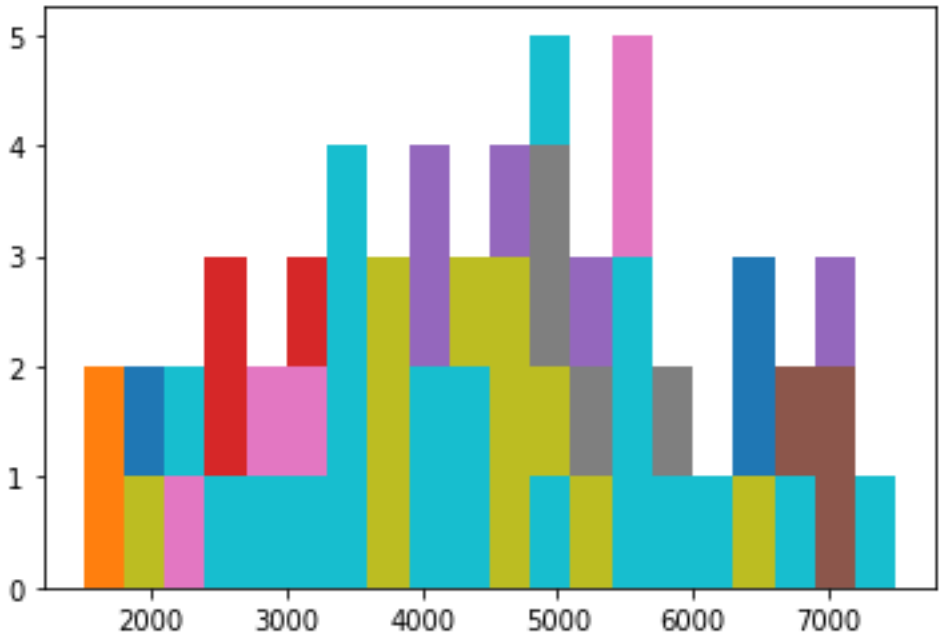


Figure 2 - Normal Distribution

C) Διαδικασία επιδιόρθωσης

Μετά την επιλογή του κάθε ατόμου, όπως περιεγράφηκε παραπάνω, εφαρμόζεται το προτεινόμενο κάτω όριο επιλεγμένων λέξεων, δηλαδή το ελάχιστο πλήθος άσσων που μπορεί να πάρει ένα άτομο είναι 1000.

i) Απορρίπτοντας μια μη νόμιμη λύση ή μια λύση που μπορεί να οδηγήσει σε μια μη νόμιμη λύση, «καλυτερεύουμε» έμμεσα την τιμή της συνάρτησης αξιολόγησης. Έχουμε δύο επιλογές, προκειμένου να παραμείνει το πλήθιος των ατόμων ίδιο, είτε να την αντικαταστήσουμε με μια καινούργια λύση (η οποία παράγεται με τον ίδιο τυχαίο τρόπο, όπως ο αρχικός πληθυσμός, είτε μπορούμε να επιλέξουμε το άτομο με την καλύτερη απόδοση και να το ξαναπροσθέσουμε στον πληθυσμό (ελιτισμός).

ii) Επιδιόρθωση: είναι παρόμοιος τρόπος λύσης με τον ελιτισμό. Η διαφορά της επιδιόρθωσης με τον ελιτισμό, είναι ότι στην επιδιόρθωση επιλέγεται τυχαία ένα άτομο του πληθυσμού για να αντικαταστήσει το απερχόμενο άτομο.

iii) Εφαρμογή ποινής: Καμία λύση δεν απορρίπτεται με άμεσο τρόπο, όμως, εφαρμόζεται μια ποινή στην τιμή της συνάρτησης καταλληλότητας του απερχόμενου ατόμου, το οποίο έχει ως αποτέλεσμα να εξαφανιστεί (σταδιακά) το άτομο από τον πληθυσμό (εφόσον υπάρχουν πάνω από ένα αντίγραφο του ατόμου αυτού στον πληθυσμό)

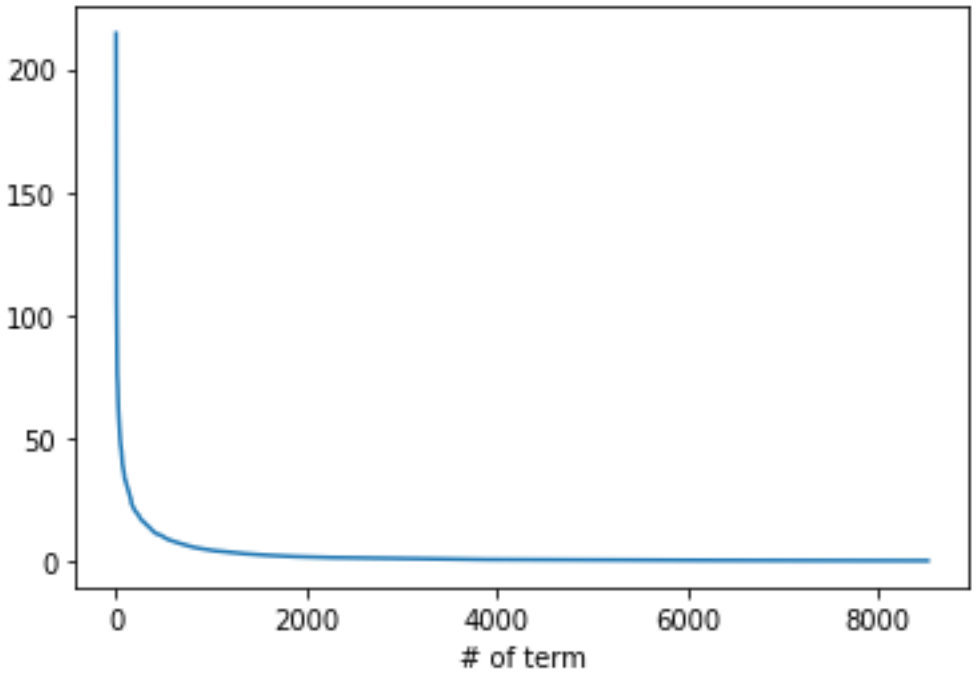
D) Υπολογισμός tf-idf

Η μετρική tf-idf αξιοποιήθηκε κατά την αξιολόγηση της σημαντικότητας μιας λέξης στο κείμενο.

- **αρχείο** tf_idf.py
- **συναρτήσεις**: όλες οι συναρτήσεις του αρχείου

Σημειώσεις πάνω στα δεδομένα:

Αφότου έχουν υπολογιστεί οι ποσότητες tf-idf για όλους τους όρους, τρέχοντας το script test_tf_idfs.py, παράγεται το παρακάτω γράφημα:



Παρατηρώ ότι η μεγαλύτερες τιμές των tf-idf τιμών, βρίσκονται συγκεντρωμένες σε ~1000 λέξεις. Ταυτόχρονα, παρατηρώ ότι πολλές λέξεις έχουν μια τιμή tf-idf πολύ κοντά στο μηδέν.

Για τον λόγο αυτό, θα εφαρμόσω μια περαιτέρω πίεση στον πληθυσμό (στην συνάρτηση Evaluate()) ώστε, τα άτομα με μεγάλο αριθμό άσσων (δηλαδή, πλήθος επιλεγμένων λέξεων > 2500) να δέχονται ποινή στο score τους.

Πειραματικά, από τα δεδομένα του dataset, παρατηρήθηκε μέγιστη τιμή του tf-idf ~= 214

Ε) Συνάρτηση καταλληλόλητας

Μια εκτίμηση του άνω ορίου, για την **μέγιστη τιμή** που μπορεί να έχει η συνάρτηση καταλληλότητας, θα είναι ο μέγιστος Μ.Ο. της τιμής αξιολόγησης από όλα τα γονίδια λέξεις που έχουν επιλεγεί. Δηλαδή, η μέγιστη τιμή θα είναι αυτή που έχει το μέγιστο δυνατό πλήθος 1 (δηλαδή 8520) επί την μέγιστη τιμή της μετρικής tf-idf (δηλαδή ~214) ~= 1.823.000 διά το ελάχιστο επιτρεπτό πλήθος επιλεγμένων λέξεων δηλαδή 1000 -> 1823. Όμως αυτό είναι ένα “χαλαρό” άνω όριο διότι δεν είναι πιθανό να βρεθούν 8520 λέξεις με την ίδια (μέγιστη) τιμή για το tf-idf διότι τότε καταρρίπτεται η ιδιότητά της μετρικής αυτής (όπως περιγράφεται στην εκφώνηση, δηλαδή δεν είναι δυνατό να έχουμε 8520 όρους με τιμή tf-idf ~= 214, εξ ορισμού).

Όπως αναφέρθηκε παραπάνω, η πειραματική μέγιστη τιμή της μετρικής tf-idf που παρατηρήθηκε είναι 214.

F) Γενετικοί Τελεστές

Επιλογή

1. Επιλογή με ρουλέτα με βάση το κόστος

Παραθέτω την λειτουργία του τελεστή επιλογής:

- αρχείο GA.py
- συναρτήσεις Select(), Choose()

Για κάθε άτομο, υπολογίζουμε την απόδοσή του (συνάρτηση Evaluate()), καθώς και τη συνολική απόδοση των ατόμων του πληθυσμού.

Για κάθε άτομο, διαιρούμε την απόδοσή του με την συνολική απόδοση.

Για κάθε άτομο του πληθυσμού υπολογίζουμε τις αθροιστικές πιθανότητες.

Τέλος, περιστρέφουμε τη ρουλέτα POP_SIZE φορές, και σε κάθε περιστροφή επιλέγεται και ένα άτομο. Τα άτομα που έχουν επιλεγεί, προχωράνε στην εφαρμογή του επόμενου γενετικού τελεστή (διασταύρωση).

Αυτός ο τύπος επιλογής, ασκεί την μεγαλύτερη πίεση επιλογής (με βάση τη θεωρία).

2. Επιλογή με ρουλέτα με βάση την κατάταξη

Ακολουθούμε την ίδια διαδικασία με πριν, όμως πριν υπολογίσουμε τις αθροιστικές πιθανότητες, ταξινομούμε τα άτομα του πληθυσμού, με βάση την απόδοσή τους, έτσι ώστε να διασταυρώνονται άτομα με παρόμοια απόδοση.

3. Τουρνουά me N άτομα

Με τυχαίο τρόπο επιλέγουμε N άτομα από τον πληθυσμό, και συγκρίνουμε τις αποδόσεις τους. «Νικάει» το άτομο που έχει την υψηλότερη απόδοση. Επαναλαμβάνουμε αυτή τη διαδικασία POP_SIZE φορές, έτσι ώστε ο πληθυσμός να παραμείνει σταθερός στην επόμενη γενιά.

Διασταύρωση

Αρχικά, για κάθε άτομο που έχει επιλεγεί, περιστρέφουμε μια ρουλέτα που έχει δύο σχισμές, οι οποίες είναι ανάλογες της πιθανότητας διασταύρωσης και της συμπληρωματικής της. Επιλέγουμε ουσιαστικά, έναν τυχαίο αριθμό στο διάστημα [0, 1] και εάν ο αριθμός αυτός, είναι μικρότερος ή ίσος με την πιθανότητα διασταύρωσης, τότε το άτομο αυτό θα διασταυρωθεί.

Επαναλαμβάνουμε την διαδικασία αυτή για κάθε άτομο που έχει επιλεγεί.

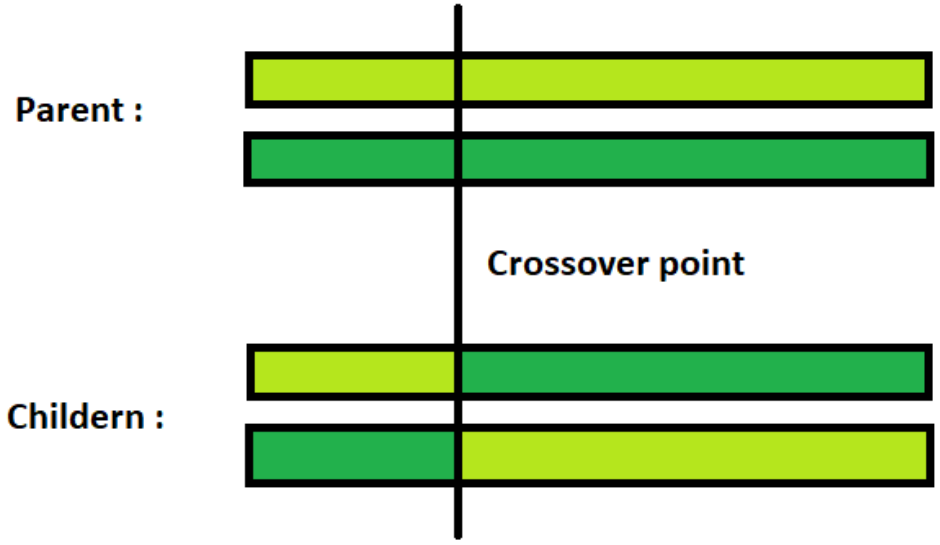
Τέλος, δημιουργούμε ζευγάρια είτε με τυχαίο τρόπο, είτε με ακολουθιακό τρόπο.

1. Διασταύρωση μονού σημείου

Κατά την διαδικασία της διασταύρωσης μονού σημείου, επιλέγεται με τυχαίο τρόπο ένα bit από τα BIT_NUM bit ενός ατόμου.

Προκύπτουν δύο απόγονοι από την διασταύρωση δύο γονέων. Στο παραπάνω bit που μόλις επιλέξαμε, θα γίνει η διασταύρωση. Δηλαδή, ο ένας απόγονος, θα κληρονομήσει το πρώτο μισό του χρωμοσώματος του γονέα Α και το δεύτερο μισό του χρωμοσώματος του γονέα Β.

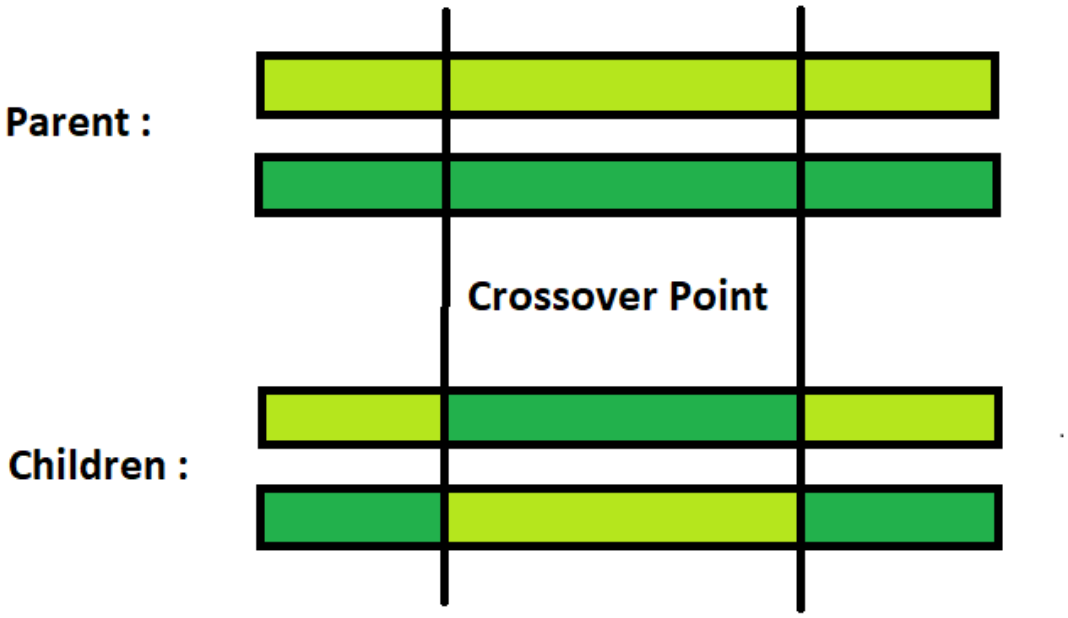
Ο άλλος απόγονος θα προκύψει με τον αντίστροφο τρόπο, δηλαδή, θα κληρονομήσει το πρώτο μισό χρωμόσωμα από τον γονέα Β και το δεύτερο μισό χρωμόσωμα από τον γονέα Α.



2. Διασταύρωση πολλαπλού σημείου

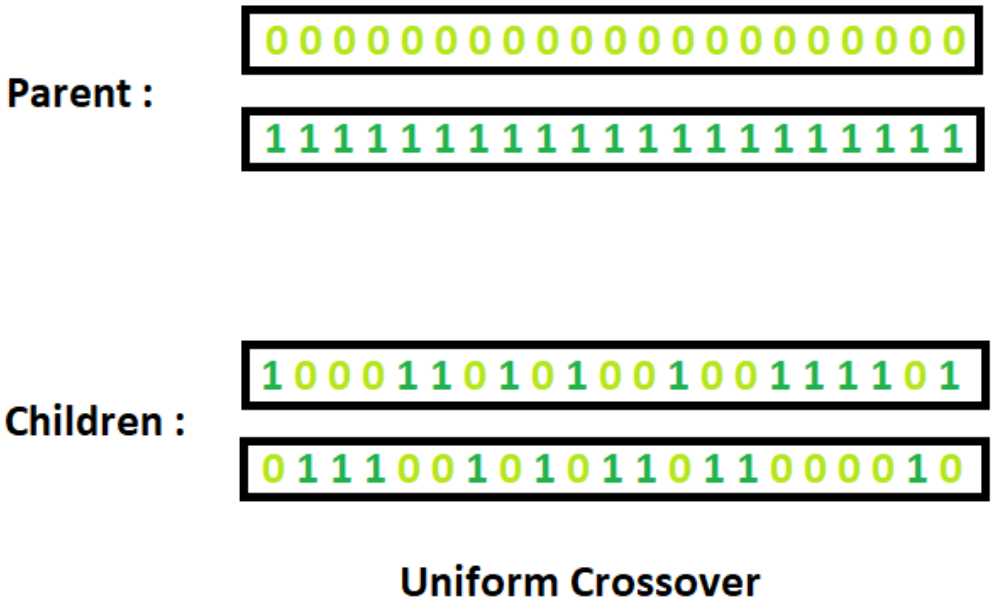
(Εξετάστηκε πειραματικά η διασταύρωση διπλού σημείου)

Με αντίστοιχο τρόπο, γίνεται και η διασταύρωση διπλού σημείου:



3. Ομοιόμορφη Διασταύρωση

Η ομοιόμορφη διασταύρωση μοιάζει με την μετάλλαξη, διότι ουσιαστικά, περιστρέφουμε μια ρουλέτα με δυο σχισμές οι οποίες είναι ανάλογες της πιθανότητας διασταύρωσης και της συμπληρωματικής της, για κάθε bit. Αν ο τυχαίος αριθμός που θα επιλέξουμε, είναι μικρότερος ή ίσος από την πιθανότητα διασταύρωσης, τότε ο απόγονος κληρονομεί το bit του γονέα Α, ενώ σε αντίθετη περίπτωση κληρονομεί το bit του γονέα Β:



Μετάλλαξη

Ο γενετικός τελεστής της μετάλλαξης, μοιάζει πολύ με την ομοιόμορφη διασταύρωση, έχει όμως μια διαφορά, εφαρμόζεται σε κάθε άτομο της νέας γενιάς που έχει προκύψει από τα προηγούμενα βήματα και περιστρέφοντας (ομοίως) μια ρουλέτα για κάθε bit, αν το αποτέλεσμα της ρουλέτας είναι μικρότερο ή ίσο από την πιθανότητα μετάλλαξης, τότε το bit γίνεται flip (αν ήταν 1 γίνεται 0 και αντίστροφα) αλλιώς δεν γίνεται κάποια αλλαγή.

- Ελιτισμός

Η λειτουργία που προσθέτει ο ελιτισμός, ουσιαστικά είναι ότι το άτομο της νέας γενιάς με την καλύτερη απόδοση, περνάει στη νέα γενιά χωρίς να υποστεί μετάλλαξη. Αυτό γίνεται σε μια προσπάθεια να διατηρηθούν αυτούσιες οι καλύτερες λύσεις στον καινούργιο πληθυσμό που προκύπτει.

Με βάσει τα πειραματικά δεδομένα (έχουν υλοποιηθεί όλοι οι παραπάνω τρόποι επιλογής) το τουρνουά φαίνεται να παράγει τα πιο ικανοποιητικά αποτελέσματα με Uniform crossover

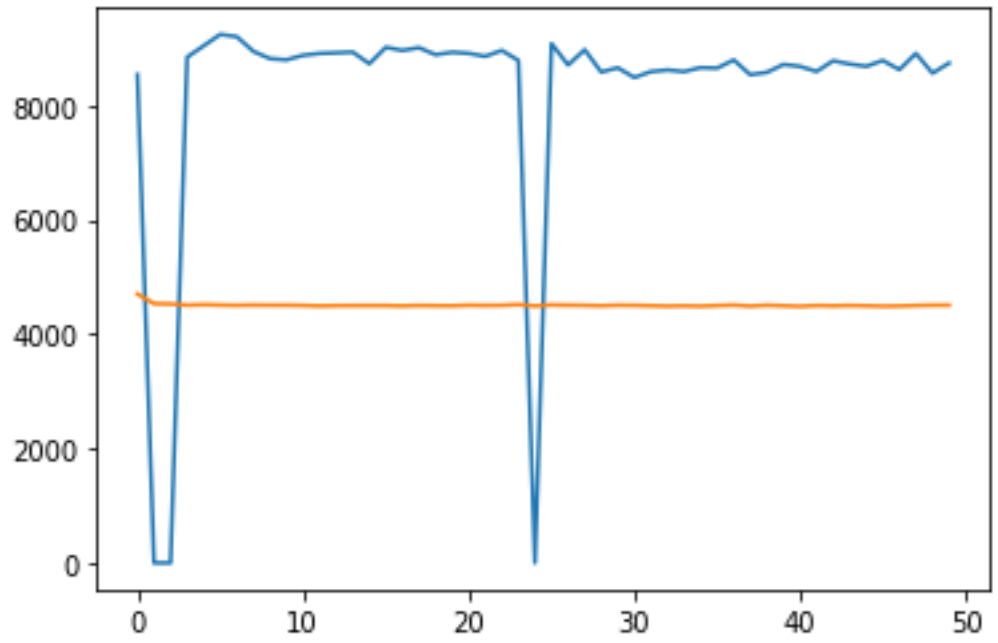
B2. Υλοποίηση ΓΑ

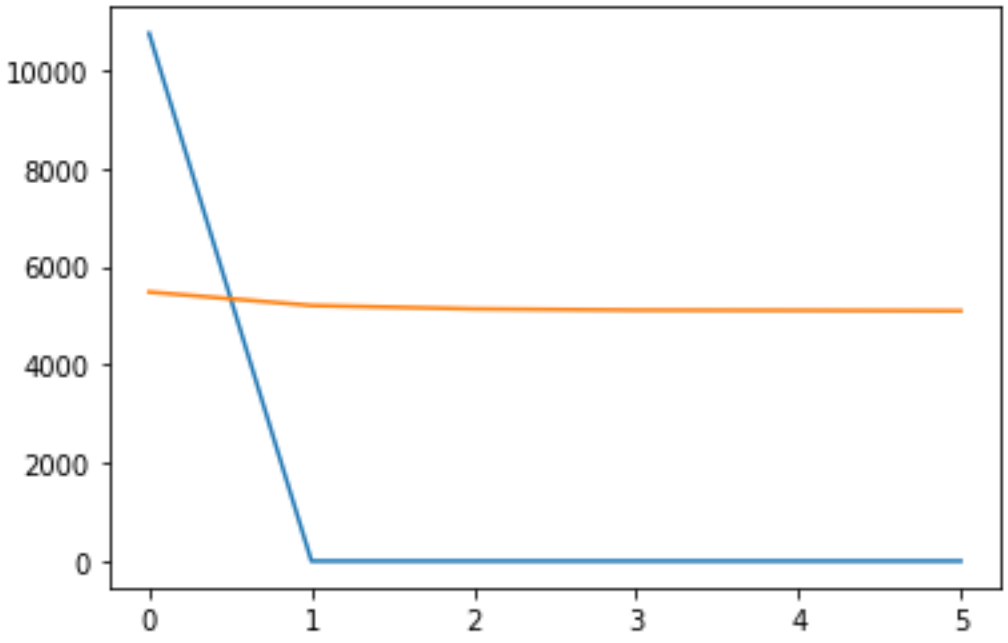
Υλοποιήθηκαν πλήρως τα ζητούμενα της εκφώνησης, παρακαλώ δείτε τα συνημμένα αρχεία κώδικα για περισσότερες πληροφορίες. (ή στο repo)

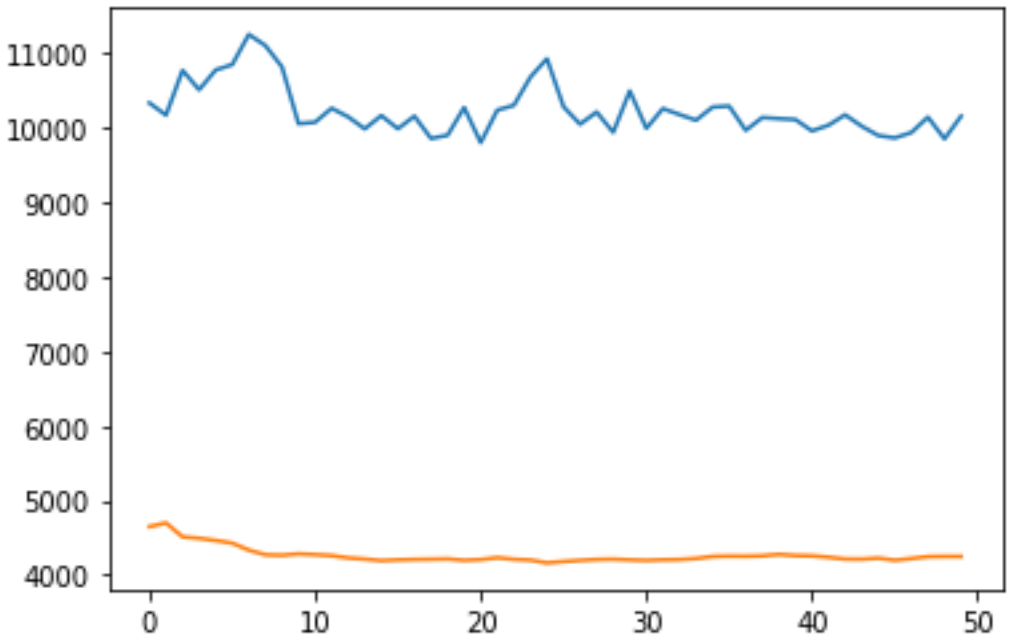
B3. Αξιολόγηση και Επίδραση Παραμέτρων

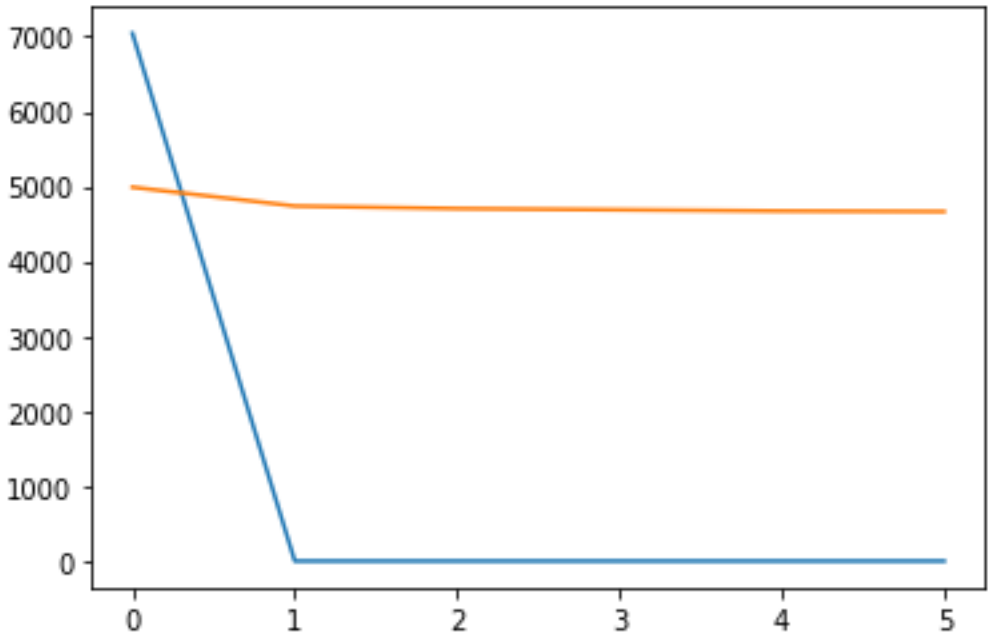
Α) Εφαρμόστηκαν τα κριτήρια τερματισμού που ζητούνται.

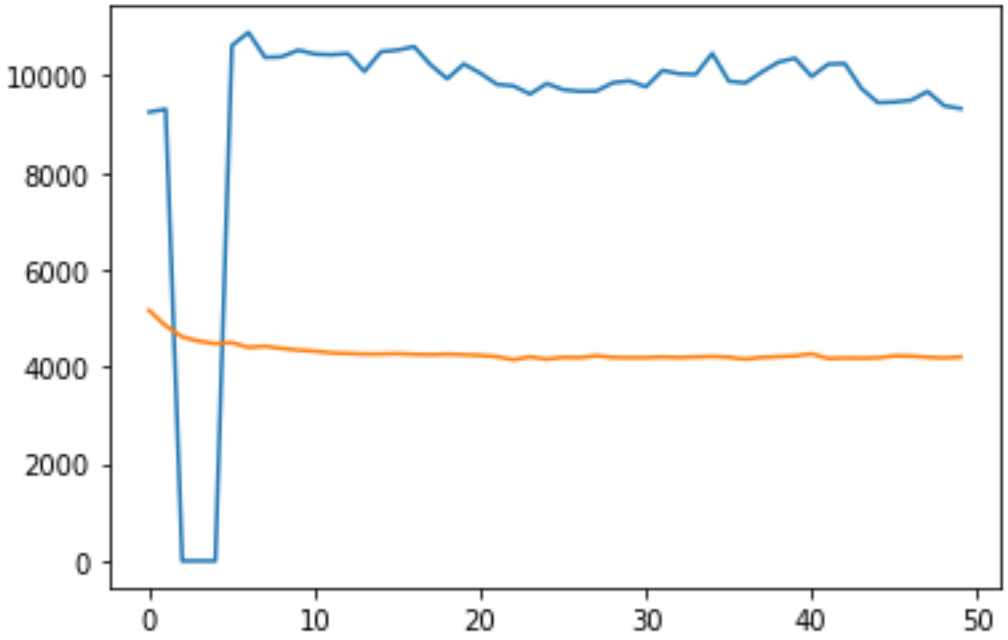
Β) Παραθέτω τα αποτελέσματα των επαναληπτικών εκτελέσεων του testbench για κάθε συνδυασμό (με τη σειρά, όπως ζητούνται) – Οι μπλε γραμμές είναι η μέση τιμή του max(scores) για κάθε μια από τις 10 επαναληπτικές εκτελέσεις του ΓΑ, και οι πορτοκαλί γραμμές είναι το μέσο πλήθος άσσων (επιλεγμένων λέξεων) – στον x άξονα φαίνονται οι γενιές:

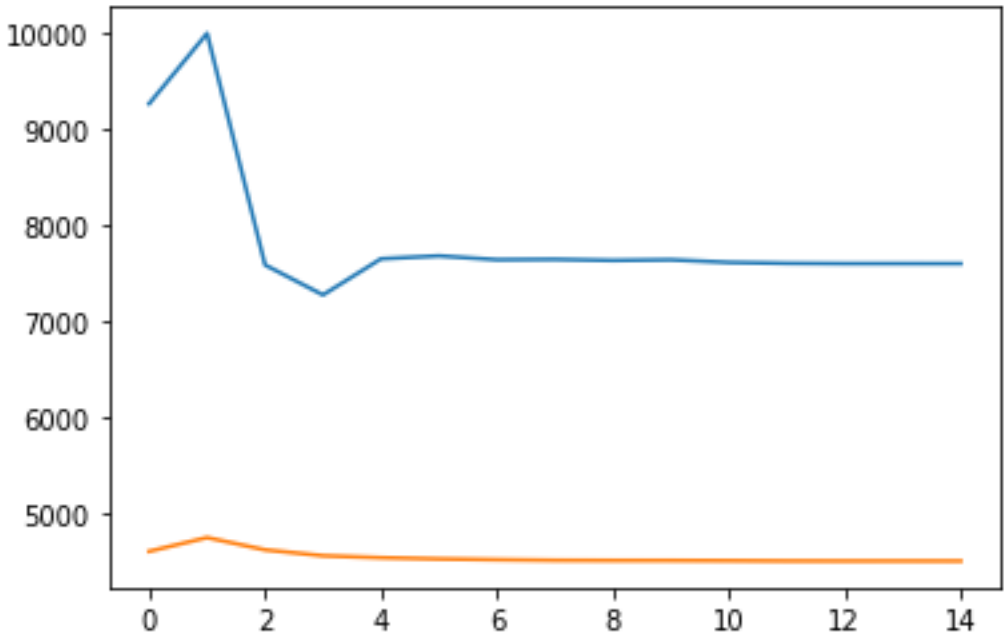


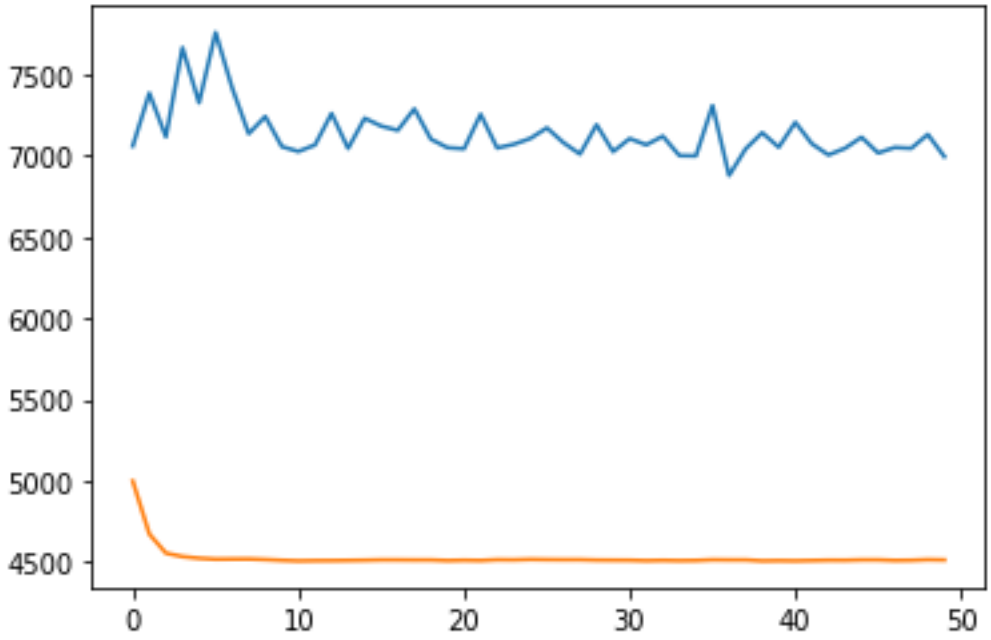

10000
8000
6000
4000
2000
0
0
1
2
3
4
5

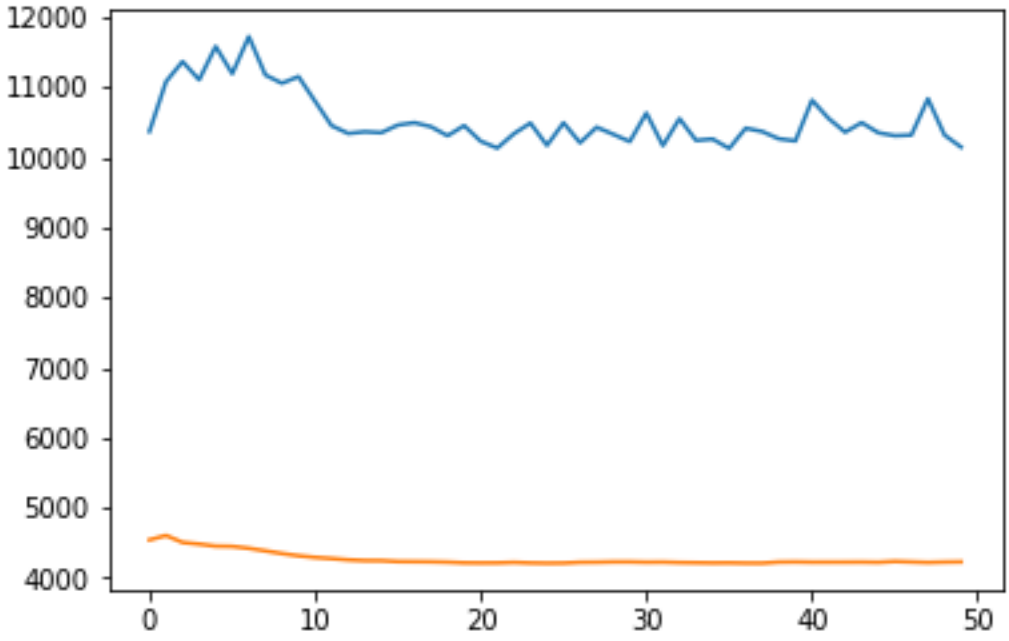

11000
10000
9000
8000
7000
6000
5000
4000
0
10
20
30
40
50

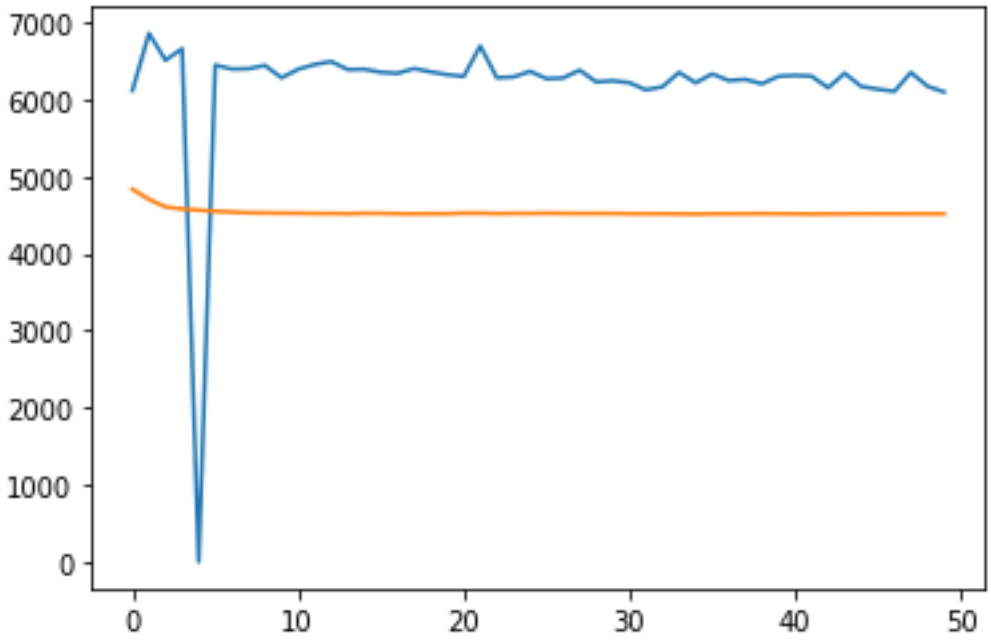

7000
6000
5000
4000
3000
2000
1000
0
0
1
2
3
4
5

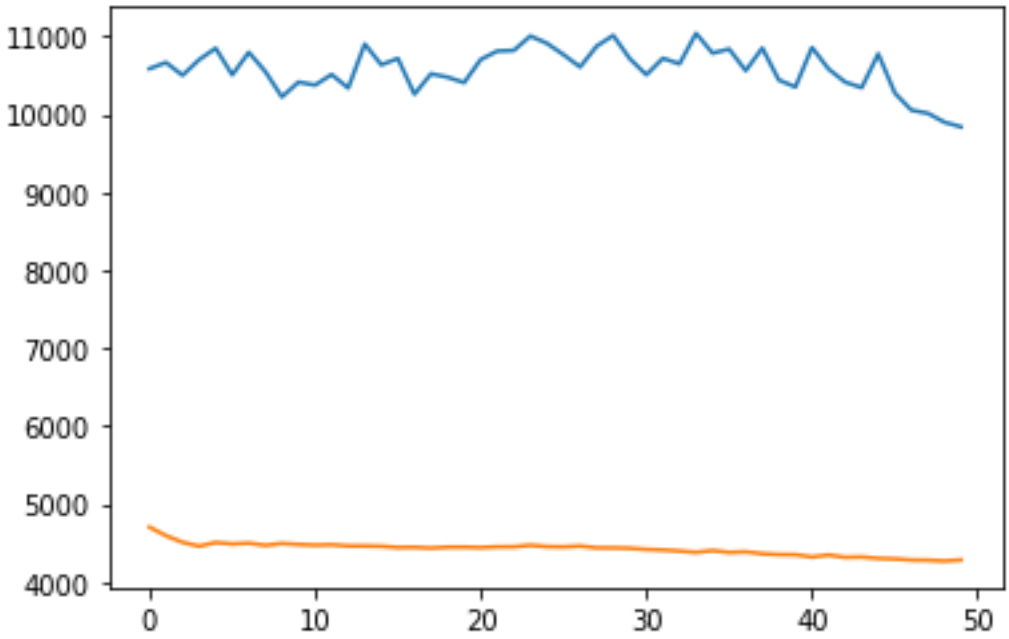

10000
8000
6000
4000
2000
0
0
10
20
30
40
50


10000
9000
8000
7000
6000
5000
0
2
4
6
8
10
12
14


7500
7000
6500
6000
5500
5000
4500
0
10
20
30
40
50


12000
11000
10000
9000
8000
7000
6000
5000
4000
0
10
20
30
40
50


7000
6000
5000
4000
3000
2000
1000
0
0
10
20
30
40
50

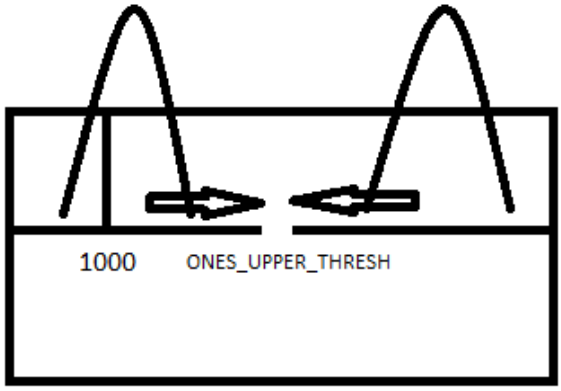


Γ. Συμπεράσματα

Με βάση τα παραπάνω, εξήγαγα το συμπέρασμα, ότι:

1. Υπάρχει ένα κάτω όριο (1000 – ONES_THRESH) στο πλήθος των επιλεγμένων λύσεων. Αν κάποιο άτομο πέσει κάτω από αυτό το όριο, τότε αντικαθίσταται από κάποιο άλλο με τυχαίο τρόπο.
2. Υπάρχει ένα άνω όριο (πειραματικά 4500 – ONES_UPPER_THRESH) στο πλήθος των επιλεγμένων λύσεων. Αν ο μέσος όρος του πλήθους αυτού ανέβει πάνω από το όριο, τότε εφαρμόζεται μεγάλο penalty στο score.
3. Έχει κάποια σημασία η κατανομή του αρχικού πληθυσμού και η διασπορά της, καθώς με αυτή
4. Ο συνδυασμός των παραπάνω επηρεάζει ως ένα βαθμό την ποιότητα των λύσεων που θα πάρουμε στην έξοδο του αλγορίθμου.

Σχηματικά:



αν σκεφτούμε τον αρχικό πληθυσμό σαν μια Normal κατανομή από «μπίλιες» με διάφορες τιμές για την διασπορά τους, τότε μπορούμε να καταλάβουμε ότι μετακινώντας την «σχισμή» (την σταθερά ONES_UPPER_THRESH) επηρεάζουμε το μέγιστο score που θα παραχθεί, διότι ο ΓΑ συμπεριφέρεται άπληστα και προσπαθεί να προσθέσει όσο περισσότερες επιλεγμένες λέξεις μπορεί

Β4. Αξιολόγηση και Επίδραση Παραμέτρων

Για την διαδικασία του feature selection, έγινε τροποποίηση της διαδικασίας preprocess. Συγκεκριμένα, προστέθηκε ένα βήμα κατά το οποίο εκτελείται ο γενετικός αλγόριθμος, και η έξοδος του (που είναι ένας αριθμός από επιλεγμένες λέξεις) εφαρμόζεται στο train dataset του νευρωνικού δικτύου. Η εφαρμογή του συνόλου λέξεων αυτού, είναι ουσιαστικά, ο περιορισμός του train dataset πάνω στις λέξεις που έχει επιστρέψει ο Γενετικός – δηλαδή αφαιρούνται όλες οι λέξεις που δεν έχουν επιλεγεί. Αυτή η διαδικασία έχει ως αποτέλεσμα να παραμείνει ο αριθμός εισόδων ο ίδιος (και ίσος με τον αριθμό δειγμάτων = 8251), όμως έχουν μειωθεί τα features στο επόμενο επίπεδο (το κρυφό επίπεδο του νευρωνικού δικτύου – αξιολογώντας τον χρόνο εκπαίδευσης αναλογικά με την απόδοση, χρησιμοποιήθηκε το νευρωνικό δίκτυο με 1 κρυφό επίπεδο με αριθμό νευρώνων = Ο).

Έγινε και hyperparameter tuning με το keras_tuner, οι καλύτερες τιμές του οποίου ήταν οι εξής:

alpha: 0.23, learning rate: 0.08, momentum: 0.16000000000000003, decay: 0.03

Έτρεξαν δύο testbenches (ένα με decay και ένα χωρίς). Παραθέτω τα αποτελέσματα:

Χωρίς decay:

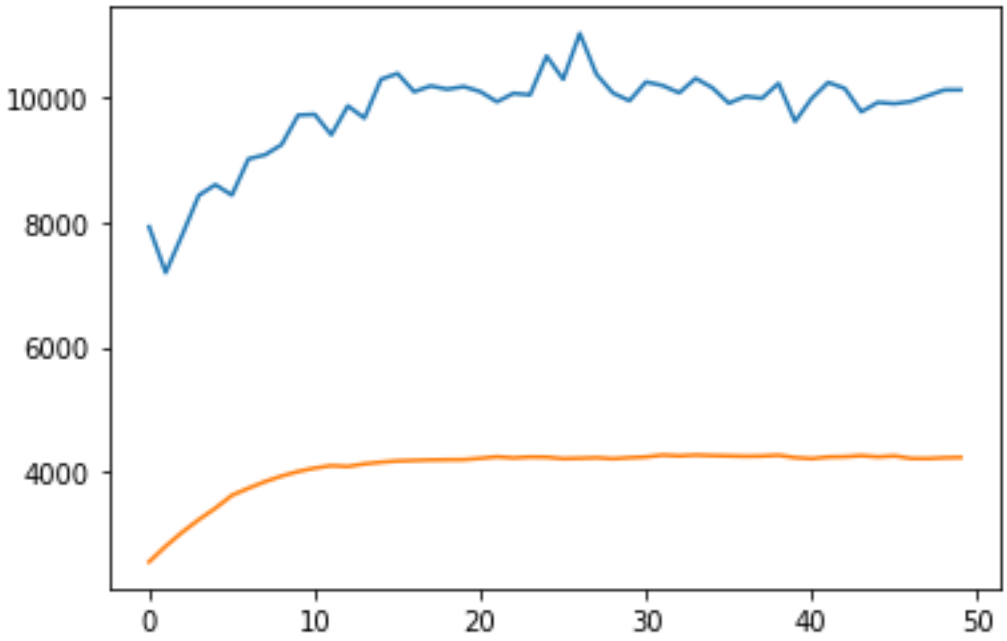


Figure 3- η εκτέλεση του γενετικού

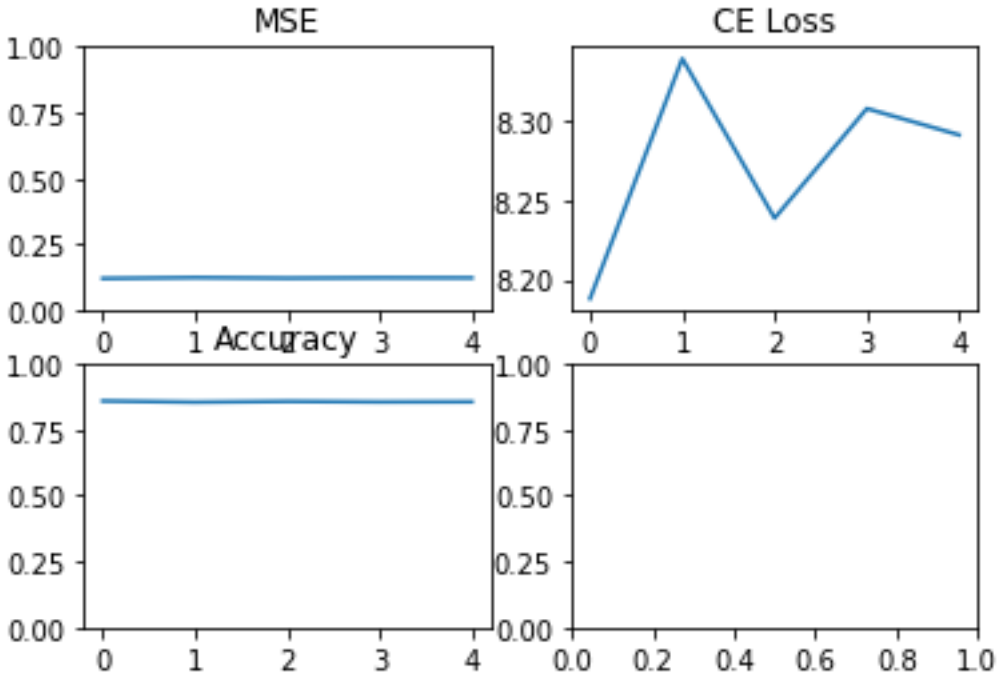
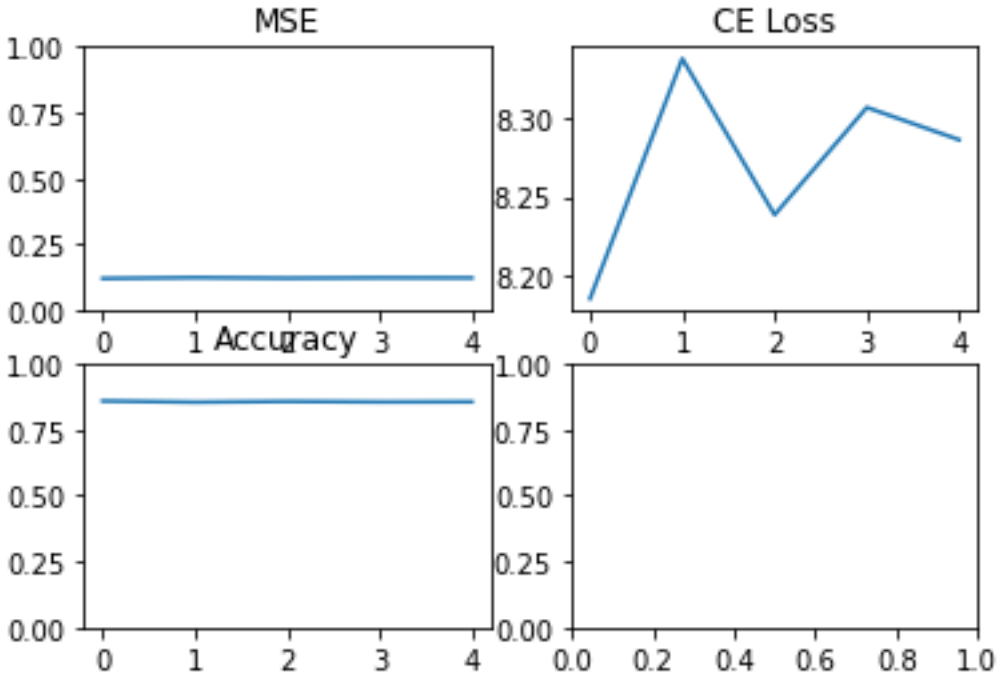


Figure 4- η απόδοση του νευρωνικού

Με decay:



Κατά την εκτέλεση του Hyperparameter tuning, αναφέρθηκε (για την καλύτερη εκτέλεση) η ελάχιστη τιμή για τη Loss function: 8.257484436035156 η οποία είναι καλύτερη από την ελάχιστη που πετύχαινε το Νευρωνικό δίκτυο (με τις αντίστοιχες παραμέτρους) στο μέρος Α: 8.289506912231445. Άρα έχουμε μια πολύ μικρή βελτίωση στην απόδοση του νευρωνικού δικτύου.

Άρα:

i) Η γενικευτική ικανότητα του δικτύου αυξήθηκε (οριακά)

ii) Το feature selection είχε ως αποτέλεσμα την ταχύτερη εκπαίδευση του Νευρωνικού δικτύου, την μείωση των εισόδων του κρυφού επιπέδου και τέλος, καλύτερη γενικευτική ικανότητα του Νευρωνικού δικτύου.

Πηγές

- https://www.geeksforgeeks.org/crossover-in-genetic-algorithm/
- Διαφάνειες μαθήματος