ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΥΠΟΛΟΓΙΣΤΙΚΗ ΝΟΗΜΟΣΥΝΗ

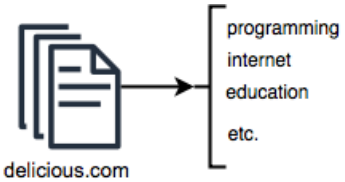
Διδάσκων: Δ. Κουτσομητρόπουλος

Ακαδημαϊκό Έτος 2021-2022

Εργαστηριακή Άσκηση
Μέρος Α΄

Α. Αναγνώριση κειμένου με Χρήση Νευρωνικών Δικτύων

Συστήματα επεξεργασίας φυσικής γλώσσας (natural language processing[1]), με την κωδικοποίηση λέξεων που προσφέρουν, συμβάλλουν καθοριστικά στην κατανόηση κειμένου από υπολογιστικά συστήματα. Η αξιοποίηση των συστημάτων αυτών βοήθησε στην αποδοτικότερη διαχείριση δεδομένων κειμενικής μορφής, με ανάπτυξη πεδίων έρευνας όπως ανάλυση συναισθήματος, αναγνώριση ομιλίας, κ.α. Μια χαρακτηριστική περίπτωση με ιδιαίτερο επιστημονικό ενδιαφέρον αποτελεί και η προσπάθεια ανάθεσης ετικετών σε ένα σώμα κειμένου. Επομένως, στην παρούσα εργασία, σας δίνονται δεδομένα κειμένου και σας ζητείται υλοποίηση που να αναθέτει μια ή περισσότερες ετικέτες σε αυτά. Στην ακόλουθη εικόνα έχουμε παράδειγμα με ανάθεση ετικετών σε ιστοσελίδες από χρήστες του site delicious.com.



Ειδικότερα, πρέπει θα εξετάσετε τη χρήση ενός πολυεπίπεδου ΤΝΔ για την πρόβλεψη των ετικετών (αρχείο labels.txt) που αντιστοιχούν σε κάθε κείμενο. Για τον σκοπό αυτό θα αξιοποιηθεί το σύνολο δεδομένων *DeliciousMIL*[2] που συγκεντρώνει *8.251 δείγματα* για εκπαίδευση και *3.983 δείγματα* για έλεγχο. Τα δεδομένα κειμένου περιέχονται στο αρχείο train-data.dat και test-data.dat, ενώ αντίστοιχα οι ετικέτες στα αρχεία train-label.dat και test-label.dat. Η κωδικοποίηση λέξεων του κειμένου δίνεται έτοιμη με την μορφή λεξικού, μεγέθους 8.520 (αρχείο vocabs.txt). Στο λεξικό έχει ήδη γίνει αποκοπή καταλήξεων (stemming) και αφαίρεση των διακοπτουσών λέξεων (stop words).

Η κάθε εγγραφή - γραμμή περιλαμβάνει το πλήθος των προτάσεων στην αρχή του σώματος κειμένου, το πλήθος των λέξεων στην αρχή της κάθε πρότασης και τέλος την κωδικοποίηση της κάθε λέξης. Οπότε έχουμε τις παρακάτω μη προκαθορισμένες στήλες, χωρισμένες με κενό:

<# sentences> <# words> word_1 word_2 … <# words> word_1 …

Οι ετικέτες προσφέρονται σε διανύσματα των 20 θέσεων, με 1 και 0 στις θέσεις που υπάρχει ή όχι ετικέτα. Για παράδειγμα, η πρώτη γραμμή των δεδομένων κειμένου εκπαίδευσης έχει ως εξής:

<2> <8> 6705 5997 8310 3606 674 8058 5044 4836 <4> 4312 5154 8310 4225

και οι αντίστοιχές ετικέτες: 1 0 1 1 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0

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

Για την υλοποίηση των αλγορίθμων μπορείτε να χρησιμοποιήσετε οποιοδήποτε περιβάλλον, βιβλιοθήκη ή γλώσσα προγραμματισμού κρίνετε σκόπιμο. Ενδεικτικά αναφέρονται *MatLab, WEKA, Azure ML Studio, Google Colaboratory, TensorFlow, Keras, SciKit-Learn.*

Το ζητούμενο στην εργασία αυτή είναι να κατασκευαστεί και να εκπαιδευτεί ένα ΤΝΔ που να ταξινομεί εισόδους (σώματα κειμένου) σε είκοσι διαφορετικές κλάσεις.

Α1. Προεπεξεργασία και Προετοιμασία δεδομένων [*20 μονάδες*]

Προσοχή: Ό,τι μετασχηματισμοί εφαρμοστούν στα δεδομένα του συνόλου εκπαίδευσης, οι ίδιοι θα πρέπει να εφαρμοστούν και στα δεδομένα του συνόλου ελέγχου ή εναλλακτικά να αντιστραφούν πρώτου μετρηθούν οι μετρικές αξιολόγησης παρακάτω.

α) *Κωδικοποίηση εισόδων:* Οι τιμές της κωδικοποίησης των λέξεων, όπως αναφέρθηκε, είναι ακέραιες και κινούνται στο διάστημα [0,8519]. Ωστόσο, η είσοδος του δικτύου είναι ένα συνολικό σώμα κειμένου, δηλαδή μια παράγραφος που αποτελείται από πολλές φράσεις και λέξεις, το πλήθος των οποίων είναι διαφορετικό κάθε φορά. Χρειάζεται επομένως να καθοριστούν τα χαρακτηριστικά που θα δίνονται ως είσοδος στο δίκτυο. Μια απλουστευμένη κωδικοποίηση χαρακτηριστικών κειμένου δίνεται από το μοντέλο Bag-of-Words (BoW). Με βάση αυτό, ένα σώμα κειμένου μπορεί να αναπαρασταθεί ως ένα διάνυσμα διάστασης όσο το μέγεθος του λεξικού (8520) και σε κάθε θέση του αποθηκεύεται ένας ακέραιος αριθμός που αντιστοιχεί στη συχνότητα εμφάνισης της κάθε λέξης στο κείμενο. Για παράδειγμα, με βάση το συγκεκριμένο λεξικό, η φράση: 6705 5997 8310 3606 674 8058 5044 4836 4312 5154 8310 4225. Μπορεί να κωδικοποιηθεί ως το παρακάτω αραιό διάνυσμα 8520 θέσεων:

Τιμή	[0	0	…	1	…	1	…	1	…	2	…	0]
Θέση	0	1	…	674	…	4225	…	8058	…	8310	…	8519

Κωδικοποιήστε επομένως τα δεδομένα εισόδου σύμφωνα με το μοντέλο BoW. [5]

β) Ως συνέπεια τις ανωτέρω κωδικοποίησης κάθε διάνυσμα εισόδου θα περιέχει μη μηδενικές τιμές, όπου όμως κάποιες μπορεί να είναι συστηματικά πολύ υψηλές αν αφορούν λέξεις που εμφανίζονται συχνά σε ένα κείμενο και κάποιες άλλες όχι. Με δεδομένη την συγκεκριμένη αποτύπωση και την πιθανή ανάγκη προσαρμογής των τιμών αυτών σε διαφορετική κλίμακα, ιδιαίτερα αν υπάρχουν ακραίες τιμές (για εξάλειψη ενδεχόμενης πόλωσης), παρουσιάζονται οι εξής τρείς μέθοδοι:

- *Κεντράρισμα (Centering):* Με την μέθοδο αυτή αφαιρούμε τον μέσο όρο των εμφανίσεων όλων των λέξεων από κάθε τιμή του διανύσματος.
- *Κανονικοποίηση (Normalization):* Με την μέθοδο αυτή μεταφέρουμε το εύρος τιμών των κωδικοποιήσεων των λέξεων σε νέα κλίμακα πχ [0,1].
- *Τυποποίηση (Standardization):* Με την μέθοδο αυτή παρέχουμε στο δείγμα ιδιότητες όπως μηδενική μέση τιμή και μοναδιαία διακύμανση (Gaussian).

Εξετάστε τη χρησιμότητα των ανωτέρω μεθόδων για το συγκεκριμένο πρόβλημα και εφαρμόστε τη/τις στα δεδομένα εκπαίδευσης, αν κρίνετε σκόπιμο. [10]

γ) Διασταυρούμενη Επικύρωση *(cross-validation):* Βεβαιωθείτε ότι έχετε διαχωρίσει τα δεδομένα σας σε σύνολα εκπαίδευσης και ελέγχου, ώστε να χρησιμοποιήσετε 5-fold CV για όλα τα πειράματα. [5]

Α2. Επιλογή αρχιτεκτονικής [*50 μονάδες*]
Όσον αφορά την τοπολογία των ΤΝΔ για την εκπαίδευση τους με τον Αλγόριθμο Οπισθοδιάδοσης του Σφάλματος (back-propagation), θα χρησιμοποιήσετε ΤΝΔ με *ένα κρυφό*

επίπεδο και θα πειραματιστείτε με τον αριθμό των κρυφών κόμβων. Για την εκπαίδευση του δικτύου χρησιμοποιήστε αρχικά ρυθμό μάθησης η = 0.001.

α) Η εκπαίδευση και αξιολόγηση των μοντέλων σας μπορεί να γίνει με χρήση *Cross-Entropy* (CE), *Μέσου Τετραγωνικό Σφάλματος* (MSE), καθώς και *Accuracy* (Acc) για τις τιμές που περιέχονται στα σύνολα εκπαίδευσης και ελέγχου[3]. Να εξηγήσετε με απλά λόγια ποια είναι η σημασία των παραπάνω μετρικών για το συγκεκριμένο πρόβλημα. [5]

β) Πόσους νευρώνες θα χρειαστείτε στο επίπεδο εξόδου, δεδομένου του ζητούμενου της ταξινόμησης σε πολλαπλές διαφορετικές κλάσεις (multilabel multiclass classification); [5]

γ) Να επιλέξετε κατάλληλη συνάρτηση ενεργοποίησης για τους κρυφούς κόμβους και να τεκμηριώσετε την επιλογή σας. [5]

δ) Ποια συνάρτηση ενεργοποίησης θα χρησιμοποιήσετε για το επίπεδο εξόδου; Σιγμοειδή, Softmax ή κάποια άλλη; [5]

ε) Πειραματιστείτε με 3 διαφορετικές τιμές για τον αριθμό των νευρώνων του κρυφού επιπέδου και συμπληρώστε τον παρακάτω πίνακα. Εμπειρικά ενδεδειγμένες τιμές για τον αριθμό των κρυφών κόμβων βρίσκονται στο διάστημα [*O, I+O*] (*I αριθμός εισόδων, O αριθμός εξόδων, H αριθμός κόμβων στο κρυφό επίπεδο)*. Να συμπεριλάβετε και τις γραφικές παραστάσεις σύγκλισης (Μ.Ο.) ανά κύκλο εκπαίδευσης. Διατυπώστε τα συμπεράσματά σας σχετικά με (i) τον αριθμό των κρυφών κόμβων, (ii) την επιλογή της συνάρτησης κόστους και (iii) την ταχύτητα σύγκλισης ως προς τις εποχές εκπαίδευσης. [15]

Αριθμός νευρώνων στο κρυφό επίπεδο	**CE loss**	**MSE**	**Acc**
$H_1 = O$			
$H_1 = (I+O)/2$			
$H_1 = I+O$			

στ) Με τον βέλτιστο αριθμό κρυφών κόμβων που βρήκατε στο ε) δοκιμάστε να προσθέσετε ένα ακόμα κρυφό επίπεδο H_2 στο δίκτυο. Πειραματιστείτε με τον αριθμό των κόμβων του H_2. Περιγράψτε μια λογική για τη στοίχιση των κρυφών επιπέδων (είναι καλό να έχουν τον ίδιο αριθμό κόμβων; Μειούμενο; Αυξανόμενο;). Να αναφέρετε CE, MSE και Acc και να διατυπώσετε τα συμπεράσματα σχετικά με την προσθήκη κρυφών επιπέδων. [10]

Αριθμός νευρώνων στο κρυφό επίπεδο	**CE loss**	**MSE**	**Acc**
$H_2 =$			
$H_2 =$			
$H_2 =$			

[3] Έστω δύο διανύσματα **p** και **q** διάστασης K, όπου K ο αριθμός των κλάσεων, **p** επιθυμητό και **q** πραγματικό διάνυσμα εξόδου. Έστω **x** το διάνυσμα εισόδου και n το πλήθος των δειγμάτων. $CE = \frac{1}{n}\sum_{\boldsymbol{x}} \sum_{i=1}^{K} p_i(\boldsymbol{x})\log\left(q_i(\boldsymbol{x})\right)$ και επιστρέφει μόνο θετικές τιμές. Όσες από αυτές είναι μικρές, θα αναφέρονται σε μεγάλη ομοιότητα. Επίσης, $MSE = \frac{1}{2n}\sum_{\boldsymbol{x}} \sum_{i=1}^{K}(p_i(\boldsymbol{x}) - q_i(\boldsymbol{x}))^{\boldsymbol{2}}$.Το MSE ισοδυναμεί με το τετράγωνο της ευκλείδειας απόστασης των δύο διανυσμάτων, με μικρότερες τιμές να αναφέρονται σε μικρότερο σφάλμα. Τέλος, για p, q δυαδικά διανύσματα, η ακρίβεια (*Acc)* στην περίπτωση πολλαπλών ετικετών μπορεί να υπολογιστεί ως $Acc = \frac{1}{n}\sum_{\boldsymbol{x}} \frac{|p(x) \wedge q(x)|}{|p(x) \vee q(x)|}$, όπου $\wedge$, $\vee$ bitwise AND, OR.

ζ) Κριτήριο τερματισμού. Επιλέξτε και τεκμηριώστε κατάλληλο κριτήριο τερματισμού της εκπαίδευσης κάθε φορά (για κάθε fold). Μπορεί να χρησιμοποιηθεί η τεχνική του πρόωρου σταματήματος (early stopping); [5]

Προσοχή: σε όλα τα πειράματα θα χρησιμοποιήσετε 5-fold cross validation (5-fold CV).

Α3. Μεταβολές στον ρυθμό εκπαίδευσης και σταθεράς ορμής [15 *μονάδες*]
Επιλέγοντας την τοπολογία που δίνει το καλύτερο αποτέλεσμα βάσει του προηγούμενου ερωτήματος, να πραγματοποίησετε βελτιστοποίηση των υπερπαραμέτρων ρυθμού εκπαίδευσης η και σταθεράς ορμής m με χρήση CV και να συμπληρώστε τον παρακάτω πίνακα. Να συμπεριλάβετε και τις γραφικές παραστάσεις σύγκλισης (Μ.Ο.) ως προς τους κύκλους εκπαίδευσης που θα χρειαστούν. Να τεκμηριώσετε θεωρητικά γιατί m < 1.

η	**m**	**CE loss**	**MSE**	**Acc**
0.001	0.2			
0.001	0.6			
0.05	0.6			
0.1	0.6			

Να διατυπώσετε σύντομα τα συμπεράσματα που προκύπτουν από τα 4 πειράματα.

Α4. Ομαλοποίηση [*15 μονάδες*]
Μια μέθοδος για την αποφυγή υπερπροσαρμογής του δικτύου και βελτίωση της γενικευτικής του ικανότητας είναι η ομαλοποίηση του διανύσματος των βαρών (regularization). Να εφαρμόσετε *L2* ομαλοποίηση (φθορά βαρών) και να επανεκπαιδεύσετε το δίκτυό σας, όπως προέκυψε από το Α3, αξιολογώντας διάφορες τιμές για τον συντελεστή φθοράς *r*.

i) r = 0.1 ii) r = 0.5 iii) r = 0.9
Συμπληρώστε τον παρακάτω πίνακα για κάθε μία από τις παραπάνω περιπτώσεις με χρήση 5-fold CV. Να συμπεριλάβετε και τις γραφικές παραστάσεις σύγκλισης (Μ.Ο.) ανά κύκλο εκπαίδευσης.

Συντελεστής φθοράς	**CE loss**	**MSE**	**Acc**
0.1			
0.5			
0.9			

Διατυπώστε τα συμπεράσματά σας σχετικά με την επίδραση της μεθόδου στη γενικευτική ικανότητα του δικτύου.

Α5. Ενσωματώσεις Λέξεων *[προαιρετικό ερώτημα - 10 μονάδες bonus]*
Οι ενσωματώσεις λέξεων (word embeddings) μπορούν να χρησιμοποιηθούν για τη διανυσματική αναπαράσταση λέξεων και κειμένων, εναλλακτικά προς το απλοϊκό μοντέλο BoW. Για την εκμάθηση αυτών των αναπαραστάσεων χρησιμοποιείται ένα επίπεδο ενσωματώσεων (embedding layer). Το επίπεδο αυτό μπορεί να χρησιμοποιηθεί ώστε να μάθει διανυσματικές αναπαραστάσεις λέξεων αλλά και να αξιοποιήσει υπάρχουσες (που έχουν προκύψει από ήδη εκπαιδευμένα μοντέλα). Σε μια άλλη χρήση του, μπορεί να αποτελέσει μέρος ενός μοντέλου βαθιάς μάθησης όπου θα μαθαίνει αναπαραστάσεις λέξεων μαζί με το ίδιο το μοντέλο. Η προσθήκη του εξασφαλίζει την μετατροπή ενός αραιού διανύσματος ακέραιων αριθμών σε ένα πυκνό διάνυσμα σταθερού μεγέθους.

α) Αξιοποιήσετε το επίπεδο αυτό, διαμορφώνοντας κατάλληλα τόσο την είσοδο όσο και τις τιμές των ορισμάτων[4]. Ως είσοδος μπορούν να χρησιμοποιηθούν οι φράσεις των κειμένων όπως είναι κωδικοποιημένες στο dataset: δηλαδή, διανύσματα ακεραίων μεταβλητού μήκους στα οποία μπορεί να γίνει padding. Τα embeddings που προκύπτουν στην έξοδο αντιστοιχούν στα χαρακτηριστικά του κειμένου που εξάγονται. Ανάλογα με το μέγεθος του λεξικού, η διάσταση ενός embedding είναι καλό να μην είναι πολύ μεγάλη, π.χ. μικρότερη από 100.

β) Χρησιμοποιήστε αυτό το layer ως είσοδο στην καλύτερη αρχιτεκτονική που προέκυψε από το Α2. Να αναφέρετε τιμές για CE, MSE και Acc, σχολιάστε τα αποτελέσματα και συγκρίνετε με την προηγούμενη υλοποίηση.

γ) Τα Long short-term memory (LSTM) αποτελούν επαναλαμβανόμενα μοντέλα μάθησης που βρίσκουν εφαρμογή σε διάφορους τομείς, χαρακτηριστικό παράδειγμα και οι εφαρμογές επεξεργασίας κειμένου. Τα LSTM, με τον μηχανισμό μνήμης που ενσωματώνουν, διαχειρίζονται αποτελεσματικά κάθε είδους ακολουθιακά δεδομένα. Η ικανότητα αυτή βοηθά στην ανακάλυψη εξαρτήσεων – σχέσεων που υπάρχουν μεταξύ των λέξεων ενός κειμένου. Σας ζητείται να υλοποιήσετε μοντέλα αποτελούμενα από: i) το επίπεδο ενσωματώσεων που υλοποιήσατε στο (α), ii) LSTM κρυφό επίπεδο και iii) έξοδο για την ταξινόμηση των κειμένων, όπως πριν. Πειραματιστείτε με διαφορετικές αρχιτεκτονικές (τουλάχιστον δύο) τροποποιώντας το κρυφό επίπεδο. Συγκεκριμένα, επιλέξτε έως δύο κρυφά επίπεδα και εξετάστε διαφορετικούς αριθμούς νευρώνων. Να αναφέρετε τιμές για CE, MSE και Acc, σχολιάστε τα αποτελέσματα και συγκρίνετε με την υλοποίηση στο (β).

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 link προς τον κώδικα που έχετε χρησιμοποιήσει (σε κάποια file sharing υπηρεσία ή code repo).
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, 2/5/2022, στις 23:59.
2. Για οποιαδήποτε διευκρίνηση / ερώτηση μπορείτε να χρησιμοποιείτε το σχετικό forum στο eclass του μαθήματος.

[4] Δείτε για παράδειγμα την υλοποίηση εδώ: https://keras.io/api/layers/core_layers/embedding/