Χωρικός και Σημασιολογικός Εμπλουτισμός, Αναζήτηση και Οπτικοποίηση Αδόμητου Κειμένου

Παπαδιάς Ευάγγελος

ΔΜΠΣ Γεωπληροφορική
Μεταπτυχιακή εργασία

Ζωγράφου, Μάρτιος 2018

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
Σχολή ΑΓΡΟΝΟΜΩΝ ΤΟΠΟΓΡΑΦΩΝ ΜΗΧΑΝΙΚΩΝ

Χωρικός και Σημασιολογικός Εμπλουτισμός, Αναζήτηση και Οπτικοποίηση Αδόμητου Κειμένου

Επιβλέπων: Δρ. Μαργαρίτα Κόκλα

Επιτροπή

Μ. Κόκλα

Μ. Κάβουρας

Ν. Δουλάμης

Στη Βουλίτσα και στο μικρό μου Μάριο!

.

Ευχαριστίες

Θερμές ευχαριστίες οφείλονται στη Μαργαρίτα Κόκλα (Λέκτορα), την Ελένη Τομαή (μέλος ΕΔΙΠ) και τον Μαρίνο Κάβουρα (Καθηγητή), για την εμπιστοσύνη, την καθοδήγηση, την άριστη και εποικοδομητική συνεργασία, καθώς και τις δημιουργικές τους ιδέες.

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

Περίληψη

Ανέκαθεν ο άνθρωπος χρησιμοποιούσε το γραπτό λόγο για να αποθηκεύσει και να μεταδώσει τη γνώση, την οποία αποκτούσε κάθε φορά, στις επόμενες γενεές. Πάνω στην υπάρχουσα γνώση κατάφερε και "έχτισε" την καινούρια οδηγώντας την επιστήμη στο επίπεδο που βρίσκεται σήμερα. Ως άνθρωποι κατανοούμε την τάξη των πραγμάτων γύρω μας από τη σχετική τους θέση, δηλαδή τη γεωγραφία τους. Θα ήταν μεγάλη πρόοδος να μπορεί ο πλούτος γνώσης που περιέχεται στον γραπτό λόγο να ταυτιστεί με τον φυσικό χώρο, στον οποίο αναφέρεται, ώστε να μπορεί να αναζητηθεί με χωρικά κριτήρια.

Μεγάλη πρόοδος θα ήταν επίσης και η εξεύρεση μεθόδων και τεχνικών με τις οποίες η τεχνολογία της πληροφορικής θα μπορούσε να δημιουργήσει εργαλεία για την κατανόηση της σημασίας και του νοήματος που περιέχεται στα κείμενα, όπως ακριβώς θα έκανε ένας άνθρωπος αν τα διάβαζε.

Στην παρούσα εργασία έγινε μια προσπάθεια να εξαχθούν ονομασίες τοποθεσιών από το σώμα αδόμητων κειμένων με τεχνολογικά μέσα. Οι τοποθεσίες αυτές μεταφράστηκαν σε πραγματικές θέσεις στο χώρο και οργανώθηκαν ιεραρχικά με άλλες γεωγραφικές οντότητες. Κείμενα εμπλουτίστηκαν χωρικά και απέκτησαν αποτύπωμα στο χώρο που δεν είχαν πριν, καθιστώντας την αναζήτησή τους με χωρικά κριτήρια εφικτή. Από την ανάλυση κατέστη δυνατό επίσης, να εξαχθεί και σημασιολογική πληροφορία εμπλουτίζοντας έτσι την αναζήτηση και με σημασιολογικά κριτήρια.

Όλα τα αποτελέσματα της ανάλυσης παρουσιάζονται μέσα από μια διαδικτυακή διαδραστική εφαρμογή που δομήθηκε για το σκοπό αυτό. Για την ανάλυση και παρουσίαση των δεδομένων χρησιμοποιήθηκε αποκλειστικά λογισμικό δωρεάν διαθέσιμο και ανοικτού κώδικα.

Λέξεις κλειδιά :
Named Entity Recognition, Geocoding, Semantic Network, Semantic Enrichment, R.

Ε. Παπαδιάς, Αθήνα, 2018.

Abstract

Maybe the greatest achievement of mankind and the base of evolution through time, is language and written word. In order to transmit the knowledge acquired from one generation to another, man has always utilised the written word to contain and preserve it. New knowledge can't be succeeded without any prior one and the science that led today's living conditions to a higher level, has been built upon the foundations of previous efforts found in literature. As people we understand the order of things around us from their relative position towards us and that is their geography. It would be a great advance for science to take the wealth of knowledge contained in the written scientific discord, ground it to the physical space and be able to search it using spatial criteria.

A great advance would also be the development of methods, tools and techniques, based in modern information technology, so as computers to be able to understand and extract the meaning contained in unstructured text just as a person would do if he read it.

This dissertation presents work on how to extract spatial entities by name and transform them to coordinates as well as spatial concepts from user-generated educational scenarios. The spatial and semantic enrichment procedure was applied to a corpus of total 159 scenarios and the results of the analysis improved significantly overviewing and searching processes.

The knowledge extracted by the analysis was used to built an interactive web application that allows for better understanding of the processes and datasets while providing advanced reviewing of the user-generated scenarios and the GEOTHNK semantic network. The complete process and presentation is open-source oriented while performed on a single-platform.

Keywords :
Named Entity Recognition, Geocoding, Semantic Network,Semantic Enrichment, R.

E. Papadias, Athens, 2018.

Εισαγωγή

Το μεγαλύτερο επίτευγμα του ανθρώπινου είδους, που ταυτόχρονα είναι και η βάση για οτιδήποτε άλλο έχει αναπτυχθεί, αδιαμφισβήτητα είναι η γλώσσα και ο γραπτός λόγος. Χωρίς αυτά δεν θα υπήρχε επικοινωνία μεταξύ των ανθρώπων ώστε να μπορούν να συνεργαστούν, το ανθρώπινο είδος ίσως να μην είχε κυριαρχήσει και ο πολιτισμός δεν θα ήταν έτσι όπως τον γνωρίζουμε σήμερα. Δεν θα ήταν λάθος επίσης να πούμε πως και η αίσθηση της όρασης διαδραμάτισε σπουδαίο ρόλο στην ανάπτυξη του πολιτισμού όπως τον ξέρουμε, καθώς οι εικόνες της πραγματικότητας διαμεσολαβούν την κατανόηση του φυσικού κόσμου. Φιλόσοφοι-σημειωτικοί όπως ο Roland Barthes στο έργο του "Μυθολογίες", έχουν τονίσει πως μια εικόνα τη στιγμή που αποκτά σημασία για τον άνθρωπο μετατρέπεται αυτόματα σε γραφή στον εγκέφαλο.

Η εικόνα και η λεκτική περιγραφή που τη συνοδεύει έδωσαν στον άνθρωπο την ικανότητα να προσανατολίζεται και να επικοινωνεί με άλλους ανθρώπους. Πολύ πριν δημιουργηθούν οι πρώτοι χάρτες αλλά ακόμα και σήμερα που η τεχνολογία του διαδικτύου και της ψηφιακής απεικόνισης χαρτών κυριαρχεί στην καθημερινότητά μας, οι άνθρωποι χρησιμοποιούν τα τοπωνύμια για να προφέρουν και να κατανοήσουν τις χωρικές αναφορές που ακούν καθημερινά. Λεκτικές περιγραφές δηλαδή μιας τοποθεσίας οι οποίες προσδίδουν χωρικότητα στον προφορικό λόγο και δημιουργούν εικόνες που ίσως ανακαλούνται ή συνθέτονται αυτοστιγμής από τη μακρόχρονη μνήμη του εγκεφάλου. Με αυτό τον τρόπο τα άτομα κατανοούν τη σειρά των πραγμάτων γύρω τους και η χωρική κατανόηση γίνεται πιο εύκολη, γρήγορη και πιο σαφής.

Όπως και στον απλό προφορικό λόγο έτσι και στον γραπτό, οι συγγραφείς χρησιμοποιούν τις λεκτικές αναφορές για να περιγράψουν την περιοχή ή τις περιοχές στις οποίες αναφέρεται το αντικείμενο της εργασίας τους. Ένα κείμενο όμως περικλείει εκτός από απλές αναφορές τοποθεσιών και πιο πολύπλοκες. Αυτές είναι οι εννοιολογικές αναφορές που χρησιμοποιεί κάθε φορά ο συγγραφέας για να προσδώσει νοηματικό περιεχόμενο στα γραφόμενά του και να αναφερθεί σε καθορισμένες έννοιες. Όμως το ίδιο λεκτικό περιεχόμενο μιας πρότασης αλλά με άλλη σύνταξη, σημαίνει και κάτι διαφορετικό κάθε φορά. Ακόμα και η ίδια έννοια για την οποία χρησιμοποιείται το ίδιο λεκτικό μπορεί να έχει πολλαπλή σημασιολογική υπόσταση ανάλογα με τη θέση της μέσα σε μια πρόταση ή την ύπαρξη άλλων συνοδευτικών προσδιοριστικών πριν ή και μετά από αυτή.

Ένα νέο σχετικά ερευνητικό πεδίο στο χώρο της επεξεργασίας φυσικής γλώσσας (Natural Language Proccesing), μιας και η ανάπτυξη λογισμικού και υπολογιστικής ισχύος ικανής

να αντιμετωπίσει σύνθετα προβλήματα είναι σχετικά πρόσφατη, είναι αυτό της αυτοματοποιημένης γεωκωδικοποίησης κειμένων στα οποία περιέχονται τοπωνύμια. Είναι πρόκληση η δημιουργία ενός αλγορίθμου ο οποίος θα έχει τη δυνατότητα και ικανότητα, μαζικά και με τη δύναμη μιας υπολογιστικής μηχανής, να αναγνωρίζει την ονομασία μιας τοποθεσίας ή τοποθεσιών οι οποίες αναφέρονται σε ένα μήνυμα ή ένα σώμα κειμένου και να της αναθέτει χωρίς σφάλματα το σωστό ζεύγος γεωγραφικών συντεταγμένων του φυσικού χώρου που αντιστοιχούν σε αυτήν, όπως θα έκανε ένας άνθρωπος. Πρόκληση επίσης είναι και η ανάπτυξη μεθόδων, τεχνικών και υπολογιστικών διαδικασιών, ικανών να εξάγουν από απλό κείμενο εκτός από χωρική αλλά και σημασιολογική πληροφορία εμπλουτίζοντας έτσι την επισκόπηση εγγράφων με σκοπό την γρήγορη και προσανατολισμένη αναζήτησή τους. Η μεγαλύτερη πρόκληση ωστόσο για όλα τα παραπάνω είναι η ακρίβεια του προσδιορισμού των εννοιών αυτών μιας και μέθοδοι υπάρχουν χωρίς όμως να αντιμετωπίζουν το πρόβλημα με μονοσήμαντα αποτελέσματα.

Η αναγνώριση χωρικής και σημασιολογικής πληροφορίας με συμβατικές μεθόδους, όπως η ανάγνωση από ένα εκπαιδευμένο άτομο κειμένων ώστε εκείνος να αποφασίσει, παίρνει χρόνο και σε μερικές περιπτώσεις είναι αδύνατο να καταφέρει ο άνθρωπος όσα μια αυτόματη μηχανή. Το κόστος ώστε άνθρωποι να αναλύουν κείμενα που μπορεί να είναι και χιλιάδες ή και εκατομμύρια σελίδες, είναι απαγορευτικό στη σύγχρονη οργανωμένη κοινωνία με τα οικονομικά συστήματα που γνωρίζουμε. Ψυχολογικές έρευνες όπως αυτή των Baumeister and Tierney το 2011 (Leetaru and Schrodt, 2013), έδειξαν ότι η παρατεταμένη και συνεχής λήψη αποφάσεων από ανθρώπους όταν προσπαθούν να κωδικοποιήσουν κάτι επηρεάζεται από την κούραση, την έλλειψη προσοχής και μια τάση αυτοί να εφευρίσκουν μεθόδους ώστε να απλοποιούν την εργασία τους μιας και είναι επαναλαμβανόμενη, με επίπτωση στην ψυχολογία τους αλλά και στο τελικό αποτέλεσμα.

Είναι πολύ σημαντική λοιπόν η ανάπτυξη μεθόδων και αυτοματοποιημένων τεχνικών που θα διεκπεραιώνονται από υπολογιστές και οι οποίες θα μπορούν να εκμεταλλευτούν με ακρίβεια τον σημασιολογικό πλούτο γνώσης, όπως επίσης και το χωρικό αποτύπωμα που περιέχεται σε κείμενα.

Αντικείμενο της παρούσας εργασίας είναι η εξαγωγή ονομασιών τοποθεσίας από το σώμα κειμένων και η ταύτισή τους με ζεύγη συντεταγμένων. Επιπρόσθετα αναγνωρίστηκε η ύπαρξη χωρικών εννοιών του δικτύου GEOTHNK και προσδιορίστηκαν νοηματικές φράσεις που χαρακτηρίζουν αυτά. Η διαδικασία εφαρμόστηκε σε ένα σύνολο 159 εκπαιδευτικών σεναρίων τα οποία σκοπό έχουν την ενίσχυση της χωρικής σκέψης στη δευτεροβάθμια εκπαίδευση.

Στην πρώτη ενότητα περιλαμβάνεται μια βιβλιογραφική επισκόπηση των μεθόδων και τεχνικών που χρησιμοποιούνται από την επιστημονική κοινότητα για την εξαγωγή ονομασιών τοποθεσίας και προσδιορισμού των συντεταγμένων αυτών στο φυσικό χώρο, όπως επίσης και για την εξαγωγή νοηματικού και σημασιολογικού περιεχομένου από το σώμα κειμένων.

Στη δεύτερη ενότητα παρουσιάζονται τα δεδομένα που συλλέχθηκαν καθώς και η δια-

δικασία που ακολουθήθηκε και τα αποτελέσματα από την ανάλυση αυτών. Διερευνώνται και παρουσιάζονται οι δυνατότητες ελεύθερου λογισμικού και διαδικτυακών υπηρεσιών, ελεύθερα διαθέσιμων στην επιστημονική κοινότητα. Συγκεκριμένα διερευνάται η χρήση των πακέτων λογισμικού openNLP του Apache Foundation, του coreNLP του Stanford University και της βιβλιοθήκης spacy από τη γλώσσα python. Διερευνώνται και παρουσιάζονται οι δυνατότητες και η απόδοση των geocoders της Google, του OpenStreetMap (nominatim) και του Geonames.

Η τρίτη ενότητα περιλαμβάνει την περιγραφή της λειτουργικότητας της διαδικτυακής εφαρμογής που δημιουργήθηκε για την καλύτερη κατανόηση της απόδοσης των διαδικασιών και των εργαλείων που χρησιμοποιήθηκαν, όπως επίσης και για την παρουσίαση των αποτελεσμάτων του χωρικού και σημασιολογικού εμπλουτισμού αναζήτησης κειμένου που πραγματοποιήθηκε.

Η τέταρτη και τελευταία ενότητα περιλαμβάνει τα συμπεράσματα από την χρήση των παραπάνω διαδικασιών.

Όλα τα βήματα της ανάλυσης των δεδομένων και της παρουσίασης των αποτελεσμάτων ολοκληρώθηκαν χρησιμοποιώντας τη γλώσσα προγραμματισμού R μέσα από το περιβάλλον εργασίας της πλατφόρμας Rstudio. Καταβλήθηκε επίσης αρκετή προσπάθεια έτσι ώστε τα βήματα εργασίας να είναι αυτοτελείς προγραμματιστικές διαδικασίες με σκοπό να μπορούν να επανεκτελεστούν μελλοντικά σε παρόμοιες εργασίες, είτε να διορθωθούν σφάλματα που μπορεί να εντοπιστούν.

Κεφάλαιο 1

Βιβλιογραφική επισκόπηση

1.1 Εξαγωγή πληροφορίας

Η σημασιολογική ανάλυση στοχεύει στον προσδιορισμό και την εξόρυξη νοηματικού περιεχομένου, το οποίο βρίσκεται κρυμμένο στο σώμα ενός κειμένου, ανάμεσα στις λέξεις και το συντακτικό που χρησιμοποιεί ο συγγραφέας και το οποίο τελικά το χαρακτηρίζει και το ξεχωρίζει από τα υπόλοιπα. Για την εκτέλεση της διαδικασίας έχει αναπτυχθεί πληθώρα τεχνικών και μεθόδων οι οποίες όμως στο σύνολό τους βασίζονται σε τεχνικές επεξεργασίας φυσικής γλώσσας NLP. Οι τεχνικές αυτές έχουν αναπτυχθεί και χρησιμοποιούνται με σκοπό να μπορούν οι υπολογιστές να αναγνωρίσουν σε ένα σύνολο λέξεων, ή ακόμα και σε μια φωνητική καταγραφή, περίπου ότι και ένας άνθρωπος όταν διαβάζει ή ακούει κάτι. Να μπορούν δηλαδή να ξεχωρίσουν παραγράφους, προτάσεις, σημεία στίξης, αλλά και πιο πολύπλοκες πληροφορίες όπως έννοιες και τη σημασία τους. Οι τεχνικές αυτές απασχόλησαν τον άνθρωπο και την επιστήμη από την πρώτη στιγμή που έκαναν την εμφάνισή τους οι ηλεκτρονικοί υπολογιστές (Jurafsky and Martin, 2008), τις δεκαετίες δηλαδή 1940 και 1950.

Πράγματι ανατρέχοντας στον πλούσιο κατάλογο της ταινιοθήκης του παγκόσμιου κινηματογράφου μπορούμε να βρούμε πληθώρα παραδειγμάτων, αρκετές δεκαετίες πριν από τη σημερινή εποχή, στις οποίες φαίνεται η επιδίωξη και η διαρκής προσπάθεια του ανθρώπου να δημιουργήσει μηχανές που θα κατανοούν την ανθρώπινη ομιλία και γραφή και θα προβαίνουν σε διάφορες ενέργειες όπως για παράδειγμα η απάντηση ερωτημάτων ή η διεκπεραίωση εργασιών με τη λήψη φωνητικών εντολών. Κάτι τέτοιο όμως δεν έχει καταστεί πλήρως εφικτό, σύμφωνα με την τεχνολογία που γνωρίζουμε στις μέρες μας και τις ταινίες αυτές μπορούμε να τις βρούμε καταχωρημένες μόνο στην κατηγορία *επιστημονικής φαντασίας*.

1.1.1 Ορισμοί εννοιών

Ο όρος αδόμητο ή αμορφοποίητο κείμενο (unstructured text) αναφέρεται σε κείμενο στο οποίο η πληροφορία που περιέχεται και έχει τη μορφή λεκτικών αναφορών, δεν έχει συ-

νταχθεί με καθορισμένο τρόπο, δεν είναι γνωστό σε τι αναφέρεται σημασιολογικά και δεν μπορεί να ταξινομηθεί σε πεδία βάσης δεδομένων με κανονικοποιημένη μορφή. Τέτοια κείμενα παράγονται καθημερινά κάθε φορά που γράφουμε ένα email, ένα σύντομο μήνυμα, ένα έγγραφο, όταν κοινοποιούμε κάτι στα μέσα κοινωνικής δικτύωσης κ.α. Έστω το παρακάτω τμήμα ενός πίνακα δεδομένων :

location
Italy, Great Britain, Norway

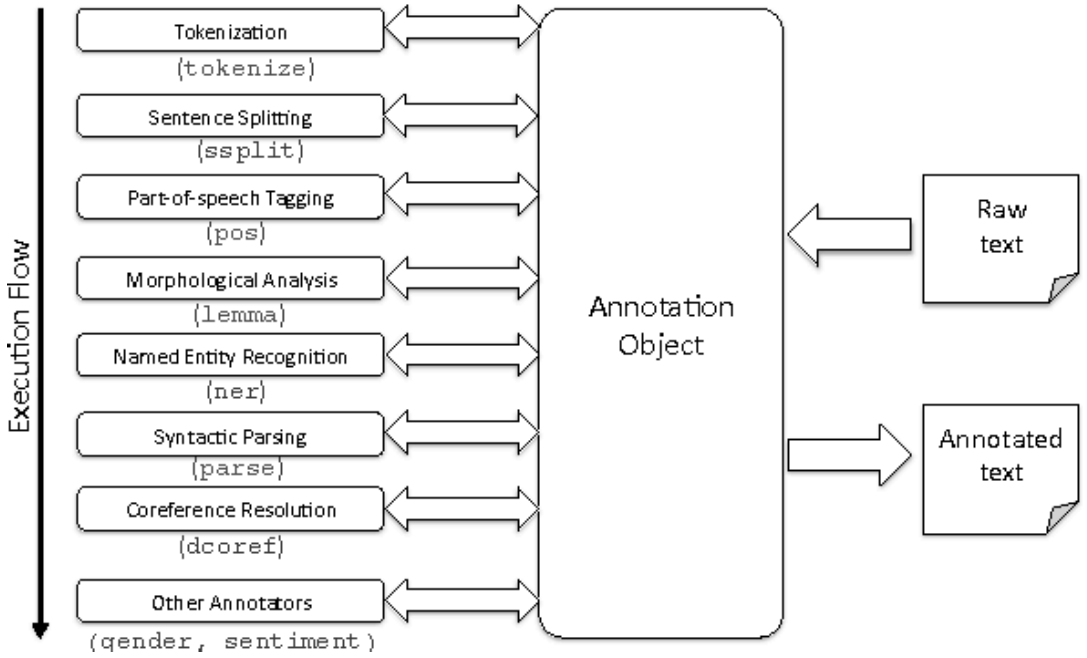
Οι λεκτικές αναφορές που περιέχονται στην εγγραφή του πίνακα γνωρίζουμε ότι αναφέρονται σε τοποθεσίες και ότι το όνομα κάθε τοποθεσίας είναι μια ενιαία οντότητα που διαχωρίζεται από μια άλλη με κόμμα(,). Σε αυτή την περίπτωση η πληροφορία του κειμένου δεν είναι αδόμητη, αλλά έχει δομημένη μορφή.

Η διαδικασία αυτόματης μετατροπής συγκεκριμένου και προκαθορισμένου τύπου πληροφορίας που περιέχεται σε κείμενο, σε δεδομένα κανονικοποιημένης μορφής, συμπληρώνοντας έτσι τα περιεχόμενα μιας σχεσιακής βάσης δεδομένων, ονομάζεται εξαγωγή πληροφορίας (Information Extraction (IE)), ανήκει στο ευρύτερο πεδίο της επεξεργασίας φυσικής γλώσσας (Natural Language Processing (NLP)) και υπόκειται στους περιορισμούς της.

Ο όρος annotation (σχολιασμός) αναφέρεται στην επισήμανση των στοιχείων ενός συνόλου δεδομένων, όπως ένα κείμενο, με επισημειώσεις μεταδεδομένων (metadata tags). Τα μεταδεδομένα παρέχουν επιπρόσθετες πληροφορίες για τα στοιχεία που επισημαίνονται με αυτά και έτσι οι υπολογιστικές διαδικασίες που εκτελούνται είναι σε θέση να προσδιορίσουν πρότυπα και μοτίβα και να εξάγουν συμπεράσματα. Μια υπολογιστική διαδικασία Annotation ή ένα λογισμικό Annotator, όπως το coreNLP του Stanford University, περιλαμβάνει ένα σύνολο διαδικασιών επιμέρους σχολιασμού (annotation) όπως φαίνεται στο σχήμα που ακολουθεί.

Από τα παραπάνω προκύπτει πως για κάθε προσπάθεια εξαγωγής οποιασδήποτε πληροφορίας (IE) προηγείται πάντα μια διαδικασία Annotation. Η εξαγωγή σημασιολογικής πληροφορίας (semantic information extraction) είναι υποπεδίο της IE και αναφέρεται στην εξαγωγή πληροφορίας αναγνωρίζοντας έννοιες (concepts), οντότητες (entities) καθώς και σχέσεις μεταξύ όρων και εννοιών μέσα ένα κείμενο. Συνήθως υποβοηθείται από μια καλά δομημένη οντολογία.

Ο όρος εμπλουτισμός (enrichment) αναφέρεται στη βελτίωση, ποιοτικά και ποσοτικά, των συνοδευτικών δεδομένων που χαρακτηρίζουν ένα κείμενο και το ξεχωρίζουν από τα υπόλοιπα. Σκοπός της διαδικασίας είναι η βελτίωση της επισκόπησης του περιεχομένου ενός εγγράφου καθώς και η πιο στοχευμένη αναζήτησή με βάση αυτό.



Σχήμα 1.1: Αρχιτεκτονική συστήματος Annotation
Πηγή: *Manning et. al., 2014*

1.1.2 Διαδικασία annotation

Για να καταστεί δυνατή η κατανόηση ενός κειμένου από έναν υπολογιστή έχει αναπτυχθεί η διαδικασία που ονομάζεται annotation (σχολιασμός) και προηγείται κάθε μεθόδου εξόρυξης σημασιολογικού περιεχομένου από αυτό. Κατά τη διαδικασία αυτή αναγνωρίζονται και σημαίνονται οι λέξεις, οι αριθμοί, τα σημεία στίξης, τα κενά, τα σύμβολα χαρακτήρων όπως οι παρενθέσεις και τα εισαγωγικά και προσδιορίζονται οι θέσεις τους (αρχή και τέλος) μέσα στο σώμα του κειμένου. Η διαδικασία αυτή είναι γνωστή ως tokenization. Ακολουθεί η αναγνώριση και ο διαχωρισμός των προτάσεων (sentense spliting). Μαζί με τις δυο προαναφερθείσες διαδικασίες η γλωσσική ανάλυση περιλαμβάνει και τη μορφοσυντακτική επισημείωση μερών του λόγου (Part of speech tagging, POS). Κατά τη διάρκεια αυτής αναγνωρίζεται για κάθε token το μέρος του λόγου με το οποίο αυτό χρησιμοποιείται μέσα στην κάθε πρόταση. Στη συνέχεια αναγνωρίζεται για κάθε token (τεκμήριο) η λεξική ρίζα από την οποία προέρχεται (lemma), για παράδειγμα η ρίζα του επιθέτου Cities είναι το city και του ρήματος showing το show. Η διαδικασία αυτή είναι γνωστή ως λημματοποίηση (lemmantization). Ακολουθεί η συντακτική ανάλυση (parsing) καθώς και άλλες διαδικασίες κατά περίπτωση. Στη συνέχεια ακολουθεί ένα παράδειγμα της διαδικασίας αυτής (πίνακας 1.1), η οποία στο σύνολό της, όπως προαναφέρθηκε, είναι γνωστή ως Annotation και το λογισμικό το οποίο την εκτελεί Annotator.

Κείμενο προς επεξεργασία : *"The teacher prepares a role play game in the classroom before visiting the complex. The learners are divided into 2 groups – visitors and hosts."*

Αποτέλεσμα διαδικασίας annotation με το λογισμικό Stanford's coreNLP:

id	sid	tid	word	lemma	upos	pos	cid
doc1	1	1	The	the	DET	DT	0
doc1	1	2	teacher	teacher	NOUN	NN	4
doc1	1	3	prepares	prepare	VERB	VBZ	12
doc1	1	4	a	a	DET	DT	21
doc1	1	5	role	role	NOUN	NN	23
doc1	1	6	play	play	NOUN	NN	28
doc1	1	7	game	game	NOUN	NN	33
doc1	1	8	in	in	ADP	IN	38
doc1	1	9	the	the	DET	DT	41
doc1	1	10	classroom	classroom	NOUN	NN	45
doc1	1	11	before	before	ADP	IN	55
doc1	1	12	visiting	visit	VERB	VBG	62
doc1	1	13	the	the	DET	DT	71
doc1	1	14	complex	complex	NOUN	NN	75
doc1	1	15	.	.	.	.	82
doc1	2	1	The	the	DET	DT	84
doc1	2	2	learners	learner	NOUN	NNS	88
doc1	2	3	are	be	VERB	VBP	97
doc1	2	4	divided	divide	VERB	VBN	101
doc1	2	5	into	into	ADP	IN	109
doc1	2	6	2	2	NUM	CD	114
doc1	2	7	groups	group	NOUN	NNS	116
doc1	2	8	–	–	.	:	123
doc1	2	9	visitors	visitor	NOUN	NNS	125
doc1	2	10	and	and	CONJ	CC	134
doc1	2	11	hosts	host	NOUN	NNS	138
doc1	2	12	.	.	.	.	143

Πίνακας 1.1: Παράδειγμα διαδικασίας annotation

Στον πίνακα 1.2 που ακολυθεί παρατίθεται η επεξήγηση των συντομογραφιών POS όπως προκύπτουν από την εφαρμογή με το λογισμικό Stanford's coreNLP.

pos	description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word

IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

Πίνακας 1.2: Επεξήγηση συντομογραφιών διαδικασίας POS

1.1.3 Υπάρχουσες προσεγγίσεις

Η πιο απλή μέθοδος εξαγωγής κάποιου νοήματος από ένα σύνολο λέξεων που συνθέτουν ένα κείμενο, όπως μια πρόταση, ένα μήνυμα ή ένα κείμενο, θα λέγαμε πως είναι η αντιμετώπιση όλων των λέξεων ως μεμονωμένες και ανεξάρτητες μεταξύ τους οντότητες

(bag of words), με σκοπό την καταμέτρηση της συχνότητας εμφάνισής τους. Με αυτό τον τρόπο το νόημα που χαρακτηρίζει αυτό το σύνολο λέξεων, πολύ απλά προκύπτει από τις λέξεις που χρησιμοποιούνται πιο συχνά. Όμως μια τέτοια αντιμετώπιση δεν λαμβάνει σε καμία περίπτωση υπόψη την ύπαρξη άλλων λεκτικών αναφορών, που συνυπάρχουν εντός του κειμένου και μπορεί να προηγούνται μιας λέξης και της προσδίδουν διαφορετική σημασία κάθε φορά. Ένα κείμενο περικλείει εκτός από απλές αναφορές σε γνωστά σε όλους αντικείμενα και πιο πολύπλοκες. Αυτές είναι οι εννοιολογικές αναφορές που χρησιμοποιεί κάθε φορά ο συγγραφέας για να προσδώσει νοηματικό περιεχόμενο στα γραφόμενά του και να αναφερθεί σε καθορισμένες έννοιες. Όμως το ίδιο λεκτικό περιεχόμενο μιας πρότασης αλλά με άλλη σύνταξη, σημαίνει και κάτι διαφορετικό κάθε φορά. Ακόμα και η ίδια έννοια για την οποία χρησιμοποιείται το ίδιο λεκτικό μπορεί να έχει πολλαπλή σημασιολογική υπόσταση ανάλογα με τη θέση της μέσα σε μια πρόταση ή την ύπαρξη άλλων συνοδευτικών προσδιοριστικών πριν ή και μετά από αυτή. Η σημασιολογική πληροφορία που περιέχεται σε κείμενα που είναι γραμμένα σε φυσική γλώσσα, η οποία είναι από τη φύση της αδόμητη, χαρακτηρίζονται από λεξική ασάφεια και πολυσημία.

Άλλη μια απλή μέθοδος η οποία προσπαθεί να αντιμετωπίσει τον προσδιορισμό του περιεχομένου που χαρακτηρίζει ένα κείμενο, είναι η εξαγωγή των n-grams. Τα n-grams είναι ακολουθίες δυο λέξεων (bigrams), τριών λέξεων (trigrams) κ.ο.κ. Και πάλι σε αυτή την περίπτωση υπολογίζοντας τη συχνότητα εμφάνισης ακολουθιών λέξεων εντός του κειμένου, μπορεί κάπως να καθοριστεί το νοηματικό περιεχόμενο αυτού, κάνοντας την υπόθεση ότι κάτι είναι σημαντικό, αφού ο συγγραφέας του το χρησιμοποιεί πολύ συχνά. Όμως και αυτή η τεχνική δε λαμβάνει υπόψη τη συντακτική δομή που χρησιμοποιεί ο συγγραφέας κάθε φορά και μπορεί να διαφοροποιεί το νόημα.

Τα n-grams όμως είναι μια τεχνική που χρησιμοποιούμε καθημερινά, οι περισσότεροι, καθώς μια από τις εφαρμογές της είναι η αυτόματη διόρθωση και συμπλήρωση προτάσεων όταν συντάσσουμε κείμενο σε φυσικό ή εικονικό πληκτρολόγιο στις συσκευές των smartphone και περιγράφεται στην πατέντα των Caskey et. al., της Google με αριθμό US9779080B2.

Μια άλλη μέθοδος που χρησιμοποιείται συχνά είναι αυτή που πρότειναν οι Justeson and Katz το 1995 και οι οποίοι διέκριναν πως οι φράσεις ουσιαστικών (noun phrases ή NPs), δηλαδή οι συνδυασμοί επιθέτων τα οποία ακολουθούνται από ουσιαστικά, αποτελούν το κεντρικό νόημα σε μια πρόταση. Το νόημα αυτό, τονίζουν, μπορεί και να μην δηλώνεται άμεσα από τις λέξεις που χρησιμοποιούνται για να τη σχηματίσουν. Χρησιμοποίησαν αυτή την τεχνική για να δημιουργήσουν έναν αλγόριθμο και να εξάγουν τεχνική ορολογία από κείμενα στην Αγγλική.

Το ουσιαστικό είναι ένα μέρος του λόγου που αναφέρεται σε πράγματα, τόπους, πρόσωπα αλλά και σε έννοιες που δηλώνουν ιδιότητα, ενέργεια και κατάσταση, όπως για παράδειγμα η λέξη *στύλος*. Τα επίθετα είναι μέρη του λόγου που προσδίδουν μια ιδιότητα ή ένα χαρακτηριστικό σε ένα ουσιαστικό όπως πχ η λέξη *κεντρικός*. Ο συνδυασμός του επιθέτου με το ουσιαστικό, πχ *κεντρικός στύλος* είναι μια noun phrase (NP) η οποία είναι

σημαίνουσα και χαρακτηριστική για την πρόταση που την περιέχει. Οι φράσεις ουσιαστικών χρησιμοποιούνται αρκετά για την εξαγωγή νοηματικού περιεχομένου από αδόμητο κείμενο.

Οι Rusu et.al το 2007 χρησιμοποίησαν τεχνικές POS για να προσδιορίσουν NPs και σχέσεις μεταξύ υποκειμένου και αντικειμένου μέσα σε προτάσεις. Εξήγαγαν συνδυασμούς υποκείμενου-κατηγορούμενου-αντικείμενου από ένα σύνολο προτάσεων στην Αγγλική γλώσσα. Ως υποκείμενο μιας πρότασης θεώρησαν μια φράση ουσιαστικού (NP) και ως κατηγορούμενο μια ρηματική φράση (VP). Χρησιμοποίησαν τον αλγόριθμο που ανέπτυξαν για τη δημιουργία περιλήψεων και εξαγωγή γεγονότων από κείμενα στην Αγγλική.

Οι Reiter και Frank το 2010 αξιοποιώντας επιβλεπόμενες τεχνικές μηχανικής μάθησης με επιλεγμένα γλωσσικά χαρακτηριστικά, προσδιόρισαν γενικές φράσεις ουσιαστικών οι οποίες εκφράζουν γνώση για είδη και γεγονότα. Η μέθοδός τους ήταν πολύ σημαντική για την αυτόματη κατασκευή βάσεων γνώσης, έτσι ώστε να αντληθεί νέα γνώση από την οργάνωση της πληροφορίας που προσφέρουν.

Οι Handler et. al., το 2016, προσπάθησαν να αντιμετωπίσουν το κενό που δημιουργείται από την αντιμετώπιση μιας πρότασης ή ενός κειμένου ως το σύνολο των ξεχωριστών απλών λέξεων (bag of words) ή ζευγών που τις αποτελούν (n-grams). Από το σύνολο αυτό υπολογίζοντας τη συχνότητα εμφάνισής τους μπορούν να εξαχθούν οι λέξεις κλειδιά που χαρακτηρίζουν σημασιολογικά το κείμενο στο οποίο περιέχονται. Τέτοιες προσεγγίσεις, καθώς και εκείνες που υπολογίζουν τη συχνότητα εμφάνισης ζευγών συνεχόμενων λέξεων, δεν λαμβάνουν σε καμία περίπτωση το συντακτικό και τη γραμματική η οποία χρησιμοποιήθηκε για τη σύνταξη μιας πρότασης και με τον τρόπο αυτό, τα ίδια ουσιαστικά η σημασία των οποίων αλλάζει από το επίθετο που χρησιμοποιείται κάθε φορά, αντιμετωπίζονται ως όμοιας σημασίας. Σημαντικές και σημαίνουσες νοηματικές προτάσεις δεν καθίσταται εφικτό να προσδιοριστούν.

Για την αντιμετώπιση αυτού του προβλήματος πρότειναν μια νέα μέθοδο καθορισμού νοηματικών φράσεων, με σκοπό τον εμπλουτισμό του συνόλου των απλών λέξεων κλειδιών, λαμβάνοντας υπόψη το συντακτικό και τη γραμματική αυτή τη φορά. Για το σκοπό αυτό χρησιμοποίησαν έναν μετατροπέα πεπερασμένων καταστάσεων (finite state transducer) και εισήγαγαν τον καθορισμό της απλής νοηματικής φράσης.

Ως απλή φράση η οποία μπορεί να θεωρηθεί σημαντική, σημαίνουσα και χαρακτηριστική μιας πρότασης, ενός μηνύματος ή ενός κειμένου καθόρισαν μια ακολουθία επιθέτων, ουσιαστικών (κοινά και συγκεκριμένα), προθέσεων καθώς και άρθρων ορισμού, η οποία περιγράφεται από το regular expression που ακολουθεί :

$$regex = "(A|N) * N(PD * (A|N) * N) * "$$

, όπου A : adjective, D : determiner, P : preposition, N : common/proper noun.

Η μέθοδός τους, όπως δηλώνουν, πέτυχε χαμηλή κατανάλωση υπολογιστικών και ανθρώπινων πόρων, προσδιόρισε τις πιο σχετικές και σημαντικές φράσεις, όπως θα έκανε ένας άνθρωπος και οι φράσεις που ανακτήθηκαν χαρακτηρίζονται από υψηλή ερμηνευ-

σιμότητα και χαμηλή ασάφεια. Η μέθοδος μπορεί να χρησιμοποιηθεί με το πακέτο λογισμικού phrasemachine για τη γλώσσα R, ενώ διατίθεται και για τους χρήστες της γλώσσας python. Χρησιμοποιεί τον Part of speech tagger του λογισμικού openNLP του Apache Foundation και είναι λειτουργικό μόνο για την Αγγλική γλώσσα. Ως λογισμικό ανοικτού κώδικα μπορεί κάλλιστα κάποιος να χρησιμοποιήσει τη μέθοδο αυτή και με άλλους POS-taggers (coreNLP, spacy) και συντακτικά μοντέλα σε άλλες γλώσσες πέραν της Αγγλικής, μιας και αυτός του openNLP δεν χαρακτηρίζεται από μεγάλη ακρίβεια.

Μια πιο σύνθετη μέθοδος καθορισμού νοηματικού περιεχομένου, είναι η Latent Semantic Analysis (LSA). Κατά τη μέθοδο αυτή συγκρίνονται κείμενα τα οποία έχουν αναπαρασταθεί ως διανύσματα που αποτυπώνουν το σημασιολογικό τους περιεχόμενο. Η μέθοδος στηρίζεται στον υπολογισμό της ομοιότητας μεταξύ ζευγών διανυσμάτων αξιοποιώντας το συνημίτονο της γωνίας που αυτά σχηματίζουν. Με αυτό τον τρόπο ξεκίνησε η λειτουργία της μεθόδου, το 1988, για τον εντοπισμό και την επιλογή κειμένων που συσχετίζονται περισσότερο με μια έννοια, μέσα από μια μεγάλη βάση δεδομένων. Σύμφωνα με τους Wiemer-Hastings et. al. η μέθοδος αποτελείται από 4 στάδια :

- Το πρώτο στάδιο περιλαμβάνει τη συλλογή των κειμένων προς συσχέτιση. Ως ξεχωριστά κείμενα μπορούν να θεωρηθούν και οι διαφορετικές παράγραφοι στις οποίες μπορεί να οργανώνεται ένα κείμενο.

- Κατά το δεύτερο στάδιο δημιουργείται μια μήτρα συσχέτισης κειμένων και περιεχόμενων όρων (document-term matrix). Δηλαδή ένας πίνακας όπου για κάθε κείμενο έχουν αντιστοιχηθεί οι περιεχόμενοι όροι (terms) καθώς και η συχνότητα εμφάνισής τους. Επισημαίνεται πως η μέθοδος θεωρεί ως όρους (terms) μόνο εκείνες τις λέξεις οι οποίες είναι κοινές μεταξύ των κειμένων, χωρίς καμία μορφολογική επέμβαση, όπως λημματοποίηση ή stemming για παράδειγμα.

- Το τρίτο στάδιο περιλαμβάνει τη στάθμιση των τιμών κάθε τιμής στη μήτρα (matrix). Μια κοινή μέθοδος που χρησιμοποιείται για τον υπολογισμό αυτό, είναι η log entropy.

- Έπειτα κατά το τέταρτο στάδιο, εφαρμόζεται η αλγεβρική μέθοδος SVD (Singular Value Decomposition) κατά την οποία επαναπροσδιορίζονται και ταξινομούνται οι θέσεις κάθε όρου, στο χώρο που έχει δημιουργηθεί από τις διαστάσεις της μήτρας, με αποτέλεσμα να μπορούν να προσδιοριστούν οι πιο σημαντικές.

Μια ακόμη μέθοδος εξαγωγής πιο εξειδικευμένης σημασιολογικής πληροφορίας που περιλαμβάνεται σε κείμενα, είναι και η Ontology-Based Information Extraction (OBIE). Η μέθοδος αυτή αποτελεί υποκατηγορία του γενικότερου επιστημονικού πεδίου IE. Στηρίζεται στην αξιοποίηση των συσχετίσεων που δημιουργεί η οργανωμένη δόμηση μιας οντολογίας, που αφορά ένα συγκεκριμένο επιστημονικό πεδίο, όπως για παράδειγμα το δίκτυο χωρικών εννοιών του GEOTHNK. Σύμφωνα με τον Gruber, μια οντολογία είναι μια τυπική και ρητή εξειδίκευση μιας εννοιολογικής σκέψης ή αντίληψης. Ειδικότερα μια οντολογία

είναι θα λέγαμε μια αναπαράσταση γνώσης, ένα λεξικό εκείνων των εννοιών που χρησιμοποιούνται για την αναπαράσταση ενός συγκεκριμένου τομέα του λόγου, οργανωμένο σύμφωνα με ορισμούς σε κλάσεις με καθορισμένες σχέσεις και λειτουργίες μεταξύ των εννοιών αυτών (Gruber 1993). Έτσι η μέθοδος αυτή στοχεύει σε εξαγωγή πληροφορίας συγκεκριμένου επιστημονικού πεδίου και μπορεί να συνδυαστεί με μια από τις άλλες μεθόδους που περιγράφηκαν στην παρούσα ενότητα.

Στην παρούσα εργασία επιχειρήθηκε ο συνδυασμός της μεθόδου OBIE αξιοποιώντας το δίκτυο χωρικών εννοιών του GEOTHNK, με την εξαγωγή νοηματικών φράσεων ουσιαστικών όπως τις όρισαν οι Handler et. el.

1.2 Γεωκωδικοποίηση

Ως γεωκωδικοποίηση θα μπορούσε να οριστεί η διαδικασία μετατροπής είτε της ταχυδρομικής διεύθυνσης είτε της ονομασίας μιας τοποθεσίας, από κείμενο σε ένα ζεύγος συντεταγμένων σε κάποιο σύστημα αναφοράς. Ως σύστημα αναφοράς, για την ανάπτυξη κοινής "γλώσσας επικοινωνίας" των θέσεων σε παγκόσμιο επίπεδο αλλά και τη συνεργασία με άλλα επίπεδα πληροφορίας, συνηθίζεται να χρησιμοποιείται το σύστημα γεωδαιτικών συντεταγμένων WGS 84 (φ,λ).

Ο συχνότερος τρόπος γεωγραφικών περιγραφών εντός των πόλεων είναι οι διευθύνσεις (Clodoveu and Fonseca, 2007). Μια από τις πρώτες προσπάθειες και ονομασίες της γεωκωδικοποίησης είναι το address geocoding και είναι αυτή που έδωσε η Google και ο Ge Xianping στην πατέντα με την οποία κατοχύρωσε την εφεύρεσή της (US 6934634 B1). Το βασικό τμήμα του συστήματος αυτού, όπως περιγράφεται στην πατέντα, ήταν ένας πίνακας με σειρές που αντιστοιχούσαν σε μια ή περισσότερες διευθύνσεις και ανάλογα με τους όρους αναζήτησης (στήλες), το σύστημα μπορούσε εύκολα να εντοπίσει την εγγραφή που ταίριαζε τουλάχιστον με έναν από αυτούς. Η εφεύρεση μπορούσε επίσης να αναγνωρίσει την ταχυδρομική διεύθυνση, αν αυτή περιεχόταν σε ένα κείμενο στο διαδίκτυο, εφαρμόζοντας όμως προκαθορισμένους κανόνες. Οι προκαθορισμένοι κανόνες δεν ήταν τίποτα άλλο παρά ένας κατάλογος με ορισμούς γνωστών διευθύνσεων τους οποίους το σύστημα προσπαθούσε να ταυτίσει με τους πιθανούς υποψήφιους ορισμούς διευθύνσεων που εντόπιζε στο σώμα ενός κειμένου. Κάθε εγγραφή στον κατάλογο περιείχε και ένα ζεύγος γνωστών συντεταγμένων και με αυτό τον τρόπο η γεωκωδικοποίηση του κειμένου γινόταν εφικτή. Οι απλοί αυτοί κατάλογοι σήμερα έχουν εξελιχθεί σε πολύπλοκες βάσεις γνώσεις και πέρα από απλή αντιστοίχηση της ονομασίας μιας τοποθεσίας με ένα ζεύγος γεωγραφικών συντεταγμένων, παρέχουν και πολλές άλλες πληροφορίες που βοηθούν τα σύγχρονα συστήματα στην γρήγορη αναζήτηση και αποσαφήνιση μιας ονομασίας και είναι γνωστά ως gazetteers.

Από την αυγή της τεχνολογίας του διαδικτύου φάνηκε το επιστημονικό ενδιαφέρον για την ταύτιση λεκτικών περιγραφών με το χώρο, όπως επίσης διαγνώστηκαν και τα προβλήματα και οι δυσκολίες που είχαν οι προσπάθειες αυτές. Πράγματι υπάρχουν πολλοί τρόποι να αναγραφεί μια διεύθυνση ή μια τοποθεσία και συνήθως ποτέ δεν είναι πλήρης αλλά ούτε και περιγράφεται από όλους με τον ίδιο τρόπο. Άλλες ονομασίες που χρησιμοποιούνται για την γεωκωδικοποίηση (geocoding) στη βιβλιογραφία, είναι το localising και το grounding (Patullo, 2008).

Το βασικό πρόβλημα της γεωκωδικοποίησης ενός κειμένου (ενός συνόλου λέξεων οργανωμένο σε προτάσεις), είναι πρωτίστως η αναγνώριση μιας ή πολλών λέξεων που συνθέτουν την ονομασία μιας περιοχής, ή ενός συνόλου λέξεων και αριθμών που συνθέτουν μια ταχυδρομική διεύθυνση ή ένα τοπωνύμιο στο σώμα του κειμένου και στη συνέχεια, η ταύτιση αυτού με ένα ζεύγος συντεταγμένων στο παγκόσμιο σύστημα αναφοράς. Οι παραδοσιακές μέθοδοι, όπως εκείνη που χρησιμοποίησε η Google όταν ξεκίνησε να χτίζει

το σύστημά της, στηρίζονται στην αξιοποίηση καταλόγων με γνωστές γεωγραφικές ονομασίες και συντεταγμένες που αντιστοιχούν σε αυτές. Συγκρίνεται μια υποψήφια λέξη ή ένα σύνολο λέξεων, οι οποίες αποτελούν την ονομασία μιας περιοχής και περιέχονται σε ένα κείμενο, με όλες τις εγγραφές στον γεωγραφικό κατάλογο και επιλέγεται αυτή που ταιριάζει καλύτερα. Έτσι γίνεται και αντιστοίχηση με ένας ζεύγος συντεταγμένων.

Όμως η γεωκωδικοποίηση μιας πρότασης, ενός τίτλου, της περίληψης μιας εργασίας ή ακόμα και πολλών σελίδων ενός βιβλίου, είναι μια πιο σύνθετη διαδικασία που απαιτεί αρκετά βήματα εργασίας που πρέπει να υλοποιηθούν ώστε να αντιμετωπιστούν πρακτικές δυσκολίες που σχετίζονται πρώτα με την αναγνώριση της λεκτικής περιγραφής μιας τοποθεσίας μέσα σε από ένα σύνολο λέξεων και στη συνέχεια με την ποιότητα του προσδιορισμού της πραγματικής θέσης στον φυσικό χώρο αλλά και χρόνο, στην οποία αυτή αναφέρεται.

1.2.1 Στάδια

Σε έναν αλγόριθμο μιας διαδικασίας γεωκωδικοποίησης μπορούμε να διακρίνουμε 4 βασικά στάδια. Το πρώτο στάδιο περιλαμβάνει τη γλωσσολογική ανάλυση και προεπεξεργασία του κειμένου όπως περιγράφηκε σε προηγούμενο κεφάλαιο (διαδικασία annotation).

Κατά το δεύτερο στάδιο πρέπει να αναγνωριστούν οι πιθανές ονομασίες που υποδεικνύουν τοποθεσίες μέσα στο σώμα του κειμένου. Το στάδιο αυτό είναι γνωστό ως Named Entity Recognition (NER) και περιλαμβάνει μεθόδους οι οποίες αναγνωρίζουν εκείνες τις λέξεις που υποδεικνύουν τοποθεσία. Οι αλγόριθμοι NER μπορούν να αναγνωρίσουν συνήθως ημερομηνία, τοποθεσία, χρηματικά ποσά, οργανισμούς, ποσοστά και πρόσωπα κατατάσσοντας ότι μένει στα λοιπά (date, location, money, organization, percentage, person, misc). Μέρος των τεχνικών NER είναι και το geotagging το οποίο αναφέρεται μόνο στην αναγνώριση της τοποθεσίας και όχι όλων των οντοτήτων που προαναφέρθηκαν. Ο προσδιορισμός των ονομασιών στο σώμα κειμένων προσεγγίζεται με δυο τρόπους. Με σύγκριση του κειμένου με ένα σύνολο γνωστών ονομασιών τοποθεσίας (bag of words) και με τεχνικές μηχανικής μάθησης. Στη δεύτερη κατηγορία τεχνικών, εξελιγμένοι αλγόριθμοι έχουν εκπαιδευτεί σε αλληλουχία γλωσσικών και συντακτικών μοντέλων και από την ανάλυση μερών του λόγου (Part of Speech Tagging) σε κάθε πρόταση είναι σε θέση να αναγνωρίσουν τις οντότητες που προαναθέρθηκαν. Τα πακέτα λογισμικού που ενσωματώνουν τέτοιους αλγόριθμους στο περιβάλλον εργασίας της γλώσσας R και δοκιμάστηκαν είναι το coreNLP του Stanford Univercity, το openNLP του Apache Foundation και το spacy.

Τα δύο αυτά πρώτα στάδια περιλαμβάνονται σε ένα σύνολο τεχνικών που είναι γνωστό στην επιστήμη της Πληροφορικής ως Natural Language Processing (NLP). Οι τεχνικές NLP χρησιμοποιούνται με σκοπό οι υπολογιστές να μπορούν να αναγνωρίσουν σε ένα σύνολο λέξεων, παρότι αυτό είναι ανέφικτο με την τεχνολογία που γνωρίζουμε σήμερα, περίπου ότι και ένας άνθρωπος όταν διαβάζει ή ακούει κάτι (παραγράφους, προτάσεις, σημεία στίξης κλπ).

Το τρίτο στάδιο της διαδικασίας περιλαμβάνει το πιο δύσκολο κομμάτι, αυτό της αποσαφήνισης, γνωστό στη βιβλιογραφία ως disambiguation. Ασάφεια στο όνομα μιας τοποθεσίας προκαλείται όταν η ονομασία της έχει πολλαπλά νοήματα, όταν χρησιμοποιείται με τον ίδιο τρόπο για πάνω από μια θέσεις και όταν χρησιμοποιούνται άλλες λέξεις για να προσδιορίσουν την τοποθεσία και της προσδίδουν διαφορετικό χωρικό αποτύπωμα. Η συνήθης αντιμετώπιση της αποσαφήνισης πραγματοποιείται χρησιμοποιώντας στοιχεία από τους γεωγραφικούς καταλόγους και το κείμενο που περιέχει το τοπωνύμιο.

Η δυσκολότερα αντιμετωπίσιμη ασάφεια είναι οι πολλαπλές τοποθεσίες οι οποίες χρησιμοποιούν την ίδια ονομασία. Για παράδειγμα στον Ελλαδικό χώρο με το όνομα Αγία Μαρίνα θα βρούμε πολλές περιοχές. Το πρόβλημα περιπλέκεται ακόμα περισσότερο αν συνυπολογιστεί η σημασιολογική ερμηνεία της ονομασίας Αγία Μαρίνα. Είναι ο Δήμος της Αγίας Μαρίνας ή η εκκλησία που βρίσκεται εντός των ορίων αυτού ; Ή μήπως το κείμενο που αναλύεται είναι ένα εκκλησιαστικό κείμενο το οποίο αναφέρεται στο πρόσωπο της Αγίας Μαρίνας και δεν περιέχει καμία χωρική πληροφορία ; Το ίδιο συμβαίνει και με ονομασίες όπως η Αρχαία Κόρινθος. Η περιγραφή αναφέρεται στην αρχαία Κόρινθο ως έννοια ή στην περιοχή που σήμερα ονομάζεται Αρχαία Κόρινθος και είναι συγκεκριμένη τοποθεσία.

Τα προθέματα και οι επιθετικοί προσδιορισμοί που χρησιμοποιούνται συχνά μαζί με τις ονομασίες τοποθεσιών δυσκολεύουν ακόμα περισσότερο την αποσαφήνιση και τον προσδιορισμό της σωστής θέσης. Για παράδειγμα η τοποθεσία South Wales υποδεικνύει την περιοχή της Νότιας Ουαλίας. Όμως ποια ακριβώς είναι η περιοχή αυτή ; Πού σταματάει ο νότος και αρχίζει ο βοράς ; Το πρόβλημα γίνεται ακόμα πιο σύνθετο αν το όνομα της τοποθεσίας είναι New South Wales, η οποία βρίσκεται στην Αυστραλία και όχι στην Ουαλία. Η περιοχή 400μ ανατολικά του ποταμού Πηνειού πως μπορεί να προσδιορισθεί με ακρίβεια ; Ποιες λέξεις είναι αυτές που μαζί θα απαρτίζουν την τοποθεσία για την οποία θα πρέπει να προσδιορισθεί ένα σημείο;

Μεγάλο πρόβλημα επίσης είναι και η αποσαφήνιση των ονομασιών τοποθεσιών που έχουν μεταφραστεί σε άλλη γλώσσα από αυτή της χώρας που βρίσκονται. Πολλές τοποθεσίες πολύ συχνά αποτελούνται από 2 λέξεις και αναφέρονται και ως ενιαία μορφή και ως ξεχωριστή.

Όλα τα παραπάνω προβλήματα μπορούμε να τα συνοψίσουμε στην έννοια που ο Patullo ονομάζει γεωπολυπλοκότητα (geo-complexity). Προκύπτει όταν απαιτείται ειδική πολιτισμική γνώση για να κατανοηθεί η χωρικότητα μιας ονομασίας στο περιεχόμενο ενός κειμένου και όταν η χωρική αναφορά δεν είναι δομημένη με κάποιο πλήρη ή γνωστό τρόπο όπως μια διεύθυνση ή ένας Ταχυδρομικός Κώδικας που είναι κάτι συγκεκριμένο και εύκολο να προσδιορισθεί η θέση στην οποία αναφέρεται σε οποιαδήποτε γλώσσα.

Στη βιβλιογραφία συναντάται και ο όρος toponym resolution. Ο όρος αυτός αναφέρεται στην ανάλυση που μπορεί να διαθέτει ή εκείνη που απαιτείται να χτιστεί για ένα τοπωνύμιο. Στην ανάλυση περιλαμβάνονται όλα εκείνα τα συμπληρωματικά διαθέσιμα δεδομένα που μπορούν να ανακτηθούν από το κείμενο ή από άλλες πηγές και συνεπικουρούν στην

σωστή επιλογή της τοποθεσίας όταν μια ονομασία χρειάζεται αποσαφήνιση (Wei Zhang and Judith Gelernter, 2014). Οι Moncla et al χρησιμοποιούν τον πιο πάνω όρο για την διαδικασία ταύτισης μιας ονομασίας με το σωστό ζεύγος συντεταγμένων στο φυσικό χώρο μετά από αποσαφήνιση.

Το τέταρτο και τελευταίο στάδιο είναι αυτό της ανάθεσης του ζεύγους συντεταγμένων που παρέχεται από έναν γεωγραφικό κατάλογο με απλό τρόπο ανάλογο με αυτό που χρησιμοποίησε ο Ge Xianping το 2005 όταν κατοχύρωνε την πατέντα για λογαριασμό της Google. Η διαδικασία αυτή σε συνδυασμό με την αποσαφήνιση αναφέρεται και ως geocoding.

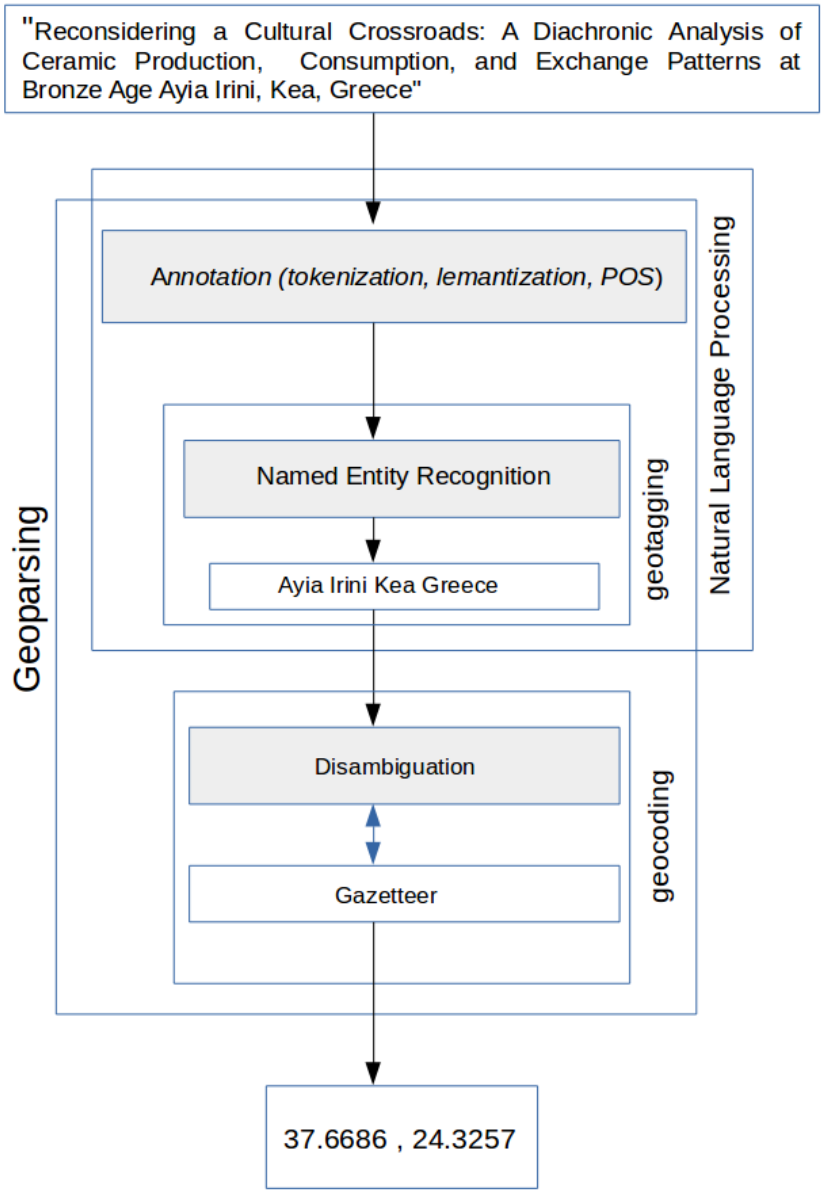
Συχνά στη βιβλιογραφία όλη η παραπάνω διαδικασία αναφέρεται και ως geoparsing (NLP + geocoding). Ένας σχετικός με το geoparsing όρος ο οποίος μπορεί να δημιουργήσει σύγχυση είναι η διαδικασία Geographic Information Retrieval (GIR), η οποία όμως δεν περιορίζεται στην εξαγωγή γεωγραφικής πληροφορίας θέσης από μια απλή αναφορά τοποθεσίας σε ένα κείμενο, αλλά επιπλέον την καταλογοποιεί και επιτρέπει την χωρική αναζήτηση με αυτήν μέσα στο κείμενο. (Leetarou, 2012).

Στο σχήμα 1.2 που ακολουθεί φαίνονται τα βήματα και οι διαδικασίες μιας τυπικής διαδικασίας που συνήθως ακολουθείται.

1.2.2 Υπάρχουσες προσεγγίσεις

Μια αξιόλογη και πετυχημένη προσπάθεια είναι αυτή του Kalev H. Leetaru το 2012, ο οποίος ανέπτυξε έναν αλγόριθμο με σκοπό να αναγνωρίσει τις τοποθεσίες που αναφέρονται στο σύνολο του κειμένου των κειμένων της αγγλικής έκδοσης της Wikipedia (4 εκατομμύρια σελίδες τότε). Προσπάθησε να ανακαλύψει αν τα άρθρα της εστιάζονται σε κάποια περιοχή του πλανήτη περισσότερο και να δείξει ότι τα αποτελέσματα των μελετών που στηρίζονται στη γεωγραφική πληροφορία που παρέχουν οι συγγραφείς, οδηγούν σε εσφαλμένη εκτίμηση της πραγματικότητας, καθώς αυτή είναι ελλιπής. Σύγκρινε δηλαδή τα αποτελέσματά της μαζικής και αυτοματοποιημένης διαδικασίας γεωκωδικοποίησης την οποία εκτελεί ένα υπολογιστικό σύστημα, με τα αποτελέσματα της συγκέντρωσης των τοποθεσιών από την χειροκίνητη διαδικασία της επικόλλησης γεωγραφικής πληροφορίας (συντεταγμένες) στα αντίστοιχα σημεία ενός άρθρου από τον συγγραφέα του κατά τη δημοσίευση. Ο αλγόριθμος εντόπισε 10 φορές περισσότερες ονομασίες τοποθεσιών από αυτές που παρέχουν οι συγγραφείς. Εστίασε στην εξόρυξη της λεκτικής περιγραφής μιας ονομασίας τοποθεσίας από το σώμα ενός κειμένου και όχι σε πολυσύνθετη αποσαφήνιση περιγραφών όπως “κοντά σε”, “25Χλμ βορειοδυτικά από” κλπ.

Οι πρακτικές που συνήθως εφαρμόζονταν τότε είχαν ως πρώτο βήμα τον εντοπισμό των πιθανών υποψήφιων λεκτικών αναφορών τοποθεσιών κάνοντας την υπόθεση ότι αυτές ξεκινούν με κεφαλαία γράμματα και συνήθως προηγείται κάποια λέξη κλειδί που χρησιμοποιείται στην Αγγλική γλώσσα όπως in, at, near κλπ. Σε αντίθεση με αυτές εκείνος χρησιμοποίησε διάφορα λεξικά της Αγγλικής γλώσσας για να δημιουργήσει μια λίστα με όλες τις γνωστές λέξεις της. Έπειτα σύγκρινε τη λίστα με τους gazetteers GNS και



Σχήμα 1.2: workflow γεωκωδικοποίησης

GNIS και προσδιόρισε τις λέξεις οι οποίες δεν περιέχονται σε αυτούς και συνεπώς δεν χρησιμοποιούνται για τον προσδιορισμό καμίας από τις γνωστές ονομασίες τοποθεσίας. Έτσι κατάφερε να εξαλείψει όλες τις λέξεις από τα κείμενα οι οποίες σίγουρα δεν χρησιμοποιούνται στις τοποθεσίες και προσπάθησε να εντοπίσει στο υπόλοιπο τους πιθανούς υποψήφιους ορισμούς, χρησιμοποιώντας τους GNS και GNIS. Δημιουργήθηκε επίσης ένας κατάλογος με όλες τις λέξεις που περιέχονται και στα λεξικά και στους gazetteers και συγκρίθηκε με τα κείμενα της Wikipedia ώστε να καταρτιστεί μια λίστα με τις συχνότητες εμφάνισης των λέξεων αυτών με κεφαλαίο αρχικό γράμμα και με μικρό.

Αρχικά ελέγχθηκε αν οι αναφορές ταυτίζονται με τις ανώτερες διοικητικές μονάδες όπως ονόματα χωρών, πρωτευουσών ή μεγάλων πόλεων και όσες ταυτίζονταν αποτελούσαν πιθανούς υποψήφιους. Οι υπόλοιπες ελέγχθηκε αν ταυτίζονται με κάποιο άλλο τοπωνύμιο και αν ναι αποτελούσαν έναν πιθανό υποψήφιο. Για ότι δεν ίσχυε κάτι από τα παραπάνω χρησιμοποιήθηκε ο πίνακας με τις συχνότητες εμφάνισης ώστε να ελεγχθούν οι συνδυασμοί των λέξεων που απαρτίζουν την τοποθεσία. Αρχίζοντας από το τέλος, εάν η συχνότητα εμφάνισης μιας λέξης του πιθανού υποψήφιου εμφανιζόταν συχνότερα να αρχίζει με μικρό γράμμα από ότι με κεφαλαίο απορρίπτονταν και συνέχιζε με τις υπόλοιπες λέξεις μέχρι η συχνότητα με κεφαλαίο να είναι μεγαλύτερη, οπότε και ταυτιζόταν με κάποια εγγραφή στον gazetteer. Απορρίφθηκαν οι πιθανοί υποψήφιοι των οποίων οι ξεχωριστές λέξεις όταν ενώνονταν σχημάτιζαν λέξεις με μόνο 4 ή λιγότερους χαρακτήρες. Δημιουργήθηκε επίσης και μια λίστα με ονόματα προσωπικοτήτων τα οποία χρησιμοποιούνται και για την ονομασία πόλεων αλλά και άλλες ονομασίες και λέξεις που διαπιστώθηκε ότι οδηγούν σε ασάφεια. Όλοι οι πιθανοί υποψήφιοι που περιείχαν λέξη που ταυτιζόταν με τα περιεχόμενα της λίστας αυτής απορρίφθηκαν. Όλος ο κατάλογος των πιθανών υποψήφιων ονομασιών προωθήθηκε στο επόμενο στάδιο που αποτελεί και την κρισιμότερη και δυσκολότερη διαδικασία από όλες, αυτή της αποσαφήνισης. Δηλαδή το σε ποια από όλες τις διαφορετικές τοποθεσίες που έχουν το ίδιο όνομα αναφέρεται η περιγραφή.

Η μέθοδος που υλοποιήθηκε ήταν αυτή που εφαρμόζεται συχνότερα και περιλαμβάνει την συνεκτίμηση των υπόλοιπων αναφορών στο κείμενο οι οποίες υποδεικνύουν τη σωστή. Αν δηλαδή στο κείμενο που περιέχει το τοπωνύμιο υπάρχει η ονομασία μιας χώρας ή μιας μεγάλης διοικητικής μονάδας, τότε το πιο πιθανό είναι το σωστό τοπωνύμιο να είναι αυτό που περιέχεται στον κατάλογο της συγκεκριμένης χώρας. Αν υπάρχουν πάνω από μια τέτοιες αναφορές τότε λαμβάνεται υπόψη αυτή που βρίσκεται πιο κοντά της μέσα στο κείμενο. Αν υπάρχουν πάνω από μια ονομασίες στην ίδια χώρα χωρίς άλλα στοιχεία τότε επιλέγεται εκείνη που βρίσκεται στην μεγαλύτερη βαθμίδα της διοικητικής διαίρεσης, σύμφωνα με την υπόθεση ότι αυτή είναι πιθανότερο να αναφερθεί χωρίς επιπρόσθετα διευκρινιστικά στοιχεία από μια μικρότερη. Η μελέτη των Pouliquen et al το 2006, έδειξε ότι αυτή η απλή τεχνική επιτυγχάνει σε ποσοστό που φτάνει το 76%. Άλλες μέθοδοι που στηρίζονται στην μηχανική μάθηση απαιτούν πολύ μεγάλο κόστος σε υπολογιστικές διαδικασίες για να αυξήσουν τελικά το ποσοστό επιτυχίας μόνο κατά 1%.

Ο πιο πάνω αλγόριθμος χρησιμοποιήθηκε στην δημιουργία του προγράμματος GDELT

(Global Data on Events, Location and Tone). Το πρόγραμμα αυτό ξεκίνησε το 2013 από τους Leetaru and Schrodt οι οποίοι έφτιαξαν ένα dataset με 200 εκατομμύρια εγγραφές γεγονότων που αναφέρθηκαν οπουδήποτε στον πλανήτη από το 1979 μέχρι και τότε μαζί με τις τοποθεσίες που συνέβησαν. Τα δεδομένα δημιουργήθηκαν από αναφορές στο διεθνή τύπο από πολλές και διάφορες διαδικτυακές πηγές. Σήμερα διαθέτει συστοιχίες δεδομένων που περιέχουν τρισεκατομμύρια σημεία με γεγονότα που συμβαίνουν οπουδήποτε στον πλανήτη και ανανεώνονται κάθε περίπου 15 λεπτά, σε 65 διαφορετικές γλώσσες προσφέροντας ελεύθερη και ανοικτή πρόσβαση σε αυτά, είτε με downloading είτε με on-line ανάλυση μέσω κατάλληλης πύλης που έχει αναπτυχθεί (http://www.gdeltproject.org/).

Οι Wei Zhang and Judith Gelernter το 2014, προσπάθησαν να γεωκωδικοποιήσουν μηνύματα από το Twitter στην Αγγλική γλώσσα. Το βασικό πρόβλημα που αντιμετώπισαν ήταν το μικρό μέγεθος του μηνύματος το οποίο δεν περιείχε αρκετό πλήθος πληροφοριών ώστε να μπορεί να χρησιμοποιηθεί για την αποσαφήνιση της ονομασίας της τοποθεσίας που εντόπιζαν. Ένα επιπρόσθετο πρόβλημα σε αυτή την περίπτωση είναι και η ελεύθερη γραφή χωρίς κάποια δομημένη σύνταξη που ακολουθείται από τους συγγραφείς σε τέτοια σύντομα μηνύματα στα μέσα κοινωνικής δικτύωσης. Για να επιλύσουν το πρόβλημα της αποσαφήνισης χρησιμοποίησαν στοιχεία από άλλες τοποθεσίες που αναφέρονται στα μηνύματα, καθώς και στοιχεία που περιέχονται στους γεωγραφικούς καταλόγους. Με αυτό τον τρόπο “έχτισαν’ αυτό που ονομάζεται ανάλυση τοπωνυμίου (toponym resolution). Τα στοιχεία αυτά συνδυάστηκαν με στοιχεία που έχουν χωρική αναφορά όπως είναι η τοποθεσία του χρήστη, η ζώνη ώρας, η περιγραφή του χρήστη(το προφίλ) και οι συντεταγμένες που δείχνουν το από που δημοσιεύτηκε το μήνυμα και το συνοδεύουν κατά τη δημοσίευση. Ο αλγόριθμος που ανέπτυξαν ήταν σε θέση να ανακαλεί από τα μεταδεδομένα μόνο τις χρήσιμες ανά περίπτωση συμπληρωματικές πληροφορίες και να αγνοεί όσες οδηγούσαν σε αντιθέσεις.

Παρότι ο Buscaldi αναγνωρίζει τρεις γενικά μεθόδους για το “χτίσιμο” της ανάλυσης ενός τοπωνυμίου για την αποσαφήνιση του, οι συγγραφείς προσθέτουν μια παραπάνω (No3):

1 Με επεκτάσεις δεδομένων. Δηλαδή αντιστοίχηση με άλλα σύνολα δεδομένων ή κείμενα τα οποία συσχετίζονται με το προς εξέταση κείμενο και περιέχουν στοιχεία που βοηθούν στην αποσαφήνιση. Τα στοιχεία αυτά μπορεί να είναι εξωτερικές βάσεις γνώσης ή δεδομένα που ενυπάρχουν στο κείμενο. Στην περίπτωση των Tweets μπορούν να χρησιμοποιηθούν για παράδειγμα τα σχόλια άλλων χρηστών σε ένα μήνυμα ή άλλα μηνύματα του ίδιου χρήστη είτε τα μεταδεδομένα που συνοδεύουν το μήνυμα.

2 Με ανάπτυξη γλωσσικών μοντέλων. Δηλαδή μια βάση γνώσης όπου έχουν υπολογιστεί οι συχνότητες εμφάνισης συγκεκριμένων τοποθεσιών που αναγράφονται ακολουθώντας συγκεκριμένο τρόπο ή μορφή και συνεπώς είναι δυνατό να αναγνωριστεί ταύτιση μεταξύ τους.

3 Με ευρετικά μέσα και γραφικές μεθόδους. Με συγκεκριμένες μεθόδους που έχουν δοκιμαστεί από την επιστημονική κοινότητα για συγκεκριμένες περιπτώσεις και εμπλέκουν στοιχεία όπως ο πληθυσμός, η θέση στη διοικητική ιεραρχία, η ελαχιστοποίηση της απόστασης μεταξύ των τοπωνυμίων που αλιεύονται μέσα από ένα κείμενο.

4 Τεχνικές μηχανικής μάθησης. Υπολογιστικές τεχνικές όπου αλγόριθμοι προσαρμόζονται και εκπαιδεύονται κατά τη διάρκεια μιας διαδικασίας και δεν ακολουθούν αυστηρά καθορισμένες προγραμματιστικές ενέργειες. Στην περίπτωση των συγγραφέων προσδιορίστηκαν από τους ίδιους οι συντεταγμένες μιας ομάδας αντιπροσωπευτικών μηνυμάτων και το σετ δεδομένων αυτό αποτέλεσε εκπαιδευτικό υλικό για τη διαδικασία που ακολουθήθηκε.

Χρησιμοποίησαν κανόνες και εκπαίδευσαν έναν αλγόριθμο με δεδομένα από Tweets ώστε με Natural Language Processing να αναγνωρίσουν ποιες λέξεις μέσα σε ένα μήνυμα είναι τοπωνύμιο, χρησιμοποιώντας γνώση από τον γεωγραφικό κατάλογο (gazetteer). Χρησιμοποίησαν από τον κατάλογο Geonames την πληροφορία του πληθυσμού θεωρώντας ότι οι περιοχές που κατοικούνται περισσότερο είναι πιο πιθανό να αναφερθούν, καθώς επίσης το πλήθος των διαφορετικών ονομασιών με τις οποίες είναι γνωστό ένα μέρος κάνοντας την υπόθεση ότι τα πιο γνωστά μέρη έχουν περισσότερες ονομασίες. Έπειτα με τεχνικές επιβλεπόμενης μηχανικής μάθησης απέδωσαν βάρη στα διάφορα πεδία ενός μηνύματος στο Twitter και στις καταγραφές παγκόσμιων γεωγραφικών καταλόγων ώστε να προτιμηθεί μια τοποθεσία από τον κατάλογο και να αντιστοιχηθεί με την υποψήφια τοποθεσία του μηνύματος. Πέτυχαν, όπως οι ίδιοι ισχυρίζονται, καλύτερα αποτελέσματα από τους υψηλής τεχνολογίας ανταγωνιστές, με τον αλγόριθμο που ανέπτυξαν, παρότι όπως οι ίδιοι αναφέρουν η ταύτιση μιας τοποθεσίας από τον κατάλογο με έναν υποψήφιο από το μήνυμα δεν ήταν πάντα η σωστή.

Οι Moncla et al το 2014 ανέπτυξαν έναν αλγόριθμο ο οποίος με μη επιβλεπόμενο τρόπο μπορούσε να αποσαφηνίσει ονομασίες τοποθεσιών που εντόπιζε και περιλαμβάνονται στους gazetteers καθώς και να εκτιμήσει τη θέση άλλων ονομασιών, πολύ μεγάλης ανάλυσης (fine grain toponyms), που δεν περιλαμβάνονται σε αυτούς, με τη βοήθεια συστάδων που δημιούργησαν από τα στοιχεία που συνοδεύουν τα τοπωνύμια στους καταλόγους. Δοκίμασαν την τεχνική τους σε οδηγίες περιγραφών πεζοπορίας στην Γαλλική, Ισπανική και Ιταλική γλώσσα για μια μικρή περιοχή. Οι περιγραφές ήταν κείμενα που περιέγραφαν οδηγίες μετακινήσης περιηγητών στο χώρο και οι οποίες χρησιμοποιούσαν τοπωνύμια, χωρικές σχέσεις, μικρές φυσικές μορφές όπως λίμνη, ή κατασκευές όπως εκκλησία, γέφυρα, αγροικία κλπ που έχουν μεν χωρική αναφορά αλλά δεν υπάρχουν στους gazetteers και μπορούν να θεωρηθούν ως συνοδευτικά ενός τοπωνυμίου. Οι περιγραφές συνοδεύονταν από καταγεγραμμένα ίχνη συντεταγμένων GPS. Για την εκτίμηση της ποιότητας των αποτελεσμάτων επιλέχθηκαν 30 περιγραφές για κάθε γλώσσα και αναλύθηκαν από τους ερευνητές για να δημιουργηθεί ένα σύνολο δεδομένων αναφοράς.

Αξιοποίησαν με τη βοήθεια τεχνικών NLP την αναγνώριση του μέρους του λόγου που ανήκει κάθε λέξη στο κείμενο. Έπειτα με μια αλληλουχία μετατροπών με τη βοήθεια συντακτικών και σημασιολογικών προτύπων αναγνωρίστηκαν οι λέξεις που καταδεικνύουν τοπωνύμιο και χωρική σχέση. Για το σκοπό αυτό χρησιμοποιήθηκαν λεξικά και λέξεις από τοπικά γραμματικά σύνολα. Οι ερευνητές στο σημείο αυτό δημιούργησαν μια νέα κατηγορία που την ονόμασαν διευρυμένη χωρική ονομαστική οντότητα. Αυτή ήταν μια οντότητα η οποία δημιουργήθηκε από την ονομασία ενός τοπωνυμίου η οποία συσχετίστηκε με μια ή περισσότερες έννοιες που χρησιμοποιούνται για να εκφραστεί αυτή στη γλώσσα που εξετάζεται. Η διαδικασία ήταν σε θέση να ανιχνεύσει σε μια περιγραφή το τοπωνύμιο, τη χωρική σχέση καθώς και ρήματα που υποδηλώνουν κίνηση ώστε αυτά να συνδέονται με το τοπωνύμιο ή τη διευρυμένη χωρική ονομαστική οντότητα που αναγνωριζόταν στο κείμενο.

Αντιμετώπισαν τρία είδη ασάφειας. Τη δομική ασάφεια η οποία προκύπτει όταν η έκφραση μιας ονομασίας περιέχει μια υποκατηγορία μιας κύρια ονομασίας και στον κατάλογο περιέχεται μόνο η κύρια. Για να το επιλύσουν χρησιμοποίησαν τα μεταδεδομένα που περιέχονται στους καταλόγους και προσδιορίζουν τις κατηγορίες. Την ασάφεια αναφοράς, όταν στον κατάλογο περιέχονται πάνω από μια εγγραφές με το ίδιο όνομα και κατηγορία. Για την αντιμετώπισή της χρησιμοποίησαν έναν αλγόριθμο (DBSCAN) ο οποίος ομαδοποιεί τις τοποθεσίες σε γειτονιές με βάση την πυκνότητά τους στο χώρο (ελάχιστη περιοχή και αριθμός σημείων) και ο οποίος αντιστοιχίζει σε κάθε εγγραφή του καταλόγου την ένδειξη της γειτονιάς. Έτσι προσδιόρισαν τις γειτονιές των σημείων και άρα των ονομασιών, οι οποίες είναι περισσότερο πιθανό να ανήκουν στα μονοπάτια πεζοπορίας. Την ασάφεια μη αναφοράς της τοποθεσίας, όταν δηλαδή η ονομασία είναι υποπεριοχή μιας άλλης και δεν υπάρχει στον κατάλογο. Στην περίπτωση αυτή οι ερευνητές συμπέραναν την θέση από την περιοχή που σχηματίζουν οι προηγούμενες τοποθεσίες που είχαν αποσαφηνίσει, καθώς και άλλες χωρικές πληροφορίες που βρίσκονται μέσα στην περιγραφή. Στην περίπτωση που δεν υπήρχε άλλη πληροφορία η περιοχή αυτή ήταν η γειτονιά των τοπωνυμίων της προηγούμενης διαδικασίας, όταν υπήρχε χωρική ένδειξη όπως νότια, η περιοχή περιοριζόταν ως υποπεριοχή της προηγούμενης, ενώ όταν η χωρική πληροφορία ήταν περισσότερη αναπτύχθηκαν βήματα για να σημειώνουν χωρικές συσχετίσεις όπως αποστάσεις και τοπολογικές σχέσεις, ώστε αυτές τελικά να μεταφραστούν σε συντεταγμένες.

Για να συμπεράνουν-υπολογίσουν τη θέση των τοπωνυμίων υψηλής ανάλυσης προτείνουν την αξιοποίηση του ελάχιστου κυρτού πολυγώνου και του περιγεγραμμένου κύκλου στα σημεία του περιβάλλοντος παραλληλόγραμμου που σχηματίζονται από το νέφος των σημείων της καλύτερης συστάδας για κάθε διαδρομή. Αφού προσδιόρισαν τις πραγματικές θέσεις των τοπωνυμίων που δεν περιέχονται στους καταλόγους, διαπίστωσαν ότι η πλειοψηφία αυτών περιέχεται στον περιγεγραμμένο κύκλο. Χρησιμοποίησαν κυρίως τοπικούς γεωγραφικούς καταλόγους για κάθε γλώσσα και συμπληρωματικά τον Geonames και OSM και διαπίστωσαν ότι οι 2 τελευταίοι καλύπτουν τα κενά των τοπικών. Διαπί-

στωσαν ότι το 45-70% των τοπωνυμίων που εντόπισαν είχε πάνω από μια εγγραφές στους καταλόγους με μέσο όρο 15-20 εγγραφές ανά τοπωνύμιο. Το πρόβλημα αυτό αντιμετωπίστηκε με τις συστάδες που δημιουργήθηκαν και επιλέχθηκαν οι πιο κοντινές (βέλτιστες) στα νέφη των σημείων των διαδρομών από τις καταγραφές των GPS. Τα τοπωνύμια που αλιεύονταν από τις περιγραφές και δεν υπήρχαν στις χωρικές συστάδες σχετίζονταν με περιγραφές θέσεων υψηλής ανάλυσης, δηλαδή υποκατηγορίες των βασικών τοπωνυμιών, όπως πχ βόρεια από το ρέμα κλπ.

Οι Angel et. al. το 2008, ανέπτυξαν αλγορίθμους ικανούς να αναγνωρίσουν και να γεωκωδικοποιήσουν αριθμούς τηλεφώνων, διευθύνσεις και τοπωνύμια. Κατασκεύασαν μια επέκταση (extension) η οποία δημιουργούσε μια χαρτογραφική διεπαφή στον web browser Firefox και επέτρεπε στο χρήστη να δει και να διορθώσει το αποτέλεσμα της αυτόματης γεωκωδικοποίησης την οποία εκτελούσε η εφαρμογή στο κείμενο της ιστοσελίδας που άνοιγε, καθώς και να προσθέσει ότι δεν κατόρθωνε να εντοπίσει ο αλγόριθμος. Δημιουργούσε δηλαδή η εφαρμογή επισήμανση του εντοπιζόμενου τοπωνυμίου στο κείμενο, τοποθετούσε ένα σημείο στο χάρτη για κάθε ένα από αυτά και ο χρήστης είχε τη δυνατότητα να δει και τα 2 ταυτόχρονα και να αποφασίσει αν θα επέμβει. Πρότειναν τη δημιουργία ενός κεντρικού αποθετηρίου στο οποίο θα αποθηκεύονταν οι διορθώσεις αυτές και θα τις αξιοποιεί ο αλγόριθμος για τη γεωκωδικοποίηση. Με αυτό τον τρόπο, την ανθρώπινη παρέμβαση, θέλησαν να αντιμετωπίσουν το σύνθετο πρόβλημα της αποσαφήνισης που προκύπτει κατά τη διαδικασία μιας αυτόματης γεωκωδικοποίησης. Το δοκίμασαν για την Ελληνική γλώσσα στην Ελληνική έκδοση της Wikipedia. Επέλεξαν τη Wikipedia επειδή οι σελίδες της ανανεώνονται αργά και θα συνεχίσουν να υπάρχουν στο διηνεκές.

Αρχικά μετέτρεψαν τις ελληνικές λέξεις στις φωνητικά ισοδύναμες στην Αγγλική. Έπειτα κανονικοποίησαν τα δεδομένα σε μια ομοιόμορφη κανονική σύνταξη διεύθυνσης ή τοπωνυμίου. Για την αναγνώριση των υποψήφιων τοποθεσιών χρησιμοποίησαν τις εγγραφές διαφόρων καταλόγων τοπικών και διεθνών από διάφορες πηγές. Δεδομένου ότι κατά την εποχή σύνταξης της εργασίας δεν υπήρχαν δεδομένα με ελληνικούς χαρακτήρες, όλες οι λεκτικές περιγραφές από τα κείμενα και τους καταλόγους μετασχηματίστηκαν χρησιμοποιώντας φωνητικά κλειδιά κατακερματισμού (hash keys) σε πιο συμπυκνωμένη μορφή για εξοικονόμηση υπολογιστικών πόρων. Μετά τη σύγκριση κάθε υποψηφίου με τον κατάλογο επιστρεφόταν ένα μικρό σύνολο εγγραφών από το οποίο έπρεπε να επιλεγεί η σωστή αντιστοίχηση. Για το σκοπό αυτό χρησιμοποίησαν την απόσταση Levenshtein. Η απόσταση αυτή μεταξύ 2 λέξεων είναι το πλήθος των χαρακτήρων που πρέπει να προσθέσουμε, να αφαιρέσουμε ή να αλλάξουμε προκειμένου να πάρουμε απόλυτη ταύτιση. Θέτωντας το κατάλληλο κατώφλι απόστασης ο αλγόριθμος είναι σε θέση να επιλέξει την καλύτερη ταύτιση μεταξύ κανονικοποιημένης και φωνητικά ισοδύναμης λέξης με την εγγραφή στον κατάλογο. Μετά τη σύγκριση δηλαδή έπαιρναν μια ταξινομημένη λίστα με τις αποστάσεις Levenshtein και επέλεγαν την πρώτη με τη μικρότερη απόσταση. Η μέθοδος αυτή δεν αντιμετωπίζει την αποσαφήνιση και γιαυτό ενσωματώθηκε η δυνατότητα της ανθρώπινης παρέμβασης.

Από τις πιο πάνω εργασίες γίνεται κατανοητό πως το ζήτημα της αναγνώρισης της τοποθεσίας που αναφέρεται στο σώμα ενός κειμένου και η αντιστοίχηση της με το σωστό ζεύγος συντεταγμένων του φυσικού χώρου, είναι πολυσύνθετο, απαιτεί αλληλουχία διαφορετικών διαδικασιών για κάθε περίπτωση και αποτελεί μια ερευνητική πρόκληση στις μέρες μας.

Προσπάθειες δημιουργίας ολοκληρωμένων geoparsers, όπως αυτή της πλατφόρμας https://geoparser.io/, επιστρέφει αξιοποιήσιμα αποτελέσματα μόνο για μεγάλες γεωγραφικές οντότητες. Άλλοι αυτόματοι geoparsers που δοκιμάστηκαν από τους Gritta et.al. το 2017, παρουσιάζουν σειρά περιορισμών που καθιστούν τη χρήση τους συμπληρωματική σε διαδικασίες όπως αυτές που παρουσιάστηκαν παραπάνω και όχι αποκλειστική. Το ζήτημα της αυτόματης και μη επιβλεπόμενης αναγνώρισης και μετατροπής ονομασιών τοποθεσίας σε ζεύγη συντεταγμένων του πραγματικού χώρου με ακρίβεια που δεν αμφισβητείται, παραμένει ανοικτό κατά τη διάρκεια συγγραφής της παρούσας εργασιας.

Κεφάλαιο 2

Δεδομένα - επεξεργασία

2.1 Περιγραφή δεδομένων

Η πλατφόρμα GEOTHNK είναι μια συνεργατική ευρωπαϊκή πρωτοβουλία με στόχο την ενίσχυση της χωρικής σκέψης στη δευτεροβάθμια εκπαίδευση. Η χωρική σκέψη θεωρείται βασική προϋπόθεση για κάποιο μαθητή ώστε να προχωρήσει με επιτυχία σε τομείς της επιστήμης που σχετίζονται με την τεχνολογία, τη μηχανική και τα μαθηματικά και ορίζεται από τη σύνθεση τριών βασικών συστατικών, τις έννοιες του χώρου, τα εργαλεία αναπαράστασης και τις συλλογιστικές διαδικασίες (Kavouras et al., 2014). Φορείς και εξειδικευμένοι ερευνητές από διάφορες χώρες της Ευρώπης έχουν συνεισφέρει σε αυτήν εκπαιδευτικά σενάρια δημιουργώντας έτσι μια πολύτιμη δεξαμενή πόρων, ένα δίκτυο μονοπατιών μάθησης, προς την κατεύθυνση της ενίσχυσης της χωρικής σκέψης μέσα από καινοτόμες εκπαιδευτικές διαδικασίες και πρακτικές.

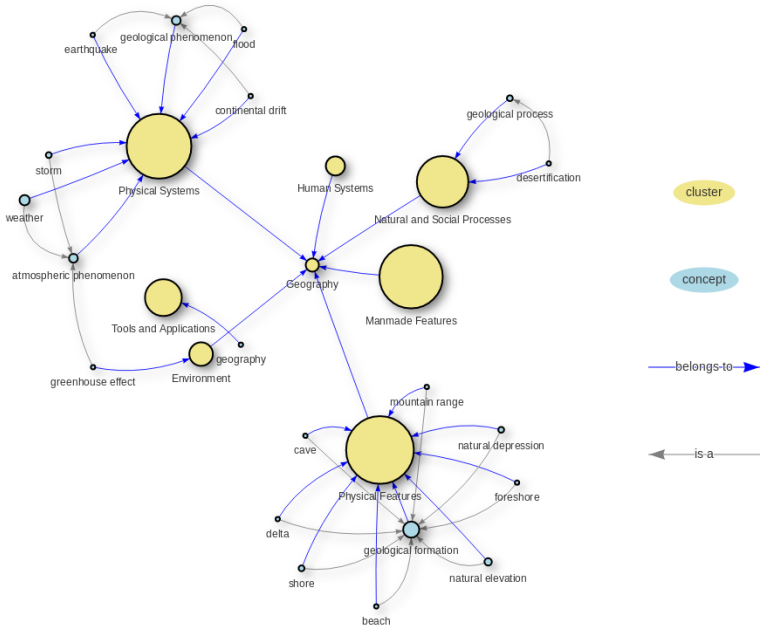
Από το αποθετήριο της πλατφόρμας[1], αντλήθηκαν συνολικά 159 μοναδικά εκπαιδευτικά σενάρια σε portable document format (.pdf) τον Οκτώβρη του 2017. Από το ίδιο αποθετήριο αντλήθηκαν και δεδομένα που συνοδεύουν και χαρακτηρίζουν κάθε σενάριο, όπως ο τίτλος, η γλώσσα στην οποία είναι γραμμένο, οι λέξεις κλειδιά, τα επιστημονικά πεδία ενδιαφέροντος, οι τοποθεσίες και οι έννοιες του χώρου στις οποίες αυτό αναφέρεται όπως επίσης και ο σύνδεσμος (uri) του τεκμηρίου στο αποθετήριο με σκοπό να ενσωματωθούν στο εργαλείο που αναπτύχθηκε. Τμήμα των δεδομένων που καταγράφηκαν φαίνεται στον πίνακα 2.1.

Το κείμενο που περιείχαν όλα τα pdf ανακτήθηκε σε μορφή string χρησιμοποιώντας τη βιβλιοθήκη Poppler από τη γλώσσα C++. Τα σενάρια ήταν γραμμένα σε 6 διαφορετικές γλώσσες, Ελληνικά, Αγγλικά, Γερμανικά, Ολλανδικά, Ρουμάνικα και Βουλγάρικα, με όλα να περιέχουν Αγγλικούς όρους. Πολλά από αυτά ήταν γραμμένα σε δυο γλώσσες. Το σύνολο των κειμένων καθώς και ο πίνακας με τα συνοδευτικά δεδομένα που συλλέχθηκαν, αποτέλεσαν 2 από τα 3 σύνολα δεδομένων εισόδου για τη συνέχεια της διαδικασίας επεξεργασίας.

[1] http://portal.opendiscoveryspace.eu/community/geothink-community-400866

aa	123
url	http://portal.opendiscoveryspace.eu/sites/default/files/authoring_tool_uploads/attachments/840438/edu_obj_840438.pdf
filename	edu_obj_840438.pdf
title	Σεισμική και ηφαιστειακή δραστηριότητα ως γεωδυναμικά φαινόμενα: η χωρική κατανομή τους
titleng	Seismic and volcanic activities as geodynamic phenomenons: their spatial distribution
keywords	earthquake, volcano, seismic activity, volcanic activity, geothermical phenomenon, tectonic plates
domain	seismology, Earthquakes, Global distribution of tectonic activity, Tectonic processes, Volcanoes
geo.instance	Vesuvius, Hellenic Republic
geo.concept	earthquake, tsunami, disaster, volcano, country, scale factor, Earth's crust, plate tectonics, class, map, Representation
language	Greek, English

Πίνακας 2.1: Δεδομένα που συνοδεύουν κάθε σενάριο



Σχήμα 2.1: network subset sample

Το τρίτο σύνολο δεδομένων (dataset) είναι το δίκτυο των εννοιών που αφορούν το χώρο που έχει αναπτυχθεί από τους Kavouras et al.(Kavouras et al., 2017) και αποτελείται από 327 χωρικές έννοιες μαζί με τους ορισμούς τους που αντλήθηκαν από το δίκτυο Wordnet. Η πληροφορία έχει οργανωθεί σε 3 επίπεδα ιεραρχίας σε 15 συστάδες με συνολικά 802 ταξονομικές σχέσεις να τα συνδέουν μετατρέποντας έτσι το σημασιολογικό δίκτυο σε μια οντολογία (σχήμα 2.1).

2.2 Λογισμικό - πλατφόρμα εργασίας

Για τη συνολική διαδικασία της επεξεργασίας, οργάνωσης και παρουσίασης των δεδομένων χρησιμοποιήθηκε η γλώσσα προγραμματισμού R μέσα από την πλατφόρμα εργασίας RStudio σε περιβάλλον GNU/Linux 64bit. Η R και το RStudio ανήκουν στην κατηγορία των λογισμικών ανοικτού κώδικα και ως τέτοια μπορούν να εγκατασταθούν σε οποιαδή-

ποτε πλατφόρμα λειτουργικού συστήματος (Windows, Linux, Mac κα). Επιπρόσθετα είναι δωρεάν, διαθέτουν μεγάλη κοινότητα χρηστών για αναζήτηση βοήθειας κατά τη διαδικασία εκμάθησης, καθώς επίσης αναπτύσσονται συνεχώς ενσωματώνοντας κάθε νέα επιστημονική γνώση και τεχνική που παράγεται από την επιστημονική κοινότητα αλλά και αυτή των χρηστών.

Η R διαθέτει ένα εκτενές οικοσύστημα πακέτων λογισμικού, 12.162 την 17/2/2018 [2], που καλύπτουν ένα ευρύ φάσμα εφαρμογών και προγραμματιστικών διαδικασιών, όπως επίσης παρέχει διεπαφές σε πολλούς δημοφιλείς και υψηλών δυνατοτήτων και απόδοσης αλγόριθμους άλλων γλωσσών που χρησιμοποιούνται ευρέως στον στατιστικό προγραμματισμό (keras, tensorflow, κα). Αν συνυπολογιστεί η υψηλής ποιότητας παραγωγή γραφικών (base, lattice, ggplot2), η δυνατότητα κατασκευής διαδικτυακών διαδραστικών εφαρμογών (shiny) αλλά και τεχνικών εκθέσεων (Rmarkdown) από την ίδια γραμμή εκτέλεσης εντολών, κάνουν το συνδυασμό R και RStudio το μοναδικό ίσως δωρεάν εργαλείο-πλατφόρμα εργασίας, που διατίθεται στην επιστημονική κοινότητα και συνδυάζει όλα τα παραπάνω και μάλιστα σε τόσο υψηλό βαθμό ποιότητας (ανάλυσης-παρουσίασης).

Σημαντική προσπάθεια καταβλήθηκε κατά την διαδικασία επεξεργασίας και παραγωγής των δεδομένων, ώστε να κατασκευαστούν, κατά το δυνατόν, συνεχείς αλληλουχίες προγραμματιστικών βημάτων με σκοπό την αξιοποίησή τους σε μελλοντικές παρόμοιες εργασίες ανάλυσης.

2.3 Επεξεργασία

2.3.1 Δοκιμές λογισμικού Annotation - Named Entity Recognition

Το στάδιο αυτό περιλαμβάνει τη γλωσσολογική ανάλυση και προεπεξεργασία του κειμένου και πιο συγκεκριμένα τις διαδικασίες tokenization, lemmantization και part of speech tagging που περιγράφηκαν σε προηγούμενο κεφάλαιο.

Για την εκτέλεση των διαδικασιών χρησιμοποιήθηκαν τρεις διαφορετικοί αλγόριθμοι από τρία διαφορετικά πακέτα λογισμικού. Αυτά είναι το coreNLP του Stanford Univercity που παρέχει γλωσσικά μοντέλα για την Αγγλική, τη Γερμανική, τη Γαλλική, την Ισπανική και την Κινεζική γλώσσα, το openNLP του Apache Foundation που διαθέτει μοντέλα για την Αγγλική, την Γερμανική, την Ισπανική και την Ολλανδική και το spacy το οποίο έχει τη φήμη του πιο γρήγορου και δυνατού parser στην αγορά και το οποίο προορίζεται για βιομηχανική χρήση στην αγορά λογισμικού και παρέχει πακέτα για την Αγγλική, τη Γαλλική, τη Γερμανική, την Ολλανδική, την Ισπανική, την Πορτογαλική και την Ιταλική. Τα δυο πρώτα πακέτα είναι γραμμένα σε γλώσσα Java και το τρίτο σε συνδυασμό των γλωσσών python και Cython. Από την δοκιμή τους πράγματι ο spacy είναι ο γρηγορότερος με διαφορά και το coreNLP είναι το αργότερο και μάλιστα το πιο απαιτητικό σε κατανάλωση υπολογιστικών πόρων (μνήμη, επεξεργαστική ισχύ). Δοκιμάζοντας την απόδοση

[2] https://cran.r-project.org/web/packages/

και των τριών παραπάνω αλγορίθμων και μοντέλων, μέσα από τις αντίστοιχες διεπαφές που προσφέρονται για τη γλώσσα R, σε ένα βιβλίο στην Αγγλική 467 σελίδων, το coreNLP χρειάστηκε 1 ώρα και 20 λεπτά και ο Spacy 9.5 λεπτά της ώρας σε μηχάνημα με επεξεργαστή Intel Core2 Duo CPU E8400 3.00GHz και μνήμη RAM 5.7GiB. Το openNLP στη συγκεκριμένη δοκιμή απέτυχε να επιστρέψει αποτέλεσμα.

Στη συνέχεια ακολούθησε η δοκιμή των παραπάνω αλγορίθμων στην εξόρυξη ονομασιών τοποθεσίας από το σώμα κειμένου, σε δυο περιλήψεις εργασιών, από τις οποίες η πρώτη περιέχει ονόματα μεγάλων γεωγραφικών οντοτήτων, όπως ονομασίες κρατών και η δεύτερη μικρότερων, όπως όνομα ελληνικού νησιού και περιοχής πάνω σε αυτό. Για το κείμενο που ακολουθεί τα αποτελέσματα φαίνονται στον πίνακα 2.2.

Εργασία 1.
Keil, Emily J, (2014), Phylogeography of Batrachospermum gelatinosum (Batrachospermales, Rhodophyta) in Europe, Ohio University Summary:

- The freshwater red alga, *Batrachospermum gelatinosum* (L.) DC., inhabits streams of Europe and North America having been collected frequently on both continents. A study of this species showed evidence of a glacial refugium in the southeastern US with little genetic variation throughout its more northern range in eastern North America. This study was initiated to investigate its phylogeography throughout Europe. Specimens were collected from Belgium, Estonia, Finland, France, Great Britain, Italy, Latvia, Lithuania, Poland and Spain. Of the 70 individuals analyzed, there were 12 cox1 haplotypes. In addition, ITS variation of 68 individuals was surveyed and showed 22 haplotypes. The haplotype network of cox1 data showed 54 individuals distributed among three common haplotypes. The other nine haplotypes only differed from the commons ones by 1-2 base pair and were represented by 1-5 individuals. For the ITS data, the network had a star appearance with common haplotype (16 individuals) and many closely related haplotypes with few individuals per haplotype. Compared to North America, there are more haplotypes present in Europe and the relationship among haplotypes is more complex. The geographic distribution of haplotypes did not appear to follow a glaciation pattern, but rather the common haplotypes were widely spread suggesting a recent expansion.

software	**locations found**
coreNLP	DC., Europe, North America, US, Belgium, Estonia, Finland, France, Great Britain, Italy, Latvia, Lithuania, Poland, Spain
spacy	L.) DC, Europe, North America, US, Belgium, Estonia, Finland, France, Great Britain, Italy, Latvia, Lithuania, Poland, Spain
openNLP	Europe, North America, Belgium, Estonia, Finland, France, Britain, Italy, Latvia, Lithuania, Poland, Spain

Πίνακας 2.2: αποτελέσματα NER για το δοκιμαστικό κείμενο 1

Για το δοκιμαστικό κείμενο που ακολουθεί και περιέχει τοπωνύμια μικρότερων γεωγραφικών οντοτήτων τα αποτελέσματα φαίνονται στον πίνακα 2.3.

Εργασία 2.

Abell, Natalie D., (2014), Reconsidering a Cultural Crossroads: A Diachronic Analysis of Ceramic Production, Consumption, and Exchange Patterns at Bronze Age Ayia Irini, Kea, Greece, University of Cincinnati. Summary:

- Although studies of exchange and interaction in the prehistoric Cyclades have become increasingly theoretically sophisticated, the mechanisms through which exchange occurred are still not fully understood, and diachronic analyses that emphasize variability between Cycladic communities are rare. This dissertation provides a new perspective, employing a micro-level approach that is focused on the details of stratigraphy, architecture, ceramics, and other objects in Area B at Ayia Irini on Kea, in order to reconsider published deposits and assess detailed questions related to changing production, consumption, and exchange patterns in this island community over time. I argue that the distinctive, multicultural nature of the ceramic assemblage at Ayia Irini is more than the result of its position as a hub between networks, connecting culturally distinct regions to each other and the metal deposits of Lavrion in Attica. Ayia Irini was also a physical locus of exchange, where people of different cultural backgrounds interacted with each other and with local residents, and where non-local people, including craftspeople, were integrated into the local community. Although participation in exotic or elite drinking or eating practices may have formed part of internal negotiations over status, it is also probable that diverse ceramic shapes in circulation at Ayia Irini were used in drinking or eating activities that connected locals and non-locals in shared practices, familiar to both groups. Participation in such activities would have reinforced social bonds between local and non-local people and enabled the development and strengthening of personal relationships, an important precondition for most types of preindustrial exchange. At Ayia Irini, a community whose entire existence seems to have been predicated on participation in exchange networks, creating and maintaining such social bonds was vital, and key to the longevity and prosperity of the settlement.

software	**locations found**
coreNLP	Cyclades, Ayia Irini, Kea, Lavrion, Attica
spacy	-
openNLP	-

Πίνακας 2.3: αποτελέσματα NER για το δοκιμαστικό κείμενο 2

Από τα παραπάνω παρατηρούμε πως τις μεγάλες γεωγραφικές οντότητες, όπως τα ονόματα κρατών, το coreNLP και ο spacy δεν έχουν κανένα πρόβλημα να τις εντοπίσουν

και επιστρέφουν σχεδόν όμοια αποτελέσματα. Αντίθετα το openNLP αποτυγχάνει να εντοπίσει την ονομασία Great Britain και βρίσκει μόνο το Britain όπως επίσης αποτυγχάνει σε συντομογραφίες όπως το U.S. Όταν όμως τα κείμενα περιέχουν μικρότερες γεωγραφικές οντότητες, όπως στη δεύτερη περίπτωση, φαίνεται ξεκάθαρα η υπεροχή των γλωσσικών και συντακτικών μοντέλων στα οποία έχει εκπαιδευτεί ο Conditional Random Field (CRF) tagger του coreNLP ο οποίος εντοπίζει τις οντότητες τη στιγμή που οι άλλοι δυο αποτυγχάνουν πλήρως.

Έτσι προτιμήθηκε για την συνέχεια της επεξεργασίας το λογισμικό του Stanford University, καθώς η δοκιμή έδειξε πως βρίσκεται σε ένα στάδιο μεγαλύτερης ωριμότητας από τους άλλους δυο, για τη συγκεκριμένη ενέργεια της εξαγωγής των τοπωνυμίων.

2.3.2 Εξαγωγή τοπωνυμίων

Η εφαρμογή του αλγορίθμου, που προκρίθηκε από την δοκιμή που προηγήθηκε, πέτυχε την αναγνώριση συνολικά 202 μοναδικών ονομασιών σε 80 από τα 159 σενάρια που συλλέχθηκαν από την πλατφόρμα του GEOTHNK. Η διαδικασία αναγνώρισε λανθασμένα 30 οντότητες ως τοποθεσία, όπως ”Description”, ”Google Earth”, ”Longitude”, ”Power Point”, ”Greek”, ”This”, ”N”, ”WGS”, ”Forest Lesson”, ”Step2”, ”PLANTELE”, ”Toate”, ”Homer’s Odyssey”, ”Tasiouli Georgia”, που είναι το όνομα ενός από τους συγγραφείς και ”Marin Sorescu”, ονομασία που είναι Ρουμάνος ποιητής.

Επιπρόσθετα 12 ονομασίες αναγνωρίστηκαν σωστά ως ονομασίες τοποθεσίας αλλά χωρίς σημασία για το επόμενο στάδιο της γεωκωδικοποίησης, μιας και δεν περιλαμβάνονται στους γεωγραφικούς καταλόγους και δεν έχει πρακτική σημασία ο προσδιορισμός της θέσης τους στο χώρο. Στις οντότητες αυτές περιλαμβάνονται ονόματα πλανητών όπως, ”Jupiter”, ”Mars”, ”Pluto”, ”Cassiopeia” και ”Neptune”, ιστορικές τοποθεσίες άλλης εποχής όπως, ”Judea”, ”Iudeea”, ”Babylon”, and ”Pangaea”, ασαφείς ονομασίες τοποθεσίας όπως, ”Great Sea” και ”The City Center”, όπως επίσης και μια ονομασία φανταστικής τοποθεσίας από τον κόσμο των παραμυθιών όπως, ”Wonderland”.

Τελικά από τις 202 μοναδικές τοποθεσίες που εντοπίστηκαν, οι 160 επιλέχθηκε να προωθηθούν στο επόμενο στάδιο της γεωκωδικοποίησης ώστε να αντιστοιχηθούν με ζεύγη συντεταγμένων του πραγματικού γεωγραφικού χώρου.

2.3.3 Δοκιμές geocoders - γεωκωδικοποίηση

Για την αντιστοίχιση των τοπωνυμίων με ζεύγη συντεταγμένων επιλέχθηκε να διερευνηθούν και να αξιοποιηθούν οι δυνατότητες 3 βάσεων γνώσης που προσφέρονται ως διαδικτυακές υπηρεσίες (API) και είναι δυνατό να ενσωματωθούν σε προγραμματιστικές διαδικασίες. Χρησιμοποιήθηκαν τα web service του Google Maps (πακέτο ggmap), του OpenStreetMap.org (πακέτο nominatim) και η υπηρεσία που προσφέρει το Geonames.org (πακέτο geonames). Και τα τρία διαθέτουν διεπαφές για πρόσβαση και χειρισμό με τη γλώσσα R και απαιτούν προμήθεια κλειδιού. Η δωρεάν υπηρεσία της Google προσφέρει

2.500 αιτήματα ανά ημέρα, του OSM 15.000 ανά μήνα και του Geonames 30.000 ανά ημέρα και όχι πάνω από 2.000 ανά ώρα.

Και οι 3 παραπάνω υπηρεσίες δοκιμάστηκαν στο σύνολο των τοποθεσιών που διέθεσαν οι συγγραφείς κατά την υποβολή κάθε σεναρίου. Οι συγγραφείς προσέφεραν 64 μοναδικές τοποθεσίες για συνολικά 25 σενάρια. Στον πίνακα 2.4 που ακολουθεί φαίνεται ο βαθμός επιτυχίας κάθε υπηρεσίας. Επισημαίνεται ότι για την εκτίμηση της επιτυχίας κάθε υπηρεσίας λήφθηκε υπόψη μόνο ο πρώτος υποψήφιος της λίστας που επέστρεφε κάθε υπηρεσία σε κάθε αναζήτηση. Τα μεγάλα ποσοστά επιτυχίας που εμφανίζονται, οφείλονται στη δομημένη μορφή που είχαν οι τοποθεσίες κατά τη διάρκεια της χειροκίνητης υποβολής από τους συγγραφείς. Οι τρεις τοποθεσίες που απέτυχε η υπηρεσία του Geonames είναι οι ”Vesuvius”, ”Rhodes” και ”Ireland”, οι οποίες γεωκωδικοποιήθηκαν χειροκίνητα.

web servive	**Successful response**	**ratio**
Geonames	61	95,31%
Google Maps	58	90,63%
OSM	53	82,81%

Πίνακας 2.4: Εκτίμηση επιτυχίας geocoding APIs

Λόγω του μεγαλύτερου ποσοστού επιτυχίας που πέτυχε η υπηρεσία του Geonames, αυτή επιλέχθηκε έναντι των άλλων, να εφαρμοστεί και στο σύνολο των 160 τοπωνυμίων που προέκυψαν από τη διαδικασία Named Entity Recognition που περιγράφηκε πιο πάνω.

Ο κατάλογος Geonames περιέχει στοιχεία για πάνω από 10 εκατομμύρια ονομασίες και θέσεις σε παγκόσμιο επίπεδο κατηγοριοποιημένα σε 9 κλάσεις, και περιέχει τα στοιχεία από άλλους καταλόγους όπως ο GEOnet Names Server (GNS) της υπηρεσίας United States National Geospatial-Intelligence Agency, που περιέχει τοποθεσίες από όλες τις περιοχές του πλανήτη εκτός των ΗΠΑ και ο Geographic Names Information System (GNIS) της υπηρεσίας United States Geological Survey, που περιέχει τοποθεσίες μόνο εντός των περιοχών δικαιοδοσίας των ΗΠΑ.

Τα 160 τοπωνύμια προωθήθηκαν στην υπηρεσία και 131 ταυτίστηκαν επιτυχώς με τον πρώτο υποψήφιο της λίστας που επιστράφηκε, με το ποσοστό επιτυχίας να φτάνει το 81.88%. Για 14 από αυτά δεν επεστράφη ταύτιση με τον κατάλογο και αυτά είναι τα, ”Balkan Peninsula Natural Zones”, ”Burgas Lake”, ”Gheorgheni Lake”, ”Atanasovsko Lake”, ”Thission”, ”Aegean Archipelago”, ”Aegean Greece”, ”Apahida village”, ”Parthenon Gallery”, ”Cyclades island”, ”Nineveh Citadel”, ”Ninive Citadel”, ”Dionysou Aeropagitou”, ”Athenian Acropolis”. Επτά (7) από αυτά ταυτίστηκαν σωστά με το δεύτερο υποψήφιο της λίστας και αυτά είναι τα ”Spain”, ”Salisbury”, ”Ithaca”, ”France”, ”Andros”, ”Ireland” and ”Salamina”. Τα υπόλοιπα οκτώ (8) ταυτίστηκαν χειροκίνητα και αυτά είναι τα ”Central Athens”, ”UK”, ”Monastiraki”, ”Parthenon”, ”US”, ”Greenland”, ”Thera” and ”Greenwich”.

Το ποσοστό επιτυχίας που φτάνει το 80% και επιτυγχάνει η αυτόματη διαδικασία χωρίς καμία απολύτως διαδικασία αποσαφήνισης, καθιστούν το web service του καταλόγου

Geonames ένα πολύτιμο εργαλείο σε προγραμματιστικές διαδικασίες όπως στην παρούσα. Η κατανομή των θέσεων που παρείχαν οι συγγραφείς σε σύγκριση με αυτές που εξήχθησαν με την διαδικασία NER φαίνεται στους χάρτες του σχήματος 2.2 .

2.3.4 Εξαγωγή ιεραρχίας τοπωνυμίων

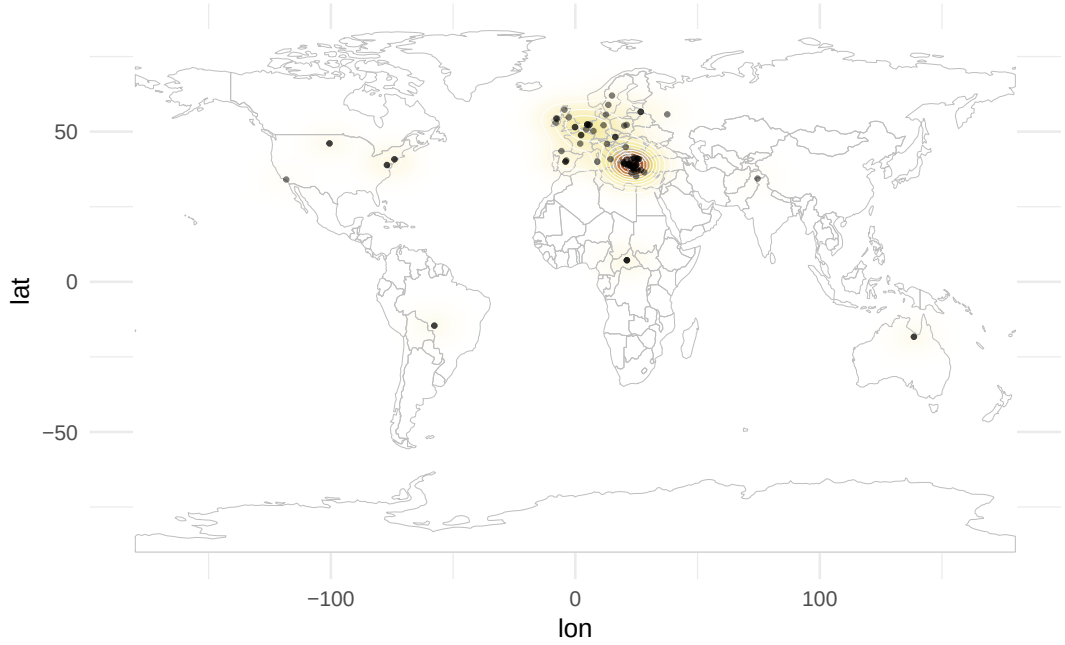
Ο γεωγραφικός κατάλογος Geonames, όπως προαναφέρθηκε είναι μια βάση γνώσης που προσφέρει για κάθε εγγραφή του (θέση), επιπρόσθετες πληροφορίες πέραν του ζεύγους συντεταγμένων, όπως υψόμετρο, την κλάση της οντότητας, τον πληθυσμό αν πρόκειται για πόλη ή οικισμό, τη χώρα στην οποία αυτή ανήκει και άλλα πολλά. Μεταξύ αυτών των πληροφοριών παρέχεται και η θέση στη διοικητική ιεραρχία που ανήκει κάθε εγγραφή στη χώρα που βρίσκεται. Μέσω της διαδικτυακής υπηρεσίας παρέχεται και η δυνατότητα εξαγωγής της ιεραρχίας κάθε τοποθεσίας από τη θέση που αυτή καταλαμβάνει στην διοικητική διαίρεση και προς τα πάνω. Το ανώτερο επίπεδο ιεραρχίας κάθε εγγραφής είναι το Earth. Για παράδειγμα για τη θέση Syntagma η ιεραρχία είναι Earth, Europe, Hellenic Republic, Attica, Nomarchía Athínas, Dimos Athens, Athens, Syntagma.

Χρησιμοποιώντας τη δυνατότητα που παρέχει το web service ανακτήθηκε η ιεραρχία κάθε τοποθεσίας, που η ταύτιση με κάποια εγγραφή του καταλόγου κρίθηκε σωστή, ξεχωριστά για το σύνολο των τοποθεσιών που παρείχαν οι συγγραφείς και εκείνων που εξήχθησαν από τη διαδικασία Named Entity Recognition. Κάθε εγγραφή ταυτίστηκε με το τεκμήριο από το οποίο προήλθε ώστε να μπορεί να χρησιμοποιηθεί μια στήλη ως κλειδί και να συνδεθεί με τους υπόλοιπους πίνακες. Η στήλη που χρησιμοποιήθηκε και συνδέει όλους τους πίνακες των δεδομένων που παρήχθησαν με τα αρχικά δεδομένα εισόδου είναι η στήλη με το όνομα του αρχείου. Τμήμα του τελικού πίνακα δεδομένων φαίνεται στον πίνακα 2.5 που ακολουθεί.

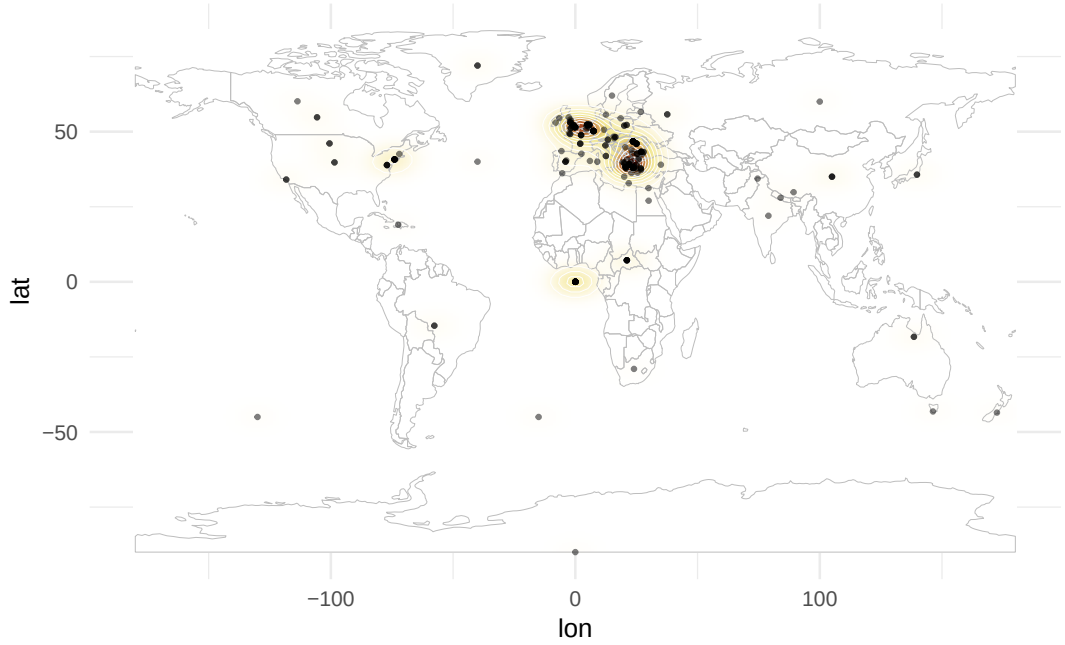
2.3.5 Εξαγωγή σημασιολογικής πληροφορίας

Η σημασιολογική ανάλυση που πραγματοποιήθηκε αναφέρεται στην εξαγωγή, χρησιμοποιώντας μεθόδους επεξεργασίας φυσικής γλώσσας, νοηματικών πληροφοριών από το σώμα των κειμένων, που δεν υπήρχαν στα συνοδευτικά δεδομένα, έτσι ώστε αυτές να μπορούν να χρησιμοποιηθούν σε μια σύνθετη αναζήτηση μαζί με εκείνες που εξήχθησαν από τη διαδικασία του χωρικού εμπλουτισμού.

Για τη σημασιολογική ανάλυση προηγήθηκε επεξεργασία των αρχικών δεδομένων κειμένου, καθώς αυτά περιείχαν πληροφορία από τους συγγραφείς, που ήδη είχε εξαχθεί και περιλαμβάνεται στον πίνακα δεδομένων που συνοδεύουν κάθε τεκμήριο, όπως ο τίτλος, οι αναφερόμενες χωρικές έννοιες, οι τοποθεσίες, τα επιστημονικά πεδία κλπ, ώστε αυτά να αφαιρεθούν. Από τα κείμενα αφαιρέθηκαν επίσης οι επικεφαλίδες των ενοτήτων, ενώ οι λεζάντες με επεξηγηματικό κείμενο εικόνων, που τυχόν ήταν ενσωματωμένες σε αυτά, επιλέχθηκε να παραμείνουν.



(i) Τοποθεσίες που παρείχαν οι συγγραφείς



(ii) Τοποθεσίες που εξήχθησαν από NER

Σχήμα 2.2: τοποθεσίες σεναρίων

toponym	**key**	**hierarchy**
Romania	edu_obj_829389.pdf	Earth,Europe,România,Earth,Europe,România
Serbia	edu_obj_829599.pdf	Earth,Europe,Serbia
Europe	edu_obj_829599.pdf	Earth,Western Europe
Vinca	edu_obj_829599.pdf	Earth,Europe,Republic of France,Occitanie,Département des Pyrénées-Orientales,Arrondissement de Prades,Vinça,Vinça
Shumen	edu_obj_829982.pdf	Earth,Europe,Republic of Bulgaria,Oblast Shumen,Obshtina Shumen,Shumen
Louvre Museum	edu_obj_829985.pdf	Earth,Europe,Republic of France,Île-de-France,Paris,Musée du Louvre
London	edu_obj_829985.pdf	Earth,Europe,United Kingdom of Great Britain and Northern Ireland,England,Greater London,London
Oxford	edu_obj_829985.pdf	Earth,Europe,United Kingdom of Great Britain and Northern Ireland,England,Oxfordshire,Oxford District,Oxford
Rome	edu_obj_829985.pdf	Earth,Europe,Repubblica Italiana,Lazio,Città metropolitana di Roma Capitale,Roma Capitale,Rome
Balkan Peninsula	edu_obj_829986.pdf	Earth,Europe,Republic of Bulgaria,Balkan Peninsula

Πίνακας 2.5: Διοικητική ιεραρχία τοπωνυμίων που ανακτήθηκε από τη βάση του Geonames (τμήμα του πίνακα)

Από την επεξεργασία προέκυψε ένα νέο σύνολο δεδομένων το οποίο επεξεργάστηκε γλωσσολογικά ώστε να αναγνωριστούν οι προτάσεις, τα μέρη του λόγου και οι ρίζες κάθε λέξης (tokenization-lemmantization-part of speech tagging). Για την διαδικασία επιλέχθηκε και χρησιμοποιήθηκε ο annotator spacy καθώς προσφέρει για τη συγκεκριμένη ενέργεια τον καλύτερο συνδυασμό ταχύτητας και ποιότητας αποτελέσματος σε σύγκριση με τους άλλους δυο. Το παραγόμενο σύνολο δεδομένων χρησιμοποιήθηκε για τον εντοπισμό των χωρικών εννοιών της οντολογίας που αναφέρονται στο σώμα του κειμένου κάθε σεναρίου, καθώς και στην εξαγωγή του κειμένου της συγκεκριμένης πρότασης στην οποία αυτές εντοπίζονταν κάθε φορά.

Στο σημείο αυτό θα πρέπει να τονιστεί πως στο δίκτυο των χωρικών εννοιών της οντολογίας, περιέχονται έννοιες με την ίδια ονομασία αλλά διαφορετική σημασία. Στον πίνακα 2.6 που ακολουθεί παρουσιάζονται 2 από αυτές με τον αντίστοιχο ορισμό τους.

concept	definition
island	a land mass (smaller than a continent) that is surrounded by water
island	a zone or area resembling an island
country	the territory occupied by a nation
country	a politically organized body of people under a single government

Πίνακας 2.6: Έννοιες με διαφορετική σημασία

Για να αντιμετωπιστεί ο καθορισμός της σωστότερης έννοιας μεταξύ εκείνων που έχουν την ίδια ονομασία, υπολογίστηκε ο δείκτης ομοιότητας cosine distance κάθε πρότασης με τον ορισμό κάθε έννοιας που εντοπιζόταν μέσα σε αυτή. Ο δείκτης που υπολογίστηκε χρησιμοποιήθηκε για να επιλεχθεί η κατάλληλη χωρική έννοια της οντολογίας που εντοπιζόταν στις περιπτώσεις που αυτές εμφανίζονταν εντός της πάνω από μια φορά. Μικρότερη τιμή σημαίνει μεγαλύτερη ομοιότητα, όπως φαίνεται στον πίνακα 2.7 που ακολουθεί.

concept	definition	sentence	cosine distance
island	a land mass (smaller than a continent) that is surrounded by water	the objective of this educational scenario be the study of the geomorphological evolution of the aegean archipelago and the cyclades island area in particular .	0.104050321405123
island	a zone or area resembling an island	the objective of this educational scenario be the study of the geomorphological evolution of the aegean archipelago and the cyclades island area in particular .	0.129789880671356
country	the territory occupied by a nation	student have to solve task , which improve knowledge about the town and village in the territory of a particular country (namely bulgaria) .	0.0936912156949105
country	a politically organized body of people under a single government	student have to solve task , which improve knowledge about the town and village in the territory of a particular country (namely bulgaria) .	0.0990052391934451

Πίνακας 2.7: επιλογή σωστότερης έννοιας (cosine similarity)

Το δεύτερο πρόβλημα που παρουσιάστηκε και αντιμετωπίστηκε είναι οι έννοιες που

περιέχουν άλλες έννοιες, όπως για παράδειγμα οι έννοιες ”map” και ”map projection”. Σε περίπτωση που εντοπιζόταν 7 φορές η έννοια map και 4 η έννοια map projection, η συχνότητα της πρώτης διαμορφωνόταν σε 7-4 = 3.

Επιπρόσθετα η ανάλυση περιλαμβάνει και εξαγωγή νοηματικού περιεχομένου (φράσεις κλειδιά) από τα σενάρια. Ως φράσεις κλειδιά καθορίστηκαν εκείνες οι φράσεις που αποτελούνται από συνδυασμούς επιθέτων, ουσιαστικών (κοινά και συγκεκριμένα), προθέσεων καθώς και άρθρων ορισμού, όπως πρότειναν οι Handler et.al., 2016. Για το σκοπό αυτό χρησιμοποιήθηκε το πακέτο λογισμικού phrasemachine που εκείνοι δημιούργησαν για την εφαρμογή της μεθόδου στη γλώσσα R. Η ανάλυση περιλαμβάνει όλα τα βήματα της διαδικασίας annotation που έχουν περιγραφεί πιο πάνω και πραγματοποιήθηκε, όπως και όλα τα άλλα στάδια ανάλυσης της παρούσας εργασίας, μόνο για την Αγγλική γλώσσα. Αποφασίστηκε πως οι νοηματικές φράσεις κλειδιά που χαρακτηρίζουν κάθε σενάριο που εξετάστηκε, είναι εκείνες που περιέχουν κάποια από τις χωρικές έννοιες της οντολογίας και εντοπίζονται στο σώμα του κειμένου τουλάχιστον δυο φορές.

Παράδειγμα του νοηματικού περιεχομένου που ανακτήθηκε από το τεκμήριο με τίτλο ”Perceptual image of an urban environment” φαίνεται στον πίνακα 2.8 που ακολουθεί.

phrase	**freq**
mental map	12
urban environment	5
urban space	5
map of Athens	4

Πίνακας 2.8: Φράσεις κλειδιά σεναρίου ”Perceptual image of an urban environment”

Τέλος η σημασιολογική ανάλυση περιλαμβάνει και την ανάκτηση της ιεραρχίας που διαμορφώνεται για κάθε έννοια, από τη θέση της μέσα στο γράφο και παράγεται από τις μεταξύ τους συνδέσεις, μιας και αυτές είναι προσανατολισμένες. Τμήμα της πληροφορίας που ανακτήθηκε φαίνεται στον πίνακα 2.9 που ακολουθεί.

concept	hierarchy
city	city, municipality, administrative division, urban area, territory, geographical area, region
closeness	closeness, distance , size
cluster	cluster, agglomeration
coast	coast, shore, geological formation
cold weather	cold weather, weather, atmospheric phenomenon
compass	compass, navigational instrument
compass point	compass point, direction
computation	computation, problem solving
continent	continent, land mass/ landmass, land
continental drift	continental drift, geological phenomenon
conversion	conversion, computation, problem solving
coordinate axis	coordinate axis, axis, line , location
country	country, administrative division, political unit, territory, region
current	current, flow, change of location, motion

Πίνακας 2.9: ιεραρχία εννοιών από κάτω προς τα πάνω

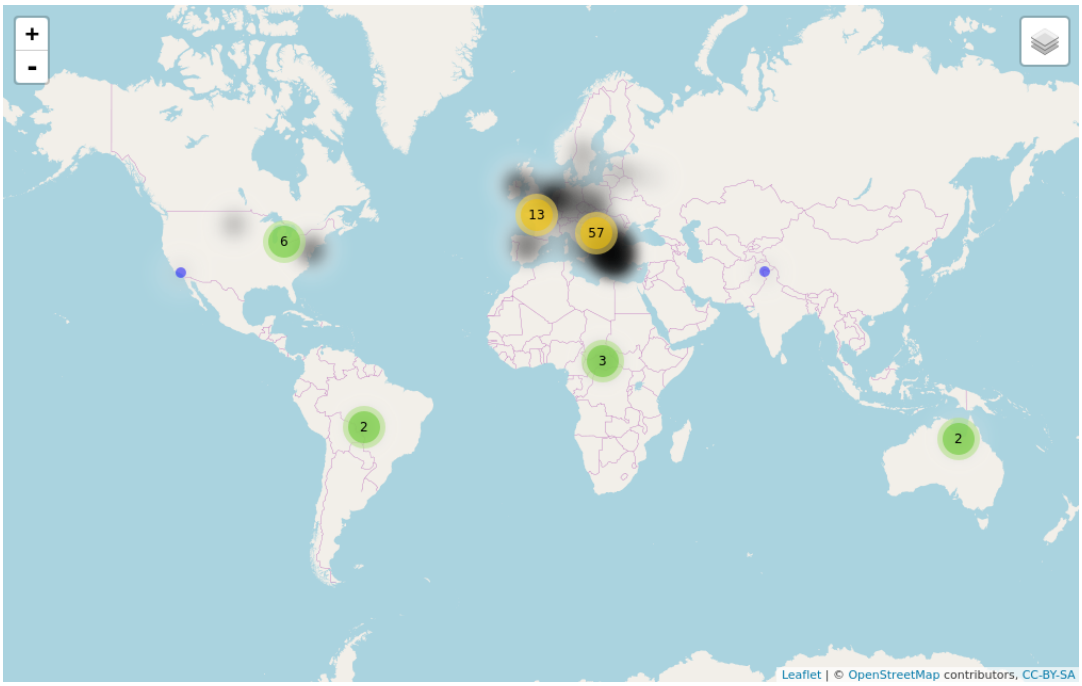
2.4 Αποτελέσματα

2.4.1 Χωρικός εμπλουτισμός

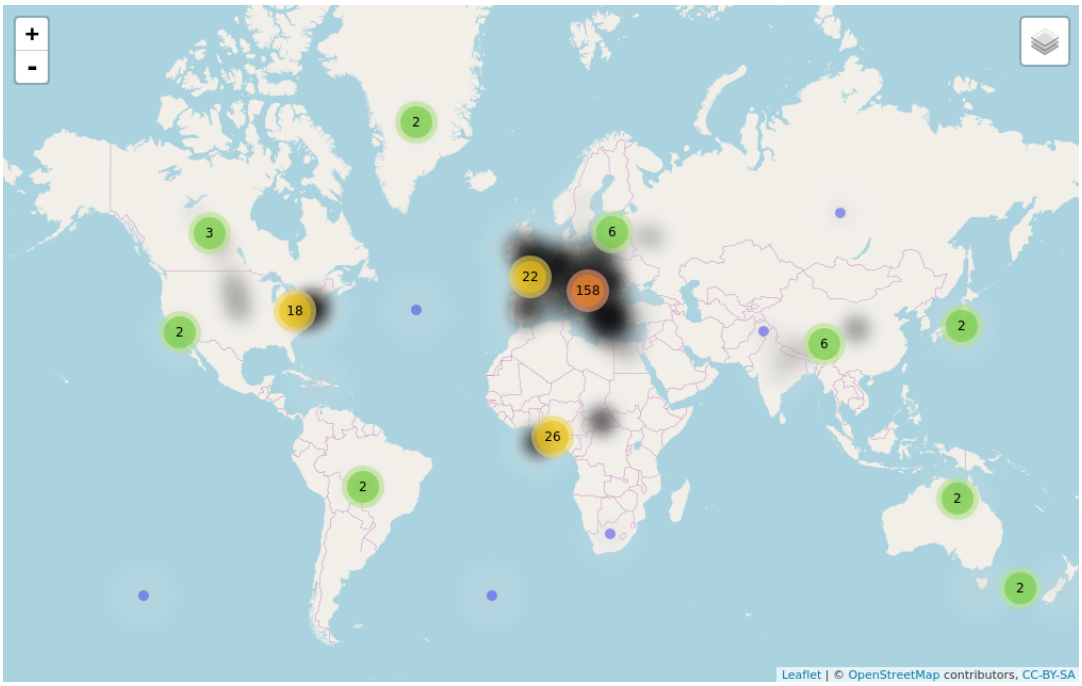
Ο χωρικός εμπλουτισμός που πραγματοποιήθηκε αναφέρεται σε δυο ενέργειες. Η πρώτη είναι η αύξηση του πλήθους των γεωγραφικών οντοτήτων στις οποίες αναφέρονται τα εκπαιδευτικά σενάρια του GEOTHNK. Πράγματι οι συγγραφείς παρείχαν χειροκίνητα κατά την διαδικασία υποβολής 64 μοναδικές τοποθεσίες σε 25 από αυτά, ενώ με τη διαδικασία που περιγράφηκε στις προηγούμενες ενότητες έγιναν διαθέσιμες 146 μοναδικές τοποθεσίες για συνολικά 80 σενάρια από τα 159 που συλλέχθηκαν. Η *χωρική* διαφορά φαίνεται πιο καθαρά στους χάρτες του σχήματος 2.3 που ακολουθούν.

Η δεύτερη ενέργεια που πραγματοποιήθηκε προς την κατεύθυνση του χωρικού εμπλουτισμού είναι η οργάνωση της διοικητικής ιεραρχίας κάθε θέσης η οποία ανακτήθηκε από τη βάση του Geonames. Κάθε ιεραρχία αντιστοιχήθηκε με το σενάριο από το οποίο εξήχθη το τοπωνύμιο και κατά συνέπεια με όλα τα υπόλοιπα διαθέσιμα δεδομένα που τα συνόδευαν, όπως αυτά περιγράφονται στον πίνακα 2.1.

Οι δυο παραπάνω ενέργειες επέτρεψαν την δόμηση μιας εφαρμογής αναζήτησης με όλα τα επίπεδα ιεραρχίας η οποία παρουσιάζεται στο επόμενο κεφάλαιο. Για παράδειγμα η ιεραρχία της πόλης Shumen είναι Earth, Europe, Republic of Bulgaria, Oblast Shumen, Obshtina Shumen, Shumen. Η αντίστοιχη για το Μουσείο του Λούβρου στο Παρίσι είναι Earth, Europe, Republic of France, Île-de-France, Paris, Musée du Louvre, όπως φαίνεται



(i) θέσεις συγγραφέων

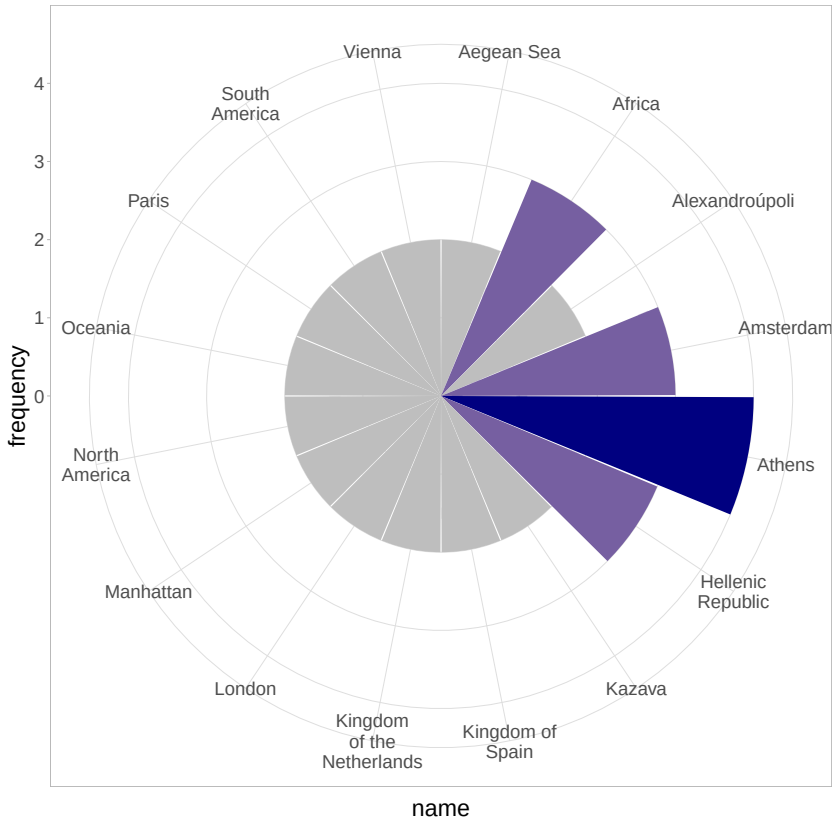


(ii) θέσεις NER.

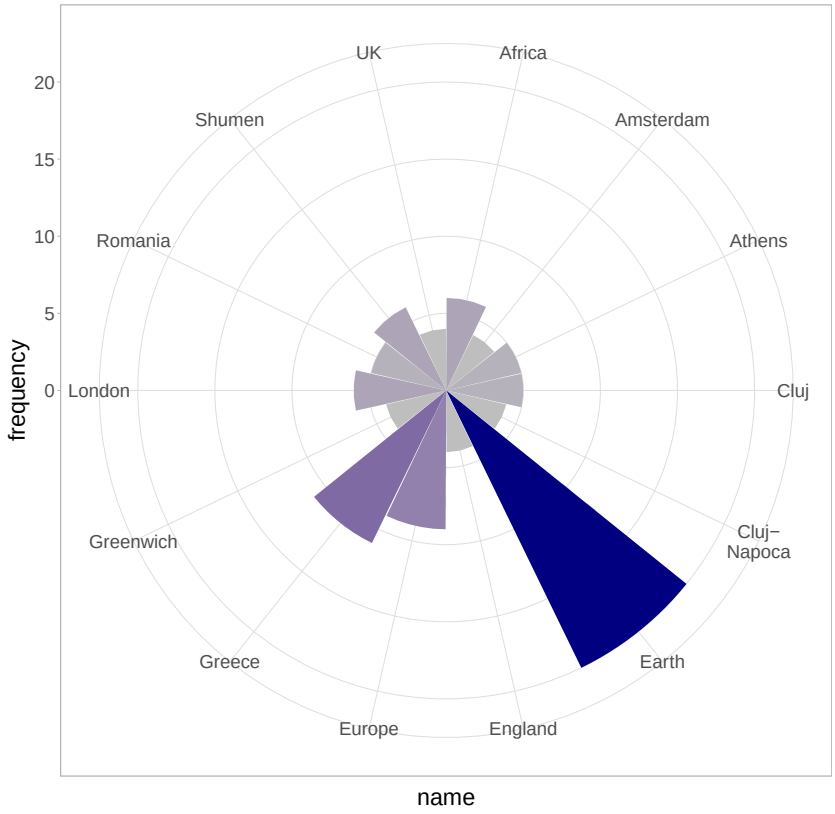
Σχήμα 2.3: 1η ενέργεια χωρικού εμπλουτισμού

στον πίνακα 2.5. Αναζητώντας στην εφαρμογή σενάρια που αναφέρονται στην Ευρώπη η εφαρμογή θα επιστρέψει στα αποτελέσματα τα σενάρια που αναφέρονται τόσο στην πόλη Shumen στη Βουλγαρία όσο και στο μουσείο του Λούβρου, αφού και οι δυο τοποθεσίες βρίσκονται στην Ευρώπη. Στο σημείο αυτό αξίζει να επισημανθεί πως η ιεραρχία που διαθέτει ο κατάλογος του Geonames για την Ευρώπη είναι Earth, Western Europe πιθανότατα επειδή οι περισσότεροι που αναφέρονται στην Ευρώπη εννοούν την Ευρωπαϊκή Ένωση η οποία γεωγραφικά βρίσκεται στο Δυτικό τμήμα της Ευρωπαϊκής ηπείρου. Παρότι το μουσείο του Λούβρου βρίσκεται στο δυτικό τμήμα της Ευρώπης, αναζητώντας σενάρια που αναφέρονται σε Western Europe, στα αποτελέσματα δεν θα περιέχεται εκείνο για το μουσείο του Λούβρου.

Από τη διαδικασία του εμπλουτισμού άλλαξε και η σημασιολογική εικόνα που προέκυπτε από την εξέταση της συχνότητας εμφάνισης των τοπωνυμίων που παρείχαν οι συγγραφείς σε σχέση με αυτές που εντοπίστηκαν στα κείμενα. Πράγματι όπως φαίνεται στο σχήμα 2.4 προκύπτει μια άλλη εικόνα. Ενώ από τα τοπωνύμια των συγγραφέων φαίνεται πως τα σενάρια είναι προσανατολισμένα σε 4 ηπείρους, η πραγματικότητα από τα τοπωνύμια που εντοπίστηκαν δείχνει ξεκάθαρα ότι ο προσανατολισμός είναι σε Ευρωπαϊκές τοποθεσίες με την Ελλάδα να κυριαρχεί και στις 2 περιπτώσεις.



(i) τοπωνύμια συγγραφέων

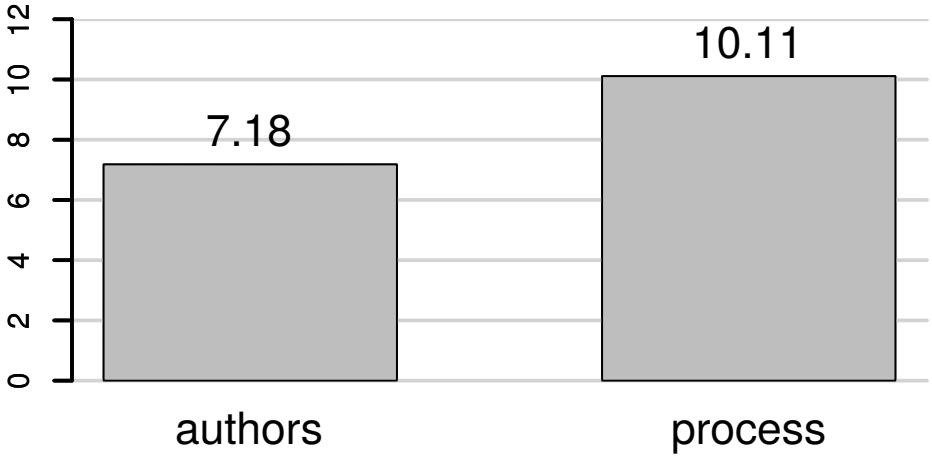


(ii) τοπωνύμια NER

Σχήμα 2.4: Διάγραμμα συχνοτήτων εμφάνισης τοπωνυμίων

2.4.2 Σημασιολογικός εμπλουτισμός

Ο σημασιολογικός εμπλουτισμός που πραγματοποιήθηκε αναφέρεται σε τρεις ενέργειες. Η πρώτη είναι ο εντοπισμός της αναφοράς εννοιών του δικτύου GEOTHNK στο σώμα του κειμένου κάθε σεναρίου και ο υπολογισμός της συχνότητας εμφάνισης τους. Τα αποτελέσματα είναι δυνατόν να αξιοποιηθούν αναλύοντας τα διαγράμματα που προκύπτουν, καθώς επίσης να αποτελέσουν έναν δείκτη σημαντικότητας για ταξινόμηση των σεναρίων σε αποτελέσματα σύνθετης αναζήτησης. Από μια πρώτη ματιά στο διάγραμμα του σχήματος 2.5 προκύπτει πως η διαδικασία πρόσθεσε κατά μέσο όρο τρεις χωρικές έννοιες ανά σενάριο που οι συγγραφείς δεν είχαν αναφέρει, ενώ αυτές εντοπίστηκαν από τη διαδικασία.



Σχήμα 2.5: μέσος όρος χωρικών εννοιών ανά σενάριο

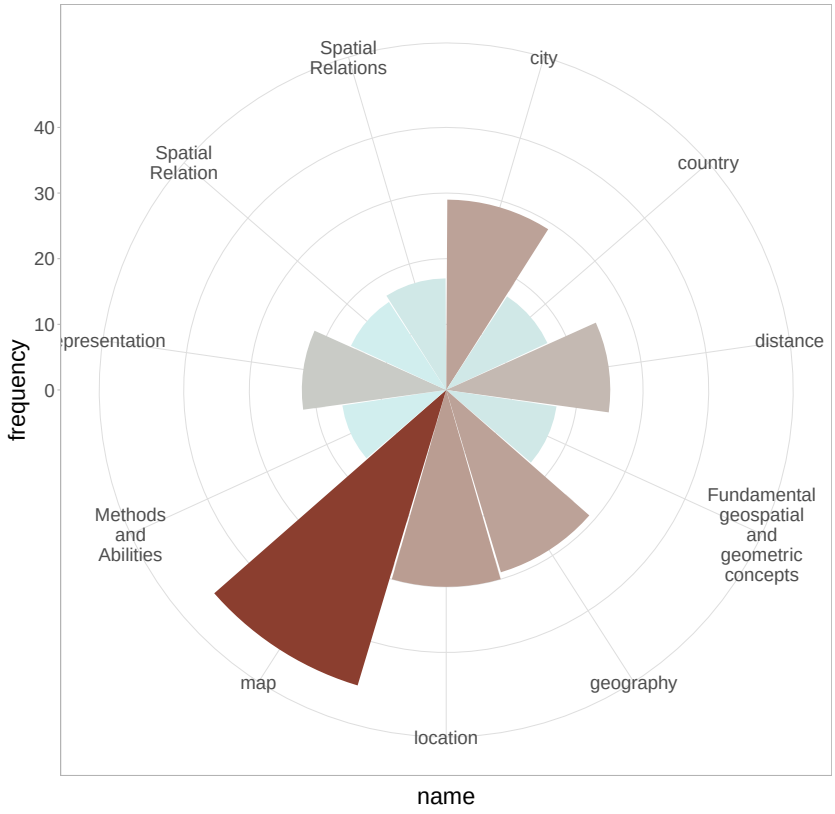
Από την ανάλυση των χωρικών εννοιών που διέθεσαν οι συγγραφείς κατά την υποβολή προκύπτει ένα αφήγημα. Πράγματι εξετάζοντας το διάγραμμα των συχνοτήτων εμφάνισης τους, σχήμα 2.6i, παρατηρούμε πως οι έννοιες που ξεχωρίζουν από όλες τις υπόλοιπες, σύμφωνα με τους συγγραφείς, είναι οι ”map”, ”location”, ”geography”, ”city” και ”distance”. Θα έλεγε κανείς πως το αφήγημα που προκύπτει είναι πως τα εκπαιδευτικά σενάρια ασχολούνται με τη χαρτογράφηση, τη γεωγραφία, τις πόλεις και τη μελέτη θέσεων και αποστάσεων. Εξετάζοντας το διάγραμμα του σχήματος 2.6ii, που παρουσιάζει τη συχνότητα εμφάνισης των χωρικών εννοιών που εντοπίστηκαν από την επεξεργασία, παρατηρούμε πως οι κυρίαρχες έννοιες είναι οι ”map”, ”place”, ”point”, ”order”, ”earth”, και ”time”. Το αφήγημα που προκύπτει αυτή τη φορά είναι πως τα εκπαιδευτικά σενάρια ασχολούνται με τη χαρτογράφηση, τη Γη και τα τοπία καθώς και το χρόνο που συμβαίνουν όλα αυτά. Σε κάθε περίπτωση η κυρίαρχη έννοια και στις δυο περιπτώσεις είναι ο

χάρτης.

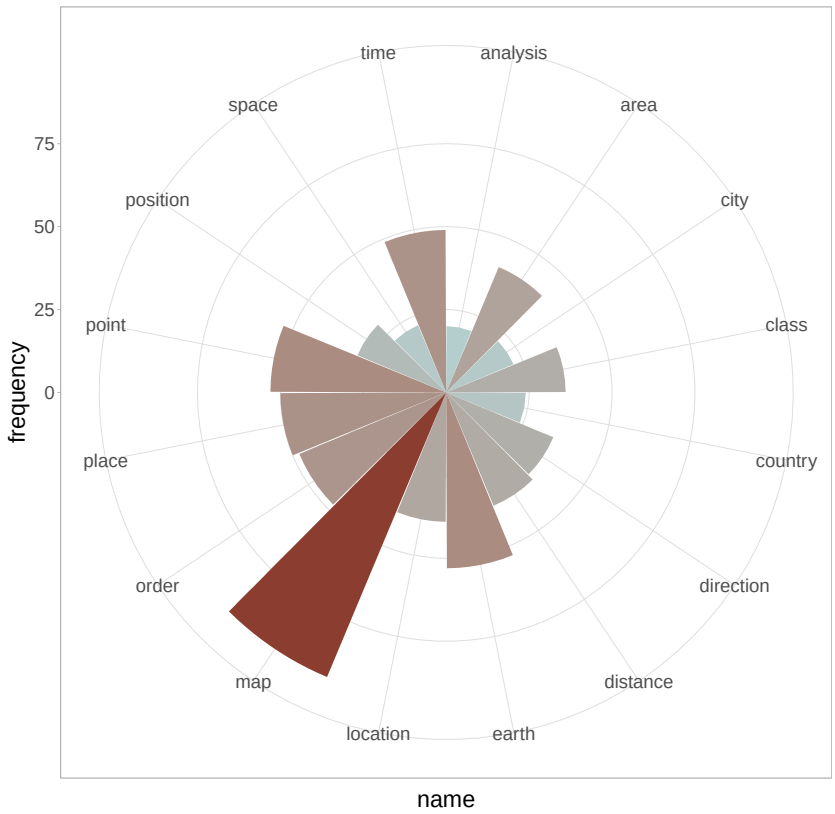
Σε αυτό το σημείο θα πρέπει να σημειωθεί πως κάθε έννοια που είτε έχει προσδιορίσει ο συγγραφέας είτε έχει εντοπιστεί από την ανάλυση, έχει ληφθεί υπόψη μία φορά για κάθε σενάριο, όσες φορές και αν αυτή εντοπίστηκε να αναφέρεται στο σώμα του κειμένου. Στο διάγραμμα του σχήματος 2.7 απεικονίζονται οι συχνότητες εμφάνισης κάθε έννοιας συνολικά για όσες φορές αυτή εντοπίστηκε στο σώμα του κειμένου κάθε σεναρίου. Οι έννοιες που ξεχωρίζουν είναι οι ”map”, ”earth” και ”direction”. Πράγματι ο χάρτης είναι το κυρίαρχο εργαλείο αναπαράστασης και κατανόησης του χώρου, επιβεβαιώνοντας πως ως άνθρωποι προσανατολιζόμαστε και κατανοούμε την τάξη των πραγμάτων γύρω μας από τη σχετική τους θέση, δηλαδή τη γεωγραφία τους.

Η δεύτερη ενέργεια στην οποία αναφέρεται ο σημασιολογικός εμπλουτισμός που πραγματοποιήθηκε, είναι η ανάκτηση της ιεραρχίας κάθε έννοιας του δικτύου GEOTHNK. Από την ανάκτηση της ιεραρχίας και την ταύτισή της με παρόμοιο τρόπο, όπως της χωρικής που περιγράφηκε σε προηγούμενη ενότητα, καθίσταται δυνατή η σημασιολογική αναζήτηση χρησιμοποιώντας τις ευρύτερες έννοιες στις οποίες κάθε μια συγκαταλέγεται. Όπως φαίνεται στο σχήμα 2.8 που ακολουθεί η αναζήτηση για σενάρια που αναφέρονται σε geological phenomenon θα επιστρέψει όλα εκείνα που αναφέρονται σε flood, continental drift και earthquake.

Η τρίτη ενέργεια είναι η εξαγωγή νοηματικού περιεχομένου. Κάθε σύνολο φράσεων κλειδιών που διαμορφώθηκε για κάθε σενάριο, όπως περιγράφηκε σε προηγούμενη ενότητα, συνοδεύει κάθε ένα από αυτά και γίνεται διαθέσιμο στο χρήστη έτσι ώστε η αναζήτηση, η επισκόπηση και επιλογή να γίνεται πιο σαφής και συγκεκριμένη.

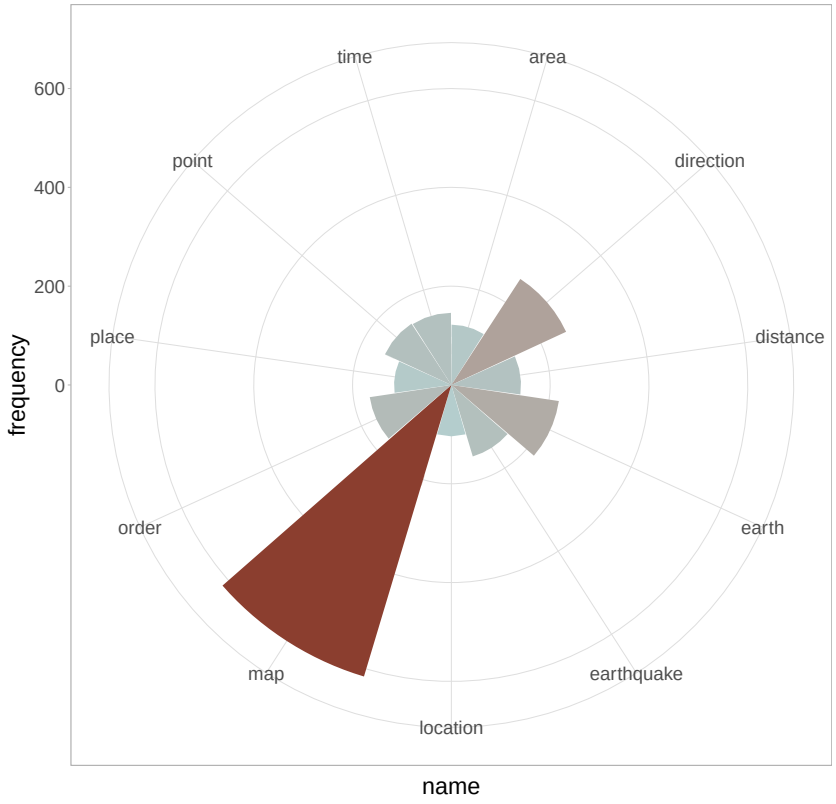


(i) έννοιες συγγραφέων

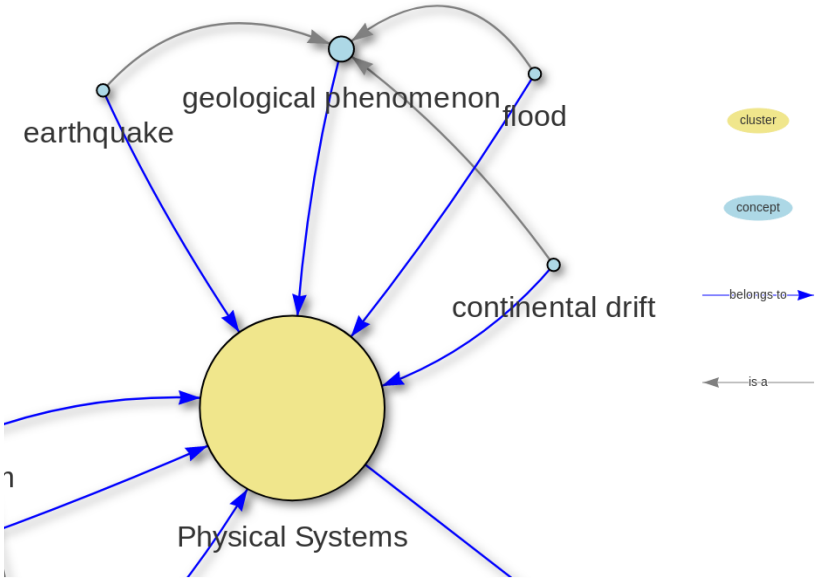


(ii) έννοιες που εξήχθησαν

Σχήμα 2.6: Συχνότητες εμφάνισης χωρικών εννοιών (μια ανά σενάριο)



Σχήμα 2.7: Συχνότητες χωρικών εννοιών που εντοπίστηκαν στο σώμα των κειμένων (συνολικά)



Σχήμα 2.8: τμήμα του δικτύου, εξαγωγή ιεραρχίας

Κεφάλαιο 3

Διαδικτυακή εφαρμογή

3.1 Λογισμικό

Όλα τα δεδομένα που συλλέχθηκαν καθώς και εκείνα που δημιουργήθηκαν από την ανάλυση αυτών, χρησιμοποιήθηκαν στη σύνθεση μιας δυναμικής διαδικτυακής εφαρμογής με σκοπό την προβολή τους, αλλά και τη μελέτη των αποτελεσμάτων αξιοποιώντας τη διαδραστικότητα. Για τη δόμηση της εφαρμογής χρησιμοποιήθηκε το πακέτο shiny που έχει δημιουργήσει η εταιρία RStudio [3] για τη γλώσσα R και όπως και το σύνολο του λογισμικού που αξιοποιήθηκε στην παρούσα εργασία, είναι ελεύθερο και ανοικτού κώδικα.

Το λογισμικό Shiny παρέχει εργαλεία για τη δημιουργία διεπαφής διαδικτυακών εφαρμογών (web framework) χρησιμοποιώντας τη σύνταξη της γλώσσας R. Παρέχει δηλαδή έτοιμες λειτουργίες (functions), τις οποίες ο χρήστης καλεί με σύνταξη R, οι οποίες τελικά δημιουργούν τον κώδικα που απαιτείται για τη δομή, την εμφάνιση και τη συμπεριφορά κάθε αντικειμένου της εφαρμογής (HTML, CSS, Javascript). Το λογισμικό παρέχει πλήρη παραμετροποίηση των αντικειμένων με τη σύνταξη των HTML, CSS, Javascript και παρότι δεν απαιτείται για τη δημιουργία εφαρμογών, η γνώση της χρήσης τους μπορεί να δημιουργήσει αρκετά ποιοτικές εφαρμογές σε εμπορικό επίπεδο.

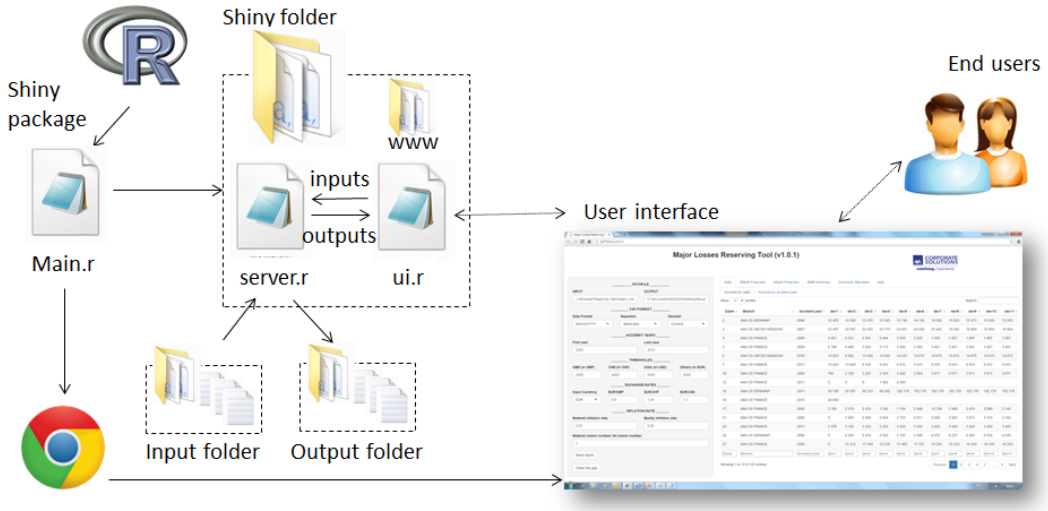
Το shiny χτίζει εφαρμογές αξιοποιώντας τη βιβλιοθήκη Bootstrap [4] για τα front-end στοιχεία (αρχείο ui) και χρησιμοποιεί την R ως back-end γλώσσα (αρχείο server). Η διαδραστικότητα μεταξύ του χρήστη (client) και του server επιτυγχάνεται εγκαθιδρύοντας επικοινωνία μεταξύ των δυο αρχείων ui.R και server.R τα οποία το ένα δημιουργεί αντικείμενα όπως πίνακες, διαγράμματα, widjets στο server και το άλλο τα παρουσιάζει στον client. Κάθε φορά που ο τελικός χρήστης (end-user) επεμβαίνει μέσω του client, το shiny αναγνωρίζει ποια αντικείμενα, από εκείνα που βρίσκονται στη server μεριά, επηρεάζονται από την αλλαγή και επανεκτελείται ο κώδικας που τα αφορά με τις καινούριες παραμέτρους και το αποτέλεσμα προωθείται ξανά στη διεπαφή [5]. Ένα τυπικό παράδειγμα δομής

[3] https://shiny.rstudio.com/

[4] https://getbootstrap.com/

[5] https://shiny.rstudio.com/articles/reactivity-overview.html

και λειτουργίας διαδικτυακής εφαρμογής με το λογισμικό shiny φαίνεται στο σχήμα 3.1 που ακολουθεί.



Σχήμα 3.1: τυπική λειτουργία εφαρμογής με το πακέτο shiny.
Πηγή: *http://littleactuary.github.io/blog/Web-application-framework-with-Shiny/*

Το shiny διατίθεται με μορφή ανοιχτού κώδικα και υποστηρίζεται μόνο για διανομές Linux, από τις οποίες έτοιμα πακέτα εγκατάστασης διατίθενται για Ubuntu 12.04 (or later), RedHat/CentOS 6 και 7, και SUSE Linux Enterprise Server 11+ [6]. Διατίθεται επίσης και με εμπορική άδεια με αυξημένες δυνατότητες απόδοσης, ασφάλειας και τεχνικής υποστήριξης. Τέλος υπάρχουν τρεις τρόποι να αναπτυχθούν (deployment) οι εφαρμογές shiny στο διαδίκτυο :

- Χρησιμοποιώντας την δωρεάν έκδοση ανοιχτού κώδικα σε ιδιόκτητο server ή λογαριασμό με άδεια χρήσης AGPLv3.

- Χρησιμοποιώντας την εμπορική έκδοση σε ιδιόκτητο server ή λογαριασμό με εμπορική άδεια χρήσης.

- Χρησιμοποιώντας έναν λογαριασμό στο server του RStudio (http://www.shinyapps.io), που έχει δημιουργηθεί για το σκοπό αυτό.

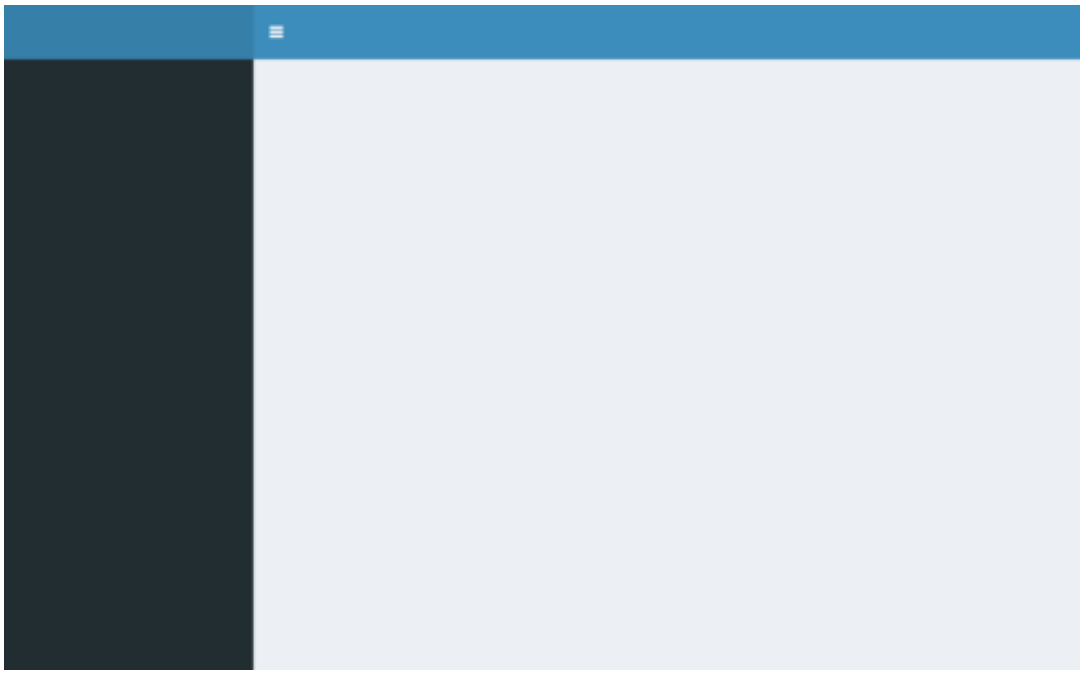
Για τις ανάγκες της παρούσας εργασίας χρησιμοποιήθηκε ένας λογαριασμός στο server του RStudio. Η εφαρμογή που δημιουργήθηκε είναι προσβάσιμη μέσω του παρακάτω συνδέσμου και οι λειτουργίες της παρουσιάζονται στις ενότητες που ακολουθούν :

https://veegee.shinyapps.io/semantics/

[6] κατά την 4/3/2018

3.2 Τμήματα - Λειτουργίες

Για τη δημιουργία της εφαρμογής χρησιμοποιήθηκε το αντικείμενο shinydashboard. Το αντικείμενο αυτό αποτελείται από 3 βασικά τμήματα, αυτό της επικεφαλίδας, μια στήλη στην αριστερή μεριά και όλο το υπόλοιπο που είναι το κύριο σώμα της εφαρμογής και στο οποίο μπορούν να προβληθούν αντικείμενα που δημιουργούνται στη server μεριά της εφαρμογής. Το αντικείμενο είναι πλήρως παραμετροποιήσιμο και προσφέρει ευελιξία, αυτόματη προσαρμογή σε κάθε μέγεθος οθόνης και πολύ καλή ποιότητα εμφάνισης-παρουσίασης. Η επιλογή της εμφάνισης/απόκρυψης της πλαϊνής στήλης πραγματοποιείται από το κομβίο menu στο τμήμα της επικεφαλίδας.



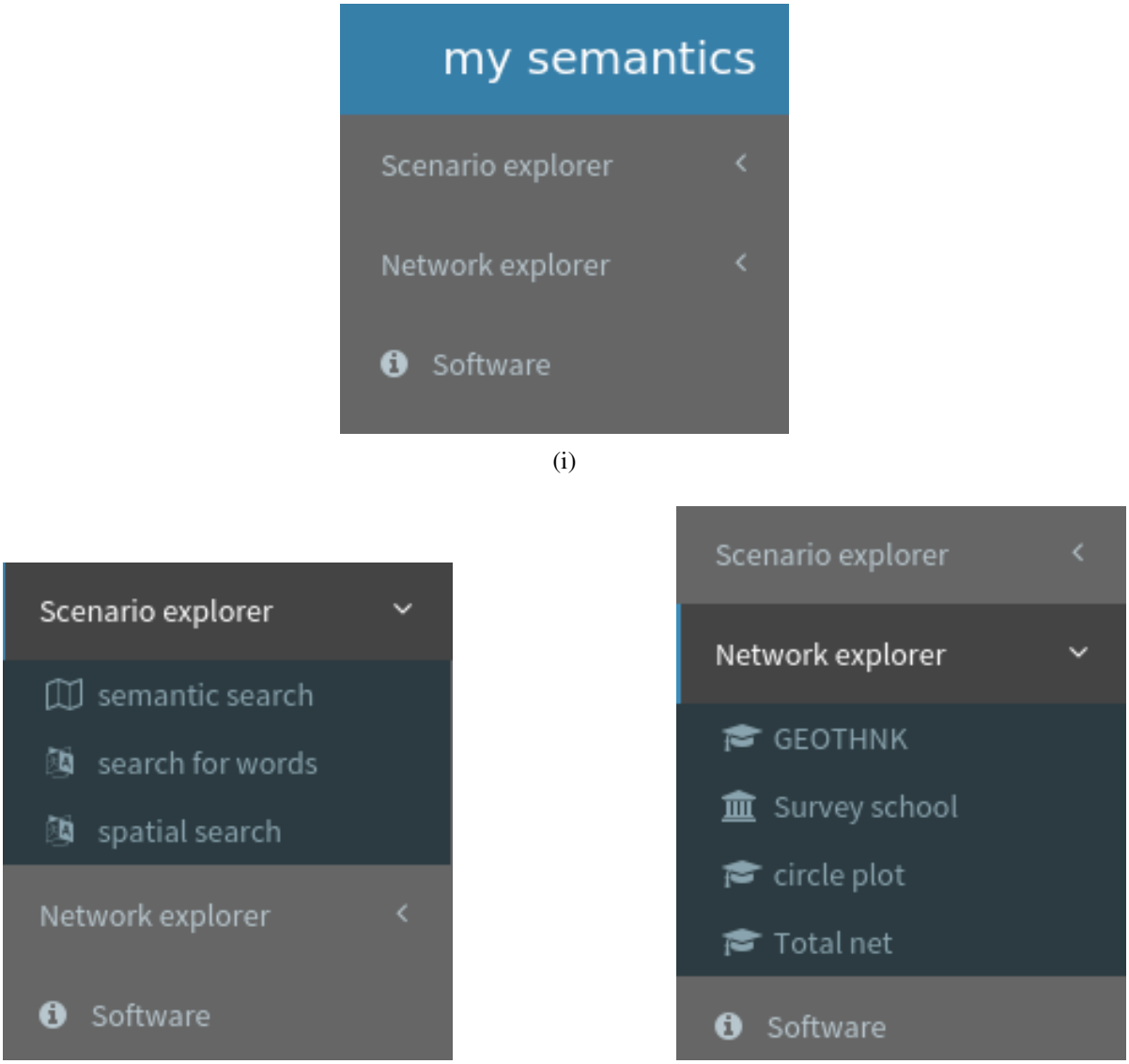
Σχήμα 3.2: Κενό αντικείμενο shiydashboard

Στο τμήμα της επικεφαλίδας είναι δυνατό να τοποθετηθούν πληθώρα κομβίων για διάφορες ενέργειες που ορίζει ο δημιουργός, όπως επίσης τίτλος ή λογότυπο. Στην εφαρμογή που δημιουργήθηκε προστέθηκε μόνο ένα κομβίο επανεκκίνησης της εφαρμογής που κάνει χρήση του ιστορικού περιήγησης του φυλλομετρητή (browser) και ξεκινά την εφαρμογή από την αρχή σε περίπτωση που για οποιοδήποτε λόγο υπάρξει αστοχία (σχήμα 3.3).

Σχήμα 3.3: Τμήμα επικεφαλίδας εφαρμογής

Στο τμήμα της πλαϊνής στήλης δημιουργούνται οι επιλογές περιεχομένου από τις οποίες ο χρήστης μπορεί να μεταβεί σε διαφορετικές συνθέσεις διεπαφής που έχει καθορίσει ο δημιουργός. Οι συνθέσεις αυτές είναι δυνατό να ομαδοποιηθούν σε ευρύτερες κατηγορίες.

Σε κάθε επιλογή περιεχομένου εκτός από κείμενο είναι δυνατή η προσθήκη εικονιδίων από τις βιβλιοθήκες Font-Awesome [7] και Glyphicons [8]. Για τις ανάγκες της παρούσας, δημιουργήθηκαν οι συνθέσεις που φαίνονται στην εικόνα 3.4 που ακολουθεί και επεξηγούνται στη συνέχεια.



(i)

(ii)

(iii)

Σχήμα 3.4: Τμήμα πλαϊνής στήλης εφαρμογής

Δημιουργήθηκαν 3 κατηγορίες παρουσίασης των δεδομένων. Η πρώτη είναι η διαδραστική εξερεύνηση των εκπαιδευτικών σεναρίων που συλλέχθηκαν και περιλαμβάνει τη σημασιολογική εξερεύνηση, την χωρική αναζήτηση και μια σύνθετη αναζήτηση αξιοποιώντας τα συνοδευτικά δεδομένα των σεναρίων όπως ο τίτλος, η γλώσσα στην οποία είναι γραμμένο, οι λέξεις κλειδιά, τα επιστημονικά πεδία ενδιαφέροντος, οι τοποθεσίες και οι έννοιες του χώρου στις οποίες αυτό αναφέρεται. Επιπρόσθετα προσφέρεται και ελεύ-

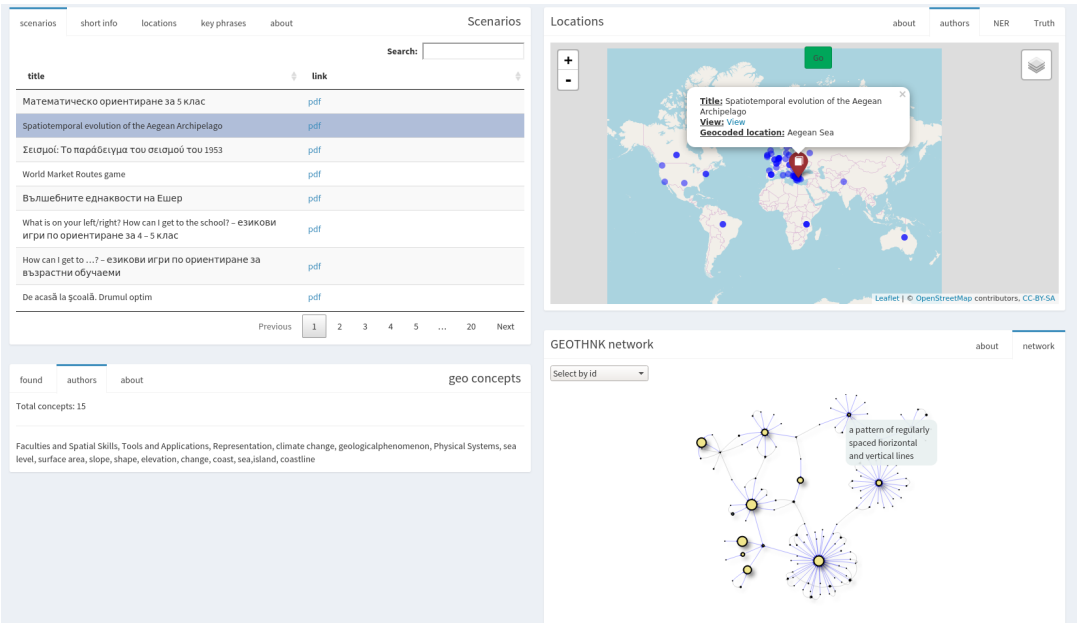
[7] https://fontawesome.com/icons?d=gallery
[8] https://getbootstrap.com/docs/3.3/components/

θερη αναζήτηση λέξεων στο σώμα των κειμένων. Η δεύτερη είναι η διαδραστική εξερεύνηση των χωρικών εννοιών του GEOTHNK και στην οποία προσφέρονται τρεις διαφορετικοί τρόποι οπτικοποίησης. Στη δεύτερη αυτή κατηγορία περιέχεται και η οπτικοποίηση του δικτύου που δημιουργείται από τα γνωστικά αντικείμενα που θεραπεύει η Σχολή των Αγρονόμων και Τοπογράφων Μηχανικών του Ε.Μ.Π. Στην τρίτη κατηγορία περιέχονται οι αναφορές στα πακέτα λογισμικού που χρησιμοποιήθηκαν για το σύνολο της ανάλυσης και της διαδραστικής εφαρμογής, καθώς και οι απαραίτητες αναφορές στους δημιουργούς τους.

3.2.1 Επισκόπηση - αναζήτηση σεναρίων

Σημασιολογική επισκόπηση

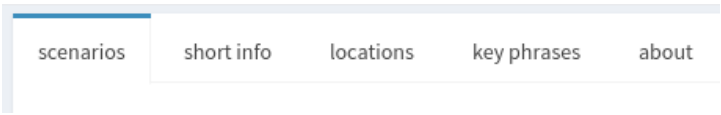
Στην εικόνα του σχήματος 3.5 που ακολουθεί παρουσιάζεται η σύνθεση της διαδραστικής σημασιολογικής εξερεύνησης των εκπαιδευτικών σεναρίων και η οποία αποτελείται από τέσσερις θεματικές περιοχές, σε κάθε μια από τις οποίες οι πληροφορίες περιέχονται σε αντίστοιχες καρτέλες. Σε κάθε ενότητα περιλαμβάνεται και μια καρτέλα (about) με απαραίτητες επεξηγήσεις για τη λειτουργικότητα.



Σχήμα 3.5: Σημασιολογική εξερεύνηση εκπαιδευτικών σεναρίων

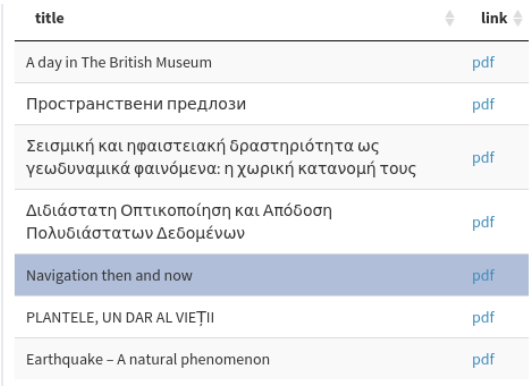
Η πρώτη θεματική περιοχή (Scenarios) αφορά πληροφορίες που σχετίζονται με τα εκπαιδευτικά σενάρια και έχουν κατηγοριοποιηθεί σε τέσσερις καρτέλες.

Η πρώτη καρτέλα (scenarios) περιλαμβάνει τον πίνακα με τους τίτλους των σεναρίων από τα οποία κάθε φορά ο χρήστης πρέπει να επιλέγει αυτό που τον ενδιαφέρει, καθώς και έναν υπερσύνδεσμο ο οποίος τον οδηγεί στην οπτικοποίηση του σεναρίου σε μορφή pdf. Αφού αυτός το επιλέξει, οι πληροφορίες που έχουν συλλεχθεί είτε δημιουργήθηκαν από

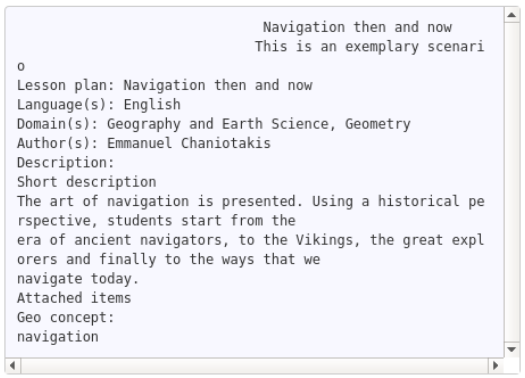


Σχήμα 3.6: Καρτέλες θεματικής ενότητας No1

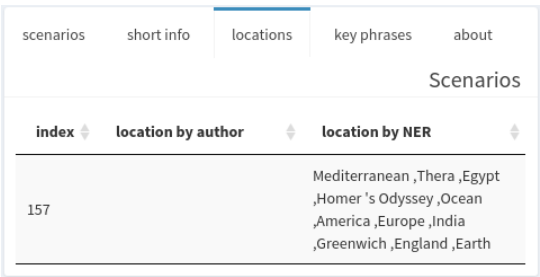
τη διαδικασία της ανάλυσης τακτοποιούνται καταλλήλως στις υπόλοιπες καρτέλες. Στην δεύτερη καρτέλα (short info) παρουσιάζεται η πρώτη σελίδα του σεναρίου που επιλέχθηκε και η οποία αναφέρει το όνομα του συγγραφέα, τη γλώσσα γραφής, τα επιστημονικά πεδία που αυτό αφορά, καθώς και σύντομη περιγραφή του περιεχομένου του. Στην τρίτη καρτέλα (locations) περιλαμβάνεται ένας πίνακας με τις τοποθεσίες που τυχόν έχει αναφέρει ο συγγραφέας κατά την υποβολή καθώς και εκείνες που εντοπίστηκαν να αναφέρονται από τη διαδικασία Named Entity Recognition. Στην τέταρτη (key phrases) περιλαμβάνεται ένας πίνακας με τις φράσεις κλειδιά που χαρακτηρίζουν το σενάριο και περιγράφηκαν στη σχετική ενότητα, οι οποίες περιέχουν χωρικές έννοιες της οντολογίας. Από την επιλογή μιας εγγραφής αυτού του κεντρικού πίνακα εξαρτάται και το περιεχόμενο όλων των άλλων θεματικών περιοχών της σύνθεσης αυτής. Τα περιεχόμενα αυτής της ενότητας φαίνονται στο σχήμα 3.7.



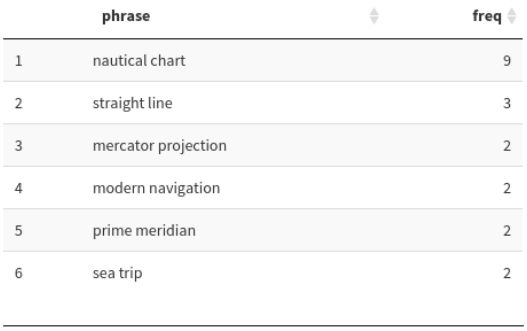
(i) scenarios



(ii) short info



(iii) locations

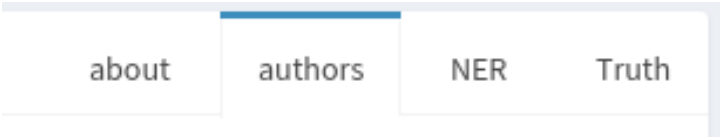


(iv) key phrases

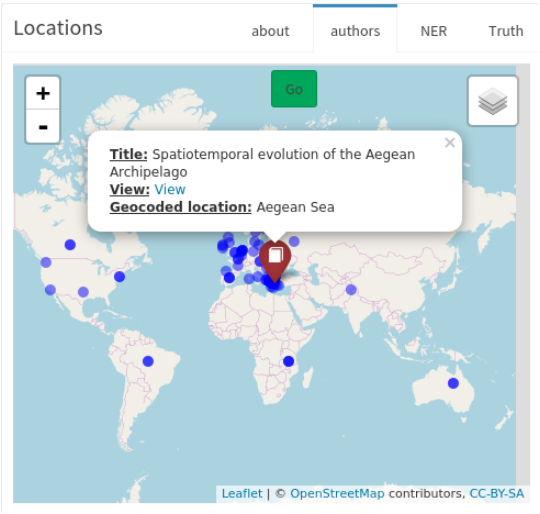
Σχήμα 3.7: Περιεχόμενα θεματικής ενότητας No1 (Scenarios)

Η δεύτερη θεματική περιοχή (Locations) αφορά τις πληροφορίες τοποθεσίας που έχουν

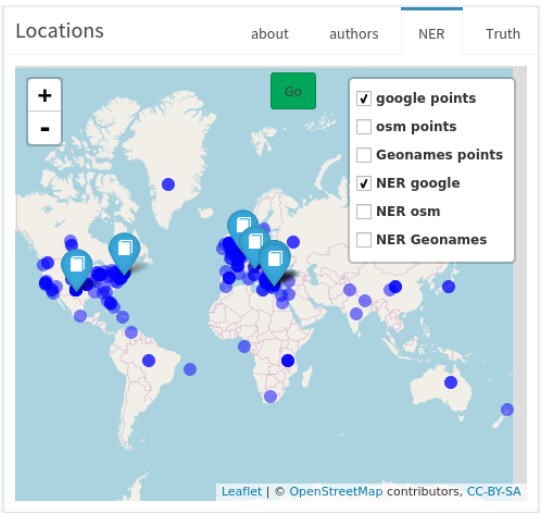
διαθέσει οι συγγραφείς καθώς και αυτές που εντοπίστηκαν να αναφέρονται από τη διαδικασία Named Entity Recognition και έχουν κατηγοριοποιηθεί σε τρεις καρτέλες.



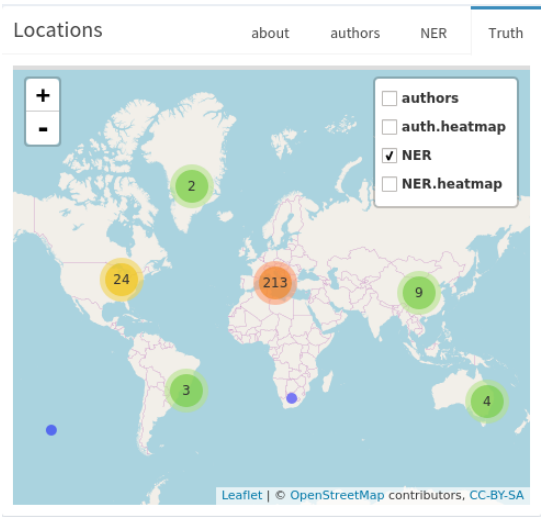
Σχήμα 3.8: Καρτέλες θεματικής ενότητας No2



(i) authors



(ii) NER



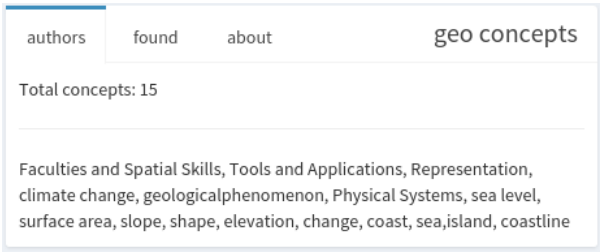
(iii) Truth

Σχήμα 3.9: Περιεχόμενα θεματικής ενότητας No2 (locations)

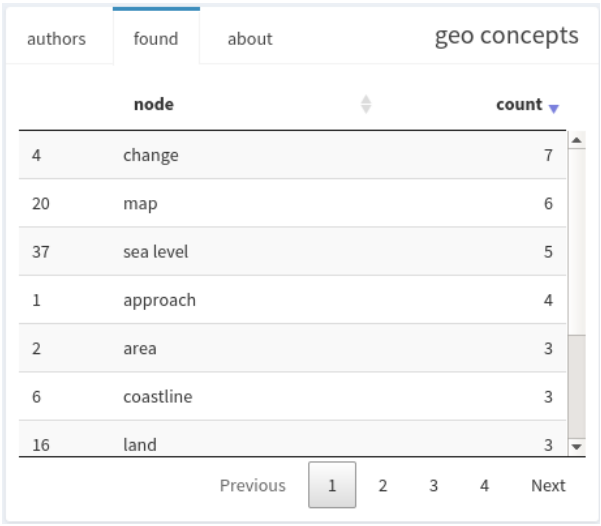
Στην πρώτη καρτέλα (authors) απεικονίζονται σε διαδραστικό χάρτη με υπόβαθρο, όλες οι θέσεις των τοποθεσιών που έχουν διαθέσει οι συγγραφείς κάθε σεναρίου. Αφού πρώτα έχει επιλέξει ο χρήστης ένα σενάριο από τον πίνακα της θεματικής ενότητας No1, πιέζοντας το κομβίο Go σχεδιάζονται οι θέσεις που αναφέρονται σε αυτό (αν υπάρχουν)

με τη δυνατότητα εμφάνισης αναδυόμενου παράθυρου (πιέζοντας πάνω στο σημείο που σχεδιάστηκε) με πληροφορίες για το σενάριο. Ομοίως στη δεύτερη καρτέλα (NER) εμφανίζονται οι θέσεις των περιοχών που εντοπίστηκαν από τη διαδικασία Named Entity Recognition. Οι θέσεις που απεικονίζονται στο συγκεκριμένο χάρτη έχουν γεωκωδικοποιηθεί και με τους τρεις geocoders που περιγράφησαν σε προηγούμενο κεφάλαιο (Google Maps, OSM και Geonames) και περιέχουν όλα τα σφάλματα και τις αδυναμίες αυτών, έτσι ώστε να μπορούν να μελετηθούν και να εξαχθούν χρήσιμα συμπεράσματα. Στην τρίτη καρτέλα (Truth) περιλαμβάνεται δυναμικός διαδραστικός χάρτης με τις διορθωμένες θέσεις των συγγραφέων και της διαδικασίας NER σε ξεχωριστά επίπεδα, χρησιμοποιώντας το γεωγραφικό κατάλογο Geonames, στον οποίο τα σημεία είναι χωρικά ομαδοποιημένα (clustering) (σχήμα 3.9).

Στην τρίτη θεματική ενότητα (geo concepts) περιλαμβάνεται η λίστα των χωρικών εννοιών που έχουν καθορίσει οι συγγραφείς καθώς και ένας πίνακας με εκείνες που εντοπίστηκαν να αναφέρονται στο σώμα του κειμένου κάθε σεναρίου μαζί με τη συχνότητα εμφάνισής τους, σύμφωνα με την ανάλυση που προηγήθηκε (σχήμα 3.10).



(i) authors

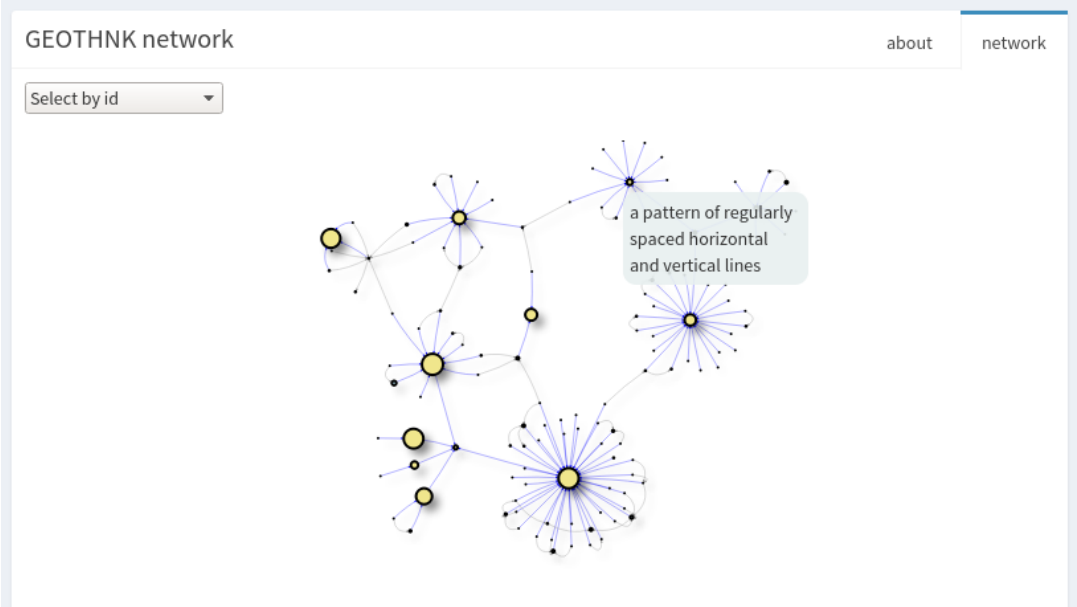


(ii) found

Σχήμα 3.10: Περιεχόμενα θεματικής ενότητας No3 (geo concepts)

Στην τέταρτη θεματική ενότητα περιλαμβάνεται μια δυναμική και διαδραστική απεικόνιση του δικτύου των χωρικών εννοιών που έχουν καθορίσει οι συγγραφείς ότι αναφέρεται το κείμενό του σεναρίου που έχει επιλεγεί από τον κεντρικό πίνακα. Το δίκτυο αυτό

είναι ένα τμήμα του συνολικού που εκτός από τις έννοιες των συγγραφέων περιέχει και τις έννοιες που συνδέονται άμεσα με αυτές καθώς και τις μεταξύ τους συνδέσεις. Ο χρήστης μπορεί να προβεί σε μεγέθυνση και μετακίνηση της απεικόνισης αλλά και κάθε κόμβου ξεχωριστά, καθώς επίσης να επιλέξει κάποια από τις έννοιες είτε με το ποντίκι είτε από τον επιλογέα και να σημανθεί αυτή μαζί με τους κοντινότερους γείτονές της. Μετακινώντας τον κέρσορα του ποντικιού πάνω από κάθε κόμβο εμφανίζεται σε αναδυόμενο παράθυρο, η περιγραφή κάθε έννοιας (σχήμα 3.11).

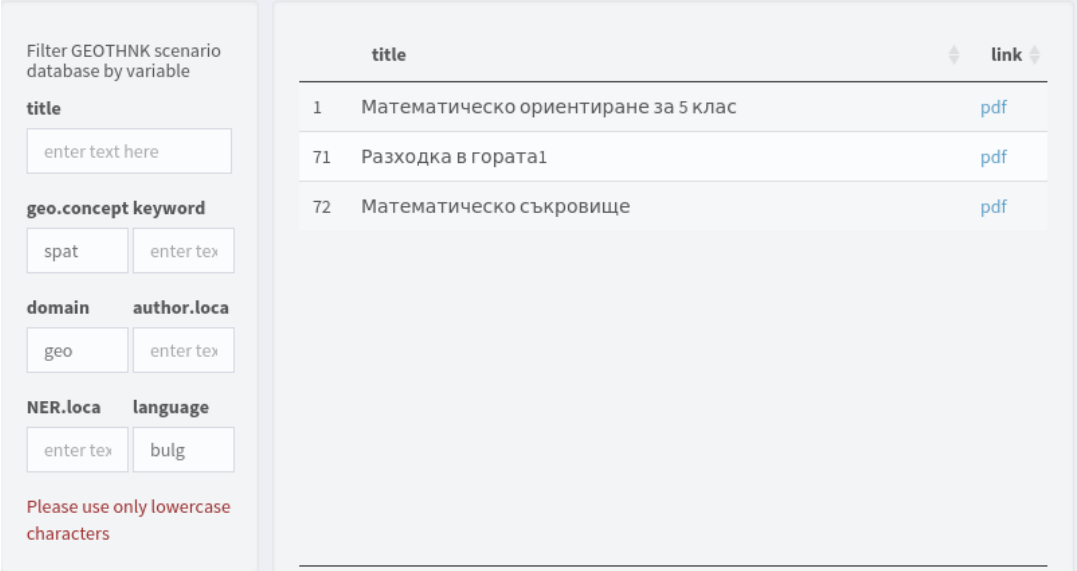


Σχήμα 3.11: Περιεχόμενα θεματικής ενότητας No4 (GEOTHNK network)

Αναζήτηση με όρους κλειδιά

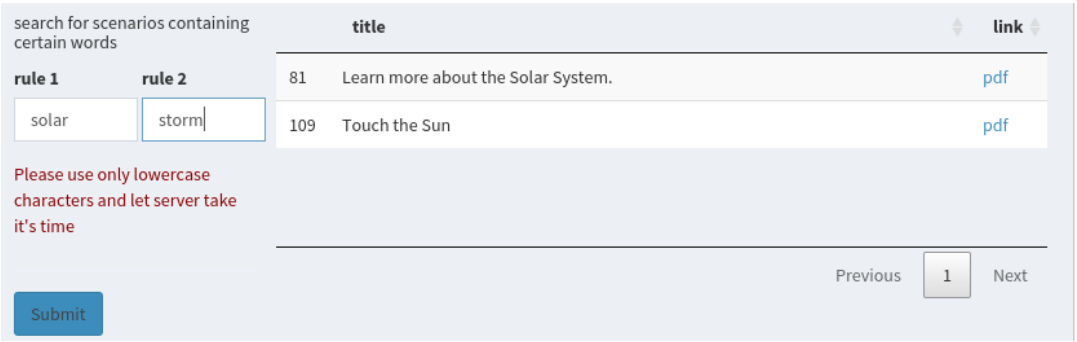
Στη σύνθεση αυτή δίνεται η δυνατότητα στο χρήστη να προβεί σε σύνθετη αναζήτηση αξιοποιώντας τα δεδομένα που συνοδεύουν και χαρακτηρίζουν κάθε σενάριο, όπως ο τίτλος, η γλώσσα στην οποία είναι γραμμένο, οι λέξεις κλειδιά, τα επιστημονικά πεδία ενδιαφέροντος, οι τοποθεσίες και οι έννοιες του χώρου στις οποίες αυτό αναφέρεται. Οι όροι αναζήτησης λειτουργούν σωρευτικά και το αποτέλεσμα της διαδικασίας επιστρέφει μία λίστα με τους τίτλους των σεναρίων που ικανοποιούν όλους τους περιορισμούς που τέθηκαν από το χρήστη. Στην εικόνα του σχήματος 3.12 φαίνεται η λίστα με τα σενάρια στα οποία οι συγγραφείς έχουν ορίσει ότι περιέχουν έννοιες που σχετίζονται με την έννοια χώρος (spat), αναφέρονται σε επιστημονικά πεδία που σχετίζονται με γεωεπιστήμες (geo) και είναι στη Βουλγάρικη γλώσσα (bulg). Η αναζήτηση μπορεί να γίνει σε οποιαδήποτε γλώσσα μόνο για τους τίτλους των σεναρίων επειδή μόνο για αυτούς έχουν ορίσει οι συγγραφείς κείμενο σε άλλη γλώσσα πέραν της Αγγλικής.

Πέραν της πιο πάνω αναζήτησης ο χρήστης μπορεί να προβεί και σε αναζήτηση σεναρίων στα οποία αναφέρονται στο σώμα του κειμένου τους συγκεκριμένες λέξεις. Για



Σχήμα 3.12: Αναζήτηση με όρους κλειδιά

παράδειγμα στο σχήμα 3.13 που ακολουθεί φαίνονται τα σενάρια στα οποία περιέχονται ταυτόχρονα οι όροι solar και storm. Ο χρήστης στην περίπτωση αυτή μπορεί να προβεί σε αναζήτηση όρων σε οποιαδήποτε γλώσσα από τις έξι που έχουν γράψει οι συγγραφείς, Ελληνικά, Αγγλικά, Γερμανικά, Ολλανδικά, Ρουμάνικα και Βουλγάρικα.



Σχήμα 3.13: Αναζήτηση με λέξεις στο σώμα του κειμένου

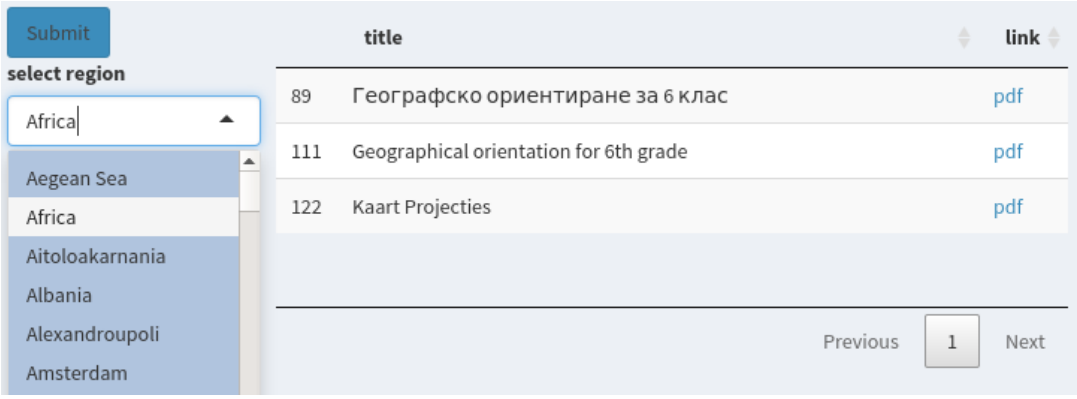
Αναζήτηση με χωρικούς όρους

Στη σύνθεση αυτή δίδεται η δυνατότητα στο χρήστη να αναζητήσει τίτλους σεναρίων με χωρικό τρόπο. Αυτή η αναζήτηση πραγματοποιείται κάνοντας χρήση της ιεραρχίας των τοπωνυμίων που εξήχθηκε από τον γεωγραφικό κατάλογο Geonames και περιγράφηκε σε προηγούμενο κεφάλαιο. Η αναζήτηση γίνεται ξεχωριστά στην ιεραρχία των τοποθεσιών που έχουν ορίσει οι συγγραφείς και αφορούν 25 σενάρια και ξεχωριστά για την ιεραρχία των τοποθεσιών που αλιεύθηκαν από τη διαδικασία Named Entity Recognition και αφορά συνολικά 80 σενάρια.

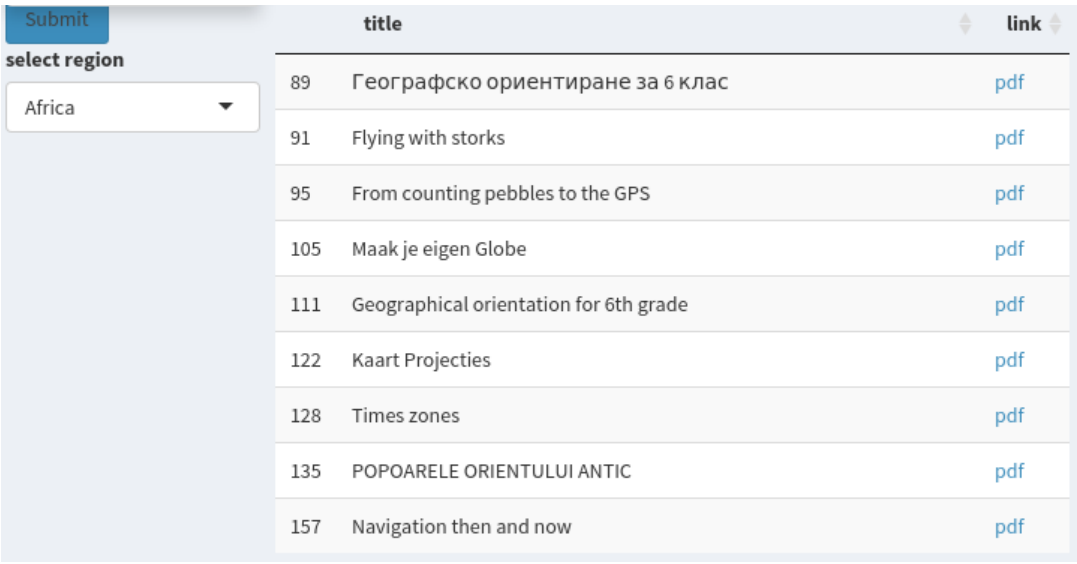
Για παράδειγμα η ιεραρχία της πόλης Shumen είναι Earth, Europe, Republic of Bulgaria,

Oblast Shumen, Obshtina Shumen, Shumen. Η αντίστοιχη για το Μουσείο του Λούβρου στο Παρίσι είναι Earth, Europe, Republic of France, Île-de-France, Paris, Musée du Louvre. Αναζητώντας στην εφαρμογή σενάρια που αναφέρονται στην Ευρώπη η εφαρμογή θα επιστρέψει στα αποτελέσματα τα σενάρια που αναφέρονται τόσο στην πόλη Shumen στη Βουλγαρία όσο και στο μουσείο του Λούβρου, αφού και οι δυο τοποθεσίες βρίσκονται στην Ευρώπη. Ο χρήστης δεν μπορεί να προβεί σε αναζήτηση με ελεύθερο κείμενο αλλά καλείται να επιλέξει κάποια γεωγραφική οντότητα από αυτές που έχουν καθοριστεί από την ιεραρχία.

Στο σχήμα 3.14 που ακολουθεί φαίνονται τα αποτελέσματα της αναζήτησης με τον όρο Africa. Από τα αποτελέσματα φαίνεται και ο χωρικός εμπλουτισμός που έχει συντελεστεί καθώς σύμφωνα με τις τοποθεσίες των συγγραφέων η αναζήτηση επιστρέφει τρεις τίτλους, ενώ σύμφωνα με τις τοποθεσίες που αλιεύθηκαν από την ανάλυση, επιστρέφονται εννέα.



(i) authors



(ii) NER

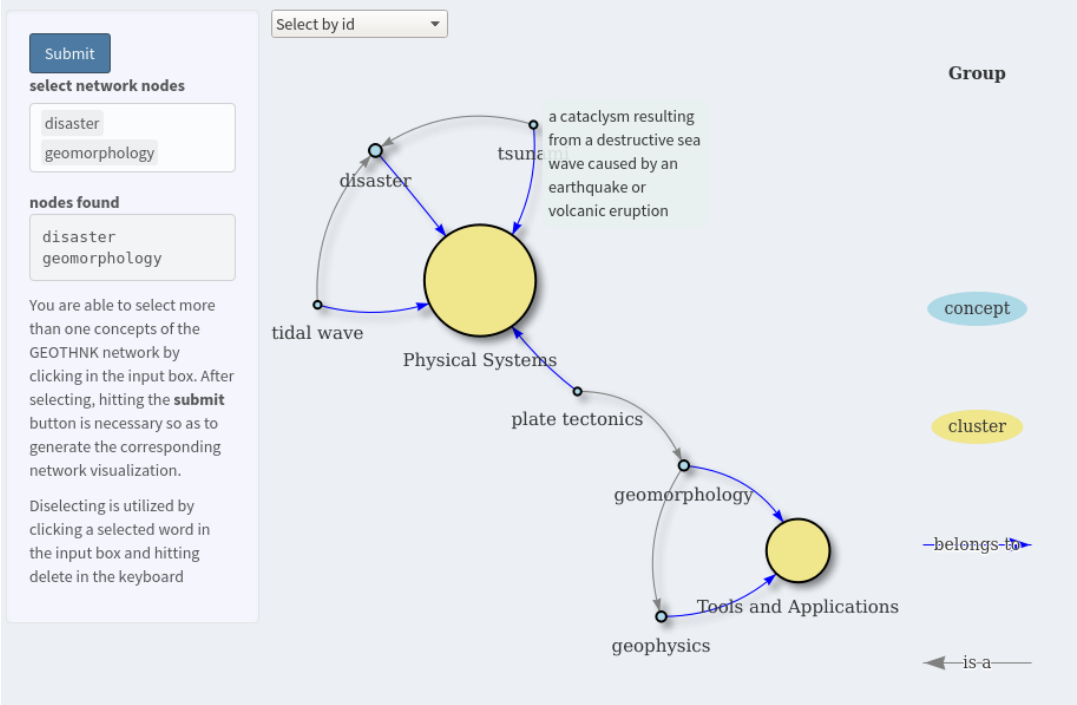
Σχήμα 3.14: Χωρική αναζήτηση σεναρίων

3.2.2 Επισκόπηση δικτύου GEOTHNK

Μπορεί για τους υπολογιστές τα διαγράμματα να μην σημαίνουν απολύτως τίποτα, αλλά για εμάς τους ανθρώπους είναι απαραίτητα. Για την καλύτερη και σε βάθος κατανόηση του δικτύου χωρικών εννοιών του GEOTHNK, παρέχονται στο χρήστη της εφαρμογής τρεις διαφορετικοί διαδραστικοί τρόποι οπτικοποίησής του.

Δίκτυο με προσανατολισμένες συνδέσεις

Η πρώτη στηρίζεται στις δυνατότητες της βιβλιοθήκης vis.js για την οποία παρέχονται εργαλεία στη γλώσσα R μέσω του πακέτου visnetwork. Το δίκτυο έχει αναπαρασταθεί ως ένας γράφος με προσανατολισμένες συνδέσεις στον οποίο οι κόμβοι έχουν χωριστεί σε δυο κατηγορίες, τις έννοιες (concepts) και τις ευρύτερες κατηγορίες που αυτές ανήκουν (clusters). Το μέγεθος κάθε κόμβου είναι ανάλογο του βαθμού που αυτός έχει μέσα στο γράφο, δηλαδή ανάλογο με το πλήθος των άλλων κόμβων που συνδέονται με αυτόν.



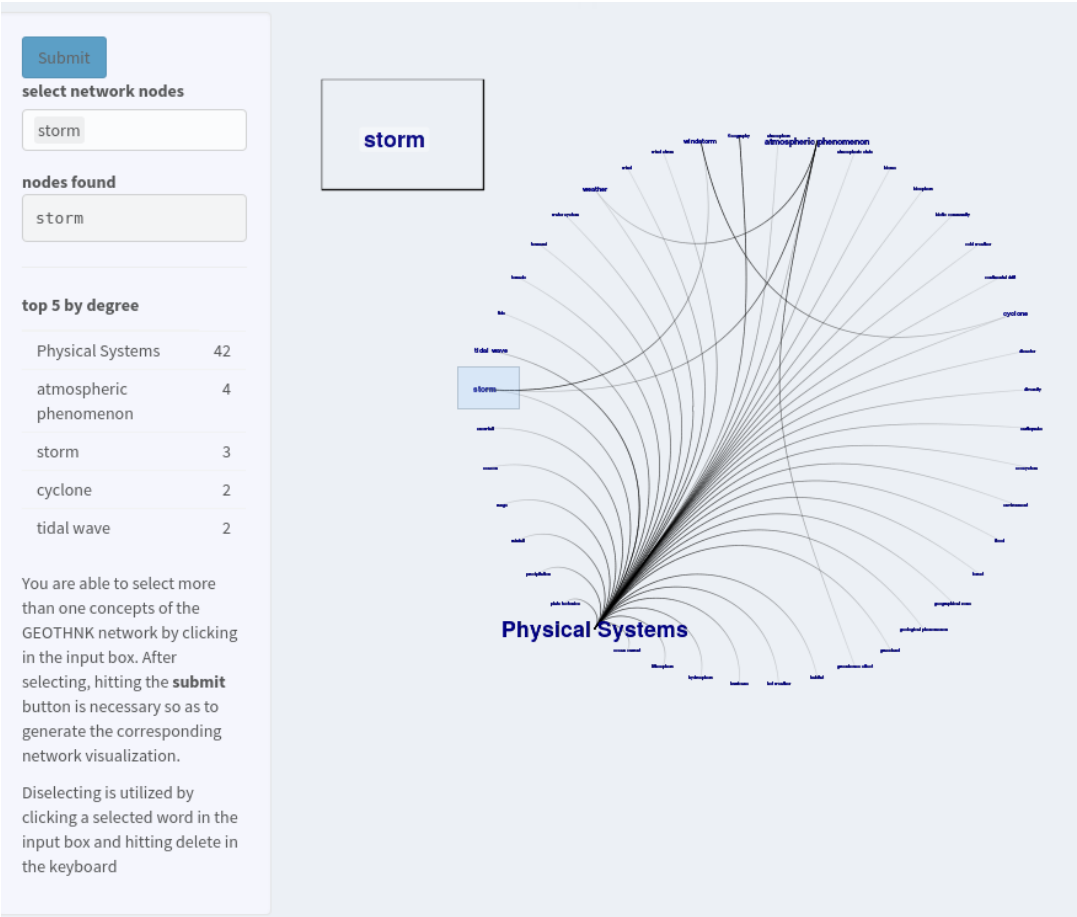
Σχήμα 3.15: Advanced graph visualization

Στο χρήστη παρέχεται η δυνατότητα να επιλέξει μια ή και παραπάνω έννοιες. Στη συνέχεια οι έννοιες αυτές ελέγχεται αν ταυτίζονται λεκτικά με άλλες έννοιες του δικτύου (πχ line με coastline) μιας και αυτές είναι πολύ πιθανό να συνδέονται εννοιολογικά. Έπειτα κατατμείται από το δίκτυο εκείνο το τμήμα των εννοιών που ταυτίστηκαν μαζί με τους κόμβους με τους οποίους συνδέονται άμεσα και τις μεταξύ τους συνδέσεις. Για παράδειγμα στο σχήμα 3.15 φαίνεται το αποτέλεσμα της αναζήτησης με τις έννοιες disaster και geomorphology.

Επιπρόσθετα ο χρήστης μπορεί να προβεί σε μεγέθυνση και μετακίνηση της απεικόνισης αλλά και κάθε κόμβου ξεχωριστά, καθώς επίσης να επιλέξει κάποια από τις έννοιες είτε με το ποντίκι είτε από τον επιλογέα και να σημανθεί αυτή μαζί με τους κοντινότερους γείτονές της. Μετακινώντας τον κέρσορα του ποντικιού πάνω από κάθε κόμβο εμφανίζεται σε αναδυόμενο παράθυρο, η περιγραφή κάθε έννοιας.

Δίκτυο κυκλικού τύπου

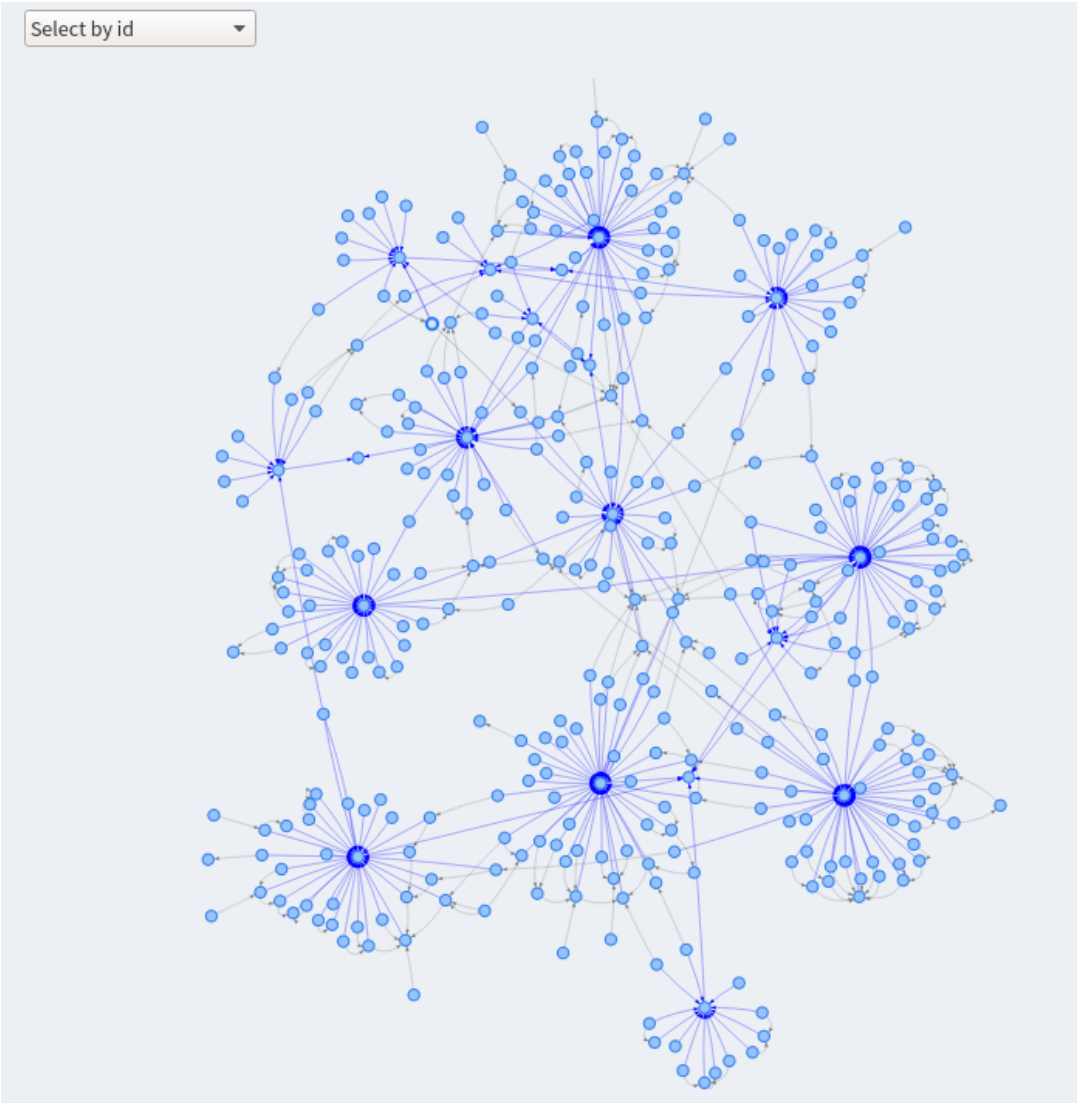
Για το δεύτερο τρόπο απεικόνισης χρησιμοποιήθηκε το πακέτο ggraph το οποίο αποτελεί μια επέκταση του ggplot2. Με αυτόν, το δίκτυο αναπαρίσταται από λεκτικές αναγραφές των εννοιών, οι οποίες είναι διατεταγμένες πάνω στην περιφέρεια ενός κύκλου, με κυρτές γραμμές να αναπαριστούν τις μεταξύ τους συνδέσεις. Ο χρήστης με ανάλογο τρόπο μπορεί να επιλέξει μια η και περισσότερες έννοιες και να επιστραφούν οι ταυτίσεις με τις υπόλοιπες του δικτύου, όπως περιγράφηκε στην πρώτη αναπαράσταση. Αυτή τη φορά το μέγεθος της γραμματοσειράς που χρησιμοποιείται στη λεκτική αναγραφή, είναι ανάλογο του βαθμού του κάθε κόμβου στο γράφο. Επιπρόσθετα σχεδιάζεται και μικρός πίνακας με τις 5 έννοιες της κατάτμησης που έχουν το μεγαλύτερο βαθμό (3.16).



Σχήμα 3.16: Circle graph visualization

Συνολικό δίκτυο

Στην τρίτη και τελευταία οπτικοποίηση το δίκτυο αναπαρίσταται ως ένας απλός γράφος με προσανατολισμένες συνδέσεις με τη διαφορά ότι αναπαρίσταται η οντολογία στην ολότητά της χωρίς παρέμβαση από το χρήστη και χωρίς πολύπλοκη μορφοποίηση (κόμβοι και συνδέσεις με ενιαίο μέγεθος και χρώμα), έτσι ώστε να είναι τεχνικά εφικτή η σχεδίαση από τον server σε χρόνο τέτοιο, που δεν θα αποτρέψει το χρήστη από το να τη χρησιμοποιήσει. Ο χρήστης μπορεί να μεγενθύνει και να μετακινήσει τη σχεδίαση καθώς και μεμονωμένους κόμβους, όπως επίσης να επιλέξει κάποιον με τον κέρσορα είτε με τον επιλογέα και να σημανθεί αυτός και οι κοντινοί του γείτονες (σχήμα 3.17).



Σχήμα 3.17: Simple graph visualization

Κατάλογος λογισμικού

Αναλυτική λίστα με τα πακέτα λογισμικού που χρησιμοποιήθηκαν όπως επίσης και οι απαραίτητες αναφορές στους δημιουργούς τους παρατίθεται στη συνέχεια. Όλα τα πακέτα λογισμικού είναι διαθέσιμα για τη γλώσσα R.

base :

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

cleanNLP :

Taylor Arnold (2017). A Tidy Data Model for Natural Language Processing using cleanNLP. The R Journal, 9(2), 1-20. URL https://journal.r-project.org/archive/2017/RJ-2017-035/index.html.

coreNLP :

Taylor Arnold and Lauren Tilton (NA). coreNLP: Wrappers Around Stanford CoreNLP Tools. R package version 0.4-2.

data.table :

Matt Dowle and Arun Srinivasan (2017). data.table: Extension of data.frame. R package version 1.10.4-3. https://CRAN.R-project.org/package=data.table.

dplyr :

Hadley Wickham, Romain Francois, Lionel Henry and Kirill Müller (2017). dplyr: A Grammar of Data Manipulation. R package version 0.7.4. https://CRAN.R-project.org/package=dplyr.

DT :

Yihui Xie (2018). DT: A Wrapper of the JavaScript Library 'DataTables'. R package version 0.4. https://CRAN.R-project.org/package=DT.

geonames :

Barry Rowlingson (2014). geonames: Interface to www.geonames.org web service. R package version 0.998. https://CRAN.R-project.org/package=geonames.

ggmap :

D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161. URL http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf.

ggplot2 :

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.

ggraph :

Thomas Lin Pedersen (2018). ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. R package version 1.0.1. https://CRAN.R-project.org/package=ggraph.

igraph :

Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. http://igraph.org.

leaflet :

Joe Cheng, Bhaskar Karambelkar and Yihui Xie (2017). leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library. R package version 1.1.0.

https://CRAN.R-project.org/package=leaflet.

leaflet.extras :

Bhaskar Karambelkar (2017). leaflet.extras: Extra Functionality for 'leaflet' Package. R package version 0.2. https://CRAN.R-project.org/package=leaflet.extras.

NLP :

Kurt Hornik (2017). NLP: Natural Language Processing Infrastructure. R package version 0.1-11. https://CRAN.R-project.org/package=NLP.

nominatim :

Bob Rudis (NA). nominatim: Tools for Working with the 'Nominatim' API. R package version 0.2.2.9000.

openNLP :

Kurt Hornik (2016). openNLP: Apache OpenNLP Tools Interface. R package version 0.2-6. https://CRAN.R-project.org/package=openNLP.

pdftools :

Jeroen Ooms (2017). pdftools: Text Extraction, Rendering and Converting of PDF Documents. R package version 1.4. https://CRAN.R-project.org/package=pdftools.

phrasemachine :

Matthew J. Denny, Abram Handler and Brendan O'Connor (2017). phrasemachine: Simple Phrase Extraction. R package version 1.1.2. https://CRAN.R-project.org/package=phrasemachine.

plyr :

Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. URL http://www.jstatsoft.org/v40/i01/.

quanteda :

Benoit K (2018). *quanteda: Quantitative Analysis of Textual Data*. doi: 10.5281/zenodo.1004683 (URL: http://doi.org/10.5281/zenodo.1004683), R package version 0.99.22, <URL: http://quanteda.io>.

readr :

Hadley Wickham, Jim Hester and Romain Francois (2017). readr: Read Rectangular Text Data. R package version 1.1.1. https://CRAN.R-project.org/package=readr.

rgdal :

Roger Bivand, Tim Keitt and Barry Rowlingson (2017). rgdal: Bindings for the 'Geospatial' Data Abstraction Library. R package version 1.2-16. https://CRAN.R-project.org/package=rgdal.

rJava :

Simon Urbanek (2017). rJava: Low-Level R to Java Interface. R package version 0.9-9. https://CRAN.R-project.org/package=rJava.

shiny :

Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2017). shiny: Web Application Framework for R. R package version 1.0.5. https://CRAN.R-project.org/package=shiny.

shinyBS :

Eric Bailey (2015). shinyBS: Twitter Bootstrap Components for Shiny. R package version 0.61.

https://CRAN.R-project.org/package=shinyBS.

shinydashboard :

Winston Chang and Barbara Borges Ribeiro (2017). shinydashboard: Create Dashboards with 'Shiny'. R package version 0.6.1. https://CRAN.R-project.org/package=shinydashboard.

spacyr :

Kenneth Benoit and Akitaka Matsuo (2018). spacyr: Wrapper to the 'spaCy' 'NLP' Library. R package version 0.9.6. https://CRAN.R-project.org/package=spacyr.

stringdist :

van der Loo M (2014). "The stringdist package for approximate string matching." *The R Journal*, *6*, pp. 111-122. <URL: https://CRAN.R-project.org/package=stringdist>.

stringr :

Hadley Wickham (2017). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.2.0. https://CRAN.R-project.org/package=stringr.

textrank :

Jan Wijffels (2017). textrank: Summarize Text by Ranking Sentences and Finding Keywords. R package version 0.2.0. https://CRAN.R-project.org/package=textrank.

tidyr :

Hadley Wickham and Lionel Henry (2018). tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions. R package version 0.8.0. https://CRAN.R-project.org/package=tidyr.

tidytext :

Silge J and Robinson D (2016). "tidytext: Text Mining and Analysis Using Tidy Data Principles in R." *JOSS*, *1*(3). doi: 10.21105/joss.00037 (URL: http://doi.org/10.21105/joss.00037), <URL: http://dx.doi.org/10.21105/joss.00037>.

tm :

Ingo Feinerer and Kurt Hornik (2017). tm: Text Mining Package. R package version 0.7-3. https://CRAN.R-project.org/package=tm Ingo Feinerer, Kurt Hornik, and David Meyer (2008). Text Mining Infrastructure in R. Journal of Statistical Software 25(5): 1-54. URL:http://www.jstatsoft.org/v25/i05/.

visNetwork :

Almende B.V., Benoit Thieurmel and Titouan Robert (2018). visNetwork: Network Visualization using 'vis.js' Library. R package version 2.0.3. https://CRAN.R-project.org/package=visNetwork.

Κεφάλαιο 4

Συμπέρασματα-διαπιστώσεις

Software

Από τα αποτελέσματα της ανάλυσης προέκυψε πως το λογισμικό coreNLP του Stanford University επιστρέφει τα καλύτερα αποτελέσματα σε εργασίες εξόρυξης οντοτήτων με βάση την ονομασία τους όπως για παράδειγμα τα τοπωνύμια. Μειονέκτημα του είναι η κατανάλωση μεγάλου μέρους από τους διαθέσιμους υπολογιστικούς πόρους του συστήματος (CPU, RAM) καθώς και η απαίτηση αρκετού χρόνου για την εκτέλεση των υπολογισμών. Το λογισμικό spacy δικαίως κατέχει τη φήμη του ταχύτερου και δυνατότερου αναλυτή κειμένου στην αγορά. Τα αποτελέσματα ανταγωνίζονται ισάξια εκείνα του coreNLP μέχρι και τη διαδικασία της ανάλυσης μερών του λόγου (sentence splitting, lemmantization, part of speech tagging) και την αναγνώριση ονομασιών τοποθεσίας μεγάλων γεωγραφικών οντοτήτων όπως οι ονομασίες κρατών. Σε τοπωνύμια μικρότερων γεωγραφικών οντοτήτων όπως για παράδειγμα η ονομασία Lavrio ή Kea ή Attica, το coreNLP δεν έχει αντίπαλο.

Από τους τρεις γεωκωδικοποιητές που δοκιμάστηκαν τα καλύτερα αποτελέσματα επέστρεψε το Geonames, το οποίο κατά γενική ομολογία παραμένει ο πιο έγκυρος γεωγραφικός κατάλογος για την ακαδημαϊκή κοινότητα, προσφέροντας πληθώρα χρήσιμων υπηρεσιών, όπως reverse geocoding, υψόμετρο, κοντινές οντότητες σε καθορισμένη απόσταση από μια θέση, εξαγωγή διοικητικής χωρικής ιεραρχίας για κάθε εγγραφή και άλλα. Από την ανάλυση διαπιστώθηκε πως το ποσοστό επιτυχίας που επιτυγχάνει η αυτόματη διαδικασία χωρίς καμία απολύτως αποσαφήνιση και λαμβάνοντας υπόψη μόνο τον πρώτο υποψήφιο της αναζήτησης, αγγίζει το 80%, κάτι που καθιστά το web service του καταλόγου Geonames ένα πολύτιμο εργαλείο σε προγραμματιστικές διαδικασίες όπως στην παρούσα.

Για την ολοκλήρωση της παρούσας εργασίας έγινε προσπάθεια και βρέθηκαν λύσεις και τεχνικές ώστε όλα τα στάδιά της να είναι δυνατό να ολοκληρωθούν χρησιμοποιώντας ελεύθερο λογισμικό και πιο συγκεκριμένα κώδικα σε γλώσσα R. Το ελεύθερο λογισμικό είναι πρώτα απ' όλα δωρεάν, υποστηρίζεται και αναβαθμίζεται δωρεάν από μια μεγάλη κοινότητα χρηστών πρόθυμη να παράσχει βοήθεια σε όσους ξεκινούν, είναι cross-platform και πάνω απ' όλα προσδίδει στο χρήστη ένα αίσθημα ελευθερίας.

Καθώς οι εργαλειοθήκες των λογισμικών γίνονται πλουσιότερες και διατίθενται ελεύθερα και δωρεάν, η ανάλυση πλέον πολύπλοκων προβλημάτων του παρελθόντος γίνεται κάθε μέρα απλούστερη, ταχύτερη και οικονομικότερη. Το μεγαλύτερο όμως πλεονέκτημα και μεγάλο κέρδος για την πρόοδο της επιστήμης, που προσφέρει το ελεύθερο λογισμικό, είναι η δυνατότητα πρόσβασης και χρήσης του από οποιονδήποτε έχει μια ιδέα και θέλει να πειραματιστεί μιας και το κόστος είναι μηδενικό και δεν υπολείπεται ποιότητας. Δεν θα ήταν λάθος να πούμε στο σημείο αυτό ότι κάθε φορά που επεξεργάζονται δεδομένα με ανοιχτό λογισμικό, η αξία τους για την επιστημονική κοινότητα πολλαπλασιάζεται.

Χωρικός εμπλουτισμός

Η διαδικασία του χωρικού εμπλουτισμού των εκπαιδευτικών σεναρίων αύξησε τον αριθμό αυτών που μπορούσαν να αναζητηθούν με χωρικά κριτήρια, από 25 σε 80 στο σύνολο των 159. Αυτό σημαίνει πως πλέον μπορούν να χαρτογραφηθούν και να προσφερθούν ηλεκτρονικές γεω-υπηρεσίες με αυτά, όπως για παράδειγμα σύνθετη αναζήτηση που θα συνυπολογίζει τον παράγοντα χώρος και θα εκμεταλλεύεται τη διοικητική ιεραρχία που συνδέει τις διάφορες θέσεις. Ο χωρικός εμπλουτισμός είναι πολύ σημαντικός όχι μόνο από ακαδημαϊκής σκοπιάς αλλά και οικονομικής καθώς η Oxera στη μελέτη που διεξήγαγε για να μετρήσει τον οικονομικό αντίκτυπο της χρήσης ηλεκτρονικών χαρτογραφικών υπηρεσιών και εφαρμογών για λογαριασμό της Google, διαπίστωσε πως ο κύκλος εργασιών αυτών των υπηρεσιών, σε παγκόσμιο επίπεδο, έφτανε σχεδόν στο μισό του κύκλου εργασιών των αεροπορικών υπηρεσιών (Oxera 2013). Τα νούμερα αυτά, δεδομένης της αύξησης της διείσδυσης της τεχνολογίας και εφαρμογών από το 2013 μέχρι σήμερα, οπωσδήποτε αναμένουμε να έχουν αναθεωρηθεί προς τα πάνω.

Σημασιολογικός εμπλουτισμός

Η διαδικασία του σημασιολογικού εμπλουτισμού πρόσθεσε κατά μέσο όρο τρεις χωρικές έννοιες σε κάθε σενάριο περισσότερες από εκείνες που προσδιόρισαν οι συγγραφείς τους. Επιπλέον υπολογίστηκε η συχνότητα εμφάνισής τους σε κάθε ένα από αυτά, δίνοντας τη δυνατότητα ταξινόμησης αποτελεσμάτων αναζήτησης με βάση συγκεκριμένες έννοιες. Τα σενάρια εμπλουτίστηκαν και με συνοδευτικά δεδομένα που καθορίζουν το νοηματικό περιεχόμενο αυτών. Έτσι βελτιώθηκε η διαδικασία επισκόπησης κάθε σεναρίου και η σωστότερη κατανόηση του περιεχομένου του χωρίς να χαθεί χρόνος ώστε να διαβαστεί το κείμενο του. Ο χρήστης οδηγείται σε μια πιο στοχευμένη διαδικασία ώστε να επιλέξει σωστά και γρήγορα σύμφωνα με τα κριτήρια που τον ενδιαφέρουν.

Προοπτικές

Οι προοπτικές που δημιουργούνται από διαδικασίες παρόμοιες με αυτές που δοκιμάστηκαν στην παρούσα εργασία είναι πράγματι πάρα πολλές και ποικίλουν ανάλογα με το είδος και τη μορφή των δεδομένων που θα χρησιμοποιηθούν. Στη συγκεκριμένη περίπτωση και με τα δεδομένα που έχουν ήδη δημιουργηθεί μπορεί να αναπτυχθεί μια πιο σύνθετη και πολύπλοκη διαδικασία αναζήτησης με βάση την χωρική ιεραρχία των τοπωνυμίων που συνοδεύουν κάθε σενάριο και τη σημασιολογική ιεραρχία των χωρικών εννοιών στο δίκτυο GEOTHNK. Επιπρόσθετα προβλέπεται να δημιουργηθεί ένα σύνολο δεδομένων με όλους τους διαφορετικούς ορισμούς των χωρικών εννοιών του δικτύου από τη βάση του Wordnet, με σκοπό τον προσδιορισμό χωρικών εννοιών που υπονοούνται ή περιγράφονται σε προτάσεις χωρίς όμως να χρησιμοποιείται το λεκτικό τους. Επιπρόσθετα σχεδιάζεται η συσχέτιση των κειμένων με τη μέθοδο LSA έτσι ώστε να αποκαλυφθεί νοηματικό περιεχόμενο και σημασιολογικές σχέσεις που αυτή τη στιγμή βρίσκονται σε λανθάνουσα κατάσταση.

Βιβλιογραφία

[1] Albert Angel, Chara Lontou, Dieter Pfoser, (2008), *Qualitative Geocoding of Persistent Web Pages*, Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems, Article No. 20, Irvine, California, November 05 - 07, 2008 , doi:10.1145/1463434.1463460

[2] Barthes R.,*Μυθολογίες, Μάθημα*, Εναρκτήρια παράδοση στη έδρα της Φιλολογικής Σημειολογίας στο College de France (7Γενάρη 1977),μτφ. Χατζηδήμου Κ., Ράλλη Ι. επιμ. Κρητικός Γ.,Αθήνα: εκδόσεις Ράππα, 1979, 201-225

[3] Buscaldi D.,*Approaches to disambiguating toponyms*, SIGSPATIAL Special, (3)2 : 16-19, Jul 2011.

[4] Caskey S., Kanevsky D., Kozloski J., Sainath T., *Text auto-correction via N-grams* , Google Patents, 2012 , US Patent US 9779080 B2, https://patents.google.com/patent/US9779080B2/en

[5] Clodoveu A. Davis Jr and Frederico T. Fonseca, *Assessing the Certainty of Locations Produced by an Address Geocoding System*, Geoinformatica (2007) 11:103–129, DOI 10.1007/s10707-006-0015-7

[6] Drymonas E., Pfoser D.,*Geospatial Route Extraction from Texts*, Proceedings of the 1st ACM SIGSPATIAL International Workshop on Data Mining for Geoinformatics (DMG) 2010, pp 29-37

[7] Goldberg, D. W.,*Advances in Geocoding Research and Practice*,Transactions in GIS, 15: 727–733., 2011, doi:10.1111/j.1467-9671.2011.01298.x

[8] Gritta, M., Pilehvar, M.T., Limsopatham, N. et al.,*What's missing in geographical parsing?*, Lang Resources and Evaluation pp 1–21, 2017, doi:10.1007/s10579-017-9385-8

[9] Gruber T., *Ontolingua: A Translation Approach to Providing Portable Ontology Specifications*, Knowledge Acquisition, 5(2):199–220, 1993.

[10] Handler, A., Denny, M. J., Wallach, H., O'Connor, B. *Bag of What? Simple Noun Phrase Extraction for Text Analysis*, (2016), In Proceedings of the Workshop on Natural Language

Processing and Computational Social Science at the 2016 Conference on Empirical Methods in Natural Language Processing

[11] Honnibal, Matthew and Johnson, Mark, *Industrial-Strength Natural Language Processing*, https://spacy.io/.

[12] Jurafsky Daniel, Martin H. James,*Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Draft 2008, Prentice-Hall, Inc.

[13] Justeson J., Katz S., *Technical terminology: Some linguistic properties and an algorithm for identification in text.*, 1995, Natural Language Engineering, 1(01):9–27.

[14] Kavouras, M., Kokla. M., Tomai, E., Darra., Pastra K., *GEOTHNK: A Semantic Approach to Spatial Thinking*, 2016, In Progress in Cartography, G. Gartner, M. Jobst and H. Huang, Eds., Springer, Berlin, 2016, pp. 319-338

[15] Kavouras, M., Kokla. M., Tomai, E., Darra, N., Baglatzi, A., Sotiriou, S., Lazoudis, A., *The GEOTHNK platform: connecting spatial thinking to secondary education*, 2014, Proceedings of 14th International Conference on Advanced Learning Technologies, Athens, 7-10 July, 754 – 758

[16] Leetaru Kalev H.,*Fulltext Geocoding Versus Spatial Metadata for Large Text Archives: Towards a Geographically Enriched Wikipedia*,2012, D-Lib Magazine, Volume 18, Number 9/10, doi:10.1045/september2012-leetaru.

[17] Leetaru Kalev, Schrodt Philip A.,*GDELT: Global Data on Events, Location and Tone, 1979-2012*, 2013

[18] Ludovic Moncla, Walter Renteria-Agualimpia, Javier Nogueras-Iso, Mauro Gaio,*Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed hiking descriptions corpus*, ACM. ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL 2014), Nov 2014, Dallas, Texas, United States. Proceedings of the 22th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2014, <10.1145/2666310.2666386>. <hal-01069625v2>

[19] Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky,*The Stanford CoreNLP Natural Language Processing Toolkit*, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2014, pp. 55-60.

[20] Oxera, *What is the economic impact of Geo services?*, 2013, http://www.oxera.com/Latest-Thinking/Publications/Reports/2013/What-is-the-economic-impact-of-Geo-services.aspx, (accessed 20/4/2017)

[21] Patullo Ian, *Improved Document Geocoding for Geo-complex Text*, 2008

[22] Pouliquen Bruno, Marco Kimler, Ralf Steinberger, Camelia Ignat, Tamara Oellinger, Ken Blackler, Flavio Fluart, Wajdi Zaghouani, Anna Widiger, Ann-Charlotte Forslund, Clive Best,*Geocoding Multilingual Texts: Recognition, Disambiguation and Visualisation*, Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006), pp. 53-58. Genoa, Italy, 24-26 May 2006.

[23] Reiter N., Frank A., *Identifying generic noun phrases*, 2010, In Proceedings of ACL, pages 40–49.

[24] Rusu D., Dali L., Fortuna B., Grobelnik M., Mladenic D., *Triplet extraction from sentences*, in Proceedings of the 10th International Multiconference" Information Society-IS, 2007, pp. 8–12.

[25] Wiemer-Hastings, P., Wiemer-Hastings, K., Graesser, A., *Latent semantic analysis*. In:Proceedings of the 16th international joint conference on Artificial intelligence. pp. 1–14. Citeseer (2004)

[26] Xianping Ge, *Address geocoding*, Google Patents, 2005 , US Patent US 6934634 B1, https://www.google.com/patents/US

[27] Zhang Wei, Judith Gelernter,*Geocoding location expressions in Twitter messages: A preference learning method*, Journal of Spatial Information Science, 9(1): 37–70, (2014) doi:10.5311/JOSIS.2014.9.170

[28] Geonames.org gazetteer, http://www.geonames.org/ (accessed 21/4/2017)

[29] The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text.(https://opennlp.apache.org)