ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΕΡΓΑΣΙΑ ΣΤΟ ΜΑΘΗΜΑ «ΓΛΩΣΣΙΚΗ ΤΕΧΝΟΛΟΓΙΑ»
ΑΚΑΔ. ΕΤΟΣ 2023-2024

ΘΕΜΑ: ΑΥΤΟΜΑΤΗ ΑΝΙΧΝΕΥΣΗ ΑΡΝΗΣΗΣ (NEGATION DETECTION)

Στόχος της παρούσας εργασίας είναι η χρήση τεχνικών μηχανικής μάθησης για την αυτόματη ανίχνευση της ύπαρξης άρνησης σε κείμενα της ελληνικής γλώσσας. Η αναγνώριση της άρνησης αποτελεί ένα πολύ σημαντικό βήμα για την αναγνώριση του συναισθήματος ή της γνώμης που κρύβει ένα κείμενο.

Η αναγνώριση άρνησης συνήθως αντιμετωπίζεται σε δυο στάδια:
1. Negation signal identification: Αναγνώριση λέξεων/φράσεων που είναι αρνητικές (πχ δεν, μην, ολότελα, καθόλου, ολοσδιόλου, εκμηδενίζω, αποτυχία)
2. Negation scope Identification: Αναγνώριση του πεδίου δράσης της άρνησης.

Παράδειγμα: [**Δεν** θα κάνω] σχόλιο.
(Με κόκκινο φαίνεται η αρνητική λέξη, με τις αγκύλες δηλώνεται το πεδίο δράσης της.)

Α. ΕΠΙΣΚΟΠΗΣΗ ΤΗΣ ΒΙΒΛΙΟΓΡΑΦΙΑΣ

Στο opencourses παρατίθενται κάποιες εργασίες που έχουν εφαρμόσει τεχνικές μηχανικής μάθησης στην αυτόματη αναγνώριση της πολυπλοκότητας ενός κειμένου.

ΒΗΜΑ 1 (ΜΕΧΡΙ 3/11): Αναζητήστε περισσότερη βιβλιογραφία σχετικά με την εφαρμογή αυτή στο scholar.google.com
Χρησιμοποιείστε για την αναζήτησή σας λέξεις-κλειδιά όπως “negation detection”, “machine learning”.

Β. ΕΠΙΛΟΓΗ ΔΕΔΟΜΕΝΩΝ

ΒΗΜΑ 2 (ΜΕΧΡΙ 6/11): Διερευνήστε σώματα κειμένων που θα μπορούσαν να χρησιμοποιηθούν, και αποφασίστε ποιο θα χρησιμοποιήσετε.

- Το ILSP/ELEFTHEROTYPIA corpus
- Κείμενα από την ελληνική Wikipedia
- Κείμενο από τον ελληνικό κοινωνικό ιστό
- Άλλο

Γ. ΕΠΙΛΟΓΗ ΚΑΙ ΥΠΟΛΟΓΙΣΜΟΣ ΧΑΡ/ΚΩΝ ΜΑΘΗΣΗΣ

ΒΗΜΑ 3 (ΜΕΧΡΙ 6/11): Βάσει της βιβλιογραφίας, μετασχηματίστε το κείμενο σε διανύσματα μάθησης (χαρακτηριστικών-τιμών). Αυτό θα αποτελεί και την βασική ερευνητική δραστηριότητα της εργασίας σας. Μερικά από τα ερωτήματα που πρέπει να απαντήσετε είναι

- Κάθε διάνυσμα μάθησης τι θα αναπαριστά (μια λέξη; μια πρόταση;)
- Τι χαρακτηριστικά μάθησης θα αναπαριστούν το κάθε διάνυσμα;
- Ποιες θα είναι οι ετικέτες της εξόδου;

Στόχος είναι να δημιουργηθεί ένα φύλλο στο excel σαν αυτό (στην περίπτωση που επιλέξετε το σχήμα επισημείωσης IOB):

~~Λέξη εστίασης~~	2η λέξη πριν	1η λέξη πριν	1η λέξη μετά	2η λέξη μετά	Επισημείωση:
Δεν	-	-	θα	κάνω	**B-NG**
θα	-	Δεν	κάνω	σχόλιο	**I-NG**
κάνω	Δεν	θα	σχόλιο	-	**I-NG**
σχόλιο	θα	κάνω	-	-	**O**

Ο αριθμός των παραδειγμάτων σας πρέπει να είναι μεγάλος τετραψήφιος/πενταψήφιος.

ΒΗΜΑ 4 (ΜΕΧΡΙ 4/12): Επισημείωση. Πρέπει χειρωνακτικά να επισημειώσετε με την τιμή της εξόδου τα παραδείγματά σας. Αφού συνεννοηθείτε πολύ καλά μεταξύ σας για τις οδηγίες (καταγράψτε τις οδηγίες ξεκάθαρα σε ένα οδηγό/tutorial) της επισημείωσης, ώστε όλοι να επισημειώνετε με τον ίδιο τρόπο και την ίδια λογική, μοιραστείτε τα παραδείγματα και επισημειώστε τα.
Δείξτε ο ένας τις επισημειώσεις του στον άλλο για να υπάρχει διασταυρωτικός έλεγχος.
Σε περιπτώσεις αμφιβολιών κουβεντιάστε τις μεταξύ σας, προκειμένου να επιλυθούν.

Δ. ΕΦΑΡΜΟΓΗ ΤΕΧΝΙΚΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

ΒΗΜΑ 5: Αποθηκεύστε το παραπάνω αρχείο σαν .csv, και φορτώστε το σε κάποιον πάγκο εργασίας μηχανικής μάθησης, όπως

- το Weka (https://www.cs.waikato.ac.nz/ml/weka/)
- το Rapidminer (https://rapidminer.com)

ΒΗΜΑ 6: Επιλέξτε την τεχνική επικύρωσης που επιθυμείτε (train/test split, cross validation), διαμορφώστε τα αρχεία δεδομένων σας ανάλογα, και τρέξτε πάνω στα δεδομένα τουλάχιστον τρεις αλγορίθμους μηχανικής μάθησης. Οι αλγόριθμοι είναι δικής σας επιλογής και βασισμένοι στην βιβλιογραφία. Πειραματιστείτε με διαφορετικά (υπο)σετ των χαρακτηριστικών σας.

ΒΗΜΑ 7 (ΜΕΧΡΙ 15/1/24): Κάντε

- καταγραφή των αποτελεσμάτων (precision, recall και μέτρο f για την κάθε τιμή της εξόδου) κάθε πειράματος που τρέχετε
- σύγκριση των πειραματικών αποτελεσμάτων (πχ ποιές τεχνικές μάθησης τα πάνε καλύτερα και γιατί, ποια σετ χαρακτηριστικών τα πάνε καλύτερα και γιατί κλπ)
- ποιοτική ανάλυση των αποτελεσμάτων σας (πχ ποια παραδείγματα ταξινομούνται λάθος και που μπορεί να οφείλεται αυτό)

- σύγκριση των αποτελεσμάτων σας με την βιβλιογραφία (πώς είναι τα αποτελέσματά σας σε σχέση με παρόμοιες δουλειές της βιβλιογραφίας; πού τα πάτε καλύτερα, πού τα πάτε χειρότερα, γιατί;)
- τι ιδιαιτερότητες έχει η ελληνική γλώσσα που επηρεάζουν την απόδοση των πειραμάτων σας;

Ε. ΣΥΝΤΑΞΗ ΕΡΓΑΣΙΑΣ

ΒΗΜΑ 8 (ΜΕΧΡΙ ΤΗΝ ΗΜΕΡΟΜΗΝΙΑ ΤΩΝ ΕΞΕΤΑΣΕΩΝ ΤΟΥ ΜΑΘΗΜΑΤΟΣ): Συντάξτε επιστημονική εργασία στα Αγγλικά (σαν τις εργασίες της βιβλιογραφίας) μέχρι 10 σελίδες το πολύ, στην οποία θα παρουσιάζετε

- το θεωρητικό υπόβαθρο της δουλειάς σας
- το σετ δεδομένων σας
- την επιλογή των χαρακτηριστικών σας
- την πειραματική σας διαδικασία και τα αποτελέσματά σας (precision και recall για κάθε τιμή της κλάσης ταξινόμησης)
- την αξιολόγηση των αποτελεσμάτων σας
- την σύγκριση των αποτελεσμάτων σας με αυτά άλλων προσεγγίσεων στη βιβλιογραφία
- τα συμπεράσματα στα οποία καταλήξατε και μελλοντικές βελτιώσεις που προτείνετε για την προσέγγισή σας
- την διαφορετικότητα στην προσέγγισή σας (τι διαφορές/καινοτομίες/συνεισφορά έχει η δική σας δουλειά σε σχέση με τις παραπλήσιες της βιβλιογραφίας)
- την λίστα των βιβλιογραφικών σας παραπομπών

ΑΛΛΑ

Η εργασία θα πραγματοποιηθεί σε ομάδες των τριών τουλάχιστον ατόμων.
Η εργασία είναι υποχρεωτική και πιάνει το 50% στον τελικό βαθμό.
Η ομάδα που θα πετύχει το μεγαλύτερο σκορ στα αποτελέσματα θα έχει

- επιπλέον βαθμολογικό bonus
- ευκαιρία να προωθήσει την εργασία της για δημοσίευση σε επιστημονικό συνέδριο