Θέματα Διπλωματικών Εργασιών Ακαδημαϊκού Έτους 2022-2023

Διαδικασία Ανάθεσης

- Θα είμαι διαθέσιμη για διευκρινίσεις/ερωτήσεις σχετικά με τα θέματα μέσω mteams από τη Δευτέρα 22 Αυγούστου και μετά.
- Αν σας ενδιαφέρει κάποιο θέμα, μπορείτε να μου στείλετε εμαιλ με το θέμα (θέματα) που σας ενδιαφέρει, την αναλυτική βαθμολογία σας και όποια άλλη πληροφορία θεωρείστε χρήσιμη.
- Η ανάθεση των θεμάτων θα γίνει τέλη Σεπτεμβρίου/αρχές Οκτωβρίου.

Περιγραφή Θεμάτων

Θέμα 1: Μελέτη διαφόρων μορφών πόλωσης στο δίκτυο Reddit

Σε πολλά κοινωνικά δίκτυα δημιουργείται πόλωση όπου κυριαρχούν δυο ακραίες απόψεις για κάποιο θέμα (π.χ., υπέρ και κατά του εμβολιασμού). Σε αυτήν την εργασία θα μελετήσουμε διαφορετικούς τύπους δομικής πόλωσης στο δίκτυο Reddit, συγκεκριμένα (α) με πρόσημο (θετικό, αρνητικό) και μη, και (β) στην ίδια ή σε διαφορετικές κοινότητες (subreddits).

Έχουμε δομική πόλωση χωρίς πρόσημο όταν δημιουργούνται δύο ομάδες, η επικοινωνία ανάμεσα στα μέλη της ίδιας ομάδας είναι συχνή, αλλά τα μέλη διαφορετικών ομάδων επικοινωνούν σπάνια. Δομική πόλωση με πρόσημο έχουμε όταν πάλι δημιουργούνται δυο ομάδες, τα μέλη μιας ομάδας συμφωνούν με τα μέλη της ίδιας ομάδας (θετικό πρόσημο) και διαφωνούν με τα μέλη της άλλης ομάδας (αρνητικό πρόσημο).

Το Reddit είναι χωρισμένο σε κοινότητες (subreddits) ανάλογα με τη θεματολογία των συζητήσεων (π.χ., Turkey, Greece). Θα εξετάσουμε τη δημιουργία πόλωσης για κάποιο θέμα μέσα σε μία κοινότητα και ανάμεσα σε δύο ή περισσότερες κοινότητες.

Η εργασία θα βασιστεί (επαναλάβει) προηγούμενη μελέτη μας στο θέμα (εργασία [1]). Για τη μέτρηση της πόλωσης, θα χρησιμοποιηθούν βιβλιοθήκες που υλοποιούν σχετικές μετρικές.

Αναφορές

[1] Chrisoula Terizi, Evaggelia Pitoura, Polarized Groups in Discussion Forums: The Case of Reddit, Unpublished, May 2021

Θέμα 2: Μελέτη μορφών προκατάληψης σε διανυσματικές αναπαραστάσεις ελληνικού κειμένου (Biases in word embeddings for Greek)

Τα word embeddings αφορούν στην αναπαράσταση λέξεων ως διανύσματα τα οποία μαθαίνουμε χρησιμοποιώντας κείμενα από διάφορες πηγές. Πολλές μελέτες έχουν δείξει ότι αυτές οι

αναπαραστάσεις συχνά κωδικοποιούν στερεότυπα όπως η διάσημη αναλογία man-computer programmer και woman-homemaker [2].

Θα μελετήσουμε πιθανές προκαταλήψεις σε word embeddings για Ελληνικά. Οι προκαταλήψεις μπορούν να αφορούν στερεότυπα σχετικά με το φύλο (για παράδειγμα, γυναικεία και αντρικά επαγγέλματα), εθνικιστικά στερεότυπα (για παράδειγμα, επίθετα που συσχετίζονται με συγκεκριμένες εθνικότητες) ή και άλλα θέματα.

Θα μελετηθούν πειραματικά διάφορα είδη embedding εκπαιδευμένα με διαφορετικές συλλογές. Επίσης, θα μελετηθεί η απόδοση γνωστών τεχνικών για αφαίρεση προκατάληψης (debiasing).

[2] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, Adam Tauman Kalai: Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. NIPS 2016: 4349-4357

Θέμα 3: Εξήγηση της έλλειψης δικαιοσύνης (Counterfactual explanations of algorithmic unfairness)

Η αλγοριθμική δικαιοσύνη προσπαθεί να εξασφαλίσει ότι τα αποτελέσματα των αλγορίθμων δεν αδικούν συγκεκριμένες ομάδες. Θα επικεντρωθούμε σε αλγορίθμους ταξινόμησης (classification). Για ευκολία, ας υποθέσουμε δυαδική ταξινόμηση και δύο κλάσεις την κλάση 0 και την κλάση 1. Ας πούμε ότι η κλάση 1 είναι η ευνοϊκή κλάση (για παράδειγμα, η κλάση που αντιστοιχεί στο ότι κάποιος παίρνει το δάνειο ή τη δουλειά). Υπάρχουν πολλοί ορισμοί της δικαιοσύνης, σε αυτήν την εργασία θα επικεντρωθούμε σε ορισμούς που βασίζονται στο λάθος ενός αλγορίθμου για κάθε ομάδα, για παράδειγμα, ποιο ποσοστό γυναικών ταξινομεί λάθος στην ομάδα 0 συγκριτικά με το αντίστοιχο ποσοστό αντρών.

Έστω ότι η είσοδος x ταξινομείτε άδικα στην κλάση 0. Με απλά λόγια, μια μη-πραγματική εξήγηση (counterfactual) για το x είναι ένα x' που έχει τη μικρότερη απόσταση από το x και αν το δώσουμε στο classifier το x' ως είσοδο αυτό θα ταξινομηθεί στην κλάση 1. Δείτε το [3] για περισσότερα σχετικά με counterfactuals. Θα χρησιμοποιήσουμε αυτή την ιδέα για να εξηγήσουμε λάθη σε συγκεκριμένες ομάδες.

Στην εργασία αυτή θα εφαρμόσουμε υλοποιήσεις τεχνικών που παράγουν countefactuals για να εξηγήσουμε την δικαιοσύνη ταξινομητών σε διάφορα datasets.

[3] Sahil Verma, John P. Dickerson, Keegan Hines: Counterfactual Explanations for Machine Learning: A Review. CoRR abs/2010.10596 (2020)

Θέμα 4: Μελέτη της δικαιοσύνης αλγορίθμων δειγματοληψίας σε γραφήματα

Οι αλγόριθμοι δειγματοληψίας επιλέγουν ένα «υποσύνολο» ενός γράφου (σύνολο ακμών, κόμβων). Αποτελούν ένα ιδιαίτερα σημαντικό κομμάτι για πολλούς γραφο-αλγόριθμους παραδοσιακούς αλλά και μηχανικής μάθησης. Δείτε το [4] για κάποιους βασικούς τέτοιους αλγόριθμους.

Σε αυτή την εργασία θα δούμε τους πιο γνωστούς αλγορίθμους δειγματοληψίας και θα τους αξιολογήσουμε όσον αφορά το πόσο δίκαιοι είναι οι γράφοι που αυτοί παράγουν.

Θα δοκιμάσουμε τους αλγορίθμους δειγματοληψίας σε διάφορα σύνολα πραγματικών και συνθετικών γράφων. Η αξιολόγηση θα γίνει με βάση διαφορετικούς ορισμούς δικαιοσύνης για γράφους. Ένα απλό παράδειγμα για γράφους που οι κόμβοι τους έχουν γνωρίσματα (πχ ηλικία, γένος), πόσους κόμβους με συγκεκριμένες τιμές στα γνωρίσματα (πχ, νέους, γυναίκες) έχουμε στο δείγμα;

[4] Jure Leskovec, Christos Faloutsos: Sampling from large graphs. KDD 2006: 631-636