ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ
ΠΡΟΗΓΜΕΝΑ ΘΕΜΑΤΑ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ
Ακ. Έτος 2022-2023

Παράδοση εργασίας : Ημερομηνία εξέτασης μαθήματος

Επιλέξτε ένα από τα παρακάτω θέματα εργασιών

Θέμα 1. Δημιουργήστε ένα σύστημα για να εντοπίσετε αναξιόπιστα άρθρα ειδήσεων

Ο στόχος της εργασίας είναι η ανάπτυξη ενός συστήματος το οποίο θα αναγνωρίζει ψευδής ειδήσεις (fake news). Στα πλαίσια της εργασίας αυτή απαιτείται η εκπαίδευση ενός μοντέλου κατηγοριοποίησης ειδήσεων σε ψευδείς ή αληθείς. Μετά την εκπαίδευση, θα πρέπει να αξιολογήσετε την αποδοτικότητα του μοντέλου σας σε σχέση με ένα σύνολο από μετρικές όπως *ακρίβεια κατηγοριοποίησης, χρόνος εκπαίδευσης και κατηγοριοποίησης*. Στη συνέχεια θα υλοποιήσετε μία εφαρμογή η οποία θα δέχεται σαν είσοδο ένα κείμενο είδησης και θα επιστρέφει την κατηγορία στην οποία ανήκει (ψευδής, αληθής).

Α) Δεδομένα

Μπορείτε να χρησιμοποιήσετε το παρακάτω σύνολο δεδομένων
https://www.kaggle.com/c/fake-news/data

ή οποιοδήποτε άλλο dataset για fake-news θέλετε

Β) Εκπαίδευση μοντέλου

Επιλέξετε την εκπαίδευση μοντέλου κατηγοριοποίησης των κειμένων κριτικής χρησιμοποιώντας δύο από τις τεχνικές κατηγοριοποίησης που συζητήσαμε (Logistic regression, SVM, νευρωνικό δίκτυο) και παρουσιάστε συγκριτικά αποτελέσματα.

Η είσοδος στο αλγόριθμο εκπαίδευσης θα είναι η αναπαράσταση των κειμένων σας σε μορφή διανύσματος. Θα πρέπει επομένως να χρησιμοποιήσετε ένα μοντέλο αναπαράστασης κειμένου όπως, TFIDF, word embedding (https://www.tensorflow.org/tutorials/representation/word2vec).

Γ) Αξιολόγηση

Θα αξιολογήσουμε την απόδοση του εκπαιδευμένου μοντέλου σε σχέση με :

- *Ακρίβεια ταξινόμησης* (Αξιολογείται στο σύνολο δεδομένων ελέγχου). Την ακρίβεια την μετράμε ως το ποσοστό των σωστά ταξινομημένων κριτικών. Πάρτε τυχαία

(π.χ.1000) κείμενα από το σύνολο δεδομένων ελέγχου τα δίνεται ως είσοδο στο μοντέλο και καταγράφετε το ποσοστό των κειμένων που ταξινομήθηκαν σωστά.

- *Χρόνος που χρειάζεται για την εκπαίδευση του μοντέλου*. Αυτή είναι η ώρα από τη στιγμή που ξεκινάμε την εκπαίδευση μέχρι να τελειώσει η διαδικασία εκπαίδευσης.
- *Χρόνος που χρειάζεται για να εκτελέσετε το μοντέλο κατηγοριοποίησης* και να λάβετε αποτελέσματα ταξινόμησης σε νέα κείμενα (από το σύνολο ελέγχου δεδομένων). Αυτός είναι ο χρόνος που απαιτείται για να ταξινομηθεί ένα κείμενο, αφού τροφοδοτήσουμε το κείμενο ως είσοδο στο μοντέλο. Ο χρόνος εκτέλεσης μπορεί να εκτιμηθεί λαμβάνοντας το μέσο όρο σε 1000 τυχαία κείμενα από το σύνολο δεδομένων.

Δ) Ανάπτυξη εφαρμογής

- Καλείστε να αναπτύξετε μία εφαρμογή όπου θα ενσωματώσετε το εκπαιδευμένο μοντέλο προκειμένου να κατηγοριοποιεί νέες κριτικές. Η εφαρμογή θα δίνει τη δυνατότητα να εισάγει κάποιος κείμενα ειδήσεων και το σύστημα θα αναγνωρίζει και θα καταχωρεί στο σύστημα ένα είναι ψευδές ή όχι.

Θέμα 2. Σύστημα συστάσεων

Στα πλαίσια αυτής της εργασίας ζητείται να υλοποιήσετε ένα σύστημα συστάσεων για πεδίο εφαρμογής της επιλογής σας.
Για την υλοποίηση του συστήματος σας μπορείτε να επιλέξετε τεχνικές memory-based, matrix factorization ή neural networks.

Α) Δεδομένα

Μπορείτε να επιλέξετε δεδομένα από τους παρακάτω συνδέσμους
https://cseweb.ucsd.edu/~jmcauley/datasets.html
https://www.kdnuggets.com/2016/02/nine-datasets-investigating-recommender-systems.html

Β) Αξιολόγηση

Επιλέξτε μέρος των δεδομένων σαν σύνολο ελέγχου και αξιολογήστε την απόδοση του μοντέλου σας με βάση το μέσο τετραγωνικό σφάλμα (RMSE).

Γ) Ανάπτυξη εφαρμογής

Θα υλοποιήσετε μία εφαρμογή (mobile or web-based) μέσω της οποία οι χρήστες θα μπορούν να λαμβάνουν προτάσεις για αντικείμενα/υπηρεσίες .

Θέμα 3. Αναγνώριση συναισθήματος από κείμενο

Ο στόχος της εργασίας είναι η ανάπτυξη ενός συστήματος το οποίο θα αναγνωρίζει το συναίσθημα του χρήστη με βάση την κριτική που δίνεται για μία ταινία. Στα πλαίσια της εργασίας αυτή απαιτείται η εκπαίδευση ενός μοντέλου ανάλυσης συναισθήματος (sentiment analysis) σε κριτικές που δίνονται σε ταινίες. Μετά την εκπαίδευση, θα πρέπει να αξιολογήσετε την αποδοτικότητα του μοντέλου σας σε σχέση με ένα σύνολο από μετρικές όπως *ακρίβεια κατηγοριοποίησης, χρόνος εκπαίδευσης και κατηγοριοποίησης*. Στη συνέχεια

θα υλοποιήσετε μία εφαρμογή η οποία θα δέχεται σαν είσοδο ένα κείμενο κριτικής και θα επιστρέφει την κατηγορία στην οποία ανήκει (θετική ή αρνητική).

Α) Δεδομένα

Χρησιμοποιήστε το σύνολο δεδομένων IMDB review dataset. Μπορείτε να το κατεβάσετε από:
https://www.kaggle.com/utathya/imdb-review-dataset
επίσης μπορείτε να δείτε πληροφορίες και να κατεβάσετε δεδομένα από
http://ai.stanford.edu/~amaas/data/sentiment/

Το σύνολο δεδομένων περιέχει κριτικές ταινιών και την αντίστοιχη κατηγοριοποίηση τους σε θετικές/αρνητικές.

Β) Εκπαίδευση μοντέλου

Η εκπαίδευση του μοντέλου κατηγοριοποίησης των κειμένων κριτικής θα γίνει με βάση SVM και νευρωνικό δίκτυο. Παρουσιάστε συγκριτικά αποτελέσματα.

Η είσοδος στο αλγόριθμο εκπαίδευσης θα είναι η αναπαράσταση των κειμένων σας σε μορφή διανύσματος. Θα πρέπει επομένως να χρησιμοποιήσετε ένα μοντέλο αναπαράστασης κειμένου όπως, TFIDF, word embedding (https://www.tensorflow.org/tutorials/representation/word2vec).

Γ) Αξιολόγηση

Θα αξιολογήσουμε την απόδοση του εκπαιδευμένου μοντέλου σε σχέση με :

- *Ακρίβεια ταξινόμησης* (Αξιολογείται στο σύνολο δεδομένων ελέγχου). Την ακρίβεια την μετράμε ως το ποσοστό των σωστά ταξινομημένων κριτικών. Πάρτε τυχαία 1000 κείμενα από το σύνολο δεδομένων ελέγχου τα δίνεται ως είσοδο στο SVM ή NN μοντέλο και καταγράφετε το ποσοστό των κειμένων που ταξινομήθηκαν σωστά.
- *Χρόνος που χρειάζεται για την εκπαίδευση του μοντέλου*. Αυτή είναι η ώρα από τη στιγμή που ξεκινάμε την εκπαίδευση μέχρι να τελειώσει η διαδικασία εκπαίδευσης.
- *Χρόνος που χρειάζεται για να εκτελέσετε το NN*/SVM και να λάβετε αποτελέσματα ταξινόμησης σε νέα κείμενα (από το σύνολο ελέγχου δεδομένων). Αυτός είναι ο χρόνος που απαιτείται για να ταξινομηθεί ένα κείμενο, αφού τροφοδοτήσουμε το κείμενο ως είσοδο στο μοντέλο. Ο χρόνος εκτέλεσης μπορεί να εκτιμηθεί λαμβάνοντας το μέσο όρο σε 1000 τυχαία κείμενα από το σύνολο δεδομένων.

Δ) Ανάπτυξη εφαρμογής

- Καλείστε να αναπτύξετε μία εφαρμογή όπου θα ενσωματώσετε το εκπαιδευμένο NN/SVM μοντέλο προκειμένου να κατηγοριοποιεί νέες κριτικές. Η εφαρμογή θα δίνει τη δυνατότητα σε κάποιον να γράφει την κριτική του και το σύστημα θα αναγνωρίζει και θα καταχωρεί στο σύστημα το συναίσθημα του χρήστη (θετικό, αρνητικό).

Θέμα 4.

Καλείστε να ορίσετε και να επιλύσετε ένα πρόβλημα ανάλυσης δεδομένων. Η εργασία θα πρέπει να αξιοποιεί τεχνικές που έχουμε συζητήσει στα μαθήματα (SVMs, Neural netowrks, text mining, graph analysis, recommender systems).

Η αναλυτική περιγραφή του προβλήματος θα πρέπει να σταλεί μέχρι τις 17/12/2022.

Παρατηρήσεις

1. Η εργασία μπορεί να γίνει σε ομάδες μέχρι 2 άτομα.
2. Δεν υπάρχει περιορισμός σε γλώσσα υλοποίησης.
3. **Η εργασία θα υποβληθεί μέσω e-class Αρίσταρχος.**
   Θα πρέπει να παραδώσετε ένα αρχείο AM1-AM2.zip (ΑΜ είναι ο αριθμός μητρώου σας) το οποίο θα περιλαμβάνει:
   - τον πηγαίο κώδικα και
   - το κείμενο της εργασίας σε μορφή pdf. Παρουσίαση όλων των βημάτων της εργασίας και των αποτελεσμάτων
   - Μία παρουσία (powerpoint or pdf) 15 λεπτών στην οποία θα παρουσιάζεται τα βασικά στοιχεία της εργασίας σας.