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 Α΄

Α. Χρονολόγηση Αρχαίων Επιγραφών με Χρήση Νευρωνικών Δικτύων

Η αρχαιολογική έρευνα μπορεί να ωφεληθεί σημαντικά από την αξιοποίηση νέων τεχνολογιών. Ιδιαίτερα στην περίπτωση των αρχαίων επιγραφών, υπάρχουν εγγενή προβλήματα, όπως μερική ή ολική καταστροφή, κατακερματισμός, δυσανάγνωστη αναγραφή ή απώλεια συμβόλων και γραμμάτων, αμφισημία κ.α. που καθιστούν δυσχερή την χρονολόγηση και την ταξινόμησή τους.

Στη συγκεκριμένη εργασία σας δίνεται ένα απόσπασμα[1] από το μεγαλύτερο διαθέσιμο σύνολο δεδομένων αρχαίων ελληνικών επιγραφών, το I.PHI[2], που προέρχεται από την επεξεργασία και ανάλυση της βάσης δεδομένων PHI (Packard Humanities Institute). Τα δεδομένα αυτά περιλαμβάνουν μεταξύ άλλων τη μεταγραφή του κειμένου των επιγραφών, τη γεωγραφική τους κατανομή και ένα εύρος χρονολόγησης. Τα δεδομένα από το I.PHI μοιάζουν όπως παρακάτω:

A	B	C	D	E	F	G	H	I	J
id	text	metadata	reg	region_main	reg	region_sub	date_str	date_min	date_max
27869	θεοπομπου.	Korinthia — Korinthos	1690	Peloponnesos (IG IV-[VI])	1667	Saronic Gulf, Corinthia, ε			
315181	[φ]ιλεταιρος ευμενου τ	Boiotia — Thespiai — mid-3rd c	1698	Central Greece (IG VII-IX)	1691	Megaris, Oropia, and Boi	mid-3rd c. B	-275	-226
201686	μαλκοιδων ηρωνος.	Crete, W. — Tarrha — 1st-3rd c	1699	Aegean Islands, incl. Crete (I	474	Crete	1st-3rd c. AI	1	300
153178	βασιλικος.	Makedonia (Bottiaia) — Pella —	1692	Northern Greece (IG X)	1485	Macedonia	3rd/2nd c. B	-300	-101
242973	ευψυχι αλεξανδρε ουδ	Syr., Antioch. — Antiochia & Ar	1693	Greater Syria and the East	1676	Syria and Phoenicia			
28582	αισκλαπιει μ [ανεθεκε	Epidauria — Epidauros — sinist	1690	Peloponnesos (IG IV-[VI])	1643	Epidauria (IG IV^2,1)	6th/5th c. B	-600	-401
333620	[---]ος αν[εθηκε δαματ	Italia — Herakleia (Policoro) —	1696	Sicily, Italy, and the West (IG	1689	Italy, incl. Magna Graecia	late 4th/earl	-350	-251

Το απόσπασμα που σας δίνεται περιλαμβάνει 2802 επιγραφές, όπου όλες διαθέτουν ένα εύρος χρονολόγησης σε έτη (date-min: -720, date-max: 1453). Συνολικά οι επιγραφές περιλαμβάνουν 24679 μοναδικές λέξεις (tokens) που απαρτίζουν το λεξικό (dictionary) του συνόλου των κειμένων.

Το ζητούμενο στην εργασία αυτή είναι να κατασκευαστεί και να εκπαιδευτεί ένα ΤΝΔ για την πρόβλεψη της ακριβούς χρονολογίας μιας επιγραφής με βάση το κείμενό της. Προαιρετικά, μπορείτε να χρησιμοποιήσετε και άλλα χαρακτηριστικά, όπως η γεωγραφική τοποθεσία της επιγραφής.

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.*

Α1. Προεπεξεργασία και Προετοιμασία δεδομένων [*30 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.

[1] Διαθέσιμο στο: https://eclass.upatras.gr/modules/document/file.php/CEID1060/iphi2802.csv
[2] Assael, Yannis, et al. "Restoring and attributing ancient texts using deep neural networks." Nature 603.7900 (2022): 280-283.

α) *Κωδικοποίηση και προεπεξεργασία δεδομένων:* Το ΤΝΔ δέχεται ως είσοδο αριθμητικές τιμές. Για το λόγο αυτό, οι επιγραφές θα πρέπει πρώτα να αναλυθούν στα δομικά τους στοιχεία (tokens), π.χ. τις λέξεις που τα απαρτίζουν (*tokenization*) και στη συνέχεια να μετατραπούν σε διανυσματική μορφή (*vectorization*). Δείτε τις παρακάτω πηγές για παραδείγματα υλοποίησης αυτών των διαδικασιών[3][4]. Για απλότητα, προτείνεται η χρήση μοντέλου BoW (Bag of Words) με unigrams και χρήση tf-idf για την κωδικοποίηση των tokens. Από όλα τα tokens που θα προκύψουν από το dataset, συνίσταται να χρησιμοποιήσετε έναν περιορισμένο αριθμό (π.χ. 1000) με βάση τη συχνότητα χρήσης τους και τη σπουδαιότητά τους. Για παράδειγμα, μπορείτε να απορρίψετε tokens που εμφανίζονται πολύ λίγες φορές στο dataset. Να τεκμηριώσετε αναλυτικά στην αναφορά σας τη συνολική διαδικασία που θα ακολουθήσετε. [20]

β) *Κανονικοποίηση (Normalization ή min-max scaling):* Με την μέθοδο αυτή μεταφέρουμε το εύρος τιμών ενός χαρακτηριστικού σε νέα κλίμακα πχ [0,1] ή [-1, 1]. Εξετάστε τη χρήση κανονικοποίησης τόσο για τις εισόδους, <u>όσο και για την έξοδο του δικτύου</u> και εφαρμόστε τη αν κρίνετε σκόπιμο. [5]

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 [*40 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 $\eta = 0.001$.

α) Η εκπαίδευση και αξιολόγηση των μοντέλων σας μπορεί να γίνει με την *Ρίζα του Μέσου Τετραγωνικού Σφάλματος* (RMSE). Το σφάλμα μπορεί να υπολογιστεί ως η απόκλιση της τιμής που προβλέπει το δίκτυο από το κοντινότερο άκρο (date-min ή date-max) του εύρους που δίνεται. Αν η τιμή που προβλέπει το δίκτυο βρίσκεται εντός του εύρους, τότε το σφάλμα είναι μηδέν. Ο υπολογισμός του σφάλματος μπορεί να γίνει με μια custom συνάρτηση, ανάλογα με το framework υλοποίησης που θα επιλέξετε. Εναλλακτικά, μπορείτε απλώς να υπολογίσετε την απόκλιση από ένα από τα δύο άκρα ή από τη μέση τιμή του διαστήματος, αλλά το σφάλμα θα είναι συστηματικά μεγαλύτερο. [5]

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

γ) Ποια συνάρτηση ενεργοποίησης θα χρησιμοποιήσετε για το επίπεδο εξόδου; Σιγμοειδή, γραμμική, Softmax ή κάποια άλλη; Είναι προφανές ότι η συνάρτηση ενεργοποίησης στο επίπεδο εξόδου θα πρέπει να είναι σε θέση να παράξει τιμές στο διαθέσιμο εύρος χρονολογιών – ενδεχομένως με κάποιο μετασχηματισμό (βλ. και Α1-β). [5]

δ) Πειραματιστείτε με 3 διαφορετικές τιμές για τον αριθμό των νευρώνων του κρυφού επιπέδου (H_1) και συμπληρώστε τον παρακάτω πίνακα. Να συμπεριλάβετε και τις γραφικές παραστάσεις σύγκλισης (Μ.Ο.) ανά κύκλο εκπαίδευσης. Διατυπώστε τα συμπεράσματά σας σχετικά με τον αριθμό των κρυφών κόμβων και την ταχύτητα σύγκλισης ως προς τις εποχές εκπαίδευσης. [10]

Αριθμός νευρώνων στο κρυφό επίπεδο	**RMSE**
H_1 =	
H_1 =	
H_1 =	

[3] https://developers.google.com/machine-learning/guides/text-classification/step-3
[4] https://www.tensorflow.org/tutorials/keras/text_classification#prepare_the_dataset_for_training

ε) Δοκιμάστε να προσθέσετε ένα έως δύο ακόμα κρυφά επίπεδα στο δίκτυο (H_2, H_3). Πειραματιστείτε με τον αριθμό των κόμβων. Περιγράψτε μια λογική για τη στοίχιση των κρυφών επιπέδων (είναι καλό να έχουν τον ίδιο αριθμό κόμβων; Μειούμενο; Αυξανόμενο;). Να αναφέρετε RMSE και να διατυπώσετε τα συμπεράσματά σας σχετικά με την προσθήκη κρυφών επιπέδων. [10]

στ) 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 πραγματοποιήσετε 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**	**RMSE**
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). Μια τεχνική για την εφαρμογή ομαλοποίησης είναι το *dropout*. Να εφαρμόστε dropout τόσο στους κόμβους εισόδου, όσο και στους κρυφούς κόμβους με πιθανότητες διατήρησης r_{in} και r_h αντίστοιχα. Να αξιολογήσετε διάφορες τιμές για τον συντελεστή r, συμπληρώνοντας τον παρακάτω πίνακα με χρήση 5-fold CV. Να συμπεριλάβετε και τις γραφικές παραστάσεις σύγκλισης (Μ.Ο.) ανά κύκλο εκπαίδευσης. Διατυπώστε τα συμπεράσματά σας σχετικά με την επίδραση της μεθόδου στη γενικευτική ικανότητα του δικτύου.

Πιθανότητες διατήρησης	**RMSE**
$r_{in} = 0.8$ $r_h = 0.5$	
$r_{in} = 0.5$ $r_h = 0.5$	
$r_{in} = 0.8$ $r_h = 0.2$	

Α5. Γλωσσικά Μοντέλα *[προαιρετικό ερώτημα - 10 μονάδες bonus]*
Χρησιμοποιήστε ένα προεκπαιδευμένο γλωσσικό μοντέλο, όπως το BERT και τις παραλλαγές του (π.χ. Ancient Greek BERT, Logion-Base) και προσπαθήστε να το ρυθμίσετε για την εργασία της χρονολόγησης των επιγραφών. Αξιολογήστε το μοντέλο σας στο σύνολο ελέγχου και συγκρίνετε με το βέλτιστο δίκτυο που καταλήξατε στα προηγούμενα ερωτήματα. Εξηγείστε τυχόν διαφορά στην επίδοση που παρατηρείτε.

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, 22/4/2024, στις 23:59.
2. Για οποιαδήποτε διευκρίνηση / ερώτηση μπορείτε να χρησιμοποιείτε το σχετικό forum στο eclass του μαθήματος.