# **Υπολογιστική Νοημοσύνη**

## **Εργασία Μέρος B’**



Τσικέλης Ιωάννης

1067407

Τμήμα Μηχανικών Η/Υ και Πληροφορικής

Πανεπιστήμιο Πατρών

2021-2022

Table Of Contents

[**Υπολογιστική Νοημοσύνη** 1](#_Toc107352203)

[**Εργασία Μέρος B’** 1](#_Toc107352204)

[Σύντομη περιγραφή 3](#_Toc107352205)

[Κώδικας Υλοποίησης: 3](#_Toc107352206)

[B1. Σχεδιασμός Γεννετικού Αλγορίθμου 4](#_Toc107352207)

[Α) Κωδικοποίηση: 4](#_Toc107352208)

[Β) Αρχικός πληθυσμός: 4](#_Toc107352209)

[Γ) Διαδικασία επιδιόρθωσης: 4](#_Toc107352210)

[Δ) Υπολογισμός tf-idf: 5](#_Toc107352211)

[E) Συνάρτηση Καταλληλότητας: 5](#_Toc107352212)

[ΣΤ) Γεννετικοί τελεστές: 6](#_Toc107352213)

[B2. Υλοποίηση Γεννετικού Αλγορίθμου 7](#_Toc107352214)

[B3. Αξιολόγηση και Επίδραση Παραμέτρων 8](#_Toc107352215)

[Α) Εκτέλεση Αλγορίθμου με συγκεκριμένες τιμές: 8](#_Toc107352216)

[Β) Διαγράμματα: 9](#_Toc107352217)

[Γ) Συμπεράσματα: 12](#_Toc107352218)

[B4. Επιλογή Χαρακτηριστικών ΤΝΔ 13](#_Toc107352219)

[Α) Εκτέλεση κώδικα: 13](#_Toc107352220)

[Β) Παρατηρήσεις: 15](#_Toc107352221)

## Σύντομη περιγραφή

Στο δεύτερο μέρος της εργασίας του μαθήματος Υπολογιστική Νοημοσύνη, κληθήκαμε να υλοποιήσουμε έναν Εξελικτικό Αλγόριθμο, ο οποίο επιλέγει το βέλτιστο σύνολο λέξεων που περιγράφουν μία συλλογή κειμένων. Στο τέλος μας ζητήθηκε να εκπαιδεύσουμε ένα από τα νευρωνικά δίκτυα της προηγούμενης υλοποίησης με το νέο, μικρότερο σύνολο λέξεων. Για τον έλεγχο των υλοποιήσεων χρησιμοποιήθκε το dataset [DeliciousMIL](https://archive.ics.uci.edu/ml/datasets/DeliciousMIL%3A+A+Data+Set+for+Multi-Label+Multi-Instance+Learning+with+Instance+Labels) με πάνω από 12.000 διαφορετικές εγγραφές για training και testing. Ο κώδικας γράφτηκε σε γλώσσα [Python](https://www.python.org/).

## Κώδικας Υλοποίησης:

Ο κώδικας της άσκησης είναι διαθέσιμος στο GitHub repository: [itsikelis/ceid-comp-int-project](https://github.com/itsikelis/ceid-comp-int-project).

## B1. Σχεδιασμός Γεννετικού Αλγορίθμου

### Α) Κωδικοποίηση:

Για την κωδικοποίηση του κάθε ατόμου-χρωμοσώματος, χρησιμοποιήθηκε μία one-hot αναπαράσταση ενός διανύσματος 8520 θέσεων (ίσες με το πλήθος των λέξεων του λεξικού):

### Β) Αρχικός πληθυσμός:

Για την δημιουργία αρχικού πληθυσμού πλήθους Ν, για κάθε άτομο-χρωμόσωμα, δημιουργείται ένα νέο διάνυσμα με τυχαία επιλογή 0 ή 1 σε κάθε θέση του, σύμφωνα με μία δοθείσα πιθανότητα p(ij = 1). Η διαδικασία επαναλαμβάνεται για κάθε άτομο που δεν έχει τουλάχιστον 1000 μη μηδενικές τιμές, αφού αυτό είναι και το κάτω όριο που δίνεται από την άσκηση. Ο πληθυσμός αναπαρίσταται ως ένα Νx8520 μητρώο:

### Γ) Διαδικασία επιδιόρθωσης:

Ως ζητούμενο της άσκησης, κάθε άτομο του πληθυσμού είνα υποχρεωμένο να έχει τουλάχιστον 1000 μη μηδενικά στοιχεία-γονίδια (άσσους). Για τον έλεγχο αυτού μπορούν να χρησιμοποιηθούν οι κάτωθι τεχνικές:

1. Απόρριψη του μη-νόμιμου ατόμου: απόρριψη αποτελέσματος και επαναδημιουργία νέου ατόμου.
2. Διόρθωση: «γέμισμα» του ατόμου με περισσότερους άσσους, ώστε να πληροί το ελάχιστο κριτήριο.
3. Εφαρμογή ποινής: εφαρμογή ποινής στο άτομο με λιγότερους από 1000 άσσους κατά την αξιολόγησή του από την συνάρτηση καταλληλότητας.

Για την παρούσα υλοποίηση, επιλέψθηκαν τα κριτήρια της απόρριψης του μη νόμιμου ατόμου κατά την αρχική δημιουτγία πληθυσμού και, δεδομένου ότι κατά τη διασταύρωση μπορεί να προκύψει απόγονος με λιγότερους άσσους, εφαρμογή ποινής (μηδενισμός) κατά την αξιολόγηση του πληθυσμού σε κάθε βήμα του Αλγορίθμου.

### Δ) Υπολογισμός tf-idf:

Μία μετρική που θα χρησιμοποιηθεί για τον ορισμό της συνάρτησης καταλληλότητας, είναι η μετρική tf-idf. Ο προσδιορισμός των όρων tf και idf έγινε ως εξής. Για τις μετρικές tf της κάθε λέξης, χρησιμοποιήθηκε η αναπαράσταση ενός διανύσματος 1x8520 θέσεων, που περιείχε τον μέσο όρο των τιμών tf για κάθε λέξη σε κάθε κείμενο.

Αντίστοιχα οι τιμές idf για την κάθε λέξη αναπαριστώνται επίσης σε ένα 1x8520 διάνυσμα.

Έτσι, το μέσο γινόμενο tf\*idf για την κάθε λέξη, είναι το γινόμενο Hadamard των παραπάνω διανυσμάτων.

### E) Συνάρτηση Καταλληλότητας:

Για την επιλογή συνάρτησης καταλληλότητας, έπρεπε να ληφθούν υπ’ όψη δύο παράγοντες για το κάθε άτομο-χρωμόσωμα:

1. η μέση τιμή των tf\*idf μετρικών των λέξεών του και
2. το πλήθος των επιλεγμένων λέξεων, ιδανικά κοντά στο κάτων όριο των 1000.

Σύμφωνα με αυτούς τους δύο περιορισμούς δημιοργήθηκε η εξής συνάρτηση καταλληλότητας:

Όπως είναι κατανοητό λοιπόν, για τον υπολογισμό του score του κάθε ατόμου, η μετρική tf\*idf παίζει θετικό ρόλο, το άνω το 1000 πλήθος λέξεων, αρνητικό. Η συνάρτηση θεωρητικά μεγιστοποιείται (απειρίζεται) όταν το άτομο i, έχει ακριβώς 1000 λέξεις, οι οποίες εκφράζουν πλήρως όλα τα κείμενα ή έχει ακριβώς 1000 λέξεις (πράγμα δεν είναι και τόσο θεμιτό) και θεωρητικά ελαχιστοποιείται (απειρίζεται αρνητικά) όταν οι επιλεχθείσες λέξεις, ανεξαρτήτου πλήθους, δεν εκφράζουν σε κανένα βαθμό κανένα κείμενο. Ο πολλαπλασιασμός με ένα βαθμωτό, έγινε για την αποφυγή πολύ μικρών αριθμών κατά τον υπολογισμό του score.

### ΣΤ) Γεννετικοί τελεστές:

Οι γεννετικοί τελεστές για την επιλογή, διασταύρωση, και μετάλλαξη επιλέχθηκαν ως εξής:΅

1. Επιλογή: για την επιλογή των ατόμων που θα περάσουν στην επόμενη γενιά, χρησιμοποιήθηκε η ρουλέτα με βάση το κόστος. Η επιλογή αυτή έγινε για την αποφυγή πολύ κακών ατόμων, που μπορεί να προέκυπταν από την χρήση άλλων τεχνικών όπως ρουλέτας με βάση το κόστος ή τουρνουά.
2. Διασταύρωση: λόγω της στατιστικής ανεξαρτησίας της κάθε λέξης με τις γειτονικές της, για την διαδικασία της διασταύρωσης, επιλέχθηκε η τεχνική της ομοιόμορφης διασταύρωης. Κατά την διασταύρωση, επιλέγεται με πιθανότητα 0.5 αν οι δύο γονείς θα ανταλλάξουν τα γονίδιά τους, για κάθε γονίδιο τους(8520).
3. Μετάλλαξη: για την όσο είναι εφικτό αποφυγή μη ένταξης ελέγχου κάποιων λέξεων, κατά την εκτέλεση του Αλγορίθμου με μικρούς πληθυσμούς, επιλέχθηκε η διαδικασία της μετάλλαξης να εφαρμόζεται σε όλα τα άτομα του εκάστοτε πληθυσμού και να μην γίνει χρήση ελιτισμού.

## B2. Υλοποίηση Γεννετικού Αλγορίθμου

Η υλοποίηση του Αλγορίθμου έγινε στη γλώσσα Python και είναι διαθέσιμη στο GitHub repository: [itsikelis/ceid-comp-int-project](https://github.com/itsikelis/ceid-comp-int-project).

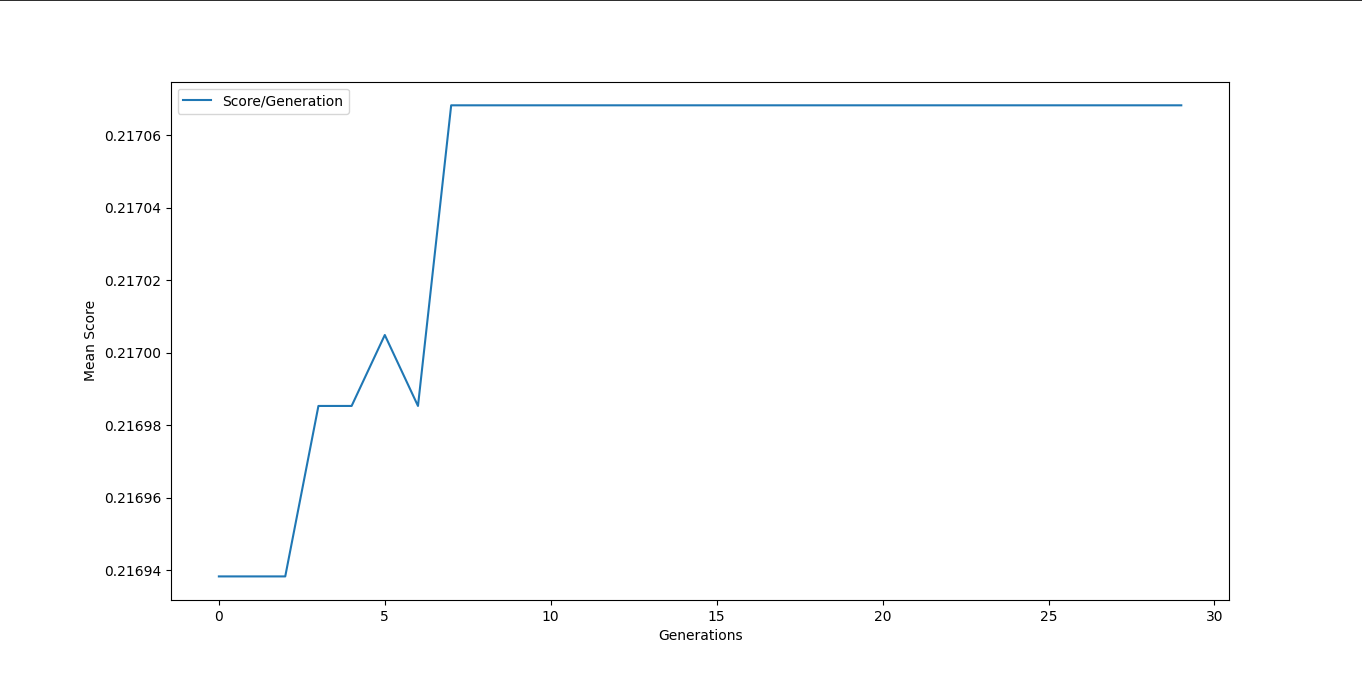
## B3. Αξιολόγηση και Επίδραση Παραμέτρων

### Α) Εκτέλεση Αλγορίθμου με συγκεκριμένες τιμές:

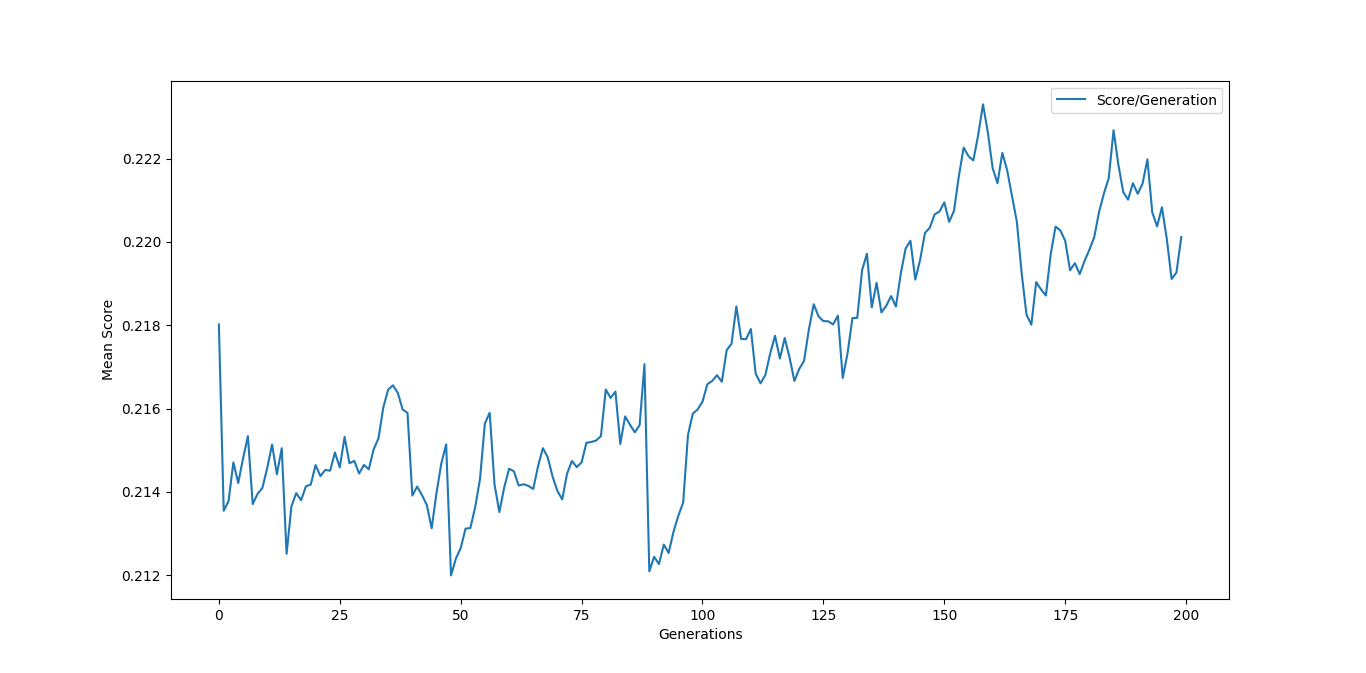
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| A/A | Μέγεθος Πληθυσμού | Πιθανότητα Διασταύρωσης | Πιθανότητα Μετάλλαξης | μ.τ Βέλτιστου | μ.α Γενεών |
| 1 | 20 | 0.6 | 0.00 | 0.21704 | 30.20 |
| 2 | 20 | 0.6 | 0.01 | 0.21696 | 198.20 |
| 3 | 20 | 0.6 | 0.10 | 0.22043 | 389.25 |
| 4 | 20 | 0.9 | 0.01 | 0.21661 | 151.25 |
| 5 | 20 | 0.1 | 0.01 | 0.21524 | 75.75 |
| 6 | 200 | 0.6 | 0.00 | 0.21641 | 30.00 |
| 7 | 200 | 0.6 | 0.01 | 0.22418 | 181.25 |
| 8 | 200 | 0.6 | 0.10 | 0.22424 | 250.50 |
| 9 | 200 | 0.9 | 0.01 | 0.22471 | 237.50 |
| 10 | 200 | 0.1 | 0.01 | 0.22523 | 260.00 |

### Β) Διαγράμματα:

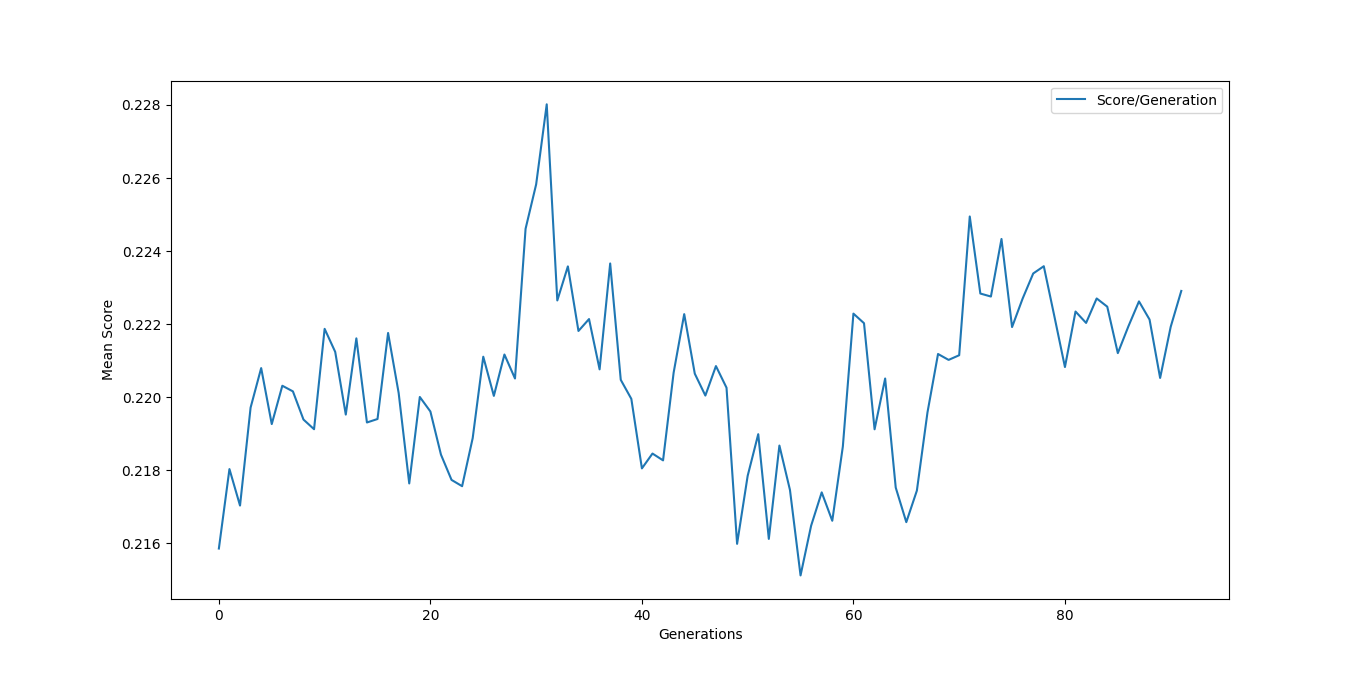
Λόγω σφάλματος κατά την αποθήκευση, ορισμένα διαγράμματα δυστυχώς απουσιάζουν.



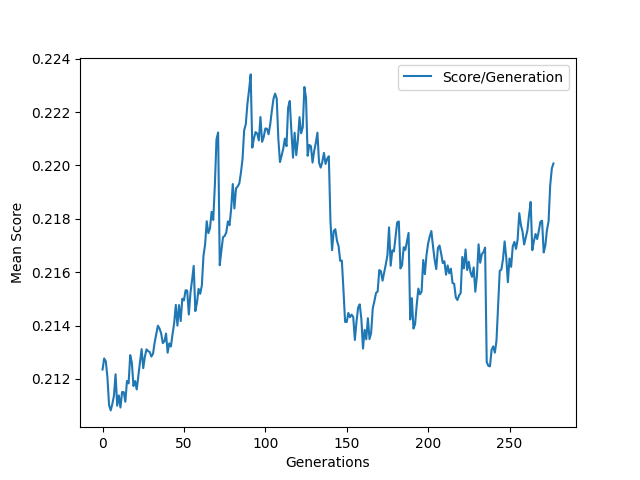
Εικόνα : Περίπτωση 1



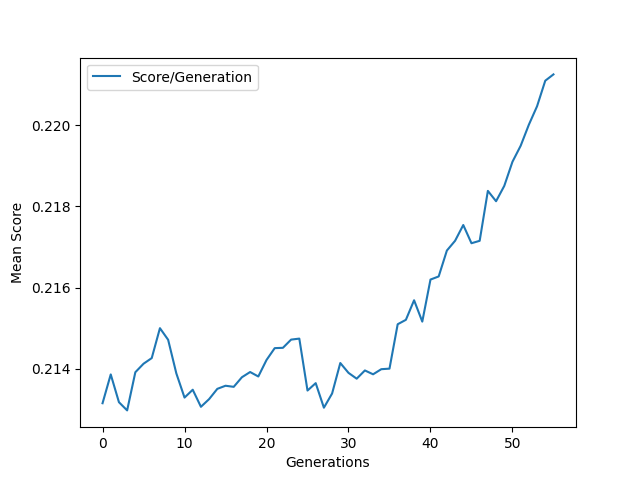
Εικόνα : Περίπτωση 2



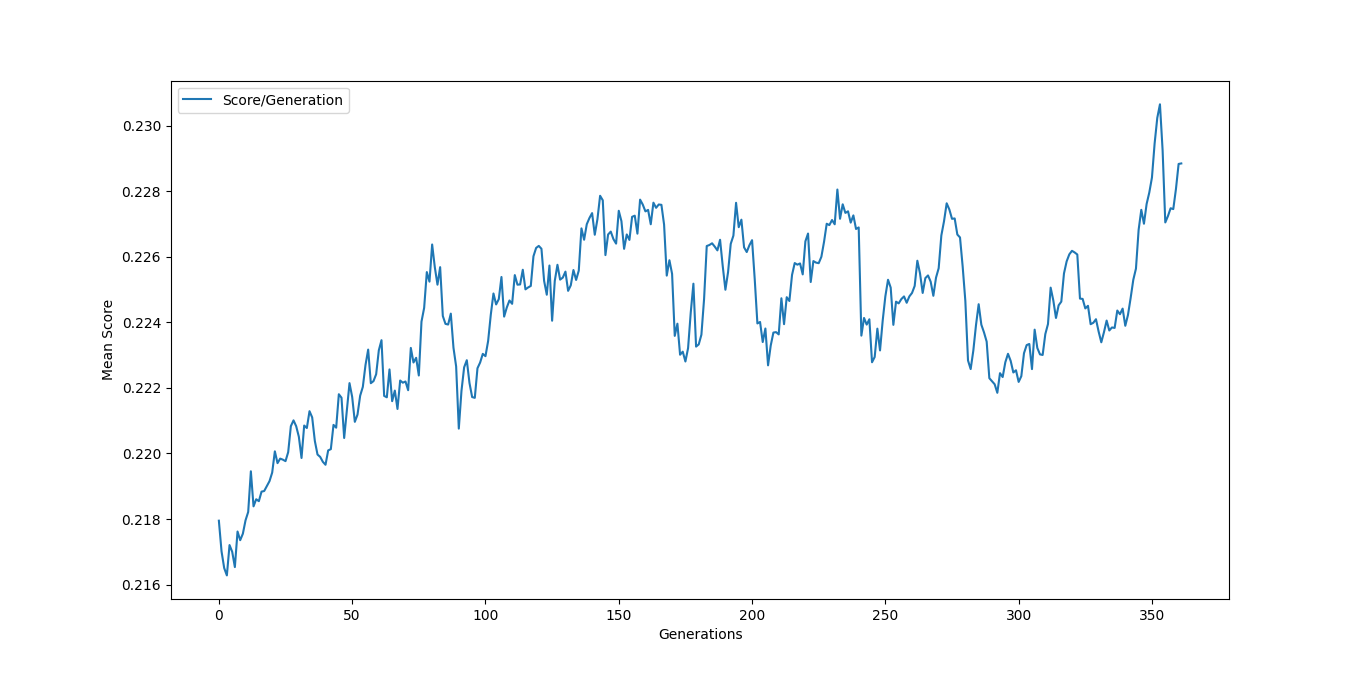
Εικόνα : Περίπτωση 3



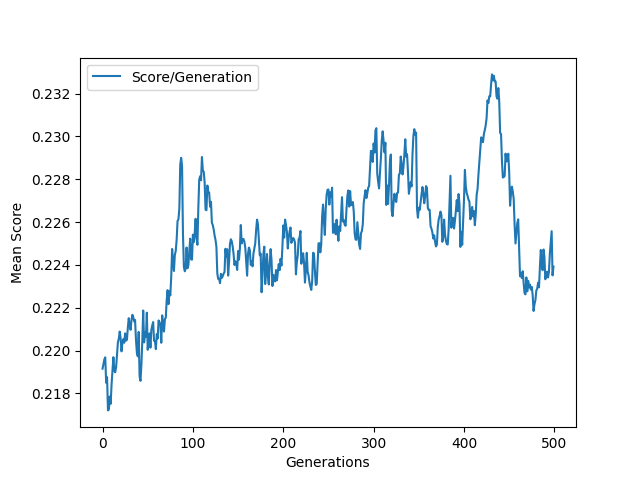
Εικόνα : Περίπτωση 4



Εικόνα : Περίπτωση 5



Εικόνα 7: Περίπτωση 7



Εικόνα 10: Περίπτωση 10

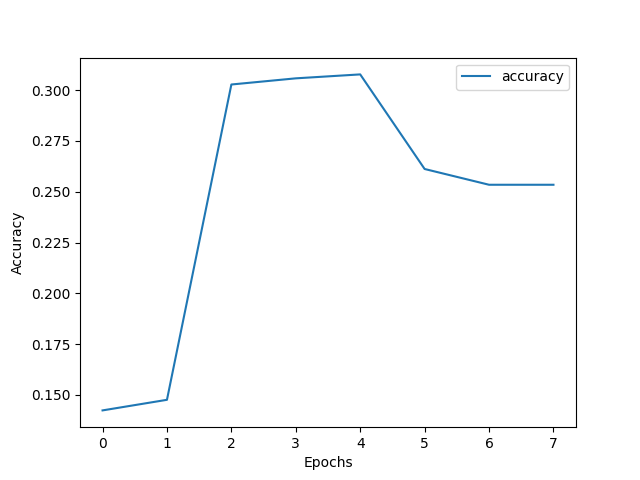
### Γ) Συμπεράσματα:

Από τα αποτελέσματα των παραπάνω εκτελέσεων, παρατηρούμε ότι η βέλτιστη περίπτωης, προκύπτει με μεγάλο αρχικό πληθυσμό, μικρή πιθανότητα διασταύρωσης και μικρή πιθανότητα μετάλλαξης. Αυτό σίγουρα οφείλεται στον τρόπο που έχει επιλεχθεί για την διασταύρωση(uniform), καθώς και στο γεγονός ότι η μετάλλαξη πρέπει να είναι σπάνιο φαινόμενο, που ενδεχομένως μπορεί να οδηγήσει σε βελτίωση.

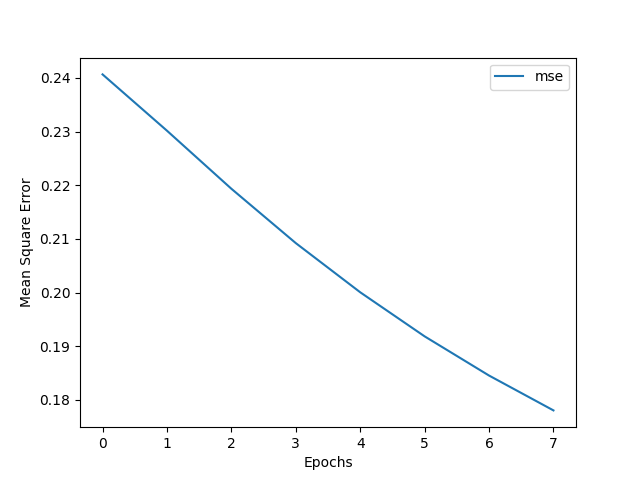
## B4. Επιλογή Χαρακτηριστικών ΤΝΔ

### Α) Εκτέλεση κώδικα:

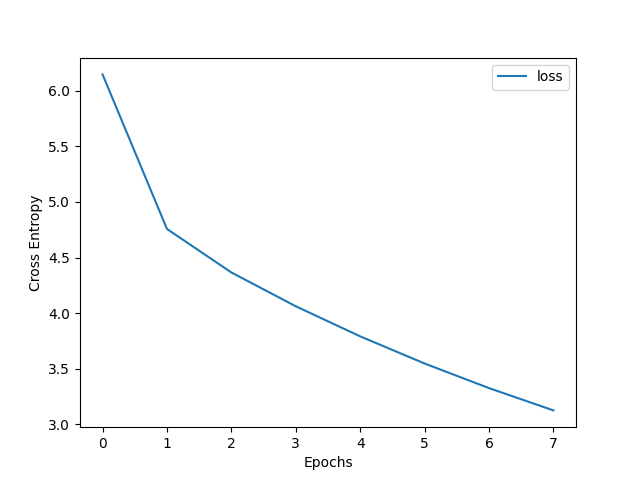
Στο τελευταίο κομμάτι της άσκησης, κληθήκαμε να χρησιμοποιήσουμε τα αποτελέσματα του Γεννετικού Αλγορίθμου και να εκπαιδεύσουμε καταλλήλως το καλύτερο νευρωνικό δίκτυο που προέκυψε από το προηγούμενο μέρος της άσκησης. Για την υλοποίηση, έπειτα από την εκτέλεση του Γεννετικού Αλγορίθμου, πάρθηκε το καλύτερο άτομο-χρωμόσωμα, από τον βέλτιστο πληθυσμό που προέκυψε από τις επαναλήψεις και χρησιμοποιήθηκε για να αφαιρεθουν οι μη επιλεγμένες από αυτό λέξεις από το σύνολο εισόδου στο νευρωνικό.



Εικόνα: Ακρίβεια/Εποχή



Εικόνα: Μέσο Τετραγωνικό Σφάλμα/Εποχή



Εικόνα: Συνάρτηση Κόστους/Εποχή

### Β) Παρατηρήσεις:

Το πρώτο πράγμα που παρατηρήθηκε κατά την εκτέλεση του νευρωνικού, ήταν ο μειωμένος απαιτούμενος χρόνος υπολογσιμού για κάθε εποχή, γεγονός αναμενόμενο, καθώς το σύνολο εισόδου είχε μειωθεί κατά πολύ. Επιπλέον, παρατηρείται ταχύτερη σύγκλιση του νευρωνικού σε λύση, σε λιγότερες εποχές (αυτό μπορεί να οφείλεται και εν μέρει στην λάθος χρήση συνάρτησης αποκοπής, κατά την εκτέλεση του κώδικα). Σε σύγκριση με τα προηγούμενα αποτελέσματα, η ακρίβεια του δικτύου έχει μειωθεί και το μέσο τετραγωνικό σφάλμα είναι αυξημένο, γεγονός που σίγουρα οφείλεται εν μέρει στην μείωση των εισόδων του δικτύου. Τέλος, η γενικευτική ικανότητα του δικτύου θεωρητικά έχει πλέον αυξηθεί, δεδομένου ότι πλέον έχουν διατηρηθεί μόνο οι πραγματικά σημαντικές λέξεις που ταξινομούν το κάθε κείμενο.