Χειμερινό
Εξάμηνο
2020-2021

Θεωρία Αποφάσεων

ΕΡΓΑΣΤΗΡΙΑΚΗ ΑΣΚΗΣΗ

ΒΑΛΛΑΤΟΣ ΑΛΕΞΑΝΔΡΟΣ Γ’ ΕΤΟΣ ΣΠΟΥΔΩΝ
ΤΖΟΥΔΑΣ ΠΑΝΑΓΙΩΤΗΣ Ε’ ΕΤΟΣ ΣΠΟΥΔΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ | Πανεπιστήμιο Πατρών

Εισαγωγή

Στα πλαίσια του μαθήματος *Θεωρία Αποφάσεων* μας ζητήθηκε να υλοποιήσουμε εργαστηριακή άσκηση πάνω σε θέματα γύρω από τη Μηχανική Μάθηση. Συγκεκριμένα, επιλέξαμε την άσκηση 8. Ζητούμενό της, να υλοποιήσουμε ένα Job recommendation system.

Πρόβλημα

Μας δίνεται ένα dataset στο οποίο περιέχονται ένα σύνολο εγγραφών σχετικές με θέσεις εργασίας. Πιο αναλυτικά, οι εγγραφές χαρακτηρίζονται από τον τίτλο της θέσης εργασίας, την περιγραφή της καθώς και τη γενικότερη κατηγορία στην οποία ανήκει. Σκοπός μας είναι η υλοποίηση αλγορίθμου ο οποίος με είσοδο την περιγραφή μιας θέσης εργασίας, θα την κατατάσσει στην αντίστοιχη κατηγορία. Με μια πρώτη ματιά λοιπόν, συμπεραίνουμε ότι έχουμε να ένα κάνουμε με ένα πρόβλημα πολλών κλάσεων κατηγοριοποίησης κειμένου (multi-class text classification problem).

Μεθοδολογία

Για να μπορέσουμε να πραγματοποιήσουμε τον text classifier θα βασιστούμε σε τεχνικές εκπαίδευσης μοντέλων. Ωστόσο, η αρχική κατάσταση στην οποία βρίσκονται τα δεδομένα εισόδου μας, δεν μας επιτρέπει να πάρουμε τα επιθυμητά αποτελέσματα μόνο με την σωστή επιλογή και διαμόρφωση του αλγορίθμου. Συνεπώς, θα πρέπει να προηγηθεί μια προ-επεξεργασία των δεδομένων εισόδου.

Η διαδικασία **προ-επεξεργασίας** (text preprocessing) αποτελείται από τα εξής βήματα:

1. Tokenization. Μετατρέπουμε το κείμενο σε λέξεις. Αυτό στην ουσία σημαίνει ότι το σύστημά μας δεν έχει «χωνέψει» ένα ολόκληρο description ως αλφαριθμητικό, αλλά πολλά διακριτά, τις λέξεις.
2. Μετατρέπουμε τα κεφαλαία γράμματα σε πεζά (λόγω του uppercase sensitivity), καταργούμε τη στίξη και αφαιρούμε τις αριθμητικά σύμβολα.
3. Stop words removal. Αφαιρούμε συνηθισμένες λέξεις οι οποίες δεν έχουν κάποια επιρροή στην κατηγοριοποίηση (πχ the, is, and etc).
4. Stemming. Συρρικνώνουμε τις λέξεις κρατώντας τη ρίζα τους.
5. Word-weighing with tf-idf. Δημιουργούμε λεξιλόγιο βάσει των διαφορετικών λέξεων στα descriptions και για κάθε λέξη του λεξιλογίου στο εκάστοτε description υπολογίζουμε τον παράγοντα tf-idf, αναπαριστώντας αυτή την πληροφορία με χρήση διανυσμάτων.

Αφότου ολοκληρώσουμε την προ-επεξεργασία, έχουμε φέρει τα αρχικά δεδομένα εισόδου σε μια μορφή που πλέον μπορούν με αποτελεσματικό τρόπο να «περάσουν» μέσα από τις **τεχνικές εκπαίδευσης**. Για το συγκεκριμένο πρόβλημα επιλέξαμε δύο τεχνικές: Naïve Bayes και Support Vector Machine. Οι δύο αυτές τεχνικές επιλέχθηκαν λόγω της κατεξοχήν προτίμησής τους (βάσει και αποδοτικότητας) σε προβλήματα τύπου text classification.

Naïve Bayes: το συγκεκριμένο μοντέλο δεν ήταν αποδοτικό στο συγκεκριμένο πρόβλημα και με τα συγκεκριμένα δοθέντα δεδομένα εισόδου. Γενικότερα αυτός ο αλγόριθμος δεν αποδίδει τόσο καλά σε πολυδιάστατα προβλήματα.

SVM: ο συγκεκριμένος αλγόριθμος αποτελεί μια από τις καλύτερες τεχνικές σε πολυδιάστατα προβλήματα. Η ευελιξία του μας δίνει τη δυνατότητα να τον προσαρμόσουμε αναλόγως των ιδιαίτερων χαρακτηριστικών του προβλήματός μας.

Όσον αφορά στη εκπαίδευση του συστήματός μας, πρόκυπτε θέμα με το testing dataset. Συγκεκριμένα, στο csv δεν περιέχονταν τα δεδομένα κατηγορίας για τις θέσεις εργασίας, πράγμα το οποίο δεν μας έδινε τη δυνατότητα να δεικτοδοτήσουμε την αποδοτικότητα του συστήματός μας (δεν μπορούσαμε να πάρουμε τις μετρικές). Έτσι, αποφασίσαμε να χωρίσουμε το training dataset σε δύο μέρη, τόσο για το train όσο και για το test (70-30).

Αποτελέσματα

Το πρώτο μοντέλο που φτιάξαμε (Naïve Bayes) μας δίνει ακρίβεια 30-35% ενώ αν μειώσουμε τα features (δοκιμάσαμε στα 100), τότε φτάνουμε μέχρι και 43-48%. Δεν υπήρχε άλλο περιθώριο βελτιστοποίησης. Το δεύτερο μοντέλο που φτιάξαμε (SVM) μας δίνει ακρίβεια 61% με C=1.0. Η επιλογή του C ως βέλτιστο προκύπτει από το GridSearch που τρέξαμε για τον σκοπό αυτό. Όσον αφορά στους πυρήνες, δοκιμάσαμε linear kernel, polynomial kernel και RBF kernel με τα καλύτερα αποτελέσματα να έρχονται από τον γραμμικό. Πράγμα αναμενόμενο δεδομένης της αποτελεσματικότητάς του σε text classification problems.

Πάτρα, 24/2/2021

Βαλλάτος Αλέξανδρος, 1067478
Τζούδας Παναγιώτης, 1054372