Πανεπιστήμιο Δυτικής Αττικής
Τμήμα Μηχανικών Πληροφορικής και Υπολογιστών

Μάθημα: ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ ΚΑΙ ΣΗΜΑΣΙΟΛΟΓΙΚΟΣ ΙΣΤΟΣ

Ακαδημαϊκό έτος 2024-2025

Διδάσκοντες: ΠΑΝΑΓΙΩΤΑ ΤΣΕΛΕΝΤΗ

Χρήστος Μακρυωνίτης
Ice21390298

Άσκηση #1

Ημερομηνία παράδοσης: 9/6/2025

Εισαγωγή

Στην παρούσα εργασία υλοποιήθηκε ένα σύστημα **ταξινόμησης ηλεκτρονικών μηνυμάτων** (emails) με στόχο τον διαχωρισμό τους σε δύο βασικές κατηγορίες: **spam (ανεπιθύμητα)** και **ham (κανονικά)**. Για τον σκοπό αυτό χρησιμοποιήθηκε ένα το αρχείο mail_data.csv και εφαρμόστηκαν διαφορετικές τεχνικές **επεξεργασίας φυσικής γλώσσας (Natural Language Processing)** και **μηχανικής μάθησης (Machine Learning)**.

Αναλυτικά, εφαρμόστηκαν και συγκρίθηκαν τρεις διαφορετικοί αλγόριθμοι ταξινόμησης:

1. **Naive Bayes**
2. **Word2Vec σε συνδυασμό με Logistic Regression**
3. **Recurrent Neural Network (RNN)** με χρήση του TensorFlow/Keras

Η προεπεξεργασία των δεδομένων περιλάμβανε τον καθαρισμό του κειμένου (αφαίρεση URLs, emojis, ειδικών χαρακτήρων), μετατροπή σε πεζά, tokenization και μετατροπή σε αριθμητικά διανύσματα (embeddings ή sequences), ώστε να μπορούν να κατανοηθούν από τα μοντέλα.

Κάθε μοντέλο εκπαιδεύτηκε και αξιολογήθηκε με βάση τέσσερις βασικές μετρικές: **accuracy, precision, recall ,F1-score και Confusion_matrix**. Τα αποτελέσματα αναλύθηκαν συγκριτικά, και κατασκευάστηκαν διαγράμματα για την οπτική απεικόνιση των διαφορών αποδόσεων μεταξύ των μοντέλων.

Στόχος της εργασίας ήταν να αναδειχθεί ποιο μοντέλο είναι πιο κατάλληλο για το πρόβλημα του spam detection, τόσο ως προς την ακρίβεια όσο και ως προς τη συνολική αξιοπιστία.

Ανάλυση και Σχολιασμός Αποτελεσμάτων

Στην παρούσα εργασία πραγματοποιήθηκε αναλυτικός σχολιασμός των αποτελεσμάτων που προέκυψαν από την εκπαίδευση και αξιολόγηση των τριών μοντέλων ταξινόμησης. Κάθε μοντέλο (Naive Bayes, Word2Vec + Logistic Regression και RNN) εξετάστηκε ξεχωριστά, και αξιολογήθηκε βάσει των μετρικών **accuracy**, **precision**, **recall F1-score και Confusion_matrix**.

Όλα τα σχόλια του κώδικά τα έχουμε στο αρχείο του .ipynb notebook.

Συμπεράσματα Αξιολόγησης Μοντέλων Ταξινόμησης

Στην παρούσα εργασία, υλοποιήθηκαν και συγκρίθηκαν τρία διαφορετικά μοντέλα μηχανικής μάθησης για την ταξινόμηση ηλεκτρονικών μηνυμάτων σε spam και ham: Naive Bayes, Word2Vec με Logistic Regression, και RNN (Recurrent Neural Network). Η αξιολόγηση των μοντέλων έγινε με βάση τέσσερις βασικές μετρικές απόδοσης: accuracy, precision, recall και F1-score.

Όσον αφορά την accuracy, όλα τα μοντέλα παρουσίασαν πολύ υψηλά αποτελέσματα, με το Naive Bayes να προηγείται οριακά (0.979), ακολουθούμενο από το RNN (0.973) και το Word2Vec (0.969). Αυτό δείχνει ότι γενικά όλα τα μοντέλα καταφέρνουν να ταξινομούν σωστά τα περισσότερα emails.

Ωστόσο, όταν εξετάζουμε το precision (δηλαδή πόσο "σωστά" είναι τα spam που προβλέπει ένα μοντέλο), το RNN ξεχωρίζει με εντυπωσιακή τιμή 0.975. Αυτό σημαίνει ότι το RNN κάνει πολύ λίγα false positives — δεν χαρακτηρίζει κατά λάθος κανονικά μηνύματα ως spam. Το Word2Vec ακολουθεί με precision 0.929, ενώ το Naive Bayes φτάνει το 0.920.

Αντίθετα, στην ικανότητα να εντοπίζονται όλα τα πραγματικά spam emails, γνωστή ως recall, το Naive Bayes ξεχωρίζει με 0.925, καταγράφοντας τη μεγαλύτερη ευαισθησία. Τα άλλα δύο μοντέλα δεν τα πηγαίνουν τόσο καλά σε recall, με το Word2Vec να φτάνει το 0.834 και το RNN το 0.818, δείχνοντας ότι χάνουν περισσότερα spam (false negatives).

Ο συνδυασμός της ακρίβειας και του recall συνοψίζεται στο F1-score, το οποίο μετράει τη συνολική "ισορροπημένη" απόδοση. Σε αυτή τη μετρική, το Naive Bayes παραμένει στην κορυφή με 0.923, αποδεικνύοντας ότι προσφέρει τον καλύτερο συνολικό συνδυασμό ακρίβειας και πληρότητας. Το RNN ακολουθεί στη δεύτερη θέση με 0.890, ενώ το Word2Vec έρχεται τρίτο με 0.879.

Τελικό συμπέρασμα

Η μέθοδος **Naive Bayes** φαίνεται να είναι η πιο ισορροπημένη επιλογή: συνδυάζει υψηλή ακρίβεια, πολύ καλή precision και την καλύτερη recall, προσφέροντας το πιο εντυπωσιακό F1-score.

Από την άλλη, το **RNN** είναι ιδανικό όταν είναι κρίσιμο να μην καταλήξουν "καλά" emails σε λάθος κατηγορία ως spam, χάρη στην εξαιρετική του precision.

Τέλος, ο συνδυασμός **Word2Vec** με **Logistic Regression** προσφέρει ικανοποιητικά αποτελέσματα, αν και ελαφρώς κατώτερα, με αξιοσημείωτη σταθερότητα.

Likn Εργασίας

https://github.com/cmakionitis/erg-NLP-ice21390298