ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ «Ανταγωνιστικότητα Επιχειρηματικότητα και Καινοτομία»

ΑΞΟΝΑΣ ΠΡΟΤΕΡΑΙΟΤΗΤΑΣ 03 «Ανάπτυξη επιχειρηματικότητας με Τομεακές προτεραιότητες»

ΔΡΑΣΗ «ΕΡΕΥΝΩ – ΔΗΜΙΟΥΡΓΩ – ΚΑΙΝΟΤΟΜΩ»

LinkGeoML: Αυτοματοποιημένη και ακριβής διασύνδεση γεωχωρικών δεδομένων με τη χρήση μεθόδων μηχανικής μάθησης

ΚΩΔΙΚΟΣ ΟΠΣ «5030745»

ΤΙΤΛΟΣ ΠΑΡΑΔΟΤΕΟΥ

Π1.2: «Προδιαγραφή εξειδικευμένων χαρακτηριστικών και κανόνων εκπαίδευσης»

Πακέτο Εργασίας	**ΠΕ1: Μοντελοποίηση και ανάλυση απαιτήσεων**
Υπεύθυνος Φορέας	**Ερατοσθένης ΑΕ**
Είδος Παραδοτέου	**Αναφορά**
Ενδεικτικός Μήνας Παράδοσης	**Μ12**
Ημερομηνία Παράδοσης	**8/7/2019 (Μ12)**

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ ΥΠΟΥΡΓΕΙΟ ΟΙΚΟΝΟΜΙΑΣ & ΑΝΑΠΤΥΞΗΣ ΕΙΔΙΚΗ ΓΡΑΜΜΑΤΕΙΑ ΕΠΑ & ΤΣ ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ ΕΠΑνΕΚ
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

ΙΣΤΟΡΙΚΟ ΑΛΛΑΓΩΝ

Έκδοση	Ημερομηνία	Εργασίες	Συγγραφείς
0.1	06/05/2019	Δομή και πίνακας περιεχομένων του παραδοτέου	Γιώργος Γιαννόπουλος (ΑΘ.), Δημήτριος Σκούτας (ΑΘ.), Νώντας Τσάκωνας (ΕΡ.)
0.2	12/05/2019	Προσθήκη υλικού στα χαρακτηριστικά εκπαίδευσης	Νώντας Τσάκωνας (ΕΡ.), Γιώργος Ευταξίας (ΕΡ.), Μιχάλης Αεράκης (GE.), Γιώργος Γιαννόπουλος (ΑΘ.)
0.3	12/05/2019	Προσθήκη υλικού στα χαρακτηριστικά εκπαίδευσης	Βασίλης Καφφές (ΑΘ.), Νίκος Κωσταγιόλας(ΑΘ.)
0.4	19/06/2019	Διάφορες προσθήκες και βελτιώσεις	Γιώργος Γιαννόπουλος (ΑΘ.), Νώντας Τσάκωνας (ΕΡ.)
0.5	21/06/2019	Προσθήκη υλικού στον ορισμό προβλημάτων και στη γνώση πεδίου	Ιωάννης Μαρακάκης (ΕΡ.), Νώντας Τσάκωνας (ΕΡ.), Μιχάλης Αεράκης (GE.)
0.6	25/06/2019	Εσωτερικά επιθεωρημένη έκδοση	Δημήτριος Σκούτας (ΑΘ.), Θοδωρής Δαλαμάγκας (ΑΘ.), Γιώργος Γιαννόπουλος (ΑΘ.)
1.0	05/07/2019	Τελική έκδοση	Δημήτριος Σκούτας (ΑΘ.), Γιώργος Γιαννόπουλος (ΑΘ.), Νώντας Τσάκωνας (ΕΡ.)

ΠΙΝΑΚΑΣ ΣΥΝΤΜΗΣΕΩΝ

ΕΛΤΑ	Ελληνικά Ταχυδρομεία
ΣΕ	Σημείο Ενδιαφέροντος
ΤΚ	Ταχυδρομικός Κώδικας
GPS	Global Positioning System

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΠΕΡΙΛΗΨΗ

Το παραδοτέο περιγράφει τα χαρακτηριστικά εκπαίδευσης που αναπτύχθηκαν στο έργο προκειμένου να ενσωματώσουν τη γνώση πεδίου γεωχωρικών δεδομένων και να αξιοποιηθούν σε αλγόριθμους μηχανικής μάθησης για διασύνδεση γεωχωρικών δεδομένων. Η ομαδοποίηση των χαρακτηριστικών εκπαίδευσης ακολουθεί τη διάκριση σε τέσσερα σενάρια χρήσης, όπως παρουσιάστηκε στο Παραδοτέο Π1.1: «Προδιαγραφή περιπτώσεων χρήσης, βασικών δεικτών απόδοσης και συνόλου αναφοράς αξιολόγησης». Συγκεκριμένα, τα ορισμένα και υλοποιημένα χαρακτηριστικά εκπαίδευσης οργανώνονται και περιγράφονται ανά σενάριο χρήσης/διαδικασία μηχανικής μάθησης που εφαρμόστηκε. Το παραδοτέο δομείται ως εξής:

Στην Ενότητα 1 περιγράφεται η διαδικασία της εξαγωγής χαρακτηριστικών εκπαίδευσης και εξηγείται η σημασία της στην εφαρμογή μεθόδων μηχανικής μάθησης για διασύνδεση γεωχωρικών δεδομένων.

Στην Ενότητα 2 παρουσιάζονται τα χαρακτηριστικά εκπαίδευσης που ορίστηκαν και υλοποιήθηκαν προκειμένου να ενσωματωθούν σε αλγόριθμους μηχανικής μάθησης για διασύνδεση, κατηγοριοποίηση και ολοκλήρωση γεωχωρικών δεδομένων. Κάθε υπο-ενότητα αντιστοιχεί σε ένα διαφορετικό σενάριο χρήσης που ομαδοποιεί τα αντίστοιχα χαρακτηριστικά εκπαίδευσης. Για κάθε τέτοια ομάδα χαρακτηριστικών, αρχικά περιγράφεται η γνώση πεδίου από την οποία προέκυψαν, ενώ στη συνέχεια ακολουθεί λεπτομερής παρουσίαση του κάθε χαρακτηριστικού.

Στην Ενότητα 3 συνοψίζεται το παραδοτέο και παρουσιάζονται μελλοντικές κατευθύνσεις για αξιοποίηση και περαιτέρω επέκταση των χαρακτηριστικών εκπαίδευσης που παρουσιάσθηκαν.

1. Εισαγωγή

Η εξαγωγή κατάλληλων χαρακτηριστικών εκπαίδευσης είναι ουσιώδης διαδικασία στα προβλήματα μηχανικής μάθησης γενικότερα, και στα προβλήματα διασύνδεσης, κατηγοριοποίησης, γεωκωδικοποίησης και ολοκλήρωσης γεωχωρικών δεδομένων που χειριζόμαστε στο έργο ειδικότερα. Συνίσταται στη μετατροπή διαφόρων ιδιοτήτων των, εν προκειμένων γεωχωρικών, οντοτήτων σε συγκεκριμένες μεταβλητές που ποσοτικοποιούνται σε έναν συνεχή ή διακριτό χώρο για κάθε οντότητα. Με αυτόν τον τρόπο, κάθε οντότητα ή στιγμιότυπο (instance) των δεδομένων μας αναπαρίσταται ως ένα διάνυσμα χαρακτηριστικών σε έναν πολυδιάστατο χώρο, και με αυτή τη μορφή δίνεται ως είσοδος σε έναν αλγόριθμο μηχανικής μάθησης. Μέσω των χαρακτηριστικών εκπαίδευσης στοχεύουμε στην αιχμαλώτιση εκείνων των ιδιοτήτων των χειριζόμενων οντοτήτων που παίζουν το μεγαλύτερο ρόλο για την απόφαση (μέσω ενός αλγορίθμου μηχανικής μάθησης) για διασύνδεση, κατηγοριοποίηση, ολοκλήρωση, κτλ. Κατά συνέπεια, όσο το δυνατόν πιο κατάλληλα-αντιπροσωπευτικά χαρακτηριστικά εξαχθούν για το εκάστοτε σενάριο χρήσης, τόσο πιο αποτελεσματικό αναμένεται να είναι το μοντέλο μηχανικής μάθησης ως προς την ακρίβεια που επιτυγχάνει.

Τα γεωχωρικά[1] δεδομένα αποτελούν σύνθετα δεδομένα, με πολλές ιδιαιτερότητες, καθώς και παραλλαγές. Συγκεκριμένα, τα γεωχωρικά δεδομένα ενδέχεται να έχουν κειμενικές, γεωχωρικές και σημασιολογικές ιδιότητες, καθώς και συσχετίσεις με άλλες οντότητες, γεωχωρικές και μη. Σε διαφορετικά σενάρια χρήσης, κάποιες ομάδες από τις παραπάνω μπορεί να περιέχουν πιο πλούσια μεταδεδομένα, ενώ κάποιες άλλες ιδιότητες ενδέχεται να λείπουν ή να μην ορίζονται. Ενδεικτικά, στην περίπτωση των Σημείων Ενδιαφέροντος (ΣΕ), οι βασικές ιδιότητες που είναι διαθέσιμες είναι το όνομα του ΣΕ, η κατηγορία ή επισημείωσή του (tag) και οι συντεταγμένες του. Προαιρετικά, μπορεί επιπλέον να υπάρχουν σκορ αξιολογήσεων, εκτενείς κριτικές, κ.α. Αντιθέτως, σε ένα γεωτεμάχιο/αγροτεμάχιο, οι βασικές ιδιότητές του είναι η λεπτομερής πολυγωνική γεωμετρία του, καθώς και τα στοιχεία ιδιοκτησίας του. Σε μία άλλη περίπτωση, μία βάση τοπωνυμίων ενδέχεται να περιέχει μόνο τα ονόματα των οντοτήτων, χωρίς καμία γεωχωρική πληροφορία (π.χ. συντεταγμένες).

Η παραπάνω ετερογένεια εντείνεται περαιτέρω από την έμφυτη ετερογένεια στα επιμέρους σενάρια διασύνδεσης, κατηγοριοποίησης και ολοκλήρωσης γεωχωρικών δεδομένων που χειριζόμαστε. Για παράδειγμα, στο σενάριο της διασύνδεσης τοπωνυμίων ή της ολοκλήρωσης γεωτεμαχίων, ένα στιγμιότυπο αποτελείται από ένα ζεύγος γεωχωρικών οντοτήτων, οπότε πολλά από τα χαρακτηριστικά εκπαίδευσης ποσοτικοποιούν κάποια συσχέτιση μεταξύ των δύο οντοτήτων του ζεύγους. Αντιθέτως, στην περίπτωση της κατηγοριοποίησης ΣΕ, ένα στιγμιότυπο αποτελείται από μία μόνο γεωχωρική οντότητα.

[1] Στην πλειονότητα των περιπτώσεων που αφορούν το έργο, αλλά και γενικότερα εφαρμογές διαχείρισης και ολοκλήρωσης γεωχωρικών δεδομένων, τα δεδομένα, πέρα από τις γεωχωρικές ιδιότητές τους, συνοδεύονται και από κάποιας μορφής κειμενική πληροφορία (όνομα, όνομα ιδιοκτήτη, περιγραφή, κτλ.). Για αυτό, οι όροι «γεωχωρικό» και «χωρο-κειμενικό» χρησιμοποιούνται εναλλάξ στο παρόν παραδοτέο, έχοντας την ίδια σημασία.

Αντιστοίχως, στην περίπτωση της διασύνδεσης διαφορετικών πηγών γεωκωδικοποίησης, ένα στιγμιότυπο αποτελείται από πολλές διαφορετικές πηγές γεωκωδικοποίησης-συντεταγμένες μίας διεύθυνσης, οπότε αντίστοιχα προσαρμοσμένα χαρακτηριστικά θα πρέπει να ορισθούν.

Συμπερασματικά, σε διαφορετικά σενάρια χρήσης, τα οποία προδιαγράφουν τη χρήση διαφορετικών συνόλων γεωχωρικών δεδομένων, καθώς και διαφορετικών αλγορίθμων μηχανικής μάθησης για την επίλυση παραλλαγών του προβλήματος της διασύνδεσης/ολοκλήρωσης γεωχωρικών δεδομένων, είναι αναγκαία η εξαγωγή διαφορετικών (ή παραλλαγμένων) χαρακτηριστικών εκπαίδευσης για την επίτευξη μέγιστης αποτελεσματικότητας.

Δεδομένων των παραπάνω, η πρακτική που ακολουθήσαμε στο έργο ήταν να ξεκινήσουμε από τα επιμέρους σενάρια χρήσης, τα οποία καλύπτουν ετερογενή σενάρια μηχανικής μάθησης για διασύνδεση και, με βάση τη γνώση πεδίου των εταίρων, να εξαγάγουμε εξειδικευμένα χαρακτηριστικά εκπαίδευσης για κάθε σενάριο/υπο-πρόβλημα. Στην επόμενη ενότητα παρουσιάζονται, ομαδοποιημένα ανά σενάριο χρήσης, τα χαρακτηρίστηκα εκπαίδευσης που ορίσθηκαν, καθώς και η πρότερη γνώση πεδίου γεωχωρικών δεδομένων στην οποία βασίστηκαν.

Σημειώνουμε ότι στο παρόν παραδοτέο περιγράφεται και τεκμηριώνεται το σύνολο των χαρακτηριστικών εκπαίδευσης που ορίστηκαν, με βάση τη γνώση πεδίου των εταίρων, κατά την πρώτη φάση μοντελοποίησης και ανάπτυξης του έργου. Συνεπώς, η αξιολόγηση της αποτελεσματικότητας των παρουσιαζόμενων χαρακτηριστικών είναι εκτός στόχου του συγκεκριμένου παραδοτέου. Η αξιολόγηση διαφορετικών (ομάδων) χαρακτηριστικών εκπαίδευσης θα παρουσιαστεί σε επόμενα παραδοτέα του έργου, όπως τα Π2.2 και Π3.1, σύμφωνα με το πλάνο εκτέλεσης και το Τεχνικό Παράρτημα του έργου.

2. Χαρακτηριστικά Εκπαίδευσης

Η ενότητα παρουσιάζει αναλυτικά τα χαρακτηριστικά εκπαίδευσης αλγορίθμων μηχανικής μάθησης που ορίσθηκαν και ομαδοποιήθηκαν ανά σενάριο χρήσης. Για κάθε σενάριο χρήσης, γίνεται μία σύντομη περιγραφή (α) του προβλήματος πραγματικού κόσμου σε διασύνδεση/κατηγοριοποίηση/ολοκλήρωση γεωχωρικών δεδομένων (βλέπε Π1.1 «Προδιαγραφή περιπτώσεων χρήσης, βασικών δεικτών απόδοσης και συνόλου αναφοράς αξιολόγησης» για αναλυτικότερη περιγραφή), (β) του προβλήματος μηχανικής μάθησης στο οποίο αντιστοιχεί και (γ) της γνώσης πεδίου των επιχειρηματικών εταίρων που καθοδήγησαν τη διαδικασία εξαγωγής των οριζόμενων χαρακτηριστικών εκπαίδευσης.

2.1. Διασύνδεση Χωρο-κειμενικών Οντοτήτων

2.1.1. Περιγραφή σεναρίου χρήσης

Στη συνέχεια, περιγράφουμε εν συντομία το συγκεκριμένο σενάριο. Πιο αναλυτική περιγραφή αυτού του σεναρίου υπάρχει στο Π1.1 (Κεφάλαιο 2.1.1).

Στο συγκεκριμένο σενάριο θεωρούμε δύο πηγές γεωχωρικών δεδομένων που δυνητικά περιέχουν κοινές οντότητες, τις οποίες θέλουμε να αναγνωρίσουμε. Συγκεκριμένα, από τις δύο πηγές γεωχωρικών οντοτήτων, σχηματίζονται υποψήφια ζεύγη οντοτήτων που θέλουμε τελικά να αναγνωρίσουμε εάν όντως αντιστοιχούν στην ίδια γεωχωρική οντότητα ή όχι.

Οι γεωχωρικές οντότητες μπορεί να αφορούν ΣΕ, διευθύνσεις, τοπωνύμια, αγροτεμάχια, κ.α. Στο συγκεκριμένο σενάριο, και για την αρχική ανάπτυξη των μοντέλων μας[2], επικεντρώνουμε στην ιδιαίτερα απαιτητική εκδοχή όπου οι οντότητες είναι τοπωνύμια, δηλαδή ονόματα περιοχών, *χωρίς καμία γεωχωρική πληροφορία*. Τα τοπωνύμια μπορεί να αναφέρονται από μία μικρή περιοχή/γειτονιά, έως μία επαρχία ή χώρα. Η δυσκολία διασύνδεσης έγκειται στην έλλειψη οποιωνδήποτε άλλων ιδιοτήτων που θα μπορούσαν να χρησιμοποιηθούν για την εξαγωγή χρήσιμων χαρακτηριστικών εκπαίδευσης, και κυρίως στην έλλειψη συντεταγμένων των τοπωνυμίων. Το συγκεκριμένο πρόβλημα διασύνδεσης απαντάται σε περιπτώσεις διασύνδεσης τοπωνυμίων από μία υπάρχουσα βάση δεδομένων, με τοπωνύμια που μαζεύονται από κοινοποιήσεις χρηστών. Σε αυτές τις περιπτώσεις, οι συντεταγμένες της κοινοποίησης συνήθως εμπεριέχουν πολύ θόρυβο ή είναι τελείως λανθασμένες, αφού ο χρήστης ενδέχεται να έχει πραγματοποιήσει την κοινοποίηση αφού έχει απομακρυνθεί από το αντίστοιχο μέρος. Η δημοσίευση του [DOR+14] περιγράφει το συγκεκριμένο πρόβλημα που προκύπτει στη βάση δεδομένων του Facebook, καθιστώντας τις συντεταγμένες των κοινοποιήσεων άχρηστες και αναγκάζοντας τους ειδικούς να αξιοποιήσουν μόνο τα αντίστοιχα τοπωνύμια. Ένα άλλο παράδειγμα για

[2] Καθώς υλοποιούμε και αξιολογούμε μεθόδους μηχανικής μάθησης για περισσότερες παραλλαγές των θεωρημένων σεναρίων χρήσης, το σύνολο των χαρακτηριστικών εκπαίδευσης που ορίζουμε και υλοποιούμε θα επεκτείνεται διαρκώς. Οι συγκεκριμένες επεκτάσεις θα καταγραφούν και θα περιγραφούν λεπτομερώς σε ακόλουθα παραδοτέα των Ενοτήτων Εργασίας 2 και 3.

το παραπάνω σενάριο προκύπτει από τη διαδικασία μαζικής αναγνώρισης και εξαγωγής τοπωνυμίων από κείμενα, π.χ. τουριστικούς οδηγούς. Στο συγκεκριμένο παράδειγμα, ενώ η αναγνώριση τοπωνυμίων είναι σχετικά εύκολη υπόθεση (με μεθόδους Αναγνώρισης Ονοματικών Οντοτήτων), δεν ισχύει το ίδιο για τις συντεταγμένες που συνήθως λείπουν από τα συγκεκριμένα κείμενα. Και σε αυτήν την περίπτωση, μία εταιρία που τρέχει την παραπάνω διαδικασία θα πρέπει να βασιστεί αποκλειστικά στα τοπωνύμια για να διασυνδέσει τις εξηγμένες γεωχωρικές οντότητες με οντότητες που υπάρχουν ήδη στη βάση της.

2.1.2. Ορισμός προβλήματος μηχανικής μάθησης

Το παραπάνω πρόβλημα μπορεί να οριστεί ως ένα πρόβλημα δυαδικής κατάταξης (binary classification). Τα στιγμιότυπα του προβλήματος, όπως προαναφέρθηκε, είναι ζεύγη υποψήφιων τοπωνυμίων, τα οποία εξετάζονται ως προς το αν αντιπροσωπεύουν όντως ίδιες οντότητες ή όχι. Οι πιθανές κλάσεις που δύνανται να τους ανατεθούν από τον αλγόριθμο κατάταξης είναι {True, False}.

2.1.3. Γνώση πεδίου

Τα τοπωνύμια αποτελούν ονόματα που έχουν δοθεί σε συγκεκριμένα μέρη, είτε αυτά τα μέρη αποτελούν μικρές περιοχές, είτε ολόκληρες επαρχίες. Λόγω της παρόδου του χρόνου, των αλλαγών στην καθομιλουμένη γλώσσα, καθώς και διαφόρων άλλων κοινωνικών, πολιτισμικών και ιστορικών παραγόντων, για την ίδια γεωχωρική οντότητα (μέρος) εμφανίζονται παραλλαγές στο όνομά της, οδηγώντας σε ελαφρά ή και αρκετά διαφορετικά τοπωνύμια που την αφορούν. Ένα βασικό γνώρισμα των τοπωνυμίων αφορά ο διαχωρισμός των λέξεων που περιέχουν σε *βασικούς* και *συμπληρωματικούς* όρους. Οι βασικοί όροι είναι εκείνες οι λέξεις οι οποίες έχουν τη μεγαλύτερη σημασία στην αναγνώριση μίας γεωχωρικής οντότητας. Η έννοια του βασικού όρου δεν μπορεί να οριστεί αυστηρά, αφού εξαρτάται σε μεγάλο βαθμό από την ανθρώπινη κατανόηση των τοπωνυμίων και στη γνώση πεδίου των γεωχωρικών δεδομένων. Επομένως, δεν είναι δυνατόν να οριστεί ένα εξαντλητικό σύνολο κανόνων, συναρτήσεων ομοιότητας ή μοντέλων μηχανικής μάθησης που να μπορεί να αναγνωρίσει με απόλυτη επιτυχία τους βασικούς όρους ενός τοπωνυμίου. Παρόλα αυτά, θεωρούμε ότι η σχεδίαση εξειδικευμένων συναρτήσεων ομοιότητας και χαρακτηριστικών εκπαίδευσης που προσπαθούν να απομονώσουν, έστω και προσεγγιστικά, τους βασικούς όρους από τους συμπληρωματικούς μπορεί να βελτιώσει την ακρίβεια της διασύνδεσης.

Μία άλλη ιδιαιτερότητα των τοπωνυμίων είναι η ύπαρξη συχνών συμπληρωματικών όρων οι οποίοι αντιπροσωπεύουν «κατηγορική» πληροφορία για τις γεωχωρικές οντότητες και απαιτούν ιδιαίτερο χειρισμό σε σχέση με τους βασικούς όρους (π.χ. οι όροι «κοινότητα» ή «πλατεία»). Επιπλέον, ακόμα και βασικοί όροι ενός τοπωνύμιου μπορεί να έχουν διαφορετική βαρύτητα για τον προσδιορισμό του σε σχέση με άλλα τοπωνύμια (π.χ. οι όροι «άνω» ή «νέο», σε σχέση με το όνομα-όρο που συνήθως τους ακολουθεί).

Επίσης, ένας σημαντικός παράγοντας που εισάγει ετερογένεια είναι το γεγονός ότι οι ίδιοι βασικοί όροι απαντώνται σε διαφορετικές εκδοχές, όπως συντμημένοι, με διαφορετική ορθογραφία και γενικά παραλλαγμένοι. Επιπλέον, δύο τοπωνύμια που αναφέρονται στην ίδια γεωχωρική οντότητα ενδέχεται να περιέχουν αρκετούς κοινούς όρους, αλλά σε

διαφορετική σειρά. Το πρόβλημα ενισχύεται ακόμα περισσότερο αν σκεφτούμε ότι κάποιοι από αυτούς τους όρους, που αντιστοιχούν μεταξύ τους, μπορεί να είναι ελαφρά διαφορετικά γραμμένοι, ενώ κάποιοι άλλοι μπορεί να λείπουν από το τοπωνύμιο.

Στο σενάριο χρήσης που εξετάζουμε, οι όροι του τοπωνύμιου είναι η μόνη αξιόπιστη πληροφορία που μπορεί να χρησιμοποιηθεί για διασύνδεση τοπωνυμίων. Επομένως, τα χαρακτηριστικά εκπαίδευσης που θα εξαχθούν για κάθε υποψήφιο ζεύγος τοπωνυμίων, πρέπει να επικεντρώνονται στην ομοιότητα των δύο συγκρινόμενων τοπωνυμίων, λαμβάνοντας όμως υπόψιν τους παραπάνω παράγοντες-ιδιαιτερότητες.

2.1.4. Χαρακτηριστικά εκπαίδευσης

Στη συνέχεια περιγράφονται αναλυτικά τα χαρακτηριστικά εκπαίδευσης που σχεδιάσαμε και υλοποιήσαμε για το πρόβλημα της διασύνδεσης τοπωνυμίων μέσω δυαδικής κατάταξης. Σε αυτή τη φάση, επικεντρώσαμε σε χαρακτηριστικά εκπαίδευσης που εξετάζουν διάφορες πτυχές ομοιότητας των συγκρινόμενων τοπωνυμίων. Σημειώνουμε ότι κάθε ακόλουθη υπο-ενότητα περιγράφει μία *ομάδα* χαρακτηριστικών, και όχι ένα και μονο χαρακτηριστικό. Ο συνολικός αριθμός των μεμονωμένων χαρακτηριστικών που προκύπτουν για την κάθε ομάδα περιγράφεται στο υπο-πεδίο «Αριθμός θέσεων στο διάνυσμα χαρακτηριστικών»

2.1.4.1. Ομοιότητα των αρχικών συμβολοσειρών των τοπωνυμίων

Σύντομη περιγραφή

Αυτή η ομάδα χαρακτηριστικών εκπαίδευσης προκύπτει από την εφαρμογή διαφόρων συναρτήσεων ομοιότητας συμβολοσειρών (string similarity functions) στα δύο συγκρινόμενα τοπωνύμια. Οι συναρτήσεις ομοιότητας που υιοθετήσαμε χρησιμοποιούνται ευρέως στη σχετική βιβλιογραφία για διασύνδεση ονομάτων και τοπωνυμίων και περιλαμβάνουν από απλές μετρικές ομοιότητας, έως σύνθετες μετα-συναρτήσεις ομοιότητας [SMM17]. Συγκεκριμένα, χρησιμοποιούμε τις παρακάτω συναρτήσεις:

- *Damerau-Levenshtein*. Η μετρική αυτή ποσοτικοποιεί και μετατρέπει σε σκορ ομοιότητας τον ελάχιστο αριθμό από εισαγωγές, διαγραφές, αντικαταστάσεις και μεταθέσεις χαρακτήρων που απαιτούνται για τη μετατροπή μιας συμβολοσειράς σε μια άλλη.
- *Jaro.* Ο αλγόριθμος αυτός αποτελεί μια ευριστική συνάρτηση που βασίζεται στον υπολογισμό του αριθμού των κοινών χαρακτήρων μεταξύ δύο συμβολοσειρών και του αριθμού των μεταθέσεών τους. Συγκεκριμένα, όμοιοι χαρακτήρες που βρίσκονται σε απόσταση μικρότερη ή ίση του μισού του μήκους της μεγαλύτερης συμβολοσειράς θεωρούνται ως κοινοί, ενώ οι κοινοί χαρακτήρες που δεν ακολουθούν την ίδια διάταξη στις δύο συμβολοσειρές υπολογίζονται ως μεταθέσεις.
- *Jaro–Winkler*. Η μετρική αυτή είναι μια βελτιωμένη εκδοχή του αλγορίθμου Jaro, όπου το αρχικό σκορ ομοιότητας αυξάνεται βάσει του αριθμού των κοινών χαρακτήρων που υπάρχουν στην αρχή των δύο συμβολοσειρών πολλαπλασιασμένο με ένα σταθερό τελεστή κλιμάκωσης.

- *Jaro–Winkler Reversed*. Μια παραλλαγή της Jaro-Winkler μετρικής όπου γίνεται αντιστροφή των χαρακτήρων των δυο συμβολοσειρών πριν την εφαρμογή του αλγορίθμου.
- *Sorted Jaro–Winkler*. Η συγκεκριμένη παραλλαγή της Jaro-Winkler έχει εφαρμογή σε συμβολοσειρές που περιέχουν πάνω από μία λέξη. Αρχικά, οι λέξεις των δύο συμβολοσειρών ταξινομούνται αλφαριθμητικά. Έπειτα, υπολογίζεται το σκορ ομοιότητας με τη μετρική Jaro-Winkler στις ταξινομημένες συμβολοσειρές.
- *Cosine N-Grams*. Η μετρική αυτή υπολογίζει το κανονικοποιημένο εσωτερικό γινόμενο μεταξύ της n-gram αναπαράστασης των δύο συμβολοσειρών στο διανυσματικό χώρο. Η n-gram αναπαράσταση κάθε συμβολοσειράς γίνεται ως εξής. Σε πρώτη φάση, μετατρέπουμε κάθε συμβολοσειρά σε αλληλουχίες από n-συνεχόμενους χαρακτήρες. Έπειτα, για κάθε συμβολοσειρά φτιάχνουμε ένα διάνυσμα με μέγεθος ίσο με το συνολικό αριθμό των n-grams που έχουν προκύψει και, τέλος, θέτουμε την τιμή 1 ή 0, για κάθε n-gram, στη θέση που έχει εκχωρηθεί στο διάνυσμα ανάλογα με το αν υπάρχει ή όχι, αντίστοιχα, στην αρχική συμβολοσειρά.
- *Jaccard N-Grams*. Αρχικά, κάθε συμβολοσειρά αναπαρίσταται ως ένα σύνολο από n-gram χαρακτήρες. Έπειτα, το σκορ ομοιότητας υπολογίζεται ως το πηλίκο της διαίρεσης του αριθμού της τομής των δύο συνόλων προς τον αριθμό της ένωσης αυτών.
- *Dice Bi-Grams*. Η συνάρτηση αυτή αποτελεί παραλλαγή της μετρικής Jaccard N-Grams. Εφαρμόζεται για σύνολα διγραμμάτων (bi-grams) και ορίζεται ως δύο φορές το πηλίκο της τομής των δύο συνόλων προς το άθροισμα του πλήθους αυτών.
- *Jaccard Skip-grams*. Η συγκεκριμένη μετρική ορίζεται όπως και η Jaccard N-Grams με τη διαφορά ότι τα σύνολα διγραμμάτων δεν αποτελούνται πάντα από χαρακτήρες που γειτνιάζουν στις αρχικές συμβολοσειρές, αλλά επιτρέπεται να έχουν απόσταση 0, 1 ή 2 χαρακτήρων.
- *Monge–Elkan*. Η συνάρτηση αυτή ανήκει στις σύνθετες μετα-συναρτήσεις ομοιότητας και περιλαμβάνει τον υπολογισμό του μέσου όρου ομοιότητας μεταξύ ζευγαριών λέξεων που ανήκουν στις δύο συμβολοσειρές. Ως επιμέρους ζευγάρια επιλέγονται, από τις δύο συμβολοσειρές, οι λέξεις που έχουν το μεγαλύτερο σκορ ομοιότητας μεταξύ τους, βάσει της μετρικής Jaro-Winkler.
- *Soft–Jaccard*. Η σύνθετη αυτή μετα-συνάρτηση ακολουθεί τη φιλοσοφία της Jaccard N-Grams μετρικής με τις εξής δύο διαφοροποιήσεις. Πρώτον, εφαρμόζεται στις λέξεις των συμβολοσειρών και όχι σε n-grams και δεύτερον χρησιμοποιεί τη μετρική Jaro-Winkler για την εύρεση σχετικά όμοιων ζευγαριών λέξεων στις δύο συμβολοσειρές. Με τον τρόπο αυτό, η μετρική γίνεται πιο ευέλικτη και μπορεί να διαχειριστεί μικρές διαφορές μεταξύ των συμβολοσειρών που οφείλονται σε ορθογραφικά λάθη ή αντιστροφή γειτονικών χαρακτήρων λόγω βιασύνης.
- *Davis and De Salles*. Για τον υπολογισμό του σκορ ομοιότητας ακολουθείται η εξής διαδικασία. Αρχικά, οι δύο συμβολοσειρές χωρίζονται σε λέξεις βάσει διαφόρων διαχωριστικών όπως κενά και παύλες. Έπειτα, εντοπίζονται τυχόν γνωστές συντομεύσεις και αντικαθίστανται με το πλήρες όνομά τους. Στη συνέχεια, ελέγχεται η ύπαρξη μη συνηθισμένων συντομεύσεων ως εξής: Για κάθε μία από τις δύο συμβολοσειρές, ελέγχονται οι λέξεις που περιέχουν *τελεία*. Εάν εντοπιστεί μια τέτοια λέξη, ελέγχεται αν αποτελεί πρόθεμα σε κάποια από τις λέξεις της άλλης

συμβολοσειράς και γίνεται αντικατάστασή της. Τέλος, υπολογίζεται το τελικό σκορ ομοιότητας με γραμμικό συνδυασμό των επιμέρους σκορ που προκύπτουν από τη χρήση της μετρικής Sorted-Winkler και της μετρικής Soft-Jaccard.

- *LGM Jaro-Winkler*. Η μετρική αυτή αποτελεί προτεινόμενη από εμάς παραλλαγή της Jaro-Winkler συνάρτησης, όπου το μέγεθος κοινών προθεμάτων μεταξύ των δύο συμβολοσειρών μπορεί να προκύψει έχοντας παραλείψει μέχρι και έναν ανόμοιο χαρακτήρα, είτε στη μία από τις δύο συμβολοσειρές, είτε και στις δύο.
- *LGM Jaro-Winkler reversed*. Ο υπολογισμός του σκορ ομοιότητας γίνεται όπως και στην LGM Jaro-Winkler με τη διαφορά ότι γίνεται χρήση της μετρικής Jaro-Winkler reversed.

Αριθμός θέσεων στο διάνυσμα χαρακτηριστικών

Κάθε μία από τις παραπάνω συναρτήσεις παράγει ως έξοδο έναν πραγματικό αριθμό, ο οποίος αντιπροσωπεύει το σκορ ομοιότητας μεταξύ των δύο τοπωνυμίων. Κάθε ένα από αυτά τα σκορ χρησιμοποιείται ως ένα ξεχωριστό χαρακτηριστικό εκπαίδευσης. Συνολικά υλοποιούνται 14 χαρακτηριστικά εκπαίδευσης.

Τύπος και εύρος τιμών

Στις παραπάνω συναρτήσεις, το κάθε σκορ (που αντιστοιχεί σε ένα διαφορετικό χαρακτηριστικό) είναι κανονικοποιημένο στο διάστημα [0,1].

Κανονικοποίηση τιμών

Όλα τα σκορ ομοιότητας που παράγονται από τις παραπάνω συναρτήσεις βρίσκονται στο διάστημα [0,1] και, επομένως, δεν απαιτείται περαιτέρω κανονικοποίηση.

Κατηγορία

Κειμενικά χαρακτηριστικά

2.1.4.2. Ομοιότητα των ταξινομημένων συμβολοσειρών των τοπωνυμίων

Σύντομη περιγραφή

Η συγκεκριμένη ομάδα χαρακτηριστικών εκπαίδευσης προκύπτει από την εφαρμογή διαφόρων συναρτήσεων ομοιότητας συμβολοσειρών στα δύο συγκρινόμενα τοπωνύμια, αφού έχει προηγηθεί η αλφαριθμητική ταξινόμηση των όρων τους. Η διαδικασία αυτή επιτρέπει ίδιους/παρόμοιους όρους (λέξεις) στις δύο συμβολοσειρές, που βρίσκονται όμως σε διαφορετική θέση, να στοιχίζονται με αποτέλεσμα να επέρχεται μεγαλύτερη ακρίβεια των μετρικών ομοιότητας ως προς τη διασύνδεση των τοπωνυμίων. Συγκεκριμένα, η αλφαριθμητική ταξινόμηση των όρων-λέξεων κάθε τοπωνυμίου-συμβολοσειράς, έχοντας αφαιρέσει όλα τα κενά διαστήματα, γίνεται υπό την προϋπόθεση ότι το σκορ ομοιότητας μεταξύ τους είναι μικρότερο από ένα κατώφλι. Με τη συνθήκη αυτή μετριάζουμε προβληματικές περιπτώσεις όπου η ύπαρξη κενού διαστήματος στα τοπωνύμια έχει προκύψει είτε από λάθος είτε λόγω κάποιου τοπικού ιδιωματισμού, με αποτέλεσμα η ταξινόμηση να δυσχεραίνει περαιτέρω τη διασύνδεσή τους. Στη συνέχεια, κάθε μία από τις μετρικές που παρουσιάστηκαν στην Ενότητα 2.1.4.1, εκτός της Sorted Jaro-Winkler που ήδη ενσωματώνει μηχανισμό ταξινόμησης, εφαρμόζονται στις ταξινομημένες συμβολοσειρές για να προκύψει ένα σκορ ομοιότητας.

Αριθμός θέσεων στο διάνυσμα χαρακτηριστικών

Κάθε μία από τις παραπάνω συναρτήσεις παράγει ως έξοδο έναν πραγματικό αριθμό ο οποίος αντιπροσωπεύει το σκορ ομοιότητας μεταξύ των δύο τοπωνυμίων. Κάθε ένα από αυτά τα σκορ χρησιμοποιείται ως ένα ξεχωριστό χαρακτηριστικό εκπαίδευσης. Συνολικά υλοποιούνται 13 χαρακτηριστικά εκπαίδευσης.

Τύπος και εύρος τιμών

Στις παραπάνω συναρτήσεις, το κάθε σκορ (που αντιστοιχεί σε ένα διαφορετικό χαρακτηριστικό) είναι κανονικοποιημένο στο διάστημα [0,1].

Κανονικοποίηση τιμών

Όλα τα σκορ ομοιότητας που παράγονται από τις παραπάνω συναρτήσεις βρίσκονται στο διάστημα [0,1] και, επομένως, δεν απαιτείται περαιτέρω κανονικοποίηση.

Κατηγορία

Κειμενικά χαρακτηριστικά.

2.1.4.3. Ομοιότητα των εξειδικευμένα προεπεξεργασμένων συμβολοσειρών των τοπωνυμίων

Σύντομη περιγραφή

Αυτή η ομάδα χαρακτηριστικών ανήκει στις σύνθετες μετα-συναρτήσεις ομοιότητας. Στόχος της εξειδικευμένης μετα-συνάρτησης *LGM-Sim* που προτείνουμε και υλοποιούμε είναι να ενσωματώσει χαρακτηριστικές ιδιομορφίες των τοπωνυμίων, όπως διαφορετική ορθογραφία, κοινοί όροι σε διαφορετική σειρά ή/και έλλειψη αυτών στο ένα από τα δύο τοπωνύμια κ.α., ώστε να αποφασίσουμε με μεγαλύτερη ακρίβεια για τη διασύνδεσή τους. Συγκεκριμένα, ο υπολογισμός του σκορ ομοιότητας γίνεται ως εξής. Αρχικά, στη φάση της προ-επεξεργασίας, εξάγουμε μια λίστα με τους συχνότερους όρους για το σύνολο των τοπωνυμίων που έχουμε διαθέσιμο. Η διαδικασία αυτή εκτελείται στην αρχή της επεξεργασίας ενός νέου συνόλου τοπωνυμίων και δεν επηρεάζει την απόδοση της μετα-συνάρτησης.

Έπειτα, ακολουθεί η φάση της κύριας επεξεργασίας, η οποία επαναλαμβάνεται για κάθε ζευγάρι τοπωνυμίων. Πρώτα αφαιρούνται σημεία στίξης και τονισμού από τα δύο τοπωνύμια-συμβολοσειρές. Επιπλέον, γίνεται αλφαριθμητική ταξινόμηση των λέξεων κάθε συμβολοσειράς με τη διαδικασία που περιγράφηκε στην Ενότητα 2.1.4.2. Στη συνέχεια, για κάθε συμβολοσειρά, δημιουργούμε τρεις διακριτές λίστες, με κάθε μια από αυτές να περιέχει όρους-λέξεις με συγκεκριμένα σημασιολογικά χαρακτηριστικά. Πρώτα, εντοπίζουμε τυχόν συχνούς όρους που υπάρχουν στις δύο συμβολοσειρές, βάσει της λίστας που έχουμε δημιουργήσει στη φάση της προ-επεξεργασίας, και τους τοποθετούμε σε δύο λίστες, μία για κάθε συμβολοσειρά. Έχοντας απομακρύνει τους συχνούς όρους, κατηγοριοποιούμε τους υπόλοιπους όρους σε βασικούς, δηλαδή κοινούς και στις δυο συμβολοσειρές, και σε συμπληρωματικούς, δηλαδή όρους που υπάρχουν μόνο στη μια από τις δύο συμβολοσειρές. Αυτό γίνεται εξετάζοντας τη σχετική ομοιότητα μεταξύ ζευγαριών όρων από τις δύο συμβολοσειρές μέσω κάποιας υιοθετούμενης βασικής μετρικής ομοιότητας. Όσα ζευγάρια δίνουν σκορ ομοιότητας μεγαλύτερο από ένα σχετικά χαμηλό κατώφλι χαρακτηρίζονται ως βασικοί όροι και αποθηκεύονται στις αντίστοιχες

λίστες. Οι υπόλοιποι μεταφέρονται στις λίστες συμπληρωματικών όρων. Για την αποδοτική εκτέλεση του παραπάνω διαχωρισμού των όρων κάθε συμβολοσειράς σε δύο κατηγορίες λιστών, με βασικούς και συμπληρωματικούς όρους, ακολουθείται η εξής διαδικασία. Πρώτα, επιλέγουμε τον πρώτο όρο-λέξη από κάθε συμβολοσειρά, μέσω δεικτοδότησης, για σύγκριση. Εάν οι δύο όροι εμφανίζουν μια σχετική ομοιότητα, αποθηκεύουμε τον κάθε όρο στην αντίστοιχη, για κάθε συμβολοσειρά, λίστα βασικών όρων και προχωράμε μία θέση, δηλαδή όρο, κάθε δείκτη. Διαφορετικά μεταφέρουμε τον αλφαριθμητικά μικρότερο όρο στη λίστα συμπληρωματικών όρων που σχετίζεται με τη συμβολοσειρά που ανήκει και δεικτοδοτούμε τον αμέσως επόμενο όρο. Η παραπάνω διαδικασία συνεχίζεται μέχρι όλοι οι όροι από κάθε συμβολοσειρά να βρίσκονται είτε στη λίστα βασικών είτε στη λίστα συμπληρωματικών όρων για κάθε συμβολοσειρά ξεχωριστά. Με αυτόν τον τρόπο πετυχαίνουμε να βρούμε αποδοτικά όλα τα ζευγάρια όρων από τις δύο συμβολοσειρές που είναι σχετικά όμοια με ένα αριθμό συγκρίσεων που είναι το πολύ ίσος με τον μεγαλύτερο αριθμό όρων μεταξύ των δύο συμβολοσειρών. Τέλος, εφαρμόζουμε κάποια υιοθετούμενη βασική μετρική μεταξύ των όρων των σχετικών λιστών που έχουν προκύψει από κάθε συμβολοσειρά με αποτέλεσμα να προκύψουν τρία διαφορετικά σκορ ομοιότητας. Για να πάρουμε το τελικό σκορ, συνδυάζουμε γραμμικά και με διαφορετικούς συντελεστές-βάρη κάθε ένα από αυτά τα σκορ. Η χρήση βαρών καθορίζει τη σημαντικότητα του κάθε σκορ στην απόφαση για ομοιότητα ή μη μεταξύ ζευγαριών τοπωνυμίων. Να σημειώσουμε ότι, κάθε φορά που απαιτείται η χρήση κάποιας συνάρτησης ομοιότητας στην παραπάνω διαδικασία, επιλέγουμε οποιαδήποτε από τις μετρικές που παρουσιάστηκαν στην Ενότητα 2.1.4.1 (εκτός φυσικά της Sorted Jaro-Winkler), επιλέγοντας και τα αντίστοιχα κατώφλια τα οποία έχουν βρεθεί με ευριστικό τρόπο.

Αριθμός θέσεων στο διάνυσμα χαρακτηριστικών

Η παραπάνω μετά-συνάρτηση παράγει ως έξοδο έναν πραγματικό αριθμό για κάθε μία μετρική της Ενότητας 2.1.4.1 που χρησιμοποιείται «εσωτερικά» και ο οποίος αντιπροσωπεύει το σκορ ομοιότητας μεταξύ των δύο τοπωνυμίων. Κάθε ένα από αυτά τα σκορ χρησιμοποιείται ως ένα ξεχωριστό χαρακτηριστικό εκπαίδευσης. Συνολικά υλοποιούνται 13 χαρακτηριστικά εκπαίδευσης.

Τύπος και εύρος τιμών

Στις παραπάνω εκδοχές της μετα-συνάρτησης που προκύπτουν με τη χρήση διαφορετικών «εσωτερικών» μετρικών, το κάθε σκορ (που αντιστοιχεί σε ένα διαφορετικό χαρακτηριστικό) είναι κανονικοποιημένο στο διάστημα [0,1].

Κανονικοποίηση τιμών

Όλα τα σκορ ομοιότητας που παράγονται από τις παραπάνω διαφορετικές εκδοχές της μετα-συνάρτησης βρίσκονται στο διάστημα [0,1] και, επομένως, δεν απαιτείται περαιτέρω κανονικοποίηση.

Κατηγορία

Κειμενικά χαρακτηριστικά.

2.1.4.4. Ομοιότητα των εξειδικευμένα προεπεξεργασμένων συμβολοσειρών ανά επιμέρους τμήματα των τοπωνυμίων

Σύντομη περιγραφή

Η ομάδα αυτή χαρακτηριστικών αποτελεί παραλλαγή της σύνθετης μετα-συνάρτησης ομοιότητας LGM-Sim, που παρουσιάστηκε στην Ενότητα 2.1.4.3. Συγκεκριμένα, για την υλοποίηση της συγκεκριμένης παραλλαγής, εκτελούμε με τον ίδιο τρόπο και σειρά τα βήματα που περιγράφονται για την LGM-Sim εκτός του τελευταίου, όπου και παράγεται ως έξοδος ένα τελικό σκορ και κάνοντας χρήση της Damerau-Levenshtein ως «εσωτερικής» μετρικής. Συνεπώς, η έξοδος είναι τρία σκορ ομοιότητας που προκύπτουν από την κατηγοριοποίηση των όρων των τοπωνυμίων σε βασικούς, συμπληρωματικούς και συχνούς. Στόχος της προτεινόμενης ομάδας χαρακτηριστικών είναι να επιτραπεί στους αλγορίθμους μηχανικής μάθησης να μάθουν και να αναζητήσουν τη σημαντικότητα της κάθε λίστας όρων ξεχωριστά, καθώς και σε συνδυασμό με το συνολικό σκορ που προκύπτει από την προηγούμενη ομάδα χαρακτηριστικών.

Αριθμός θέσεων στο διάνυσμα χαρακτηριστικών

Η παραπάνω μετά-συνάρτηση παράγει ως έξοδο τρεις πραγματικούς αριθμούς, καθένας από τους οποίους αντιπροσωπεύει το σκορ ομοιότητας για τις τρεις κατηγοριοποιήσεις των όρων των δύο τοπωνυμίων σε βασικούς, συμπληρωματικούς και συχνούς. Κάθε ένα από αυτά τα σκορ χρησιμοποιείται ως ένα ξεχωριστό χαρακτηριστικό εκπαίδευσης. Συνολικά υλοποιούνται 3 χαρακτηριστικά εκπαίδευσης.

Τύπος και εύρος τιμών

Τα παραπάνω σκορ είναι κανονικοποιημένα στο διάστημα [0,1].

Κανονικοποίηση τιμών

Όλα τα ενδιάμεσα σκορ ομοιότητας που παράγονται από τη μετα-συνάρτηση LGM-Sim βρίσκονται στο διάστημα [0,1] και, επομένως, δεν απαιτείται περαιτέρω κανονικοποίηση.

Κατηγορία

Κειμενικά χαρακτηριστικά.

2.2. Κατηγοριοποίηση Σημείων Ενδιαφέροντος

2.2.1. Περιγραφή σεναρίου χρήσης

Στη συνέχεια, περιγράφουμε εν συντομία το συγκεκριμένο σενάριο. Πιο αναλυτική περιγραφή αυτού του σεναρίου υπάρχει στο Π1.1 (Κεφάλαιο 2.2.1).

Η συλλογή και κατηγοριοποίηση Σημείων Ενδιαφέροντος (ΣΕ) αποτελεί βασικό επιχειρησιακό αντικείμενο των εταιρειών Geodata και Ερατοσθένης. Ως σημείο ενδιαφέροντος μπορεί να χαρακτηριστεί οποιοδήποτε σημείο στο γεωγραφικό χώρο το οποίο μπορεί να αναπαρασταθεί με βάση τις συντεταγμένες του (x, y) και διάφορα πεδία κατηγοριών (categories), που το χαρακτηρίζουν και το ταυτοποιούν. Ο καθορισμός των κατηγοριών (κατηγοριοποίηση) σε ένα σύνολο ΣΕ έχει να κάνει συνήθως με την επιχειρησιακή τους χρήση και τις διάφορες γεωχωρικές αναλύσεις που χρησιμοποιούνται στα πλαίσια μιας μελέτης ή ενός γενικότερου έργου. Συνήθως, τηρείται μια βασική (master) έκδοση ΣΕ με κάποια σταθερά πεδία κατηγοριών, η οποία εμπλουτίζεται, ανάλογα με τις ανάγκες και τις απαιτήσεις που προκύπτουν, μέσα από τα διάφορα έργα και μελέτες που αναλαμβάνουν να διεκπεραιώσουν οι δύο εταιρείες. Συνοπτικά οι εργασίες που επιτελούνται σε σχέση με τα ΣΕ, είναι:

- Συλλογή, είτε με αποτύπωση στο πεδίο (με χρήση GPS), είτε με χρήση έτοιμων ανοιχτών πηγών δεδομένων.
- Επεξεργασία, όσον αφορά την διόρθωση των υφιστάμενων πεδίων κατηγοριών τους αλλά και εμπλουτισμός με επιπλέον πεδία για τις ειδικές ανάγκες ενός έργου. Συνήθως αυτή η περίπτωση αφορά βάσεις δεδομένων ΣΕ πελατών που απαιτούν διόρθωση και εμπλουτισμό των πεδίων κατηγοριών με νέα.

Το επιχειρησιακό πρόβλημα που προκύπτει για τις εταιρείες, όσον αφορά στην επεξεργασία των κατηγοριών των ΣΕ, έχει να κάνει με την ανάπτυξη ενός εργαλείου το οποίο θα υποβοηθά και θα ελαχιστοποιεί την ανθρώπινη παρέμβαση στη διόρθωση και ανάπτυξη νέων πεδίων κατηγοριοποίησης σε ένα σύνολο ΣΕ.

2.2.2. Ορισμός προβλήματος μηχανικής μάθησης

Το παραπάνω πρόβλημα μπορεί να οριστεί ως ένα πρόβλημα κατάταξης με πολλαπλές κλάσεις (multiclass classification). Τα στιγμιότυπα του προβλήματος είναι μεμονωμένα ΣΕ. Οι πιθανές κλάσεις που δύνανται να τους ανατεθούν από τον αλγόριθμο κατάταξης είναι το σύνολο των κατηγοριών που περιέχονται στην κατηγοριοποίηση του εκάστοτε συνόλου δεδομένων ΣΕ.

2.2.3. Γνώση πεδίου

Τα ΣΕ είναι γεωχωρικές (για την ακρίβεια, χωρο-κειμενικές) οντότητες, οι οποίες χαρακτηρίζονται κατ' ελάχιστον από ένα όνομα, από ένα σύνολο συντεταγμένων που καθορίζει τη γεωγραφική τους θέση και, ιδανικά, από μία ή περισσότερες κατηγορίες οι οποίες χαρακτηρίζουν τη λειτουργία/χρήση τους. Πέρα από αυτές τις ουσιώδεις ιδιότητες, ένα ΣΕ μπορεί να χαρακτηρίζεται από επιπλέον κειμενικές, γεωχωρικές ή σημασιολογικές ιδιότητες, όπως: σύντομη/εκτενής περιγραφή του ΣΕ, πληροφορίες του ιστορικού του,

ώρες λειτουργίας, πληροφορίες προσβασιμότητας, κριτικές και αξιολογήσεις χρηστών. Επίσης, λόγω της γεωχωρικής του φύσης, ένα ΣΕ χαρακτηρίζεται έμμεσα από τις γεωχωρικές συσχετίσεις του με άλλα ΣΕ (και γεωχωρικές οντότητες γενικότερα), όπως: αριθμός και κατηγορίες γειτονικών ΣΕ σε κάποια ακτίνα, γειτονικά ΣΕ τα οποία βρίσκονται στον ίδιο δρόμο με το ΣΕ, περιοχές πυκνότητας ΣΕ που σηματοδοτούν περιοχές με ΣΕ παρόμοιας ή συμπληρωματικής κατηγορίας, κ.α.

Δεδομένων των παραπάνω, μπορούμε να θεωρήσουμε τέσσερις τύπους χαρακτηριστικών εκπαίδευσης: κειμενικά, γεωχωρικά, γειτνίασης και σημασιολογικά. Η ύπαρξη ενός συγκεκριμένου όρου στο όνομα ενός ΣΕ αποτελεί παράδειγμα κειμενικού χαρακτηριστικού εκπαίδευσης. Το εμβαδόν του πολυγώνου ενός ΣΕ μπορεί να αποτελέσει ένα γεωχωρικό χαρακτηριστικό, ενώ ο αριθμός των ΣΕ σε συγκεκριμένη ακτίνα γύρω από ένα ΣΕ αποτελεί παράδειγμα χαρακτηριστικού γειτνίασης. Τέλος, κατηγορίες με τις οποίες είναι ήδη επισημειωμένο ένα ΣΕ θα μπορούσαν να αποτελέσουν παράδειγμα σημασιολογικού χαρακτηριστικού.

Στο σενάριο που εξετάζουμε, το οποίο αποτελεί μία συνήθη περίπτωση σε διαδικασίες πραγματικού κόσμου, τα διαθέσιμα ΣΕ χαρακτηρίζονται μόνο από το όνομά τους και τις συντεταγμένες τους. Οπότε, είναι αδύνατη η εξαγωγή εκλεπτυσμένων κειμενικών χαρακτηριστικών εκπαίδευσης από εκτενείς κειμενικές περιγραφές των ΣΕ ή γεωχωρικών χαρακτηριστικών, αφού οι πολυγωνικές γεωμετρίες των ΣΕ δεν είναι διαθέσιμες. Αντιθέτως, μπορούμε να εκμεταλλευτούμε τη δυνατότητα διασύνδεσης των ΣΕ του εκάστοτε συνόλου δεδομένων ΣΕ, με γειτονικά ΣΕ του ίδιου συνόλου ή/και με ΣΕ προερχόμενα από ανοικτά σύνολα γεωχωρικών δεδομένων, όπως το OpenStreetMap[3], για την εξαγωγή χαρακτηριστικών εκπαίδευσης γειτνίασης. Η λογική πίσω από τη θεώρηση χαρακτηριστικών γειτνίασης στηρίζεται στην εμπειρική γνώση ότι, για αρκετές κατηγορίες ΣΕ, η χωρική εγγύτητα συνεπάγεται συσχέτιση στις κατηγορίες των ΣΕ. Για παράδειγμα, είναι αναμενόμενο αρκετά ΣΕ της κατηγορίας «τράπεζα» να βρίσκονται σε περιοχές που περιέχουν πολλά ΣΕ της κατηγορίας «επιχείρηση», ενώ «μπαρ» και «εστιατόρια» πολύ συχνά συγκεντρώνονται σε κοντινές μεταξύ τους αποστάσεις.

Επιπλέον, εμπειρικά, αρκετές κατηγορίες ΣΕ εμπεριέχουν στο όνομά τους όρους οι οποίοι είναι στενά-συχνά σχετιζόμενοι με την αντίστοιχη κατηγορία. Δεδομένου αυτού, είναι δυνητικά χρήσιμη η κατασκευή χαρακτηριστικών εκπαίδευσης που θα συσχετίζουν κατηγορίες με συγκεκριμένους όρους που αναγνωρίζονται στα ονόματα των ΣΕ.

2.2.4. Χαρακτηριστικά εκπαίδευσης

2.2.4.1. Ομοιότητα ονόματος ΣΕ με κειμενικές αναπαραστάσεις κλάσεων

Σύντομη περιγραφή

Το σύνολο αυτών των χαρακτηριστικών αποτελείται από *C* χαρακτηριστικά, όπου *C* ο αριθμός των διαφορετικών κατηγοριών στο διαθέσιμο σύνολο δεδομένων. Για κάθε

[3] https://www.openstreetmap.org

κατηγορία συγκεντρώνονται τα ονόματα των ΣΕ που ανήκουν σε αυτή, ώστε να δημιουργηθεί μια κειμενική αναπαράσταση για την κατηγορία αυτή, υπό τη μορφή ενός σάκου όρων (bag of words). Με αυτήν την προσέγγιση, η κειμενική αναπαράσταση κάθε κατηγορίας αποτελείται από το σύνολο των όρων που συναντώνται στα ονόματα των ΣΕ που ανήκουν σε αυτήν, χωρίς να λαμβάνεται υπόψιν η σειρά εμφάνισής τους. Στη συνέχεια, λαμβάνεται η ομοιότητα κάθε ΣΕ με την κάθε κατηγορία που βρίσκεται στο σύνολο δεδομένων μέσω τεχνικών διανυσματικής απόστασης των κειμενικών αναπαραστάσεών τους. Οι ομοιότητες αυτές χρησιμοποιούνται ως χαρακτηριστικά εκπαίδευσης, όπου το κάθε διάνυσμα χαρακτηριστικών που αντιστοιχεί σε κάθε ΣΕ αποτελείται από τις ομοιότητες της κειμενικής αναπαράστασης του ονόματός του με τις επιμέρους κειμενικές αναπαραστάσεις της κάθε κατηγορίας.

Αριθμός θέσεων στο διάνυσμα χαρακτηριστικών

Το είδος των χαρακτηριστικών αυτών καλύπτει $|C|$ θέσεις, όπου $|C|$ ο αριθμός των διαφορετικών κατηγοριών στο σύνολο δεδομένων που χρησιμοποιείται.

Τύπος και εύρος τιμών

Τύπος: Αριθμητικές τιμές

Εύρος τιμών πριν τη χρήση μεθόδων κανονικοποίησης: [0.0, 28.365].

Εύρος τιμών μετά τη χρήση μεθόδων κανονικοποίησης: [0.0, 1.0].

Κανονικοποίηση τιμών

Οι τιμές των συγκεκριμένων χαρακτηριστικών κανονικοποιήθηκαν με τη χρήση του κανονικοποιητή MinMaxScaler της βιβλιοθήκης scikit-learn[4], ο οποίος κλιμακώνει τις τιμές των αρχικών χαρακτηριστικών ώστε να βρίσκονται εντός του πεδίου τιμών [0.0, 1.0].

Κατηγορία

Κειμενικά χαρακτηριστικά

2.2.4.2. Αναγνώριση συχνών όρων στο όνομα του ΣΕ

Σύντομη περιγραφή

Το σύνολο των χαρακτηριστικών αυτό αποτελείται από διανύσματα τιμών τύπου Boolean (λογικές τιμές 0 ή 1) που δηλώνουν την ύπαρξη ή όχι καθενός εκ των συχνών όρων (μετρημένων στο σύνολο των διαθέσιμων δεδομένων-ονομάτων ΣΕ) σε κάθε επιμέρους όνομα. Οι συχνοί όροι λαμβάνονται ποσοστιαία εκ μίας αρχικής, ταξινομημένης λίστας συχνότητας όρων, με τη χρήση μιας υπερπαραμέτρου που αντιστοιχεί στο επιλεγμένο ποσοστό.

Αριθμός θέσεων στο διάνυσμα χαρακτηριστικών

Το είδος των χαρακτηριστικών αυτών καλύπτει $w_t \times |N_t|$ θέσεις, όπου $|N_t|$ το πλήθος των διακριτών όρων-λέξεων που απαντώνται στο σύνολο των ονομάτων των ΣΕ του συνόλου

[4] https://scikit-learn.org/stable/

δεδομένων και w_t μια υπερπαράμετρος που αντιστοιχεί στο ποσοστό τους που θέλουμε να ληφθεί υπόψιν κατά την εξαγωγή τους, θεωρούμενο ως «συχνοί όροι».

Τύπος και εύρος τιμών

Τύπος: Boolean τιμές.

Εύρος Τιμών: {0, 1}.

Κανονικοποίηση τιμών

Δεν χρησιμοποιήθηκε κανονικοποίηση, μιας και τα χαρακτηριστικά λαμβάνουν Boolean τιμές.

Κατηγορία

Κειμενικά χαρακτηριστικά

2.2.4.3. Αναγνώριση συχνών ν-γραμμάτων όρων στο όνομα του ΣΕ

Σύντομη περιγραφή

Το σύνολο αυτών των χαρακτηριστικών αποτελείται από διανύσματα τιμών τύπου Boolean (λογικές τιμές 0 ή 1) που δηλώνουν την ύπαρξη ή όχι καθενός εκ των συχνών ν-γραμμάτων (n-grams) όρων των ονομάτων ΣΕ που βρίσκονται στο σύνολο δεδομένων σε κάθε επιμέρους όνομα. Τα συχνά ν-γράμματα όρων λαμβάνονται ποσοστιαία εκ μίας αρχικής, ταξινομημένης λίστας συχνότητας ν-γραμμάτων, με τη χρήση μιας υπερπαραμέτρου που αντιστοιχεί στο επιλεγμένο ποσοστό.

Αριθμός θέσεων στο διάνυσμα χαρακτηριστικών

Το είδος των χαρακτηριστικών αυτών καλύπτει $w_n \times |N_n|$ θέσεις, όπου $|N_n|$ το πλήθος του συνόλου των διακριτών ν-γραμμάτων όρων που βρίσκονται στα ονόματα των ΣΕ του συνόλου δεδομένων και w_n μια υπερπαράμετρος που αντιστοιχεί στο ποσοστό τους που θέλουμε να ληφθεί υπόψιν κατά την εξαγωγή τους, θεωρούμενο ως «συχνά ν-γράμματα».

Τύπος και εύρος τιμών

Τύπος: Boolean τιμές.

Εύρος Τιμών: {0, 1}.

Κανονικοποίηση τιμών

Δεν χρησιμοποιήθηκε κανονικοποίηση, μιας και τα χαρακτηριστικά λαμβάνουν Boolean τιμές.

Κατηγορία

Κειμενικά χαρακτηριστικά

2.2.4.4. Αναγνώριση συχνών ν-γραμμάτων χαρακτήρων στο όνομα του ΣΕ

Σύντομη περιγραφή

Το σύνολο των χαρακτηριστικών αυτό αποτελείται από διανύσματα τιμών τύπου Boolean (λογικές τιμές 0 ή 1) που δηλώνουν την ύπαρξη ή όχι καθενός εκ των συχνών ν-γραμμάτων χαρακτήρων των ονομάτων ΣΕ που βρίσκονται στο σύνολο δεδομένων σε κάθε επιμέρους όνομα. Τα συχνά ν-γράμματα χαρακτήρων λαμβάνονται ποσοστιαία εκ μίας αρχικής, ταξινομημένης λίστας συχνότητας ν-γραμμάτων χαρακτήρων με τη χρήση μιας υπερπαραμέτρου που αντιστοιχεί στο επιλεγμένο ποσοστό.

Αριθμός θέσεων στο διάνυσμα χαρακτηριστικών

Το είδος των χαρακτηριστικών αυτών καλύπτει $w_c \times |N_c|$ θέσεις, όπου $|N_c|$ το πλήθος του συνόλου των επιμέρους ν-γραμμάτων χαρακτήρων που βρίσκονται στα ονόματα των Σημείων Ενδιαφέροντος που βρίσκονται στο σύνολο δεδομένων και w_c μια υπερπαράμετρος που αντιστοιχεί στο ποσοστό τους που θέλουμε να ληφθεί υπόψιν κατά την εξαγωγή τους, θεωρούμενο ως «συχνά ν-γράμματα χαρακτήρων».

Τύπος και εύρος τιμών

Τύπος: Boolean τιμές.

Εύρος Τιμών: {0, 1}.

Κανονικοποίηση τιμών

Δεν χρησιμοποιήθηκε κανονικοποίηση.

Κατηγορία

Κειμενικά χαρακτηριστικά

2.2.4.5. Αναγνώριση ύπαρξης και πλήθους κατηγοριών σε Σημεία Ενδιαφέροντος εντός ακτίνας

Σύντομη περιγραφή

Το σύνολο των χαρακτηριστικών αυτό αποτελείται από $2 \times |C|$ χαρακτηριστικά, όπου $|C|$ ο αριθμός των διαφορετικών κατηγοριών στο σύνολο δεδομένων. Για κάθε ΣΕ βάσης, συγκεντρώνονται τα ΣΕ τα οποία βρίσκονται εντός δεδομένης ακτίνας R_p από αυτό και καταγράφονται οι κατηγορίες στις οποίες ανήκουν. Στη συνέχεια, δημιουργείται ένα ζεύγος χαρακτηριστικών για κάθε κατηγορία τα οποία αντιπροσωπεύουν την ύπαρξη της κατηγορίας αυτής εντός των γειτονικών ΣΕ (Boolean τιμή) και το πλήθος των γειτονικών ΣΕ που ανήκουν στην κατηγορία αυτή (αριθμητική τιμή), αντίστοιχα. Το σύνολο των ζευγών αυτών για κάθε κατηγορία αποτελεί το διάνυσμα χαρακτηριστικών για το συγκεκριμένο ΣΕ.

Αριθμός θέσεων στο διάνυσμα χαρακτηριστικών

Το είδος των χαρακτηριστικών αυτών καλύπτει $2 \times |C|$ θέσεις, όπου $|C|$ ο αριθμός των διαφορετικών κατηγοριών στο σύνολο δεδομένων.

Τύπος και εύρος τιμών

Τύπος: Boolean και αριθμητικές τιμές.

Εύρος τιμών (Boolean): {0,1}.

Εύρος τιμών πριν τη χρήση μεθόδων κανονικοποίησης (αριθμητικές): [0.0, 34.0].

Εύρος τιμών μετά τη χρήση μεθόδων κανονικοποίησης (αριθμητικές): [0.0, 1.0].

Κανονικοποίηση τιμών

Οι αριθμητικές τιμές των συγκεκριμένων χαρακτηριστικών κανονικοποιήθηκαν με τη χρήση του κανονικοποιητή MinMaxScaler της βιβλιοθήκης scikit-learn, ο οποίος κλιμακώνει τις τιμές των αρχικών χαρακτηριστικών ώστε να βρίσκονται εντός του πεδίου τιμών [0.0, 1.0].

Για τις Boolean τιμές των αντίστοιχων χαρακτηριστικών δεν χρησιμοποιήθηκε κανονικοποίηση.

Κατηγορία

Χαρακτηριστικά (γεωχωρικής) γειτνίασης.

2.2.4.6. Αναγνώριση ύπαρξης και πλήθους κατηγοριών σε Σημεία Ενδιαφέροντος εντός ακτίνας βάσει πλησιέστερης οδού

Σύντομη περιγραφή

Το σύνολο των χαρακτηριστικών αυτό αποτελείται από $2 \times |C|$ χαρακτηριστικά, όπου $|C|$ ο αριθμός των διαφορετικών κατηγοριών στο σύνολο δεδομένων. Αρχικά καθορίζεται η οδός στην οποία ανήκει το κάθε ΣΕ βάσης (η οδός με τη μικρότερη απόσταση από αυτό). Στη συνέχεια συγκεντρώνονται, από τα ΣΕ που βρίσκονται σε πάνω σε αυτόν τον δρόμο, εκείνα που βρίσκονται εντός δεδομένης ακτίνας R_{ps} από το ΣΕ βάσης και καταγράφονται οι κατηγορίες στις οποίες ανήκουν. Στη συνέχεια δημιουργείται ένα ζεύγος χαρακτηριστικών για κάθε κατηγορία τα οποία αντιπροσωπεύουν την ύπαρξη της κατηγορίας αυτής εντός των γειτονικών ΣΕ (Boolean τιμή) και το πλήθος των γειτονικών ΣΕ που ανήκουν στην κατηγορία αυτή (αριθμητική τιμή), αντίστοιχα. Το σύνολο των ζευγών αυτών για κάθε κατηγορία αποτελεί το διάνυσμα χαρακτηριστικών για το συγκεκριμένο ΣΕ.

Αριθμός θέσεων στο διάνυσμα χαρακτηριστικών

Το είδος των χαρακτηριστικών αυτών καλύπτει $2 \times |C|$ θέσεις, όπου $|C|$ ο αριθμός των διαφορετικών κατηγοριών στο σύνολο δεδομένων.

Τύπος και εύρος τιμών

Τύπος: Boolean και αριθμητικές τιμές.

Εύρος τιμών (Boolean): {0,1}.

Εύρος τιμών πριν τη χρήση μεθόδων κανονικοποίησης (αριθμητικές): [0.0, 16.0].

Εύρος τιμών μετά τη χρήση μεθόδων κανονικοποίησης (αριθμητικές): [0.0, 1.0].

Κανονικοποίηση τιμών

Οι αριθμητικές τιμές των συγκεκριμένων χαρακτηριστικών κανονικοποιήθηκαν με τη χρήση του κανονικοποιητή MinMaxScaler της βιβλιοθήκης scikit-learn, ο οποίος κλιμακώνει τις τιμές των αρχικών χαρακτηριστικών ώστε να βρίσκονται εντός του πεδίου τιμών [0.0, 1.0].

Για τις Boolean τιμές των αντίστοιχων χαρακτηριστικών δεν χρησιμοποιήθηκε κανονικοποίηση.

Κατηγορία

Χαρακτηριστικά (γεωχωρικής) γειτνίασης.

2.2.4.7. Αναγνώριση ύπαρξης και πλήθους κατηγοριών σε γειτονικά Σημεία Ενδιαφέροντος

Σύντομη περιγραφή

Το σύνολο των χαρακτηριστικών αυτό αποτελείται από $2 \times |C|$ χαρακτηριστικά, όπου $|C|$ ο αριθμός των διαφορετικών κατηγοριών στο σύνολο δεδομένων. Για κάθε ΣΕ βάσης, συγκεντρώνονται τα k γειτονικά σε αυτό ΣΕ (τα k κοντινότερα ΣΕ στο αρχικό). Στη συνέχεια, δημιουργείται ένα ζεύγος χαρακτηριστικών για κάθε κατηγορία τα οποία αντιπροσωπεύουν την ύπαρξη της κατηγορίας αυτής εντός των γειτονικών ΣΕ (Boolean τιμή) και το πλήθος των γειτονικών ΣΕ που ανήκουν στην κατηγορία αυτή (αριθμητική τιμή), αντίστοιχα. Το σύνολο των ζευγών αυτών για κάθε κατηγορία αποτελεί το διάνυσμα χαρακτηριστικών για το συγκεκριμένο ΣΕ βάσης.

Αριθμός θέσεων στο διάνυσμα χαρακτηριστικών

Το είδος των χαρακτηριστικών αυτών καλύπτει $2 \times |C|$ θέσεις, όπου $|C|$ ο αριθμός των διαφορετικών κατηγοριών στο σύνολο δεδομένων.

Τύπος και εύρος τιμών

Τύπος: Boolean και αριθμητικές τιμές.

Εύρος τιμών (Boolean): {0,1}.

Εύρος τιμών πριν τη χρήση μεθόδων κανονικοποίησης (αριθμητικές): [0.0, 4.0].

Εύρος τιμών μετά τη χρήση μεθόδων κανονικοποίησης (αριθμητικές): [0.0, 1.0].

Κανονικοποίηση τιμών

Οι αριθμητικές τιμές των συγκεκριμένων χαρακτηριστικών κανονικοποιήθηκαν με τη χρήση του κανονικοποιητή MinMaxScaler της βιβλιοθήκης scikit-learn, ο οποίος κλιμακώνει τις τιμές των αρχικών χαρακτηριστικών ώστε να βρίσκονται εντός του πεδίου τιμών [0.0, 1.0].

Για τις Boolean τιμές των αντίστοιχων χαρακτηριστικών δεν χρησιμοποιήθηκε κανονικοποίηση.

Κατηγορία

Χαρακτηριστικά (γεωχωρικής) γειτνίασης.

2.2.4.8. Αναγνώριση ύπαρξης και πλήθους κατηγοριών σε Σημεία Ενδιαφέροντος εντός ακτίνας από πλησιέστερη οδό

Σύντομη περιγραφή

Το σύνολο των χαρακτηριστικών αυτό αποτελείται από $2 \times |C|$ χαρακτηριστικά, όπου $|C|$ ο αριθμός των διαφορετικών κατηγοριών στο σύνολο δεδομένων. Αρχικά καθορίζεται η οδός στην οποία ανήκει το κάθε ΣΕ βάσης (η οδός με τη μικρότερη απόσταση από αυτό). Στη συνέχεια, συγκεντρώνονται τα ΣΕ τα οποία βρίσκονται εντός δεδομένης ακτίνας R_s από την οδό αυτή και καταγράφονται οι κατηγορίες στις οποίες ανήκουν. Στη συνέχεια δημιουργείται ένα ζεύγος χαρακτηριστικών για κάθε κατηγορία τα οποία αντιπροσωπεύουν την ύπαρξη της κατηγορίας αυτής εντός των ευρεθέντων ΣΕ (Boolean τιμή) και το πλήθος των ευρεθέντων ΣΕ που ανήκουν στην κατηγορία αυτή (αριθμητική τιμή), αντίστοιχα. Το σύνολο των ζευγών αυτών για κάθε κατηγορία αποτελεί το διάνυσμα χαρακτηριστικών για το συγκεκριμένο ΣΕ βάσης.

Αριθμός θέσεων στο διάνυσμα χαρακτηριστικών

Το είδος των χαρακτηριστικών αυτών καλύπτει $2 \times |C|$ θέσεις, όπου $|C|$ ο αριθμός των διαφορετικών κατηγοριών στο σύνολο δεδομένων.

Τύπος και εύρος τιμών

Τύπος: Boolean και αριθμητικές τιμές.

Εύρος τιμών (Boolean): {0,1}.

Εύρος τιμών πριν τη χρήση μεθόδων κανονικοποίησης (αριθμητικές): [0.0, 42.0].

Εύρος τιμών μετά τη χρήση μεθόδων κανονικοποίησης (αριθμητικές): [0.0, 1.0].

Κανονικοποίηση τιμών

Οι αριθμητικές τιμές των συγκεκριμένων χαρακτηριστικών κανονικοποιήθηκαν με τη χρήση του κανονικοποιητή MinMaxScaler της βιβλιοθήκης scikit-learn, ο οποίος κλιμακώνει τις τιμές των αρχικών χαρακτηριστικών ώστε να βρίσκονται εντός του πεδίου τιμών [0.0, 1.0].

Για τις Boolean τιμές των αντίστοιχων χαρακτηριστικών δεν χρησιμοποιήθηκε κανονικοποίηση.

Κατηγορία

Χαρακτηριστικά (γεωχωρικής) γειτνίασης.

2.3. Γεωκωδικοποίηση Διευθύνσεων

2.3.1. Περιγραφή σεναρίου χρήσης

Στη συνέχεια, περιγράφουμε εν συντομία το συγκεκριμένο σενάριο. Πιο αναλυτική περιγραφή αυτού του σεναρίου υπάρχει στο Π1.1 (Κεφάλαιο 2.3.1).

Παρεμφερές πρόβλημα της κατηγοριοποίησης ΣΕ αποτελεί η διαδικασία της γεωκωδικοποίησης διευθύνσεων. Με τον όρο γεωκωδικοποίηση (geocoding) διευθύνσεων, εννοούμε την αντιστοίχιση κάθε διεύθυνσης με ένα σημείο συντεταγμένων (x,y). Μια διεύθυνση, στην αρχική της μορφή, περιγράφεται από όλα ή μερικά από τα παρακάτω πεδία πληροφορίας:

- Διοικητική διαίρεση: Νομός, Δήμος, Δημοτικό διαμέρισμα (οικισμός)
- Τοπωνύμιο Περιοχής (location)
- Ταχυδρομικός Κωδικός (ΤΚ), Οδός, Αριθμός (από – έως)

Οι εταιρείες Geodata και Ερατοσθένης έχουν αναπτύξει εργαλεία λογισμικού για την αυτόματη/ημιαυτόματη γεωκωδικοποίηση μεγάλων αρχείων διευθύνσεων, που έχουν να κάνουν κυρίως με μεγάλα πελατολόγια (ΣΕ), πελατών τους. Τα συγκεκριμένα εργαλεία, παρόλο που ενσωματώνουν καλές πρακτικές και ευρέως αναγνωρισμένες μεθοδολογίες γεωκωδικοποίησης, επιδέχονται βελτίωση (ή τουλάχιστον μεγαλύτερο ποιοτικό έλεγχο) ως προς την ακρίβεια των παραγόμενων συντεταγμένων, λόγω των παρακάτω παραγόντων:

- Πιθανόν να υπάρχουν σφάλματα στη master database και στα γεωγραφικά επίπεδα αναφοράς,
- Πιθανόν να προκύπτουν σφάλματα καθαρισμού δεδομένων
- Πιθανόν να προκύπτουν σφάλματα εξαιτίας του χρησιμοποιούμενου αλγορίθμου γεωκωδικοποίησης, που βασίζεται στην μέθοδο της «απλής παρεμβολής» με βάση την απόσταση για γεωκωδικοποίηση επί των γραμμών οδικών δικτύου και στη μέθοδο της εύρεσης του κεντροβαρούς για γεωκωδικοποίηση επί πολυγώνων τοπωνυμιών, διοικητικών υποδιαιρέσεων και ΤΚ (όταν υπάρχει έλλειψη οδού και αριθμού στην δοθείσα διεύθυνση).

Προκειμένου να αντιμετωπισθούν τα παραπάνω προβλήματα, είναι θεμιτή η σύγκριση των αποτελεσμάτων των παραπάνω εφαρμογών γεωκωδικοποίησης με τρίτες, εξωτερικές πηγές, προκειμένου να επιλέγονται, κάθε φορά, οι βέλτιστες συντεταγμένες για κάθε διεύθυνση εισόδου. Με βάση εμπειρική γνώση των εταίρων, αλλά και αρχικό πειραματισμό με διάφορες πηγές γεωκωδικοποίησης, μέσω της ανοικτής βιβλιοθήκης γεωκωδικοποίησης geopy[5], καταλήξαμε σε δύο αρκετά αξιόπιστες, ανοικτές πηγές γεωκωδικοποίησης διευθύνσεων, τις ArcGIS[6] και OpenStreetMap[7].

[5] https://pypi.org/project/geopy/

[6] https://geocode.arcgis.com/arcgis/

[7] https://wiki.openstreetmap.org/wiki/Nominatim

2.3.2. Ορισμός προβλήματος μηχανικής μάθησης

Το παραπάνω πρόβλημα μπορεί να οριστεί ως ένα πρόβλημα κατάταξης με πολλαπλές κλάσεις (multiclass classification). Το κάθε στιγμιότυπο του προβλήματος αποτελείται από κάθε διεύθυνση προς γεωκωδικοποίηση, μαζί με το σύνολο όλων των υποψήφιων πηγών/αποτελεσμάτων γεωκωδικοποίησης για την αντίστοιχη διεύθυνση, συμπεριλαμβανομένου και του υφιστάμενου εργαλείου των Ερατοσθένης και Geodata. Οι πιθανές κλάσεις που δύνανται να ανατεθούν από τον αλγόριθμο κατάταξης είναι το σύνολο των διαφορετικών συντεταγμένων που προκύπτουν από τις αντίστοιχες πηγές γεωκωδικοποίησης. Στη συγκεκριμένη φάση, οι θεωρούμενες πηγές είναι τρεις: εσωτερική βάση Ερατοσθένης-Geodata, ArcGIS και OpenStreetMap.

2.3.3. Γνώση πεδίου

Βασική πηγή γνώσης για την κειμενική κωδικοποίηση μιας διεύθυνσης παρέχεται από τους κανόνες αναγραφής διευθύνσεων που ακολουθούν τα Ελληνικά Ταχυδρομεία (ΕΛΤΑ) και που ατύπως ακολουθούνται από όλους όσους καταγράφουν και διαχειρίζονται αρχεία διευθύνσεων. Με βάση τους κανόνες αυτούς, οι διευθύνσεις θα πρέπει να περιέχουν συμπληρωμένα μερικά βασικά περιγραφικά πεδία:

- Οδός (ή οδοί για διασταυρώσεις),
- Αριθμός (ή χιλιομέτρηση),
- Ταχυδρομικός Κώδικας (ΤΚ),
- Λεκτικό - Τοπωνύμιο (μπορεί να αφορά Πόλη, Χωριό, Δήμο, Τοπωνύμιο, κ.λπ.)
- Χώρα (για διευθύνσεις εντός Ελλάδος η χώρα παραλείπεται).

Μια κειμενική διεύθυνση προς γεωκωδικοποίηση (input), μπορεί να περιέχει συμπληρωμένα μερικά (ή όλα) από τα παραπάνω πεδία, είτε ως διακριτά πεδία (delimited text, πίνακες), είτε σε ένα ενιαίο κειμενικό (string) πεδίο. Σε πραγματικές δηλαδή συνθήκες, η αναγραφή μιας διεύθυνσης δεν είναι αυστηρά κωδικοποιημένη, ούτε ως προς τον αριθμό των χρησιμοποιούμενων πεδίων, ούτε ως προς το περιεχόμενό τους, παρουσιάζοντας διάφορες εναλλακτικές αναγραφής και κωδικοποίησης.

Ακολουθούν διάφορα παραδείγματα εναλλακτικών τρόπων αναγραφής Διευθύνσεων:

- Ιπποκράτους 141, 11472 Αθήνα
- Ιπποκράτους 141, 11472 Αθήνα, Αττική
- Ιπποκράτους & Αριανίτου, 11472 Αθήνα
- Αριανίτου & Ιπποκράτους, 11472 Αθήνα
- Ιπποκράτους και Αριανίτου, 11472 Αθήνα
- Ιπποκράτους 141 & Αριανίτου, 11472 Αθήνα
- 15ο χιλιόμετρο Εθνικής οδού Αθηνών – Θεσσαλονίκης, 12835 Καπανδρίτι
- 15ο χιλιόμετρο Εθνικής οδού Αθηνών – Λαμίας, 12835 Καπανδρίτι
- 15 χιλιόμετρο Εθνικής οδού Αθηνών – Λαμίας, 12835 Καπανδρίτι

Επίσης η ίδια διεύθυνση μπορεί να αναγράφεται εναλλακτικά με συντμήσεις όχι:

- Πλατεία Καρύτση 12 ή Πλ. Καρύτση 12
- Στρ. Παπάγου 13. Γουδή ή Στρατηγού Παπάγου 13, Γουδή

Επίσης πολλοί οδοί έχουν εναλλακτικά ονόματα:

- Πατησίων ή 28ης Οκτωβρίου

Θα πρέπει να αναφέρουμε τέλος ότι, των πιθανών ελλείψεων πεδίων και των διαφορετικών τρόπων αναγραφής μιας διεύθυνσης σε πραγματικές συνθήκες, παρουσιάζονται και ορθογραφικά λάθη που δυσκολεύουν ακόμη περισσότερο την διαδικασία διασύνδεσης με μια οποιαδήποτε master βάση δεδομένων διευθύνσεων.

Στο σενάριο που εξετάζουμε, ένας ρεαλιστικός τρόπος για να αντιμετωπιστούν, σε κάποιο βαθμό, τα παραπάνω προβλήματα, είναι η διασύνδεση με τρίτες πηγές γεωκωδικοποίησης, οι οποίες ενδέχεται να ακολουθούν διαφορετικές μεθοδολογίες γεωκωδικοποίησης οι οποίες να οδηγούν σε διαφορετικά προτερήματα και αδυναμίες. Δεδομένης της ύπαρξης διαφορετικών, ετερογενών πηγών, ένα μοντέλο μηχανικής μάθησης θα μπορεί να εκπαιδεύεται σε ένα αρχικό σύνολο δεδομένων εκπαίδευσης, με στόχο να μαθαίνει ποια πηγή γεωκωδικοποίησης είναι η προτιμότερη σε κάθε περίπτωση. Ουσιαστικά η παραπάνω μεθοδολογία προσομοιάζει τη χρήση ενός ευρέως χρησιμοποιούμενου και ιδιαίτερα αποτελεσματικού σχήματος, των ensembles, δηλαδή την εφαρμογή πολλών επιμέρους ετερογενών μοντέλων, τα οποία «ψηφίζουν» για την τελική απόφαση του αλγορίθμου μηχανικής μάθησης.

2.3.4. Χαρακτηριστικά εκπαίδευσης

2.3.4.1. Κανονικοποιημένες συντεταγμένες γεωκωδικοποίησης

Σύντομη περιγραφή

Το σύνολο αυτών των χαρακτηριστικών αποτελείται από τα κανονικοποιημένα ζεύγη συντεταγμένων που έχουν συγκεντρωθεί από κάθε πηγή γεωκωδικοποίησης για μία συγκεκριμένη διεύθυνση. Η κανονικοποίηση προκύπτει λαμβάνοντας το πηλίκο της κάθε επιμέρους συντεταγμένης, αφού πρώτα αφαιρεθεί από αυτήν η μέση τιμή των αντίστοιχων συντεταγμένων, με τη διακύμανση των αντίστοιχων συντεταγμένων.

Αριθμός θέσεων στο διάνυσμα χαρακτηριστικών

Από τη διαδικασία αυτή εξάγονται $2 \times |G|$ χαρακτηριστικά όπου $|G|$ το πλήθος των διαθέσιμων πηγών γεωκωδικοποίησης.

Τύπος και εύρος τιμών

Τύπος: αριθμητικές τιμές.

Εύρος τιμών πριν τη χρήση μεθόδων κανονικοποίησης: X:[19.92, 27.61] και Y:[35.01, 41.16].

Εύρος τιμών μετά τη χρήση μεθόδων κανονικοποίησης: X:[0.0, 1.0] και Y:[0.0, 1.0].

Κανονικοποίηση τιμών

Οι τιμές των συγκεκριμένων χαρακτηριστικών κανονικοποιήθηκαν με τη χρήση του κανονικοποιητή MinMaxScaler της βιβλιοθήκης scikit-learn, ο οποίος κλιμακώνει τις τιμές των αρχικών χαρακτηριστικών ώστε να βρίσκονται εντός του πεδίου τιμών [0.0, 1.0].

Κατηγορία

Γεωχωρικά χαρακτηριστικά

2.3.4.2. Κατά ζεύγη αποστάσεις σημείων

Σύντομη περιγραφή

Το σύνολο χαρακτηριστικών αυτό προκύπτει λαμβάνοντας τις κατά ζεύγη αποστάσεις των σημείων που αντιστοιχούν στα αποτελέσματα καθεμίας από τις πηγές γεωκωδικοποίησης.

Αριθμός θέσεων στο διάνυσμα χαρακτηριστικών

Από τη διαδικασία αυτή εξάγονται (|*G*|*(|*G*|-1))/2 χαρακτηριστικά όπου |*G*| το πλήθος των διαθέσιμων πηγών γεωκωδικοποίησης.

Τύπος και εύρος τιμών

Τύπος: Αριθμητικές τιμές.

Εύρος τιμών πριν τη χρήση μεθόδων κανονικοποίησης: [6.38 × 10^{-8}, 12.857].

Εύρος τιμών μετά τη χρήση μεθόδων κανονικοποίησης: [0.0, 1.0].

Κανονικοποίηση τιμών

Οι τιμές των συγκεκριμένων χαρακτηριστικών κανονικοποιήθηκαν με τη χρήση του κανονικοποιητή MinMaxScaler της βιβλιοθήκης scikit-learn, ο οποίος κλιμακώνει τις τιμές των αρχικών χαρακτηριστικών ώστε να βρίσκονται εντός του πεδίου τιμών [0.0, 1.0].

Κατηγορία

Γεωχωρικά χαρακτηριστικά

2.3.4.3. Κατά ζεύγη αποστάσεις μεμονωμένων συντεταγμένων

Σύντομη περιγραφή

Το σύνολο χαρακτηριστικών αυτό προκύπτει λαμβάνοντας τις κατά ζεύγη αποστάσεις των συντεταγμένων (τετμημένων-τετμημένων και τεταγμένων-τεταγμένων) σημείων που αντιστοιχούν στα αποτελέσματα καθεμίας από τις πηγές γεωκωδικοποίησης.

Αριθμός θέσεων στο διάνυσμα χαρακτηριστικών

Από τη διαδικασία αυτή εξάγονται |*G*|*(|*G*|-1) χαρακτηριστικά όπου |*G*| το πλήθος των διαθέσιμων πηγών γεωκωδικοποίησης.

Τύπος και εύρος τιμών

Τύπος: Αριθμητικές τιμές.

Εύρος τιμών πριν τη χρήση μεθόδων κανονικοποίησης: [1.549 × 10^{-7}, 8.544].

Εύρος τιμών μετά τη χρήση μεθόδων κανονικοποίησης: [0.0, 1.0].

Κανονικοποίηση τιμών

Οι τιμές των συγκεκριμένων χαρακτηριστικών κανονικοποιήθηκαν με τη χρήση του κανονικοποιητή MinMaxScaler της βιβλιοθήκης scikit-learn, ο οποίος κλιμακώνει τις τιμές των αρχικών χαρακτηριστικών ώστε να βρίσκονται εντός του πεδίου τιμών [0.0, 1.0].

Κατηγορία

Γεωχωρικά χαρακτηριστικά

2.3.4.4. Αποστάσεις μεμονωμένων συντεταγμένων από κεντροειδές

Σύντομη περιγραφή

Το σύνολο αυτών των χαρακτηριστικών προκύπτει λαμβάνοντας τις αποστάσεις των συντεταγμένων (τετμημένων-τετμημένων και τεταγμένων-τεταγμένων) των σημείων που αντιστοιχούν στα αποτελέσματα καθεμίας από τις πηγές γεωκωδικοποίησης από τις αντίστοιχες συντεταγμένες του κεντροειδούς τους σημείου. Το κεντροειδές σημείο ορίζεται ως το σημείο αυτό του οποίου οι συντεταγμένες προκύπτουν λαμβάνοντας το μέσο όρο των επί μέρους σημείων που έχουν προκύψει από τις πηγές γεωκωδικοποίησης.

Αριθμός θέσεων στο διάνυσμα χαρακτηριστικών

Από τη διαδικασία αυτή εξάγονται $2 \times |G|$ χαρακτηριστικά όπου $|G|$ το πλήθος των διαθέσιμων πηγών γεωκωδικοποίησης.

Τύπος και εύρος τιμών

Τύπος: Αριθμητικές τιμές.

Εύρος τιμών πριν τη χρήση μεθόδων κανονικοποίησης: [0.0, 8.929].

Εύρος τιμών μετά τη χρήση μεθόδων κανονικοποίησης: [0.0, 1.0].

Κανονικοποίηση τιμών

Οι τιμές των συγκεκριμένων χαρακτηριστικών κανονικοποιήθηκαν με τη χρήση του κανονικοποιητή MinMaxScaler της βιβλιοθήκης scikit-learn, ο οποίος κλιμακώνει τις τιμές των αρχικών χαρακτηριστικών ώστε να βρίσκονται εντός του πεδίου τιμών [0.0, 1.0].

Κατηγορία

Γεωχωρικά χαρακτηριστικά

2.3.4.5. Μέσες αποστάσεις μεμονωμένων συντεταγμένων από κεντροειδές

Σύντομη περιγραφή

Το σύνολο χαρακτηριστικών αυτό προκύπτει λαμβάνοντας τις μέσες αποστάσεις των σημείων που αντιστοιχούν στα αποτελέσματα καθεμίας από τις πηγές γεωκωδικοποίησης από τις αντίστοιχες συντεταγμένες του κεντροειδούς τους σημείου. Το κεντροειδές σημείο ορίζεται ως το σημείο αυτό του οποίου οι συντεταγμένες προκύπτουν λαμβάνοντας το μέσο όρο των επί μέρους σημείων που έχουν προκύψει από τις πηγές γεωκωδικοποίησης.

Αριθμός θέσεων στο διάνυσμα χαρακτηριστικών

Από τη διαδικασία αυτή εξάγονται 2 χαρακτηριστικά.

Τύπος και εύρος τιμών

Τύπος: Αριθμητικές τιμές.

Εύρος τιμών πριν τη χρήση μεθόδων κανονικοποίησης: $[2.7\times 10^{-3}, 3.58]$.

Εύρος τιμών μετά τη χρήση μεθόδων κανονικοποίησης: [0.0, 1.0].

Κανονικοποίηση τιμών

Οι τιμές των συγκεκριμένων χαρακτηριστικών κανονικοποιήθηκαν με τη χρήση του κανονικοποιητή MinMaxScaler της βιβλιοθήκης scikit-learn, ο οποίος κλιμακώνει τις τιμές των αρχικών χαρακτηριστικών ώστε να βρίσκονται εντός του πεδίου τιμών [0.0, 1.0].

Κατηγορία

Γεωχωρικά χαρακτηριστικά

2.3.4.6. Αποστάσεις από πλησιέστερη οδό

Σύντομη περιγραφή

Το σύνολο χαρακτηριστικών αυτό προκύπτει λαμβάνοντας τις αποστάσεις των επιμέρους σημείων που έχουν προκύψει από τις πηγές γεωκωδικοποίησης από την πλησιέστερη οδό σε κάθε μία από αυτές.

Αριθμός θέσεων στο διάνυσμα χαρακτηριστικών

Από τη διαδικασία αυτή εξάγονται $|G|$ χαρακτηριστικά όπου $|G|$ το πλήθος των διαθέσιμων πηγών γεωκωδικοποίησης.

Τύπος και εύρος τιμών

Τύπος: Αριθμητικές τιμές.

Εύρος τιμών πριν τη χρήση μεθόδων κανονικοποίησης: $[5.63 \times 10^{-4}, 10.165]$.

Εύρος τιμών μετά τη χρήση μεθόδων κανονικοποίησης: [0.0, 1.0].

Κανονικοποίηση τιμών

Οι τιμές των συγκεκριμένων χαρακτηριστικών κανονικοποιήθηκαν με τη χρήση του κανονικοποιητή MinMaxScaler της βιβλιοθήκης scikit-learn, ο οποίος κλιμακώνει τις τιμές των αρχικών χαρακτηριστικών ώστε να βρίσκονται εντός του πεδίου τιμών [0.0, 1.0].

Κατηγορία

Γεωχωρικά χαρακτηριστικά

2.4. Ολοκλήρωση Γεωτεμαχίων/Δρόμων

2.4.1. Περιγραφή σεναρίου χρήσης

Στη συνέχεια, περιγράφουμε εν συντομία το συγκεκριμένο σενάριο. Πιο αναλυτική περιγραφή αυτού του σεναρίου υπάρχει στο Π1.1 (Κεφάλαιο 2.4.1).

Στις μελέτες κτηματογράφησης, βασική οντότητα των παραδοτέων είναι τα γεωτεμάχια. Με τον όρο γεωτεμάχια εννοούμε το σύνολο των πολύγωνων που απαρτίζουν μια γεωγραφική περιοχή. Τα πολύγωνα αυτά, για να είναι αποδεκτά σε μια μελέτη κτηματογράφησης, πρέπει να πληρούν ορισμένα βασικά χαρακτηριστικά. Συγκεκριμένα, τοπολογικά θα πρέπει να καλύπτουν πλήρως την υπό κτηματογράφηση περιοχή χωρίς κενά (holes) και επικαλύψεις (overlaps) μεταξύ τους. Επίσης τα πολύγωνα αυτά θα πρέπει να είναι απλά και όχι νησίδες (multi polygon). Όσον αφορά τα περιγραφικά τους χαρακτηριστικά (attributes), τα βασικά είναι: το είδος γεωτεμαχίου, η προέλευση, το εμβαδόν και η διεύθυνση.

Το PST είναι το παραδοτέο αρχείο της γεωμετρίας των γεωτεμαχίων (shape file) μιας μελέτης κτηματογράφησης. Το PST, κατά τη διάρκεια μιας μελέτης κτηματογράφησης, παράγεται σε διάφορες εκδόσεις (παραδόσεις):

- Το προκαταρκτικό PST, που αποτελεί την αρχική διαμόρφωση από τις διάφορες πηγές δεδομένων. Επί του προκαταρκτικού PST γίνονται δηλώσεις ιδιοκτησίας από τους πολίτες.
- Ενδιάμεσες εκδόσεις PST (στο 35%, στο 70% και στο 100% των δηλώσεων) μετά την εφαρμογή των δηλώσεων.
- Τελικό PST μετά την ανάρτηση (δημοσιότητα) και την εξέταση των ενστάσεων από τους πολίτες.

Επιπλέον βασικό πρόβλημα στη διαδικασία της κτηματογράφησης μιας αγροτικής περιοχής (όχι δηλαδή αστικής), που μπορεί όμως να βρει εφαρμογή και σε άλλες επιχειρησιακές δραστηριότητες, είναι η αναγνώριση και η οριοθέτηση των οδών (δρόμων) σε πολυγωνική μορφή, πάνω σε ένα γεωγραφικό σύστημα γεωτεμαχίων (parcels).

Από τα παραπάνω, μπορούμε να θεωρήσουμε τρία υπο-προβλήματα προς αντιμετώπιση:

Στο πρώτο υπο-πρόβλημα, το πρόβλημα έχει να κάνει με την επιλογή της επικρατέστερης προέλευσης (ονομαζόμενης ORI_TYPE) για κάθε οντότητα (γεωτεμάχιο) του τελικού PST, σε σχέση με τις αρχικές πηγές δεδομένων. Δηλαδή, μετά την επεξεργασία στα διάφορα στάδια της κτηματογράφησης, η εύρεση της επικρατέστερης αρχικής πηγής δεδομένων, βάση της οποίας παρήχθη η τελική γεωμετρία μιας οντότητας (γεωτεμαχίου) του τελικού PST.

Στο δεύτερο υπο-πρόβλημα, αυτό που εξετάζουμε είναι η επιλογή της καταλληλότερης γεωμετρίας από τις προαναφερόμενες πηγές δεδομένων, προκειμένου να παραχθεί το προκαταρκτικό PST, επί του οποίου θα γίνουν οι δηλώσεις ιδιοκτησίας. Είναι πολύ σημαντικό να έχουμε ένα προκαταρκτικό PST που να προσομοιάζει όσο το δυνατόν στην πραγματικότητα, γιατί σε διαφορετική περίπτωση δεν θα μπορούν να αναγνωριστούν τα γεωτεμάχια από τους ενδιαφερόμενους πολίτες κατά την υποβολή της δήλωσής τους.

Στο τρίτο υπο-πρόβλημα, η ανάγκη που προκύπτει είναι η ανάπτυξη ενός εργαλείου βασισμένου σε τεχνολογίες μηχανικής μάθησης για την υποβοήθηση της διαδικασίας αναγνώρισης των ορίων των οδών, καθώς και νέων οδών, μέσω της αυτόματης διασύνδεσης των διαφορετικών πηγών δεδομένων (αναδασμοί, διανομές και LPIS). Πιο συγκεκριμένα, το εργαλείο αυτό θα πρέπει να συσχετίζει τις διανομές – αναδασμούς με το LPIS και, με βάση κάποιες παραμέτρους διασύνδεσης, να επιλέγει αυτόματα ή ημιαυτόματα το πολύγωνα των δρόμων, που εμφανίζονται να είναι διαφορετικά στη μία πηγή δεδομένων από την άλλη, έτσι ώστε να επισημειώνονται ως νέα κομμάτια δρόμων, και να εμπλουτίζουν τα υπάρχοντα σύνολα δεδομένων δρόμων. Κατά τον τρόπο αυτό θα αυξάνεται η ποιότητα της διαδικασίας αποτύπωσης των τελικών δρόμων.

Στη συγκεκριμένη φάση, και για την αρχική ανάπτυξη των μοντέλων μας[8], επικεντρώνουμε στο πρώτο υπο-πρόβλημα, δηλαδή την επιλογή της επικρατέστερης προέλευσης (ORI_TYPE) για κάθε οντότητα (γεωτεμάχιο) του τελικού PST. Τα χαρακτηριστικά που υλοποιούνται για το συγκεκριμένο υπο-πρόβλημα έχουν, στην πλειονότητά τους, άμεση εφαρμογή και στα υπόλοιπα υπο-προβλήματα, μιας και, όπως αναλύεται στη συνέχεια, αφορούν σε μεγάλο βαθμό γεωχωρικές/γεωμετρικές ιδιότητες των προς εξέταση πολυγωνικών γεωμετριών (γεωτεμαχίων και δρόμων).

2.4.2. Ορισμός προβλήματος μηχανικής μάθησης

Το παραπάνω πρόβλημα μπορεί να οριστεί ως ένα πρόβλημα κατάταξης με πολλαπλές κλάσεις (multiclass classification). Τα στιγμιότυπα του προβλήματος είναι μεμονωμένα γεωτεμάχια του PST. Οι πιθανές κλάσεις που δύνανται να τους ανατεθούν από τον αλγόριθμο κατάταξης είναι το σύνολο των διαφορετικών προελεύσεων που μπορεί να έχει μία οντότητα (γεωτεμάχιο) του τελικού PST. Συγκεκριμένα, οι τιμές που παίρνει το πεδίο της επικρατέστερης προέλευσης (ORI_TYPE) σε μια μελέτη κτηματογράφησης είναι:

- Κτηματογράφηση (δηλαδή, δεν ελήφθη υπόψη καμιά αρχική πηγή δεδομένων από τις επόμενες)
- Πράξη εφαρμογής
- Αναδασμός
- Διανομή
- Πράξη Καθορισμού αιγιαλού
- Απαλλοτρίωση

Σημειώνουμε ότι, με βάση τα αρχικά δεδομένα αξιολόγησης, η παραπάνω λίστα εκφυλίζεται σε δύο διαφορετικές κλάσεις-προελεύσεις: *Διανομή* και *Κτηματογράφηση*. Δεδομένου αυτού, το πρόβλημα μετατρέπεται σε πρόβλημα δυαδικής κατάταξης. Παρόλα αυτά, τα χαρακτηριστικά που ορίζουμε είναι στην ουσία τους ανεξάρτητα από τον αριθμό των κλάσεων και μπορούν να χρησιμοποιηθούν σε όλες τις εκδοχές του προβλήματος μηχανικής μάθησης.

[8] Καθώς υλοποιούμε και αξιολογούμε μεθόδους μηχανικής μάθησης για περισσότερες παραλλαγές των θεωρημένων σεναρίων χρήσης, το σύνολο των χαρακτηριστικών εκπαίδευσης που ορίζουμε και υλοποιούμε θα επεκτείνεται διαρκώς. Οι συγκεκριμένες επεκτάσεις θα καταγραφούν και θα περιγραφούν λεπτομερώς σε ακόλουθα παραδοτέα των Ενοτήτων Εργασίας 2 και 3.

2.4.3. Γνώση πεδίου

Τα γεωτεμάχια είναι χώρο-κειμενικές οντότητες που περιγράφονται από τα παρακάτω βασικά χαρακτηριστικά:

- Σχήμα (πολυγωνική γεωμετρία)
- Εμβαδό
- Περίμετρο
- Πιθανόν διάφορα χαρακτηριστικά πεδία κειμένου (όνομα, ταυτότητα, ταχυδρομική διεύθυνση, ιδιοκτήτης, κ.λπ.)

Οι δρόμοι είναι μια κατηγορία γεωτεμαχίων με κάποια ιδιαίτερα χαρακτηριστικά:

- Μακρόστενο Σχήμα
- Χωρική απεικόνιση και με γραμμή
- Όνομα γενικής αποδοχής ανεξαρτήτου πηγής προέλευσης
- Ιδιοκτήτης (συνήθως) το Δημόσιο

Στα πλαίσια το Ελληνικού Κτηματολογίου, που αποτελεί και το πεδίο έρευνας του έργου, ως γεωτεμάχιο ορίζεται η συνεχόμενη έκταση γης, που ανήκει εξ αδιαιρέτου κατά κυριότητα σε έναν ή περισσότερους δικαιούχους. Το γεωτεμάχιο αποτελεί τη μοναδιαία επιφάνεια αναφοράς όλων των πληροφοριών του Κτηματολογίου. Κάθε γεωτεμάχιο απεικονίζεται στα κτηματολογικά διαγράμματα και χαρακτηρίζεται με τον μοναδικό Κωδικό Αριθμό Εθνικού Κτηματολογίου (ΚΑΕΚ), ως τμήμα εδάφους. Τα γεωτεμάχια – δρόμοι χαρακτηρίζονται, στο Εθνικό Κτηματολόγιο, από ένα ειδικό συνθετικό (ΕΚ) στο ΚΑΕΚ.

Κατά την διαδικασία της Κτηματογράφησης το πρώτο πράγμα που κάνει ένας μελετητής είναι να κατασκευάσει το προκαταρκτικό υπόβαθρο, επί του οποίου οι ενδιαφερόμενοι δικαιούχοι υποβάλουν τις δηλώσεις ιδιοκτησίας. Βασικά συστατικά στοιχεία του κτηματολογικού υποβάθρου είναι:

- Το όριο της υπό κτηματογράφησης περιοχής
- Τα όρια της αστικής / αγροτικής περιοχής
- Τα όρια του εντός σχεδίου περιοχής και τα όρια οικισμών
- Τα όρια οριογραμμής αιγιαλού
- Τα όρια δασικής έκτασης
- Οι άξονες δρόμων με αντίστοιχα εύρη διευθύνσεων
- Τα τοπωνύμια ή άλλα χαρακτηριστικά σημεία της περιοχής
- Τα όρια των γεωτεμαχίων με του προσωρινούς ΚΑΕΚ

Το πρόβλημα της διασύνδεσης διαφορετικών πηγών πρωτογενούς πληροφορίας (Αναδασμοί, διανομές, πράξεις εφαρμογής, απαλλοτριώσεις, κ.λ.π) για την δημιουργία ενός προκαταρκτικού κτηματολογικού υποβάθρου, που θα εμπεριέχει και τα πολύγωνα των δρόμων, παρουσιάζει τις παρακάτω δυσκολίες:

- Σφάλματα ακρίβειας θέσης, σχήματος στις αρχικές πηγές δεδομένων
- Σφάλματα στοιχείων ιδιοκτήτη και άλλων κειμενικών χαρακτηριστικών

Ένα πρώτο βήμα στην αντιμετώπιση των παραπάνω είναι σύγκριση των επιμέρους γεωτεμαχίων προέλευσης ενός τελικού γεωτεμαχίου (PST), μέσω αλγορίθμων μηχανικής μάθησης και κατάλληλων χαρακτηριστικών εκπαίδευσης πάνω στις γεωχωρικές τους ιδιότητες, προκειμένου να επισημειώνεται, καταρχάς, το τελικό γεωτεμάχιο, με την ορθότερη προέλευση.

2.4.4. Χαρακτηριστικά εκπαίδευσης

2.4.4.1. Εμβαδόν πολυγώνου

Σύντομη περιγραφή

Το σύνολο χαρακτηριστικών αυτό αποτελείται από ένα χαρακτηριστικό ανά πολύγωνο προέλευσης το οποίο αντιστοιχεί στο εμβαδόν του.

Αριθμός θέσεων στο διάνυσμα χαρακτηριστικών

Από τη διαδικασία αυτή εξάγονται |N| χαρακτηριστικά, όπου |N| ο αριθμός των πολυγώνων προέλευσης.

Τύπος και εύρος τιμών

Τύπος: Αριθμητικές τιμές.

Εύρος τιμών πριν τη χρήση μεθόδων κανονικοποίησης: [32.645, 421947.608].

Εύρος τιμών μετά τη χρήση μεθόδων κανονικοποίησης: [0.0, 1.0].

Κανονικοποίηση τιμών

Οι τιμές των συγκεκριμένων χαρακτηριστικών κανονικοποιήθηκαν με τη χρήση του κανονικοποιητή MinMaxScaler της βιβλιοθήκης scikit-learn, ο οποίος κλιμακώνει τις τιμές των αρχικών χαρακτηριστικών ώστε να βρίσκονται εντός του πεδίου τιμών [0.0, 1.0].

Κατηγορία

Γεωχωρικά χαρακτηριστικά

2.4.4.2. Εμβαδόν Επικάλυψης Πολυγώνου

Σύντομη περιγραφή

Το σύνολο χαρακτηριστικών αυτό αποτελείται από ένα χαρακτηριστικό ανά πολύγωνο προέλευσης το οποίο αντιστοιχεί στο ποσοστό εμβαδού επικάλυψης του τελικού (PST) πολυγώνου με το πολύγωνο προέλευσης (π.χ. Διανομή) και έχει αντιστοιχισθεί με αυτό.

Αριθμός θέσεων στο διάνυσμα χαρακτηριστικών

Από τη διαδικασία αυτή εξάγονται |N| χαρακτηριστικά, όπου |N| ο αριθμός των πολυγώνων προέλευσης.

Τύπος και εύρος τιμών

Τύπος: Αριθμητικές τιμές.

Εύρος τιμών πριν τη χρήση μεθόδων κανονικοποίησης: [0.90, 1.0].

Εύρος τιμών μετά τη χρήση μεθόδων κανονικοποίησης: [0.0, 1.0].

Κανονικοποίηση τιμών

Οι τιμές των συγκεκριμένων χαρακτηριστικών κανονικοποιήθηκαν με τη χρήση του κανονικοποιητή MinMaxScaler της βιβλιοθήκης scikit-learn, ο οποίος κλιμακώνει τις τιμές των αρχικών χαρακτηριστικών ώστε να βρίσκονται εντός του πεδίου τιμών [0.0, 1.0].

Κατηγορία

Γεωχωρικά χαρακτηριστικά

2.4.4.3. Περίμετρος Πολυγώνου

Σύντομη περιγραφή

Το σύνολο χαρακτηριστικών αυτό αποτελείται από ένα χαρακτηριστικό ανά πολύγωνο το οποίο αντιστοιχεί στην περίμετρό του.

Αριθμός θέσεων στο διάνυσμα χαρακτηριστικών

Από τη διαδικασία αυτή εξάγονται |N|+1 χαρακτηριστικά, όπου |N| ο αριθμός των διαφορετικών πολυγώνων προέλευσης και η επιπλέον τιμή αντιστοιχεί στο PST πολύγωνο.

Τύπος και εύρος τιμών

Τύπος: Αριθμητικές τιμές.

Εύρος τιμών πριν τη χρήση μεθόδων κανονικοποίησης: [25.967, 6507.281].

Εύρος τιμών μετά τη χρήση μεθόδων κανονικοποίησης: [0.0, 1.0].

Κανονικοποίηση τιμών

Οι τιμές των συγκεκριμένων χαρακτηριστικών κανονικοποιήθηκαν με τη χρήση του κανονικοποιητή MinMaxScaler της βιβλιοθήκης scikit-learn, ο οποίος κλιμακώνει τις τιμές των αρχικών χαρακτηριστικών ώστε να βρίσκονται εντός του πεδίου τιμών [0.0, 1.0].

Κατηγορία

Γεωχωρικά χαρακτηριστικά

2.4.4.4. Αριθμός Κορυφών Πολυγώνου

Σύντομη περιγραφή

Το σύνολο χαρακτηριστικών αυτό αποτελείται από ένα χαρακτηριστικό ανά πολύγωνο το οποίο αντιστοιχεί στον αριθμό κορυφών του.

Αριθμός θέσεων στο διάνυσμα χαρακτηριστικών

Από τη διαδικασία αυτή εξάγονται |N|+1 χαρακτηριστικά, όπου |N| ο αριθμός των διαφορετικών πολυγώνων προέλευσης και η επιπλέον τιμή αντιστοιχεί στο PST πολύγωνο.

Τύπος και εύρος τιμών

Τύπος: Αριθμητικές τιμές.

Εύρος τιμών πριν τη χρήση μεθόδων κανονικοποίησης: [4.0, 273.0].

Εύρος τιμών μετά τη χρήση μεθόδων κανονικοποίησης: [0.0, 1.0].

Κανονικοποίηση τιμών

Οι τιμές των συγκεκριμένων χαρακτηριστικών κανονικοποιήθηκαν με τη χρήση του κανονικοποιητή MinMaxScaler της βιβλιοθήκης scikit-learn, ο οποίος κλιμακώνει τις τιμές των αρχικών χαρακτηριστικών ώστε να βρίσκονται εντός του πεδίου τιμών [0.0, 1.0].

Κατηγορία

Γεωχωρικά χαρακτηριστικά

2.4.4.5. Μέση Τιμή Μήκους Ακμών Πολυγώνου

Σύντομη περιγραφή

Το σύνολο χαρακτηριστικών αυτό αποτελείται από 1 χαρακτηριστικό ανά πολύγωνο, το οποίο αντιστοιχεί στη μέση τιμή του μήκους των ακμών του πολυγώνου.

Αριθμός θέσεων στο διάνυσμα χαρακτηριστικών

Από τη διαδικασία αυτή εξάγονται |N|+1 χαρακτηριστικά, όπου |N| ο αριθμός των διαφορετικών πολυγώνων προέλευσης και η επιπλέον τιμή αντιστοιχεί στο PST πολύγωνο.

Τύπος και εύρος τιμών

Τύπος: Αριθμητικές τιμές.

Εύρος τιμών πριν τη χρήση μεθόδων κανονικοποίησης: [2.643, 130.633].

Εύρος τιμών μετά τη χρήση μεθόδων κανονικοποίησης: [0.0, 1.0].

Κανονικοποίηση τιμών

Οι τιμές των συγκεκριμένων χαρακτηριστικών κανονικοποιήθηκαν με τη χρήση του κανονικοποιητή MinMaxScaler της βιβλιοθήκης scikit-learn, ο οποίος κλιμακώνει τις τιμές των αρχικών χαρακτηριστικών ώστε να βρίσκονται εντός του πεδίου τιμών [0.0, 1.0].

Κατηγορία

Γεωχωρικά χαρακτηριστικά

2.4.4.6. Διακύμανση Μήκους Ακμών Πολυγώνου

Σύντομη περιγραφή

Το σύνολο χαρακτηριστικών αυτό αποτελείται από 1 χαρακτηριστικό ανά πολύγωνο, το οποίο αντιστοιχεί στη διακύμανση του μήκους των ακμών καθεμίας του πολυγώνου.

Αριθμός θέσεων στο διάνυσμα χαρακτηριστικών

Από τη διαδικασία αυτή εξάγονται |N|+1 χαρακτηριστικά, όπου |N| ο αριθμός των διαφορετικών πολυγώνων προέλευσης και η επιπλέον τιμή αντιστοιχεί στο PST πολύγωνο.

Τύπος και εύρος τιμών

Τύπος: Αριθμητικές τιμές.

Εύρος τιμών πριν τη χρήση μεθόδων κανονικοποίησης: [0.047, 9900.902].

Εύρος τιμών μετά τη χρήση μεθόδων κανονικοποίησης: [0.0, 1.0].

Κανονικοποίηση τιμών

Οι τιμές των συγκεκριμένων χαρακτηριστικών κανονικοποιήθηκαν με τη χρήση του κανονικοποιητή MinMaxScaler της βιβλιοθήκης scikit-learn, ο οποίος κλιμακώνει τις τιμές των αρχικών χαρακτηριστικών ώστε να βρίσκονται εντός του πεδίου τιμών [0.0, 1.0].

Κατηγορία

Γεωχωρικά χαρακτηριστικά

3. Σύνοψη

Στο Παραδοτέο 1.2 παρουσιάσαμε ένα εκτεταμένο σύνολο χαρακτηριστικών εκπαίδευσης που ορίσθηκαν στο έργο, προκειμένου να χρησιμοποιηθούν σε αλγόριθμους μηχανικής μάθησης για διασύνδεση, κατηγοριοποίηση και ολοκλήρωση γεωχωρικών δεδομένων. Για την καλύτερη κατανόηση της σημασιολογίας τους, τα χαρακτηριστικά που ορίσαμε ομαδοποιήθηκαν και παρουσιάστηκαν ανά σενάριο χρήσης-πρόβλημα μηχανικής μάθησης, ακολουθώντας τη δομή του Παραδοτέου 1.1. Παρόλα αυτά, τα περισσότερα από τα χαρακτηριστικά είναι γενικεύσιμα και άμεσα εκμεταλλεύσιμα σε διάφορες παραλλαγές των προβλημάτων μηχανικής μάθησης που εξετάσαμε, καλύπτοντας ένα ευρύ φάσμα προβλημάτων διασύνδεσης, κατηγοριοποίησης και ολοκλήρωσης γεωχωρικών δεδομένων. Καθώς θα επεκτείνουμε τις μεθόδους μας, το σύνολο των παραπάνω ορισμένων χαρακτηριστικών θα εμπλουτίζεται με νέα χαρακτηριστικά, τα οποία θα καταγράφονται σε ακόλουθα παραδοτέα των Ενοτήτων Εργασίας 2 και 3. Επιπλέον θα επιτελείται αξιολόγηση των ορισμένων χαρακτηριστικών και θα επιλέγονται-προκρίνονται προς τελική χρήση οι ομάδες εκείνες των χαρακτηριστικών που οδηγούν σε μέγιστη ακρίβεια στα επιμέρους προβλήματα.

4. Αναφορές

[DOR+14]	Nilesh Dalvi, Marian Olteanu, Manish Raghavan, and Philip Bohannon. 2014. Deduplicating a Places Database. In Proceedings of WWW '14.
[SMM17]	Rui Santos, Patricia Murrieta-Flores, and Bruno Martins. 2017. Learning to combine multiple string similarity metrics for effective toponym matching. International Journal of Digital Earth.