UNIVERSITY OF THESSALY
SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL
& COMPUTER ENGINEERING

Νευρο-Ασαφής Υπολογιστική

Χειμερινό Εξάμηνο 2018-2019

Δημήτριος Κατσαρός

Coding project

Ημέρα ανανέωσης (με red font) ανακοίνωσης: Tuesday, January 22, 2019

Προθεσμία παράδοσης: Πέμπτη, Ιανουάριος 31, 2019

Περιγραφή προβλήματος

Την τελευταία δεκαετία η διαθεσιμότητα αποθετηρίων μεγάλων επιστημονομετρικών δεδομένων (big scholarly data) όπως το Google Scholar, το Elsevier Scopus, Aminer, κ.ά., προσφέρουν τεράστιες δυνατότητες στο πεδίο της επιστημονομετρίας για την ανάλυση της δυναμικής της επιστήμης [7] και την συνακόλουθη σχεδίαση δεικτών αποτίμησης του επιστημονικού αντικτύπου επιστημόνων [5], προγραμμάτων σπουδών [9], περιοδικών [4] και πανεπιστημίων [3]. Αυτοί οι δείκτες παράγουν μια διαβάθμιση (ranking) των επιστημόνων και πανεπιστημίων, η οποία χρησιμοποιείται για να ληφθούν αποφάσεις του τύπου: α) σε ποιο πανεπιστήμιο θα εγγραφεί κάποιος φοιτητής, β) σε ποιο πανεπιστήμιο θα εργαστεί κάποιος ακαδημαϊκός, γ) εάν θα προαχθεί κάποιος ακαδημαϊκός, δ) πού θα κατανεμηθούν τα (κρατικά ή μη) κονδύλια ενίσχυσης της έρευνας κ.τ.λ. Ειδικά για τους νέους επιστήμονες που η εξέλιξή τους είναι άγνωστη την στιγμή της κρίσης, θα ήταν επιθυμητό να έχουμε δυνατότητα πρόβλεψης του μελλοντικού αντικτύπου της έρευνάς τους με βάση το παρελθόν. Για παράδειγμα, στην εργασία [2] αναπτύχθηκε μια μέθοδος πρόβλεψης βασισμένη σε regression.

Ένας βασικός στόχος σε μια τέτοια μεθοδολογία είναι η πρόβλεψη του αριθμού των μελλοντικών αναφορών (citations) που θα λάβει κάποια εργασία. Στην παρούσα προγραμματιστική άσκηση καλείστε να χρησιμοποιήσετε νευρωνικά δίκτυα για να υλοποιήσετε προβλεπτική μοντελοποίηση. Δεν υπάρχει περιορισμός στο ποιο είδους νευρωνικού δικτύου, ποιας αρχιτεκτονικής, ποιας τοπολογίας, κ.τ.λ. θα χρησιμοποιήσετε. Οι εργασίες [1] και [8] χρησιμοποίησαν Recurrent Neural Networks, ενώ η εργασία [6] χρησιμοποίησε Feed Forward Neural Networks.

Θα χρησιμοποιήσετε δεδομένα από την υπηρεσία https://aminer.org/. Συγκεκριμένα θα χρησιμοποιήσετε τα δεδομένα Citation-network V1 (https://aminer.org/citation). Προφανώς πρέπει να δημιουργήσετε την χρονοσειρά που περιγράφει τις αναφορές ανά έτος κάθε εργασίας. Το μέτρο επίδοσης είναι φυσικά η ακρίβεια πρόβλεψης.

Η αποτίμηση της επίδοσης της αρχιτεκτονικής που θ' αναπτύξετε θα γίνει πάνω σε χρονοσειρές που θα σας δωθούν από τον διδάσκοντα κατά την αξιολόγηση της εργασίας σας.

Οι δυο αρχιτεκονικές με την ακριβέστερη πρόβλεψη θα έχουν το προνόμιο ν' αριστεύσουν στο μάθημα, ανεξάρτητα των επιδόσεων των μελών της ομάδας στις άλλες συνιστώσες του μαθήματος.

Ως παραδοτέο ζητείται, να γράψετε μια αναφορά που, εκτός των όποιων γενικών πληροφοριών, θα περιέχει τα ακόλουθα:

- λεπτομέρειες της αρχιτεκτονικής του νευρωνικού δικτύου,
- τον αλγόριθμο εκπαίδευσής του,
- τις τελικές (εκπαιδευμένες) τιμές των παραμέτρων του,
- αποτίμηση της επίδοσής του πάνω στα δεδομένα που έχετε διαθέσιμα,
- καταγραφή των χρόνων εκπαίδευσης του δικτύου ως συνάρτηση του μεγέθους των δεδομένων εισόδου (παρουσιάζοντας σταδιακά τα διαθέσιμα δεδομένα),
- καταγραφή του χρόνου απόκρισης του δικτύου.

Το test set που θα δωθεί από τον διδάσκοντα για την αποτίμηση της επίδοσης της λύσης που έδωσε η κάθε ομάδα, θα αποτελείται από **20 χρονοσειρές διαφορετικού μήκους** που η κάθε μια θα περιγράφει το αριθμό των αναφορών ενός αριθμό ανά έτος και θα ζητείται να προβλεφθεί η επίδοση του άρθρου σε αριθμό citations: α) στο επόμενο έτος, και β) μετά 5 έτη.

Για παράδειγμα μια χρονοσειρά θα μπορούσε να προκύψει από αυτή που αφορά στον αριθμό citations κάποιου πραγματικού άρθρου όπως εικονίζεται στον παρακάτω πίνακα. Έτσι, για παράγειγμα θα δινόταν η χρονοσειρά 3,7,17,8,10,9,7,8,11 και θα ζητούνταν το επόμενο σημείο της (δηλαδή στο έτος 2013) καθώς και το σημείο στο έτος 2017.

Έτος	Αριθμός citations
2004	3
2005	7
2006	17
2007	8
2008	10
2009	9
2010	7
2011	8
2012	11
2013	8
2014	5
2015	5
2016	2
2017	9
2018	1

Συνεπώς πρέπει το πρόγραμμά να μπορεί να διαβάσει ακολουθίες ακεραίων από txt αρχείο που διαχωρίζονται με κόμμα το οποίο θα αποτελεί την test είσοδο στο νευρωνικό δίκτυο που αναπτύξατε, και να τυπώνει τον προβλεπόμενο ακέραιο. Κάθε αρχείο θα περιέχει μια μόνο χρονοσειρά και το όνομά του θα είναι timeseries01.txt, timeseries02.txt, timeseries03.txt, …, timeseries19.txt, timeseries20.txt, αντίστοιχα.

Επίσης θα υπάρχουν αντιστοίχως ονοματισμένα αρχεία timeseries01-years.txt και timeseries01-contents.txt που το πρώτο θα περιέχει τα έτη στα οποία κτήθηκαν οι citations που καταγράφονται στο timeseries01.txt, και το δεύτερο θα περιέχει τον τίτλο και το abstract για το αντίστοιχο paper. Παράδειγμα input θ' αναρτηθεί στο site.

Βιβλιογραφία

[1] A. Abrishami and S. Aliakbary, "NNCP: A citation count prediction methodology based on deep neural network learning techniques", CoRR abs/1809.04365, September, 13, 2018.

[2] D. E. Acuna, S. Allesina, and K. P. Kording, "Future impact: Predicting scientific success", **Nature**, vol. 489, no. 7415, pp. 201–202, 2012.

[3] I.F. Aguillo, J. Bar-Ilan, M. Levene, and J.L. Ortega, "Comparing university rankings", **Scientometrics**, vol. 85, no. 1, pp. 243–256, 2010.
[4] B. De Sutter and A. van Den Oord, "To be or not to be cited in computer science", **Communications of the ACM**, vol. 55, no. 8, pp. 69-75, 2012.
[5] J.E. Hirsch, "An index to quantify an individual's scientific research output", **Proceedings of the National Academy of Sciences of the United States of America**, vol. 102, no. 46 pp. 16569–16572, 2005.
[6] T. Mistele, T. Price, and S. Hossenfelder , "Predicting citation counts with a neural network", CoRR abs/1806.04641, June, 18, 2018.
[7] V. Singh, A. Perdigones, J.L. Garcia, I.Canas-Guerroro, and F.R. Mazarron, "Analyzing worldwide research in hardware architecture, 1997-2011", **Communications of the ACM**, vol. 58, no. 1, pp. 76-85, 2015.
[8] S. Yuan, J. Tang, Y. Zhang, Y. Wang, T. Xiao, "Modeling and Predicting Citation Count via Recurrent Neural Network with Long Short-Term Memory", CoRR abs/ 1811.02129, November, 6, 2018.
[9] S. Vucetic, A.K. Chanda, S. Zhang, T. Bai, and A. Maiti, "Peer assessment of CS doctoral programs shows strong correlation with faculty citations", **Communications of the ACM**, vol. 61, no. 9, pp. 70-76, 2018.

Χρηστικές πληροφορίες:

Η προθεσμία παράδοσης είναι αυστηρή. Είναι όμως δυνατή η παροχή παράτασης (μέχρι 7 ημέρες), αλλά μόνο αφού δώσει ο διδάσκων την έγκρισή του, και αυτή η παράταση στοιχίζει 10% ποινή στον τελικό βαθμό της. Η παράδοση γίνεται με email (dkatsar@e-ce.uth.gr) του πηγαίου κώδικα, καθώς της αναφοράς που περιέχει την (σύντομη) περιγραφή του κώδικα, και των αποτελεσμάτων της πειραματικής αξιολόγησης. Το subject του μηνύματος πρέπει να είναι: CE418-Project: AEMx-AEMy

Ερμηνεία συμβόλου:

Απαιτεί την ανάπτυξη κώδικα σε Tensorflow (Keras, …). Εάν χρησιμοποιήσετε έτοιμο κώδικα από κάποια πηγή απαιτείται να δηλώσετε την πηγή, καθώς και σε ποιο σημείο του project τον χρησιμοποιήσατε.