ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Ανάπτυξη εργαλείου ημι-αυτόματης επισημείωσης μέσω ελεγχόμενου λεξιλογίου

Άννα Β. Μητσοπούλου

Επιβλέποντες: **Ιωάννης Ιωαννίδης**, Καθηγητής
Δρ. Ακριβή Κατηφόρη, **Μυρτώ Κουκούλη**

ΑΘΗΝΑ

ΣΕΠΤΕΜΒΡΙΟΣ 2020

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Ανάπτυξη εργαλείου ημι-αυτόματης επισημείωσης κειμένου μέσω ελεγχόμενου λεξιλογίου

Άννα Β. Μητσοπούλου
Α.Μ.: 1115201500097

ΕΠΙΒΛΕΠΟΝΤΕΣ: **Ιωάννης Ιωαννίδης**, Καθηγητής
Δρ. Ακριβή Κατηφόρη, **Μυρτώ Κουκούλη**

ΠΕΡΙΛΗΨΗ

Η παρούσα πτυχιακή εργασία επικεντρώνεται στην ανάπτυξη μιας διαδικτυακής εφαρμογής για την επισημείωση κειμένων, σχετικών με την πολιτιστική κληρονομιά. Αναλυτικότερα, η εφαρμογή προσαρτά ετικέτες σε λέξεις και σύνολα λέξεων ενός κειμένου αλλά και σε ολόκληρες ενότητες. Η επισημείωση, στο επίπεδο των λέξεων, επιτυγχάνεται με την εύρεση συνδέσεων με ελεγχόμενο λεξιλόγιο, ενώ η επισημείωση, στο επίπεδο των ενοτήτων, προκύπτει από το σύνολο των ετικετών των επισημειωμένων λέξεων της εκάστοτε ενότητας. Η πτυχιακή αυτή χρησιμοποιεί σαν περιεχόμενο για την αξιολόγηση και επίδειξη του εργαλείου το υλικό που δημιουργήθηκε στα πλαίσια του ερευνητικού έργου Pros-eleusis. Το έργο έχει στόχο την ανάδειξη πτυχών της πολιτιστικής κληρονομιάς της Ελευσίνας, μέσω κειμένων για την αρχαία, νεότερη και σύγχρονη ζωή στην πόλη.

Αρχικά, για την ανάπτυξη της εφαρμογής μελετήθηκαν ήδη υπάρχοντα ελεγχόμενα λεξιλόγια, σχετικά με την πολιτιστική κληρονομιά, αλλά και υπάρχουσες εφαρμογές επισημείωσης κειμένων γενικού περιεχομένου. Στη συνέχεια, μετά τον καθορισμό των τεχνολογιών για την υλοποίηση της εφαρμογής, ξεκίνησε ο σχεδιασμός και η ανάπτυξή της. Η αναλυτική περιγραφή της τελικής εφαρμογής αποτελεί κομμάτι αυτής της πτυχιακής εργασίας, η οποία ολοκληρώνεται με την αξιολόγηση της εφαρμογής από χρήστες, τα συμπεράσματα που προέκυψαν από την αξιολόγηση, καθώς και την παρουσίαση κάποιων ιδεών για μελλοντικές επεκτάσεις της εφαρμογής.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Επιστήμη Δεδομένων

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Επισημείωση κειμένου, Ελεγχόμενο λεξιλόγιο, Διαδικτυακή εφαρμογή

ABSTRACT

The current thesis focuses on the development of a web application for text annotation, relevant to cultural heritage. In more detail, the web application adds labels to words, groups of words, and entire units. The annotation, at the level of words, is achieved by finding links with the controlled vocabulary, while the annotation, at the level of sections, is a result of the set of tags of the annotated words of this section. This thesis uses as content, for the evaluation and the demonstration of the application, the material, which was created in the research project Pros-eleusis. The project aims to highlight aspects of the cultural heritage of Eleusis, through texts of ancient and modern life in the city.

Initially, for the development of the application, existing controlled vocabularies related to cultural heritage were studied, as well as, existing applications for annotating texts with general content. Then, after defining the technologies for the implementation of the application, its design and development began. The detailed description of the final application is part of this thesis, which concludes with the evaluation of the application by users, the conclusions that emerged from the users, as well as, the presentation of some ideas for future extensions of the application.

SUBJECT AREA: Data Science

KEYWORDS: Text annotation, Controlled vocabulary, Web application

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω όλα τα άτομα που βοήθησαν για την ολοκλήρωση της πτυχιακής μου.

Αρχικά τον καθηγητή Ιωάννη Ιωαννίδη που μου ανέθεσε την εν λόγω πτυχιακή εργασία.

Τους επιβλέποντες Ακριβή Κατηφόρη και Μυρτώ Κουκούλη για την πλήρη υποστήριξη που μου παρείχαν για την διεκπεραίωση της παρούσας πτυχιακής εργασίας.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΠΡΟΛΟΓΟΣ

Η παρούσα πτυχιακή εργασία εκπονήθηκε στο Τμήμα Πληροφορικής και Τηλεπικοινωνιών του Εθνικού Καποδιστριακού Πανεπιστημίου Αθηνών ως πτυχιακή εργασία. Η διάρκεια διεξαγωγής της μέχρι και την ολοκλήρωσή της ήταν επτά μήνες. Υπεύθυνος καθηγητής ήταν ο Κ. Ιωάννης Ιωαννίδης και επιβλέποντες η Κα. Ακριβή Κατηφόρη και η Κα. Μυρτώ Κούκουλη.

1. ΕΙΣΑΓΩΓΗ

Στο κεφάλαιο αυτό παρουσιάζονται κάποιες εισαγωγικές πληροφορίες για την πτυχιακή εργασία. Ειδικότερα, παρουσιάζεται το αντικείμενο της, ο στόχος της, καθώς επίσης και ο τρόπος με τον οποίο είναι διαρθρωμένη σε κεφάλαια.

1.1 Αντικείμενο Πτυχιακής

Το αντικείμενο αυτής της πτυχιακής εργασίας είναι η σχεδίαση και υλοποίηση ενός εργαλείου, το οποίο θα παρέχει τη δυνατότητα ημι-αυτόματης επισημείωσης κειμένου. Ειδικότερα, με την χρήση ενός ελεγχόμενου λεξιλογίου, θα πρέπει να αναγνωρίζονται συνδέσεις ανάμεσα σε λέξεις ή φράσεις του κειμένου και έννοιες του λεξιλογίου και με τον τρόπο αυτό θα επιτρέπεται η επισημείωση του κειμένου ανά ενότητα. Επιπλέον, θα δίνεται στον χρήστη η δυνατότητα προσθήκης νέων αντίστοιχων συνδέσεων αλλά και διαγραφής των υπαρχόντων.

Η πτυχιακή αυτή χρησιμοποιεί σαν περιεχόμενο για την αξιολόγηση και επίδειξη του εργαλείου τα κείμενα που δημιουργήθηκαν στα πλαίσια του ερευνητικού έργου Pros-eleusis. Το έργο έχει στόχο την ανάδειξη πτυχών της πολιτιστικής κληρονομιάς της Ελευσίνας, μέσω κειμένων για την αρχαία, νεότερη και σύγχρονη ζωή στην πόλη. Στο πλαίσιο του έργου δημιουργήθηκαν δύο τύποι διαδρομών μεταξύ σημείων ενδιαφέροντος διάσπαρτων στην πόλη: ένας για κατευθυνόμενη περιήγηση (προκαθορισμένη διαδρομή) και ένας δεύτερος για ελεύθερη περιήγηση με εξατομικευμένες προτάσεις (μέσω συστήματος που αναλύει τις επιλογές του χρήστη και του προτείνει δυναμικά τους επόμενους σταθμούς στην διαδρομή του).

Για κάθε σημείο ενδιαφέροντος μέσα στην πόλη της Ελευσίνας υπάρχουν ένα ή περισσότερα σύντομα κείμενα με πληροφορίες. Για να μπορέσει να λειτουργήσει το σύστημα εξατομίκευσης είναι απαραίτητο να έχει οριστεί ένας βαθμός ομοιότητας μεταξύ των σημείων ενδιαφέροντος. Για τον προσδιορισμό αυτής της ομοιότητας, χρησιμοποιήθηκαν επισημειώσεις, τις οποίες προσέθεσαν σε κάθε ενότητα οι συγγραφείς της, επιλέγοντας σημαντικές λέξεις κλειδιά (τοπωνύμια, ονόματα επιφανών προσώπων, γενικότερες έννοιες, κτλ) σαν χαρακτηριστικές για την ενότητα. Οι ειδικές αυτές έννοιες στη συνέχεια συνδέθηκαν με μια ευρύτερη κατηγοριοποίηση καταλήγοντας έτσι σε ένα δομημένο λεξιλόγιο, το οποίο χρησιμοποιήθηκε στα πλαίσια της πτυχιακής αυτής.

Στα πλαίσια αυτής της δραστηριότητας επισημείωσης από τους συγγραφείς, η οποία υπήρξε αρκετά χρονοβόρα, αναγνωρίστηκε η ανάγκη να υποστηριχθεί η εύκολη και γρήγορη προσθήκη επισημειώσεων στο κείμενο και προτάθηκε η δημιουργία ενός εργαλείου (ημι-)αυτόματης επισημείωσης, την ανάπτυξη του οποίου προσπάθησε να καλύψει η παρούσα πτυχιακή εργασία.

1.2 Στόχος Πτυχιακής

Ο στόχος της παρούσας πτυχιακής εργασίας είναι η ανάπτυξη ενός εύχρηστου εργαλείου, που θα ελαχιστοποιεί την προσπάθεια του χρήστη για την επισημείωση κειμένων, σχετικών με την πολιτιστική κληρονομιά. Για την επίτευξη αυτού του στόχου σχεδιάστηκε και υλοποιήθηκε μία εφαρμογή, η οποία επισημειώνει αυτόματα ένα κείμενο και στη συνέχεια επιτρέπει στο χρήστη να κάνει τις απαραίτητες αλλαγές.

Λαμβάνοντας υπόψη την όλο και αυξανόμενη χρήση ψηφιακών οδηγών για την περιήγηση των επισκεπτών σε μουσεία, πολιτιστικούς χώρους αλλά και το αστικό τοπίο, η εξατομίκευση καλείται να παίξει σημαντικό ρόλο στην αναζήτηση και εξερεύνηση του περιεχομένου από τους χρήστες, ώστε να έχουν άμεση πρόσβαση σε υλικό που τους

ενδιαφέρει κατά την περιήγησή τους. Η προετοιμασία του περιεχομένου μέσω της επισημείωσης αποτελεί σημαντικό κομμάτι για την εξασφάλιση της επιτυχίας του αλγορίθμου συστάσεων που θα χρησιμοποιηθεί και για το σκοπό αυτό είναι σημαντικό να υποστηριχθούν οι χρήστες που πραγματοποιούν την επισημείωση. Αυτός είναι και ο ρόλος του εργαλείου που αναπτύχθηκε στα πλαίσια της πτυχιακής αυτής.

1.3 Διάρθρωση Πτυχιακής

Μετά την ολοκλήρωση του παρόντος κεφαλαίου, το οποίο αποτελεί το πρώτο κεφάλαιο της πτυχιακής εργασίας, ακολουθούν τα παρακάτω κεφάλαια, για τα οποία στο σημείο αυτό προσφέρεται μία συνοπτική περιγραφή:

Στο κεφάλαιο 2 πραγματοποιείται μια επισκόπηση διαδεδομένων ελεγχόμενων λεξιλογίων και οντολογιών, σχετικών με την πολιτιστική κληρονομιά.

Στο κεφάλαιο 3 γίνεται μία ανασκόπηση του τομέα της επισημείωσης κειμένου με την παρουσίαση σχετικών εφαρμογών.

Στο κεφάλαιο 4 παρουσιάζονται τα προφίλ των χρηστών της εφαρμογής, οι ανάγκες της και οι προδιαγραφές για την ικανοποίηση των αναγκών αυτών.

Στο κεφάλαιο 5 αναφέρονται τα εργαλεία και οι τεχνολογίες που χρησιμοποιήθηκαν για την ανάπτυξη της εφαρμογής και παρουσιάζεται αναλυτικά η τελική έκδοσή της.

Στο κεφάλαιο 6 γίνεται αναφορά στον τρόπο αξιολόγησης της εφαρμογής και παρουσιάζονται τα αποτελέσματα της αξιολόγησης.

Στο κεφάλαιο 7 γίνεται μια παρουσίαση γενικών συμπερασμάτων για την επισημείωση κειμένου και παρατίθενται κάποιες μελλοντικές επεκτάσεις για την βελτίωση της εφαρμογής.

2. ΕΛΕΓΧΟΜΕΝΑ ΛΕΞΙΛΟΓΙΑ ΚΑΙ ΟΝΤΟΛΟΓΙΕΣ ΣΤΗΝ ΠΟΛΙΤΙΣΤΙΚΗ ΚΛΗΡΟΝΟΜΙΑ

Η επισημείωση του κειμένου, όπως έχει ήδη αναφερθεί, θα πραγματοποιείται με την χρήση ενός ελεγχόμενου λεξιλογίου. Στην ενότητα αυτή, γίνεται μία περιγραφή των πιο διαδεδομένων ελεγχόμενων λεξιλογίων, σχετικών με την πολιτιστική κληρονομιά, καθώς και της οντολογίας CIDOC, βάσεις για την ανάπτυξη του ελεγχόμενου λεξιλογίου που θα χρησιμοποιηθεί σε αυτήν την πτυχιακή.

2.1 Art and Architecture Thesaurus

Το Art and Architecture Thesaurus (AAT) [1] είναι ένα ελεγχόμενο λεξιλόγιο, που χρησιμοποιείται για να περιγράψει αντικείμενα τέχνης, αρχιτεκτονικής και υλικής πολιτιστικής κληρονομιάς. Η δημιουργία του AAT ξεκίνησε στα τέλη της δεκαετίας του 1980, με σκοπό την βελτίωση της πρόσβασης σε πληροφορίες σχετικά με την τέχνη, την αρχιτεκτονική και ό,τι άλλο σχετίζεται με την υλική πολιτιστική κληρονομιά. Το AAT χρησιμοποιείται από μουσεία, βιβλιοθήκες τέχνης, αρχεία αλλά και από ερευνητές της τέχνης και της ιστορίας της τέχνης. Είναι ένα σύστημα πολύπλευρης και ιεραρχικής ταξινόμησης. Το AAT υποστηρίζει σχέσεις ισοδυναμίας, συσχετιστικές και ιεραρχικές. Οι κατηγορίες οργανώνονται εννοιολογικά σε ένα σχήμα που ξεκινάει από αφηρημένες έννοιες και καταλήγει σε συγκεκριμένα φυσικά αντικείμενα. Οι αρχικές κατηγορίες και οι ιεραρχίες του AAT είναι οι:

- ASSOCIATED CONCEPTS FACET (*Hierarchy:* Associated Concepts)
- PHYSICAL ATTRIBUTES FACET (Hierarchies: Attributes and Properties, Conditions and Effects, Design Elements, Color)
- STYLES AND PERIODS FACET (*Hierarchy:* Styles and Periods)
- AGENTS FACET (*Hierarchies:* People, Organizations, Living Organisms)
- ACTIVITIES FACET (*Hierarchies:* Disciplines, Functions, Events, Physical and Mental Activities, Processes and Techniques)
- MATERIALS FACET (*Hierarchy:* Materials
- OBJECTS FACET (*Hierarchies:* Object Groupings and Systems, Object Genres, Components)
- BRAND NAMES FACET (Hierarchy: Brand Names)

Η κατηγορία Associated Concepts περιλαμβάνει αφηρημένες έννοιες και φαινόμενα, που σχετίζονται με την μελέτη και την εκτέλεση ενός ευρέος φάσματος της ανθρώπινης σκέψης και δραστηριότητας. Μερικά παραδείγματα αντικειμένων της κατηγορίας αυτής είναι η γνώση, η ελευθερία ή η ομορφιά.

Η κατηγορία Agents περιλαμβάνει όρους για προσδιορισμούς ανθρώπων, ομάδες ανθρώπων ή οργανισμούς, με βάση το επάγγελμα, της δραστηριότητές τους, φυσικά ή πνευματικά χαρακτηριστικά. Μερικά παραδείγματα αντικειμένων της κατηγορίας αυτής είναι διάφορες εταιρείες, αρχιτέκτονες, μοναστήρια και άλλα. Στην κατηγορία αυτή, και ειδικότερα, στους ζωντανούς οργανισμούς (Living Organisms) συγκαταλέγονται και ζώα ή φυτά.

Η κατηγορία Activities περιλαμβάνει δραστηριότητες, οι οποίες μπορεί να ποικίλουν από διανοητικές εργασίες εώς διαδικασίες που εκτελούνται με υλικά και αντικείμενα, αλλά και

από μεμονωμένες φυσικές δράσεις εώς πολύπλοκα παιχνίδια. Μερικά παραδείγματα αντικειμένων της κατηγορίας αυτής είναι το τρέξιμο, η ζωγραφική, η αρχαιολογία, εκθέσεις ή διάφοροι διαγωνισμοί.

Αυτή την στιγμή το AAT περιέχει περίπου 60.000 εγγραφές και 375.000 όρους και κάθε χρόνο προστίθενται και καινούργιοι όροι.

2.2 SHIC

Το SHIC(Social History and Industrial Classification) [3], είναι ένα ελεγχόμενο λεξιλόγιο, ευρέως χρησιμοποιούμενο από ιστορικά μουσεία, για να δημιουργεί συνδέσεις μεταξύ αντικειμένων, με βάση το περιεχόμενο και το ιστορικό τους. Δημιουργήθηκε στο Ηνωμένο Βασίλειο στα τέλη του 1970 με αρχές του 1980, λόγω της ανάγκης των μουσείων για κατηγοριοποίηση υλικού σχετικό με την κοινωνική ιστορία. Το SHIC κατηγοριοποιεί υλικά (βιβλία, αντικείμενα, εγγραφές κλπ) ανάλογα με την αλληλεπίδραση του ανθρώπου με αυτά. Οι κατηγορίες του μέχρι το δεύτερο επίπεδο του φαίνονται στις παρακάτω εικόνες.

Εικόνα 1: Κύριες κατηγορίες του SHIC

Στην κατηγορία Community life περιλαμβάνεται οποιοδήποτε υλικό σχετίζεται με την κοινωνία και όχι μόνο με μεμονωμένα άτομα ή οικογένειες.

Η κατηγορία Domestic and family life καλύπτει όλους τους τομείς της οικογενειακής ζωής, συμπεριλαμβανομένου του σπιτιού και όποιας δραστηριότητας σχετίζεται με το σπίτι ή γίνεται μέσα σε αυτό.

Στην κατηγορία Personal life περιλαμβάνονται όσα αντικείμενα ανήκουν, χρησιμοποιούνται ή σχετίζονται με ένα άτομο, αντί να συμμετέχουν στην οικιακή ζωή.

Η κατηγορία Working life χρησιμοποιείται για την κατηγοριοποίηση οποιασδήποτε εργασιακής δραστηριότητας. Περιλαμβάνει εμπορικές συναλλαγές, κατασκευές, βιομηχανίες υπηρεσίες και μεταφορές αλλά και οποιασδήποτε βοηθητικές δραστηριότητες, που σχετίζονται άμεσα με τις παραπάνω δραστηριότητες.

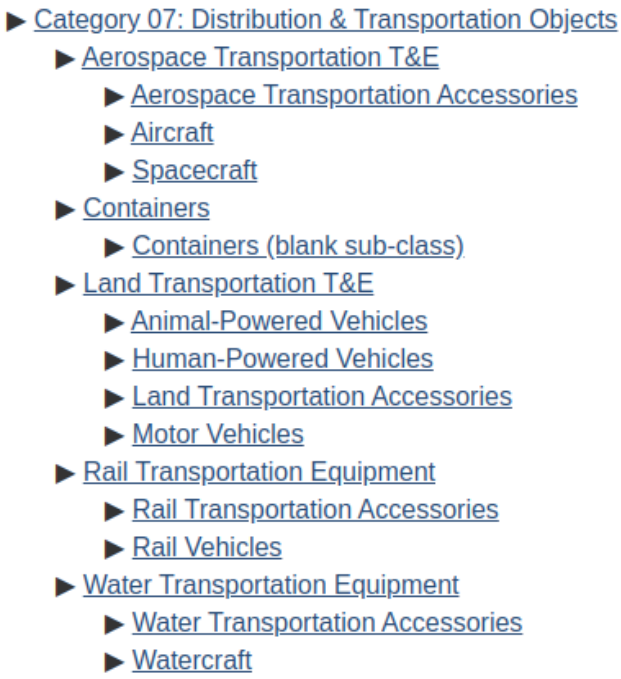
2.3 Nomenclature

Το Nomenclature [3] είναι το πιο ευρέως χρησιμοποιούμενο ελεγχόμενο λεξιλόγιο για ιστορικές και εθνολογικές συλλογές στην Βόρεια Αμερική. Από το 1978, που δημοσιεύθηκε πρώτη φορά, έχει βελτιωθεί και επεκταθεί, αλληλεπιδρώντας με την κοινότητα μουσείων, την οποία εξυπηρετεί. Περιγράφει κυρίως αντικείμενα σχετικά με την ιστορία και την τέχνη της Βόρειας Αμερικής και χρησιμοποιείται από μουσεία αλλά και άλλους οργανισμούς. Το Nomenclature αποτελεί την βάση για την κατηγορία των αντικειμένων του Getty's Art and Architecture Thesaurus.

Το Nomenclature παρέχει μια απλή δομή κατηγοριοποίησης, η οποία κατατάσσει τα αντικείμενα σύμφωνα με την λειτουργικότητά τους. Οι 10 αρχικές κατηγορίες είναι:

- Built Environment Objects
- Furnishings
- Personal Objects
- Tools & Equipment for Materials
- Tools & Equipment for Communication
- Distribution & Transportation Objects
- Communication Objects
- Recreational Objects
- Unclassifiable Objects

Το μέσο βάθος της κάθε κατηγορίας είναι 2. Η παρακάτω εικόνα παρουσιάζει ένα παράδειγμα πλήρης ανάλυσης μιας εκ των βασικών κατηγοριών.



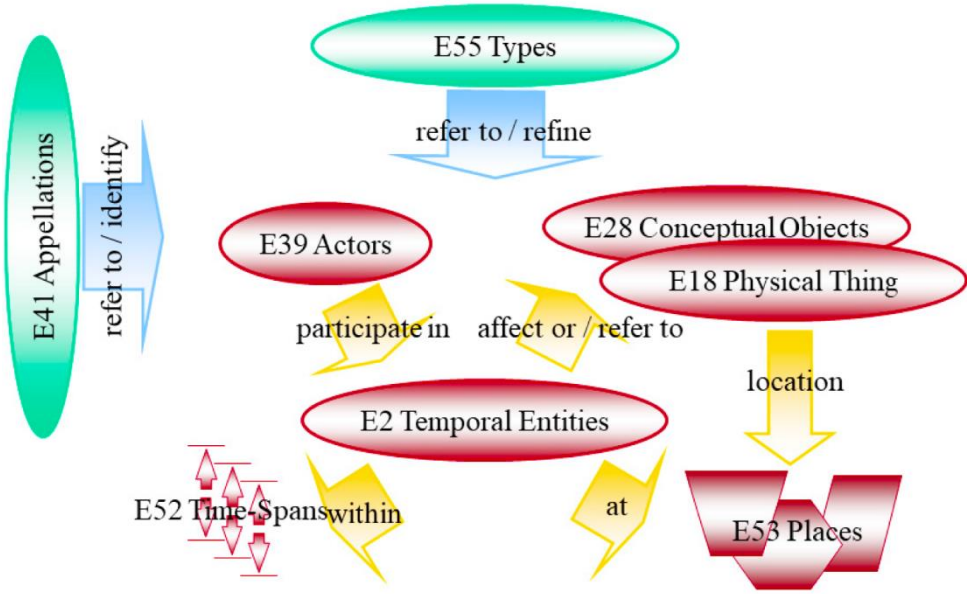
Εικόνα 2: Παράδειγμα κατηγορίας του Nomenclature

Αυτή την στιγμή το Nomenclature περιέχει περίπου 15.000 εγγραφές.

2.4 CIDOC Conceptual Reference Model (CRM)

Το CIDOC-CRM [4] είναι μία οντολογία που προσφέρει ορισμούς και επίσημη δομή για την περιγραφή άμεσων και έμμεσων εννοιών και σχέσεων, που χρησιμοποιούνται στην πολιτιστική κληρονομιά. Αποτελεί το αποτέλεσμα της πάνω από 20 χρόνια ανάπτυξης και συντήρησης, αρχικά του CIDOC Documentation Standards Working Group και, πλέον, του CIDOC CRM SIG. Το CIDOC CRM αναπτύχθηκε με σκοπό να προβάλλει μια κοινή κατανόηση για τις πληροφορίες της πολιτιστικής κληρονομιάς , παρέχοντας μία κοινή και εκτεταμένη σημασιολογική δομή για την ενσωμάτωση πληροφοριών της πολιτιστικής κληρονομιάς, βάσει αποδείξεων.

Οι βασικές κλάσεις καθώς και οι σχέσεις μεταξύ τους φαίνονται στην παρακάτω εικόνα.



Εικόνα 3: Βασικές κλάσεις του CIDOC

Η κλάση E2 αντιπροσωπεύει γεγονότα και είναι η βασική κλάση του CIDOC-CRM.

Η κλάση E53 αντιπροσωπεύει μέρη, τα οποία δεν είναι αναγκαστικά γεωγραφικές τοποθεσίες αλλά μπορεί να είναι και το εσωτερικό ενός ποτηριού ή το μπροστά μέρος ενός πλοίου.

Η κλάση E39 αντιπροσωπεύει άτομα ή ομάδες ατόμων.

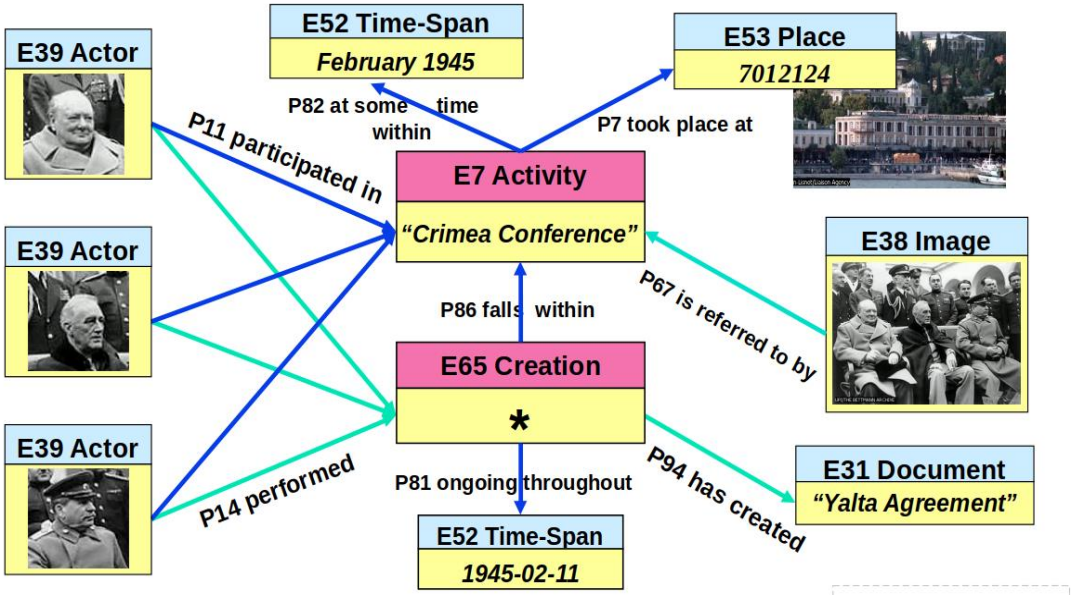
Η κλάση E18 αντιπροσωπεύει φυσικά αντικείμενα. Τα φυσικά αντικείμενα είναι αντικείμενα, τα οποία μπορούν να καταστραφούν μετατρέποντας τα σε μικρά κομμάτια, για τα οποία δεν υπάρχει ενδιαφέρον καταγραφής. Η κλάση E28 αντιπροσωπεύει εννοιολογικά αντικείμενα. Για να μπορέσουμε να καταστρέψουμε ένα εννοιολογικό αντικείμενο, θα πρέπει να καταστρέψουμε όλους του φέροντες αυτού, είτε είναι ανθρώπινα μυαλά είτε μνήμες υπολογιστών ή άλλα μέρη, στα οποία το εννοιολογικό αντικείμενο έχει αποθηκευτεί. Ο διαφορετικός τρόπος καταστροφής των αντικειμένων που ανήκουν στην κλάση E18 και E28, είναι και ο τρόπος διαχωρισμού τους.

Η κλάση E41 δίνει την δυνατότητα να δοθεί όνομα σε ένα στιγμιότυπο οποιασδήποτε κλάσης του CIDOC-CRM. Ένα στιγμιότυπο κλάσης μπορεί να έχει πολλά διαφορετικά ονόματα.

Η κλάση E55 δίνει την δυνατότητα κατηγοριοποίησης των στιγμιοτύπων των κλάσεων. Στο CIDOC-CRM ο χρήστης μπορεί να έχει ταυτόχρονα πολλές διαφορετικές κατηγοριοποιήσεις. Η κλάση E55 υπάγεται στην κλάση E28.

Οι κατηγοριοποιήσεις και οι ονομασίες είναι κομμάτι του μοντέλου, σε αντίθεση με άλλα μοντέλα, γεγονός που επιτρέπει την ευκολότερη και αποτελεσματικότερη μεταφορά του.

Ένα παράδειγμα αναπαράστασης αντικειμένου με την χρήση του CIDOC-CRM απεικονίζεται παρακάτω.



Εικόνα 4: Αναπαράσταση αντικειμένου με CIDOC

3.ΑΝΑΣΚΟΠΗΣΗ ΤΟΥ ΤΟΜΕΑ ΤΗΣ ΕΠΙΣΗΜΕΙΩΣΗΣ ΚΕΙΜΕΝΟΥ

Στο παρόν κεφάλαιο γίνεται μία αναφορά στην ανάλυση κειμένου, που αποτελεί τον τρόπο με τον οποίο επιτυγχάνεται η επισημείωση και περιγράφονται συνοπτικά κάποιες εφαρμογές για επισημείωση κειμένων γενικού περιεχομένου.

3.1 Ανάλυση κειμένου και εργαλεία

Η ανάλυση κειμένου είναι ένα υποπεδίο της γλωσσολογίας, της πληροφορικής, της τεχνολογίας της πληροφορίας και της μηχανικής μάθησης. Ασχολείται με την αλληλεπίδραση ανάμεσα στους υπολογιστές και τις ανθρώπινες γλώσσες, και ειδικότερα με τον προγραμματισμό των υπολογιστών για να επεξεργάζονται και να αναλύουν μεγάλες ποσότητες δεδομένων φυσικής γλώσσας [5].

Οι εργασίες για την ανάλυση κειμένου χωρίζονται σε δύο βασικές κατηγορίες, στις συντακτικές και τις σημασιολογικές.

Οι κυριότερες συντακτικές εργασίες για την ανάλυση ενός κειμένου είναι οι:

- Tokenization: ο χωρισμός ενός συνεχόμενου κειμένου σε λέξεις.
- Part-of-speech tagging: ο καθορισμός του μέρους του λόγου για κάθε λέξη.
- Lemmatization: η μετατροπή όλων των λέξεων στην βασική τους μορφή, γνωστή και ως lemma.
- Stemming: η αφαίρεση των καταλήξεων και η αναγωγή στην ρίζα της λέξης.
- Parsing: η κατασκευή ενός δέντρου, δεδομένης μιας πρότασης. Υπάρχουν δύο κύριες κατηγορίες parsing, το Dependency Parsing και το Constituency Parsing.
- Stopword Removal: η αφαίρεση λέξεων πολύ συχνά εμφανιζόμενων, οι οποίες δεν προσφέρουν κάποιο νέο δεδομένο και δεν είναι σημαντικές για την εξαγωγή δεδομένων.

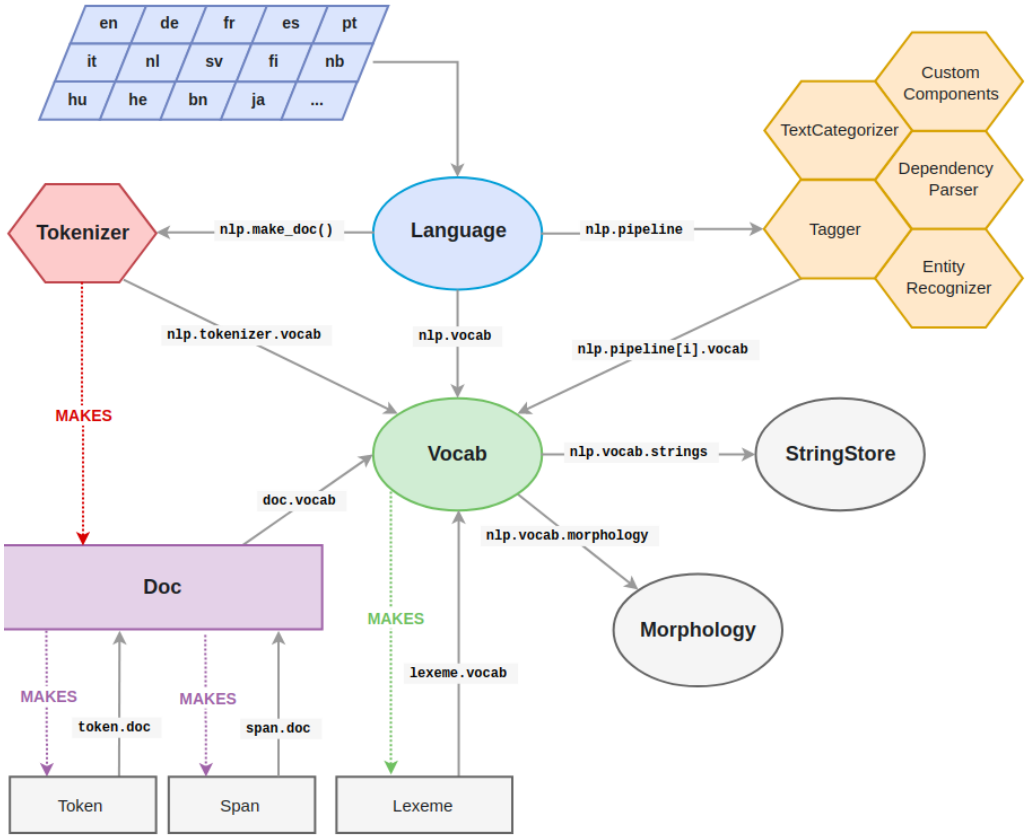
Οι κυριότερες σημασιολογικές εργασίες για την ανάλυση κειμένου είναι οι :

- Named Entity Recognition (NER): η αναγνώριση της εμφάνισης συγκεκριμένων κατηγοριών στο κείμενο. Οι βασικότερες κατηγορίες, που αναγνωρίζουν τα συστήματα NER είναι τα ονόματα ανθρώπων, οι εταιρείες, οι γεωγραφικές τοποθεσίες, τα προϊόντα, ημερομηνίες και ώρες, ποσότητες χρημάτων και ονόματα γεγονότων.
- Word Sense Disambiguation: η αναγνώριση του νοήματος μιας λέξης με διφορούμενο νόημα.
- Relationship extraction: η αναγνώριση των σχέσεων ανάμεσα στα named entities.

Υπάρχουν πολλά εργαλεία, τα οποία ασχολούνται με την ανάλυση κειμένου, όπως το NLTK, το CoreNLP του Stanford ή το AllenNLP, αλλά σε αυτήν την εργασία θα χρησιμοποιήσουμε το spaCy.

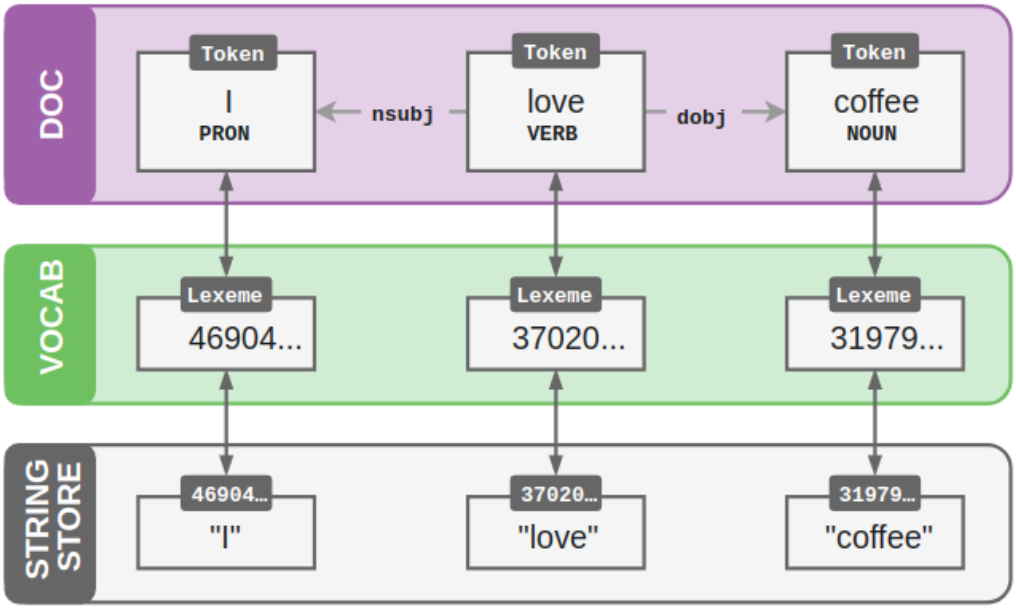
Το spaCy [7] είναι μία δωρεάν, ανοιχτού κώδικα βιβλιοθήκη για προηγμένη επεξεργασία της φυσικής γλώσσας στην Python. Είναι κατάλληλο για την δημιουργία συστημάτων για την εξαγωγή πληροφοριών και την κατανόηση της φυσικής γλώσσας, ή για την προεπεξεργασία κειμένου για μηχανική μάθηση.

Στην παρακάτω εικόνα παρουσιάζεται η αρχιτεκτονική του spaCy.



Εικόνα 5: Αρχιτεκτονική του spaCy

Οι κεντρικές δομές στο spaCy είναι το Doc και το Vocab. Το αντικείμενο Doc είναι ένας περιέκτης για την πρόσβαση στις γλωσσικές επισημειώσεις. Το αντικείμενο Vocab είναι ένας πίνακας αναζήτησης για το λεξιλόγιο, ο οποίος επιτρέπει την πρόσβαση στα αντικείμενα Lexeme. Στην παρακάτω εικόνα παρουσιάζεται ένα παράδειγμα για την καλύτερη κατανόηση των αντικειμένων Lexeme, Token, Vocab και Doc.



Εικόνα 6: Αντικείμενα του spaCy

Το αντικείμενο Span αποτελεί ένα κομμάτι του αντικειμένου Doc.

Τέλος το αντικείμενο Morphology αναθέτει γλωσσικά χαρακτηριστικά όπως λήμματα και ρήματα, βάση της λέξης και της ετικέτας, που περιγράφει το μέρος του λόγου, στο οποίο αντιστοιχεί η λέξη.

Το spaCy προς το παρόν υποστηρίζει 10 γλώσσες, εκ των οποίων και τα Ελληνικά.

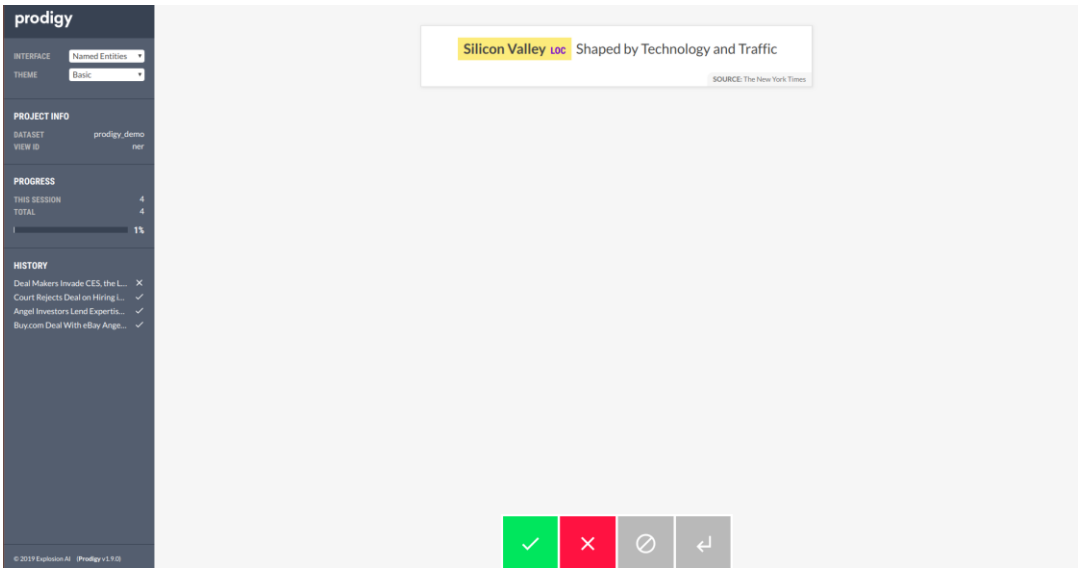
3.2 Εφαρμογές επισημείωσης

Υπάρχουν πολλές διαφορετικές εφαρμογές επισημείωσης, οι περισσότερες από τις οποίες είναι παρόμοιες μεταξύ τους, αλλά δεν εξειδικεύονται στην επισημείωση κειμένου σχετικού με την πολιτιστική κληρονομιά. Οι εφαρμογές αυτές επικεντρώνονται κυρίως στην αναγνώριση κάποιων βασικών οντοτήτων, όπως ονόματα, γεωγραφικές περιοχές, προϊόντα ή οργανισμούς, αλλά για την επισημείωση των κειμένων, που επιθυμούμε, είναι απαραίτητη η αναγνώριση και άλλων οντοτήτων. Προφανώς, και κάποια από αυτά τα εργαλεία υποστηρίζουν ημι-αυτόματη επισημείωση και πέρα των βασικών οντοτήτων, αλλά για την αποτελεσματική εκμετάλλευση της αυτοματοποίησης θα πρέπει να δοθεί από τον χρήστη ένας πολύ μεγάλος όγκος κειμένου για εκπαίδευση, τον οποίο πιθανότατα δεν διαθέτει, και πολύ μεγάλη προσπάθεια και χρόνος για την χειροκίνητη επισημείωση ή διόρθωση της επισημείωσης αυτών των κειμένων. Στη συνέχεια παρουσιάζονται κάποιες ενδεικτικές εφαρμογές επισημείωσης κειμένου.

3.2.1 Prodigy

Το Prodigy [6] είναι ένα εργαλείο χειροκίνητης αλλά και ημι-αυτόματης επισημείωσης για την δημιουργία εκπαιδευμένων δεδομένων και δεδομένων αξιολόγησης σε μοντέλα μηχανικής μάθησης. Το Prodigy δίνει την δυνατότητα στον χρήστη να επισημειώνει έγγραφα είτε χειροκίνητα είτε με την βοήθεια προτάσεων επισημείωσης, για την συλλογή δεδομένων. Ειδικότερα, η επισημείωση γίνεται είτε χειροκίνητα με την εμφάνιση κομματιών στο χρήστη, από το έγγραφο που έχει επιλέξει, είτε με την

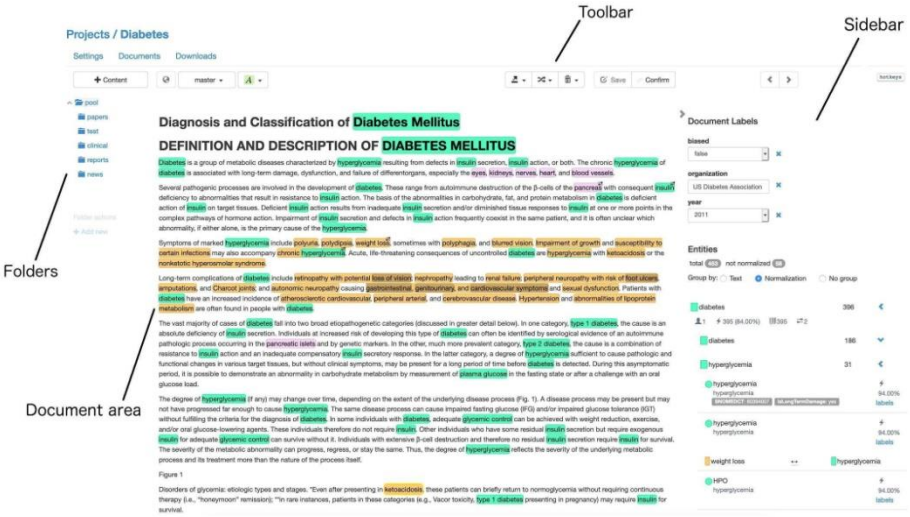
εμφάνιση προτάσεων επισημείωσης. Ο χρήστης καλείται να προσθέσει επισημειώσεις, όπου θεωρεί ότι είναι απαραίτητο, και να δεχτεί το κομμάτι του εγγράφου στην συλλογή δεδομένων του ή να το απορρίψει.



Εικόνα 7: Εφαρμογή Prodigy

3.2.2 Tagtog

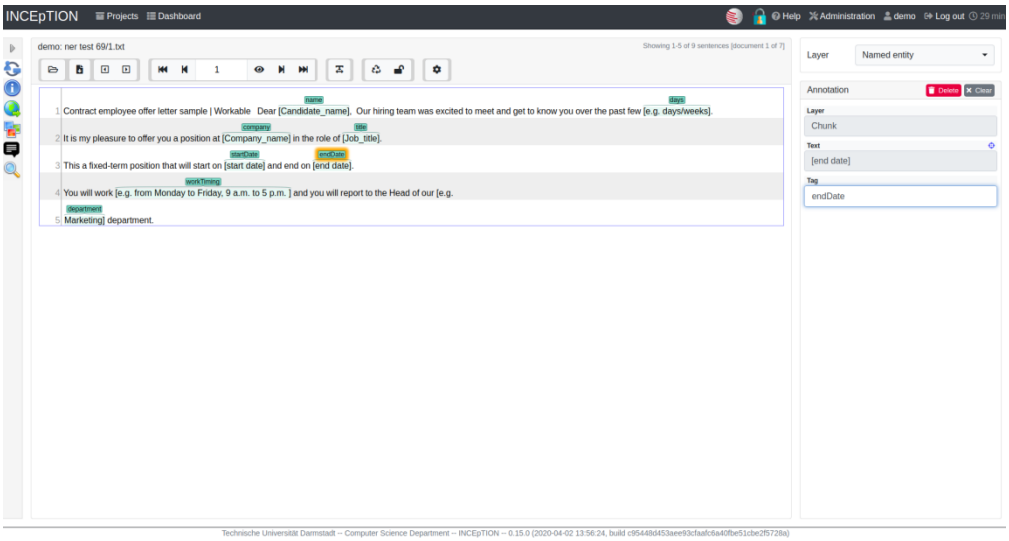
Το Tagtog [7] είναι ένα εργαλείο επισημείωσης, που προσφέρει μηχανισμούς για χειροκίνητη και αυτόματη επισημείωση εγγράφων. Η αυτόματη επισημείωση επιτυγχάνεται με την εισαγωγή λεξικού, το οποίο περιλαμβάνει συλλογές όρων, ή με μηχανική μάθηση, καθώς το Tagtog μαθαίνει συνεχώς από τις ήδη υπάρχουσες επισημειώσεις, δημιουργώντας ακριβής προβλέψεις. Επιπλέον, ο χρήστης μπορεί να προσθέσει τους δικούς του τύπους οντοτήτων, που επιθυμεί να επισημειώσει, και επίσης επιτρέπεται η επικάλυψη οντοτήτων, επιτρέποντας στο χρήστη να αξιοποιήσει στο έπακρο τα δεδομένα του.



Εικόνα 8: Εφαρμογή Tagtog

3.2.3 INCEpTION

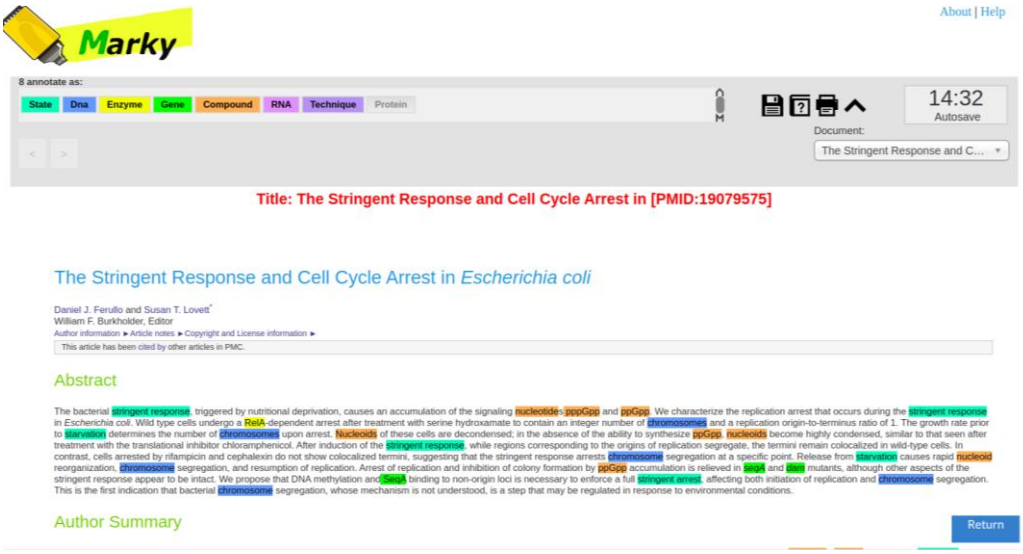
Το INCEpTION [8], είναι ένα λογισμικό ανοιχτού κώδικα, το οποίο μεταξύ άλλων προσφέρει χειροκίνητη και αυτόματη επισημείωση οντοτήτων σε έγγραφα. Οι διαθέσιμοι τύποι οντοτήτων είναι οι LOC(τοποθεσία), ORG(οργανισμός), PER(άτομο), OTH(όλα τα ονόματα, που δεν είναι άτομα, τοποθεσία ή οργανισμός). Για την βελτίωση της αποδοτικότητας της διαδικασίας της επισημείωσης, το INCEpTION προσφέρει αλγορίθμους, οι οποίοι χρησιμοποιούν μηχανική μάθηση για να παρέχουν προτάσεις επισημείωσης.



Εικόνα 9: Εφαρμογή INCEpTION

3.2.4 Marky

Το Marky [9] είναι ένα Web-based εργαλείο επισημείωσης, το οποίο προσφέρει την δυνατότητα χειροκίνητης επισημείωσης εγγράφων, από έναν ή περισσότερους χρήστες, αλλά και την λήψη στατιστικών της επισημείωσης. Ο χρήστης μπορεί να φορτώσει το έγγραφο που επιθυμεί να επισημειώσει αλλά και να κατεβάσει επισημειωμένα έγγραφα. Για να ξεκινήσει η διαδικασία της επισημείωσης, είναι απαραίτητη η προσθήκη των τύπων των οντοτήτων, καθώς δεν υπάρχουν προκαθορισμένοι τύποι. Το Marky είναι συμβατό με το Chrome, το Firefox, το IE9+, το Opera και το Safari.



Εικόνα 10: Εφαρμογή Marky

4.ΠΡΟΔΙΑΓΡΑΦΕΣ ΚΑΙ ΣΧΕΔΙΑΣΗ ΕΦΑΡΜΟΓΗΣ

Για την ανάπτυξη της εφαρμογής απαραίτητη είναι η περιγραφή των χρηστών της, καθώς και η καταγραφή των αναγκών αλλά και των προδιαγραφών που προκύπτουν. Η ανάλυση των παραπάνω αποτελεί το περιεχόμενο της ενότητας αυτής, η οποία ολοκληρώνεται με τον σχεδιασμό του προτύπου χαμηλής πιστότητας της εφαρμογής.

4.1 Περιγραφή των χρηστών

Οι αναμενόμενοι χρήστες της εφαρμογής είναι συγγραφείς περιεχομένου για εξατομικευμένες εμπειρίες σχετικές με την πολιτιστική κληρονομιά. Συνήθως, οι συγκεκριμένοι χρήστες είναι μη ειδικοί στην χρήση ηλεκτρονικών υπολογιστών και περιλαμβάνουν ιστορικούς, επιμελητές μουσείων αλλά και μεγάλο τμήμα του υπόλοιπου προσωπικού ενός μουσείου, καθώς και δημιουργούς περιεχομένου για ψηφιακούς τουριστικούς οδηγούς και εφαρμογές ξενάγησης.

Στη συνέχεια, παρουσιάζονται δύο χαρακτηριστικές περσόνες συγγραφέα, όπως αυτές ορίστηκαν στα πλαίσια του προγράμματος CHESS [12], [13] και μία ακόμα περσόνα της εφαρμογής.

- Έλλη Πέτρου, επιμελητής μουσείου.

Εικόνα 11: 1η περσόνα της εφαρμογής

"Οι νέες τεχνολογίες είναι πρόκληση για μένα αλλά αξίζουν την προσπάθεια."

Ηλικία: 51 χρονών.

Βασικά χαρακτηριστικά:

- Η Έλλη είναι ένας από τους επιμελητές του Μουσείου. Είναι υπεύθυνη, μεταξύ των άλλων καθηκόντων της, για τη δημιουργία νέων διαδραστικών εμπειριών.
- Έχει περιορισμένη εμπειρία με ψηφιακές συσκευές. Χρησιμοποιεί έναν υπολογιστή για την ανάγνωση e-mail ή τη σύνταξη εγγράφων.
- Θεωρεί τις ψηφιακές τεχνολογίες συναρπαστικές αλλά η χρήση τους αποτελεί για αυτήν πρόκληση που κάποιες φορές τη φοβίζει.
- Θα ήθελε ο συντάκτης ερωτηματολογίων να είναι αρκετά απλός για αυτήν, ώστε να μπορεί να επικεντρωθεί περισσότερο στη δημιουργική πλευρά του σχεδιασμού της ιστορίας.

Στόχοι:

- Να δημιουργήσει μία καλή εμπειρία για τον επισκέπτη του μουσείου.
- Να μπορεί να δημιουργήσει γρήγορα και αβίαστα εξατομικευμένα ερωτηματολόγια χωρίς να δυσκολευτεί.

- Laurent Boulay, μουσείο-εκπαιδευτικός.

Εικόνα 12: 2η περσόνα της εφαρμογής

"Όλα τα καινούργια εργαλεία είναι συναρπαστικά και χρήσιμα για την δουλειά μου."

Ηλικία: 30 χρονών.

Βασικά χαρακτηριστικά:

- Ο Laurent είναι ένας από τους εκπαιδευτές στο Cite de l'Espace.

Στόχοι:

- Να δημιουργήσει μία καλή εμπειρία για τον επισκέπτη του μουσείου.
- Να μπορεί να συγγράφει πολύπλοκες ιστορίες με πολλές δυνατότητες.

- Μαρία Παπαδημητρίου, δημιουργός περιεχομένου ψηφιακών εμπειριών.

Εικόνα 13: 3η περσόνα της εφαρμογής

"Μου αρέσει πολύ να δημιουργώ καινούργιες ιστορίες."

Ηλικία: 40 χρονών.

Βασικά χαρακτηριστικά:

- Η Μαρία εργάζεται σε μία εταιρεία, η οποία δημιουργεί ψηφιακές εξατομικευμένες ξεναγήσεις σε μουσεία και άλλους χώρους πολιτιστικού ενδιαφέροντος.
- Είναι πτυχιούχος του Ιστορικού και Αρχαιολογικού τμήματος του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών, αν και από μικρή ήθελε να γίνει συγγραφέας.
- Της αρέσει πολύ να δημιουργεί καινούργιες ιστορίες και είναι πολύ ευχαριστημένη με την δουλειά της, καθώς συνδυάζει το αντικείμενο των σπουδών της με την αγάπη της για την συγγραφή.
- Δεν είναι ιδιαίτερα εξοικειωμένη με τις ψηφιακές τεχνολογίες και την χρήση του ηλεκτρονικού υπολογιστή.

Στόχοι:

- Να μπορεί να συγγράφει πολύπλοκες ιστορίες, ενδιαφέρουσες για ένα μεγάλο φάσμα πιθανών πελατών.
- Να μπορεί να κατηγοριοποιεί το περιεχόμενο της ιστορίας της, για την καλύτερη εκμετάλλευσή του, από το τμήμα της δημιουργίας των ψηφιακών ξεναγήσεων.

4.2 Ανάγκες και προδιαγραφές

Σε αυτήν την υποενότητα παρουσιάζεται ο πίνακας με τις απαιτήσεις και τις προδιαγραφές της εφαρμογής.

Πίνακας 1: Προδιαγραφές Εφαρμογής

Α/Α	**Περιγραφή Ανάγκης**	**Προδιαγραφή**	**Επίπεδο Αναγκαιότητας**	**Παρατηρήσεις**
1	Να μπορούν να προσθέσουν ένα κείμενο για επισημείωση	Επιλογή για άνοιγμα νέου εγγράφου	Απαραίτητη	-
2	Να μπορούν να αποθηκεύσουν ένα επισημειωμένο κείμενο	Επιλογή αποθήκευσης	Απαραίτητη	-
3	Να μπορούν να προσθέσουν μία	Επιλογή προσθήκης επισημείωσης	Απαραίτητη	-

	επισημείωση σε ένα κείμενο			
4	Να μπορούν να αφαιρέσουν μία επισημείωση	Επιλογή αφαίρεσης επισημείωσης	Απαραίτητη	-
5	Να μπορούν να εγγραφούν στην εφαρμογή	Επιλογή για εγγραφή νέου χρήστη	Επιθυμητή	-
6	Να μπορούν να συνδεθούν στην εφαρμογή	Επιλογή για σύνδεση στην εφαρμογή, συμπληρώνοντας username και password	Επιθυμητή	-
7	Να μπορούν να αποσυνδεθούν από τον λογαριασμό τους	Επιλογή αποσύνδεσης	Επιθυμητή	-
8	Να μπορούν να επιλέγουν τον τρόπο διαχωρισμού μιας ενότητας	Checkboxes με τις διαθέσιμες επιλογές	Επιθυμητή	Το κείμενο θα μπορεί να αναλύεται σε παραγράφους ή σε μεγαλύτερες ενότητες(π.χ. κείμενο κάτω από μία κεφαλίδα).
9	Να μπορούν να επιλέξουν ποιες επισημειώσεις θα εμφανίζονται με βάση την	Checkboxes με τις διαθέσιμες κατηγορίες	Απαραίτητη	-

	κατηγορία, στην οποία ανήκουν			
10	Να μπορούν να δουν τα φίλτρα για την εμφάνιση των επισημειώσεων, τα οποία έχουν εφαρμοστεί	Καμία	Επιθυμητή	Με την επιλογή ενός φίλτρου θα εμφανίζεται αυτόματα πάνω από το κείμενο ένα κουτί με το όνομα του φίλτρου, που έχει εφαρμοστεί
11	Να μπορούν να αφαιρούν ένα φίλτρο για την εμφάνιση των επισημειώσεων	Επιλογή για αφαίρεση φίλτρου με ένα “x”, που θα υπάρχει δίπλα στο κουτάκι με το κάθε εφαρμοσμένο φίλτρο	Επιθυμητή	-
12	Να επιλέγουν το χρώμα επισημείωσης των κύριων κατηγοριών του λεξιλογίου	Φόρμα για επιλογή του επιθυμητού χρώματος	Επιθυμητή	-
13	Να μπορούν να προσθέσουν ετικέτες σε ενότητα του κειμένου	Φόρμα για εισαγωγή της ετικέτας	Απαραίτητη	-
14	Να μπορούν να αφαιρούν μία ετικέτα από ενότητα του κειμένου	Επιλογή για αφαίρεση ετικέτας με ένα “x”, που θα υπάρχει στην κάθε ετικέτα	Απαραίτητη	-

15	Να μπορούν να μετονομάζουν ένα αποθηκευμένο έγγραφο	Επιλογή κουμπιού για μετονομασία	Επιθυμητή	-
16	Να μπορούν να διαγράφουν ένα αποθηκευμένο έγγραφο	Επιλογή κουμπιού για διαγραφή	Επιθυμητή	-
17	Να μπορούν να αναζητήσουν το όρο του λεξιλογίου που επιθυμούν να εφαρμόσουν ως φίλτρο	Φόρμα αναζήτησης με αυτόματη συμπλήρωση	Επιθυμητή	-

5. ΥΛΟΠΟΙΗΣΗ ΕΦΑΡΜΟΓΗΣ

5.1 Επισκόπηση τεχνολογιών

Μετά από μία έρευνα για τις διαθέσιμες τεχνολογίες στον front-end, καταλήξαμε στις δύο παρακάτω τεχνολογίες, οι οποίες βρίσκονται ανάμεσα στις πιο δημοφιλής τεχνολογίες αυτή την στιγμή.

1) **Angular**

   Πρόκειται για ένα web application framework, ανοιχτού κώδικα, που αναπτύχθηκε από την Google και χρησιμοποιείται για την δημιουργία αποδοτικών και εκλεπτυσμένων εφαρμογών. Η Angular βασίζεται στην γλώσσα προγραμματισμού Typescript [14].

2) **React**

   Είναι μία Javascript βιβλιοθήκη, ανοιχτού κώδικα, που δημιουργήθηκε από το Facebook, για την κατασκευή διεπαφών χρήστη. Η React χρησιμοποιείται συχνά ως βάση για την ανάπτυξη εφαρμογών μιας σελίδας ή εφαρμογών για κινητά. Παράλληλα με την χρήση της React, συνήθως, απαιτείται και η χρήση πρόσθετων βιβλιοθηκών για διαχείριση κατάστασης και δρομολόγηση, καθώς η React ασχολείται αποκλειστικά με την απόδοση δεδομένων στο DOM [15].

Για την ανάπτυξη του back-end, λόγω της χρήσης της βιβλιοθήκης spaCy, καταλήξαμε στις δύο παρακάτω τεχνολογίες, οι οποίες περιλαμβάνονται στις πιο διαδεδομένες τεχνολογίες, βασισμένες στην γλώσσα προγραμματισμού Python.

1) **Django**

   Είναι ένα ανοιχτού κώδικα web framework, το οποίο ακολουθεί το μοτίβο Model-Template-View. Πρωταρχικός στόχος του Django είναι η διευκόλυνση της δημιουργίας σύνθετων ιστοτόπων, που βασίζονται σε μια βάση δεδομένων. Αυτό το framework δίνει έμφαση στην επαναχρησιμοποίηση, στην δυνατότητα μεταφοράς των στοιχείων, στην συγγραφή λιγότερου κώδικα, στη χαμηλή σύζευξη, στην ταχεία ανάπτυξη και στην αρχή της μη επανάληψης [16].

2) **Flask**

   Αποτελεί ένα micro web application framework. Κατηγοριοποιείται ως microframework, καθώς δεν απαιτεί συγκεκριμένα εργαλεία ή βιβλιοθήκες. Υποστηρίζει επεκτάσεις για επικύρωση φόρμας, χειρισμό μεταμορφώσεων, διάφορες ανοιχτές τεχνολογίες ελέγχου ταυτότητας και άλλα εργαλεία που σχετίζονται με αυτό το framework [17].

5.2 Τεχνολογίες που χρησιμοποιήθηκαν

5.2.1 Frameworks

Από τις προαναφερθείσες τεχνολογίες, για την ανάπτυξη του front-end, επιλέχθηκε η Angular, λόγω της εκτεταμένης τεκμηρίωσης, και το Flask, για την ανάπτυξη του back-end, λόγω της μικρής πολυπλοκότητας της αναπτυσσόμενης εφαρμογής.

5.2.2 Βάση δεδομένων

Για την ανάπτυξη της βάσης δεδομένων χρησιμοποιήθηκε η MySQL, ένα σύστημα διαχείρισης σχεσιακών βάσεων δεδομένων, που χρησιμοποιείται σε κάποιες από τις πιο διαδεδομένες διαδικτυακές υπηρεσίες.

5.2.3 Γλώσσες προγραμματισμού

Για την υλοποίηση της εφαρμογής τόσο στο front-end όσο και στο back-end χρησιμοποιήθηκαν οι γλώσσες προγραμματισμού που αναλύονται παρακάτω.

1) **HTML**

   Η HTML (Hypertext Markup Language, Γλώσσα Σήμανσης Υπερκειμένου) [18] είναι η κύρια γλώσσα σήμανσης για τις ιστοσελίδες και αποτελεί ένα μέσο για την δημιουργία δομημένων εγγράφων καθορίζοντας δομικά στοιχεία για το κείμενο, όπως κεφαλίδες, παραγράφους, λίστες, συνδέσμους και άλλα. Αποτελείται από ένα σύνολο ετικετών, οι οποίες αντιστοιχούν στα βασικά δομικά στοιχεία μιας ιστοσελίδας και περικλείονται μέσα στα σύμβολα “<”, “>”. Τα προγράμματα περιήγησης χρησιμοποιούν τις ετικέτες HTML για να παρουσιάσουν το περιεχόμενο μιας σελίδας που μπορεί κάποιος να διαβάσει ή να ακούσει.

2) **CSS**

   Η CSS (Cascading Style Sheets) [19] είναι μία γλώσσα στυλ φύλλων (style sheet language), που χρησιμοποιείται για την περιγραφή της αναπαράστασης ενός εγγράφου, γραμμένου σε γλώσσα σήμανσης. Έχει σχεδιαστεί για να επιτρέπει τον διαχωρισμό της παρουσίασης του περιεχομένου(χρώματα, μέγεθος γραμματοσειράς και άλλα) και του ίδιου του περιεχομένου, με σκοπό ,μεταξύ άλλων, την παροχή μεγαλύτερης ευελιξίας και ελέγχου στις προδιαγραφές των χαρακτηριστικών παρουσίασης και την βελτίωση της προσβασιμότητας του περιεχομένου.

3) **Typescript**

   Η Typescript [20] αποτελεί μία ανοιχτού κώδικα γλώσσα προγραμματισμού, που αναπτύχθηκε και συντηρείται από την Microsoft. Είναι ένα αυστηρό συντακτικό υπερσύνολο της Javascript και προσθέτει προαιρετικά ένα στατικό σύστημα τύπων στην γλώσσα. Η Typescript έχει σχεδιαστεί για την ανάπτυξη μεγάλων εφαρμογών και μπορεί να χρησιμοποιηθεί για την δημιουργία εφαρμογών τόσο σε client-side όσο και σε server-side.

4) **Python**

   Η Python [21] είναι μία διερμηνευόμενη (interpreted), γενικού σκοπού και υψηλού επιπέδου γλώσσα προγραμματισμού, που δημιουργήθηκε από τον Guido van Rossum. Η φιλοσοφία σχεδιασμού της δίνει έμφαση στην αναγνωσιμότητα του κώδικα με την αναγκαία χρήση του κενού. Είναι δυναμική γλώσσα προγραμματισμού και υποστηρίζει συλλογή απορριμμάτων (αυτόματη διαχείριση μνήμης).

5.2.4 Πρόσθετες Βιβλιοθήκες

Για την ανάπτυξη της εφαρμογής χρησιμοποιήθηκαν, επίσης και, κάποιες βιβλιοθήκες, η αναφορά και περιγραφή των οποίων, αποτελεί το περιεχόμενο αυτής της ενότητας.

1) **Angular Material**

   Η Angular Material [22] είναι μία βιβλιοθήκη, τα περιεχόμενα της οποίας βοηθούν στην κατασκευή ελκυστικών, συνεπών, λειτουργικών και πιο γρήγορων ιστοσελίδων, αναδεικνύοντας τις σύγχρονες αρχές σχεδίασης ιστοτόπων, όπως η φορητότητα ενός προγράμματος περιήγησης και η ανεξαρτησία συσκευής.

2) **Βιβλιοθήκες για το back-end**

Για την υλοποίηση του back-end χρησιμοποιήθηκαν κάποιες επεκτάσεις του Flask, όπως επίσης η βιβλιοθήκη Levenshtein για τον υπολογισμό της απόστασης μεταξύ δύο συμβολοσειρών, και οι βιβλιοθήκες time, datetime για την διαχείριση του χρόνου και των ημερομηνιών.

5.2.5 Εργαλεία που χρησιμοποιήθηκαν

Κατά την διάρκεια υλοποίησης της εφαρμογής για την εγγραφή του κώδικα, αλλά και για την ανάπτυξη της βάσης δεδομένων χρησιμοποιήθηκαν τα παρακάτω εργαλεία.

1) Visual Studio Code

Το VIsual Studio Code [23] χρησιμοποιήθηκε για την ανάπτυξη του κώδικα του front-end και είναι ένας δωρεάν επεξεργαστής πηγαίου κώδικα, που δημιουργήθηκε από την Microsoft για τα λειτουργικά συστήματα Windows, Linux, macOS. Περιλαμβάνει λειτουργίες, μεταξύ άλλων, για υποστήριξη εντοπισμού σφαλμάτων, επισήμανση σύνταξης, έξυπνη ολοκλήρωση κώδικα και αναδιαμόρφωση κώδικα. Επιπροσθέτως, προσφέρει την δυνατότητα στους χρήστες να αλλάξουν το θέμα, τις συντομεύσεις του πληκτρολογίου και να εγκαταστήσουν συντομεύσεις για περαιτέρω λειτουργικότητα.

2) Pycharm

Το Pycharm [24] χρησιμοποιήθηκε για την ανάπτυξη του κώδικα του back-end και είναι ένα ολοκληρωμένο περιβάλλον ανάπτυξης (IDE), που χρησιμοποιείται στον προγραμματισμό με γλώσσα Python και υποστηρίζεται από τα λειτουργικά συστήματα Windows, Linux και macOS. Δημιουργήθηκε από την τσέχικη εταιρεία JetBrains και παρέχει λειτουργίες ανάλυσης κώδικα, γραφικής αναπαράστασης σφαλμάτων, ενσωματωμένο unit tester και ενοποίηση, με την βοήθεια συστήματος ελέγχου έκδοσης.

3) MySQL Workbench

Το MySQL Workbench [25] είναι ένα εργαλείο οπτικής σχεδίασης βάσεων δεδομένων, που ενσωματώνει την ανάπτυξη, την διαχείριση, το σχεδιασμό, την δημιουργία και την συντήρηση βάσεων δεδομένων SQL σε ένα ενιαίο περιβάλλον ανάπτυξης για το σύστημα βάσεων δεδομένων MySQL. Όπως είναι προφανές, από τα παραπάνω, το εργαλείο αυτό χρησιμοποιήθηκε για την ανάπτυξη της βάσης δεδομένων της εφαρμογής.

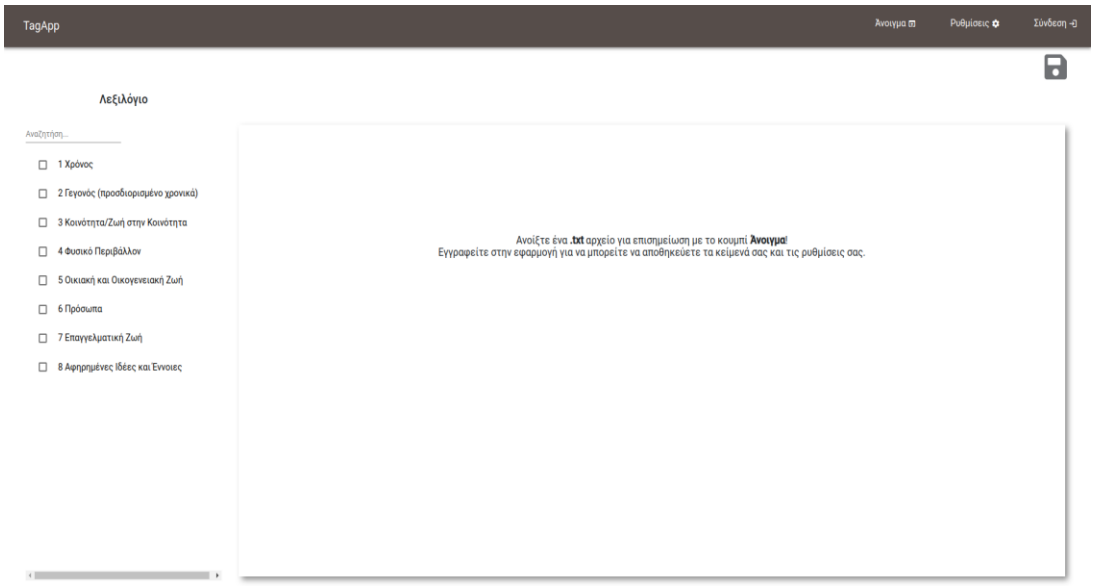
5.3 Εφαρμογή

Σε αυτήν την ενότητα παρουσιάζεται η εφαρμογή, αναλύοντας αρχικά την διεπαφή του χρήστη, στη συνέχεια το σχήμα της βάσης δεδομένων, και τέλος το back-end της εφαρμογής.

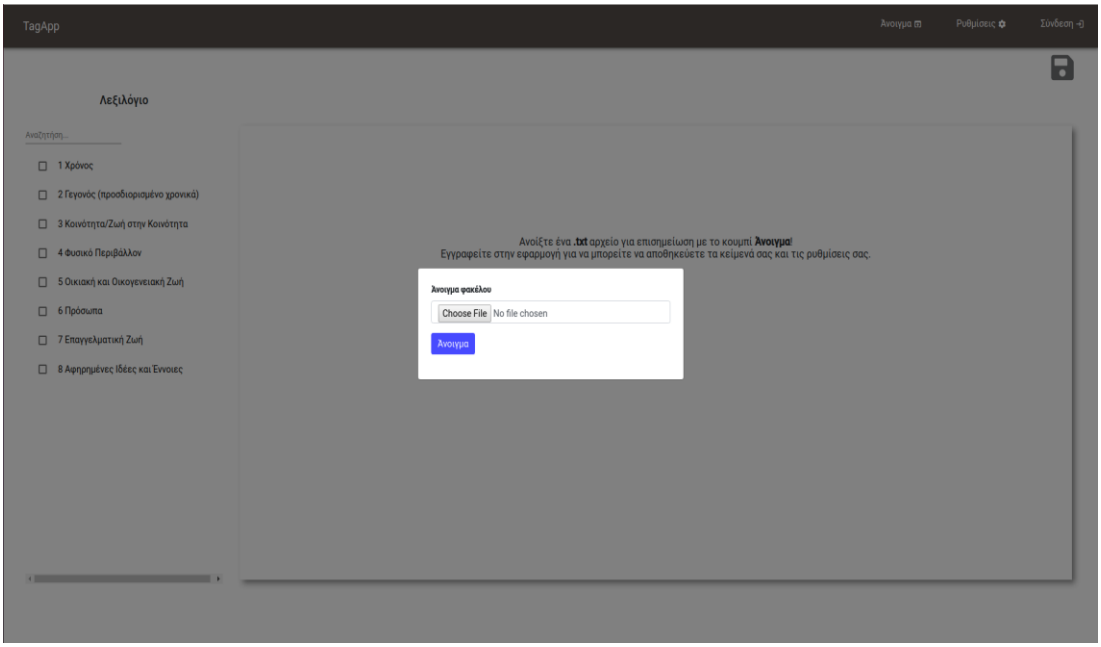
5.3.1 Διεπαφή χρήστη

Στην συγκεκριμένη υποενότητα γίνεται μία αναλυτική περιγραφή του μέρους της εφαρμογής που είναι ορατό στον χρήστη.

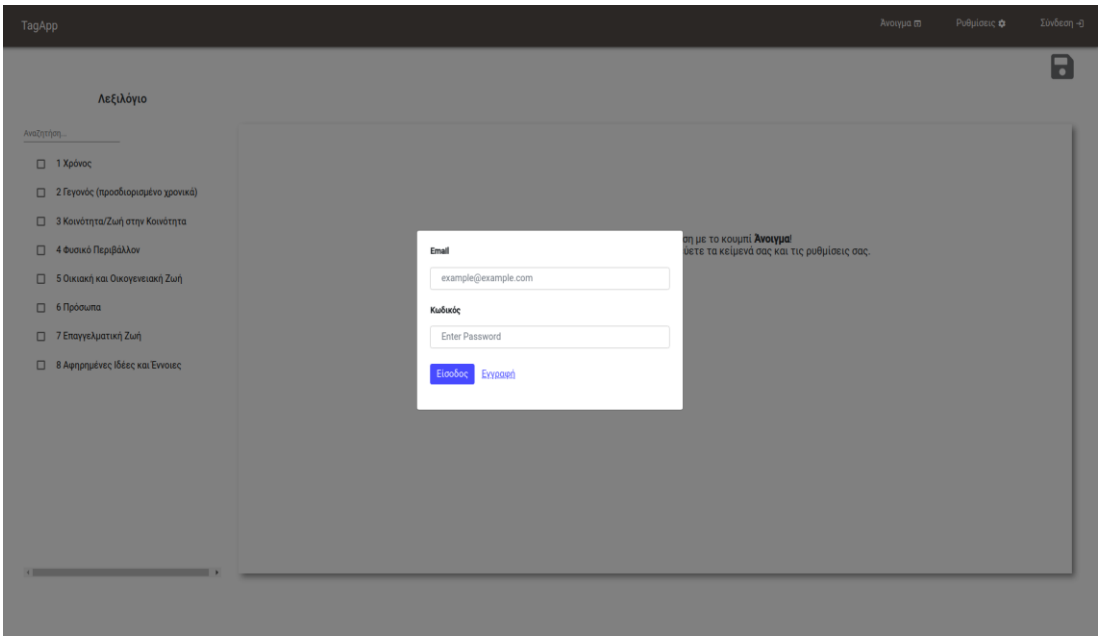
Στις παρακάτω εικόνες φαίνεται η αρχική οθόνη της εφαρμογής. Ο χρήστης από την σελίδα αυτή μπορεί να ανοίξει ένα νέο αρχείο .txt για να επισημειωθεί, μπορεί να μεταβεί στις ρυθμίσεις ή να κάνει είσοδο στην εφαρμογή, αν είναι εγγεγραμμένος χρήστης.



Εικόνα 14: Αρχική σελίδα εφαρμογής

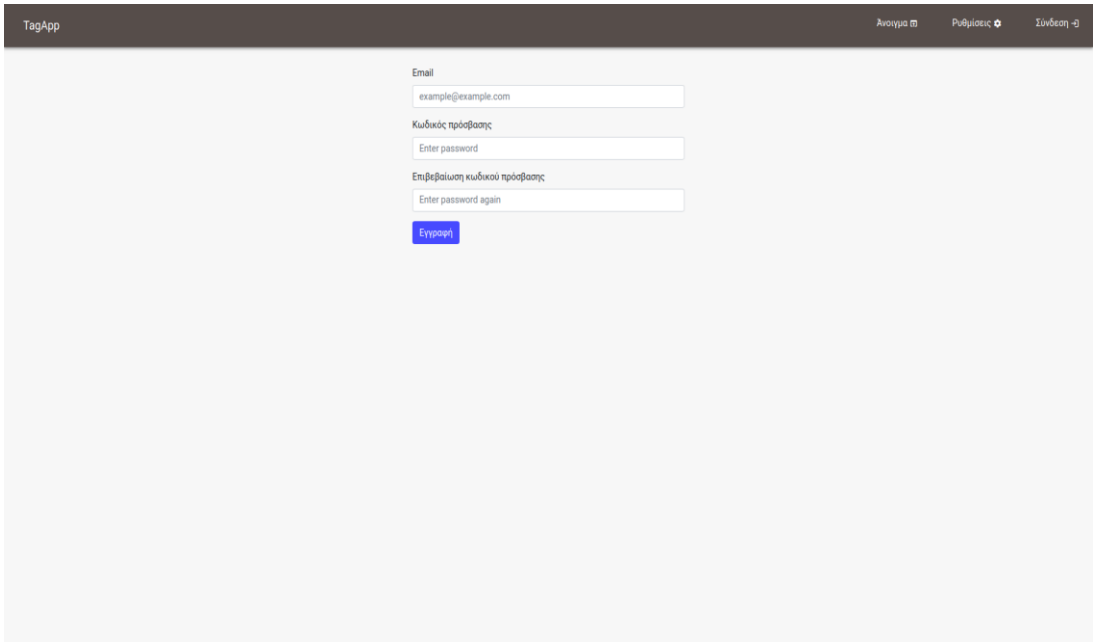


Εικόνα 15: Επιλογή ανοίγματος εγγράφου



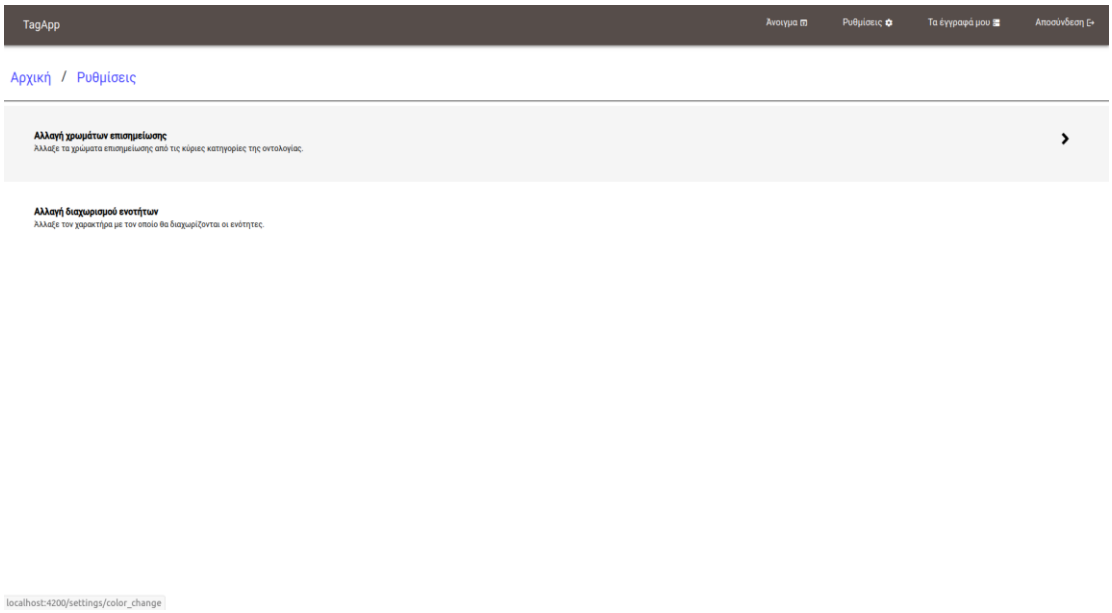
Εικόνα 16: Επιλογή εισόδου στην εφαρμογή

Με την επιλογή της εγγραφής στο αναδυόμενο παράθυρο της εισόδου μεταφέρεται στην παρακάτω σελίδα προκειμένου να εγγραφεί στην εφαρμογή. Στη σελίδα αυτή απλά συμπληρώνει τα στοιχεία που απαιτούνται για να πραγματοποιηθεί η εγγραφή του.

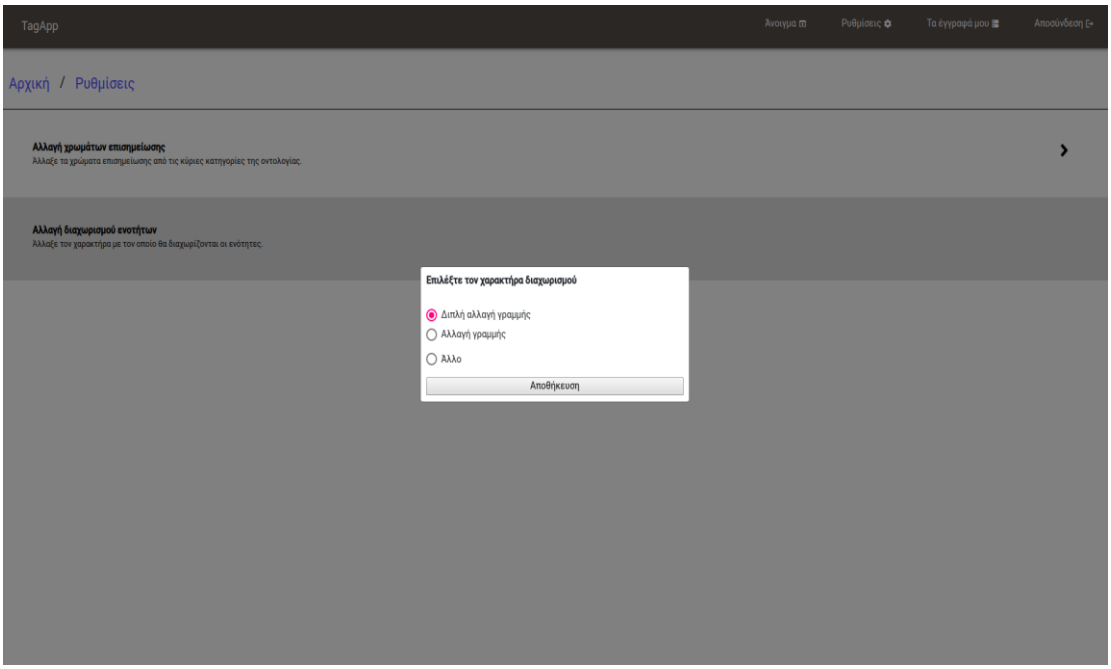


Εικόνα 17: Σελίδα εγγραφής της εφαρμογής

Με την επιλογή των ρυθμίσεων από την αρχική, ο χρήστης μεταφέρεται στην παρακάτω σελίδα. Όπως φαίνεται και στις εικόνες, από την σελίδα αυτή ο χρήστης μπορεί να αλλάξει τον τρόπο διαχωρισμού του κειμένου σε ενότητες αλλά και τα χρώματα των βασικών όρων του ελεγχόμενου λεξιλογίου.

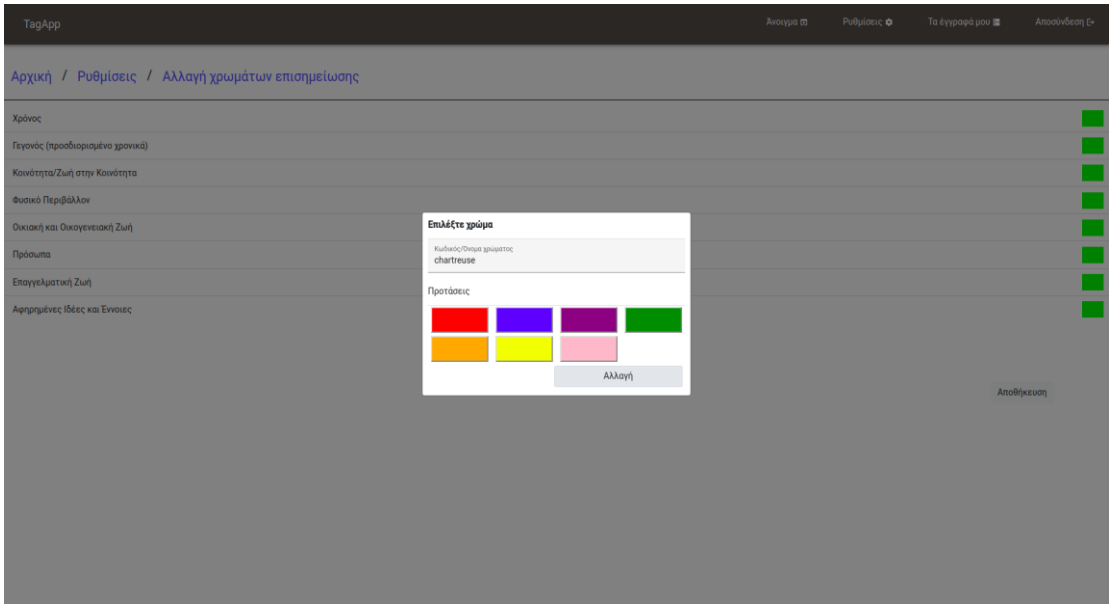


Εικόνα 18: Σελίδα ρυθμίσεων εφαρμογής



Εικόνα 19: Επιλογή για αλλαγή διαχωρισμού ενότητας

Εικόνα 20: Σελίδα αλλαγής χρωμάτων



Εικόνα 21: Επιλογή χρώματος για όρο του λεξιλογίου

Με την προσθήκη ενός κειμένου η αρχική σελίδα του χρήστη διαμορφώνεται όπως φαίνεται παρακάτω.

Εικόνα 22: Αρχική σελίδα με κείμενο

Το κείμενο το οποίο προστέθηκε έχει χωριστεί σε ενότητες και κάθε φορά εμφανίζεται μία ενότητα, η οποία αποτελείται από τις ετικέτες της και το επισημειωμένο κείμενο. Με το χρωματισμένο κείμενο δηλώνεται μία επισημείωση, ενώ με το υπογραμμισμένο, μία προτεινόμενη επισημείωση.

Ο χρήστης στο σημείο αυτό μπορεί να επιλέξει κάποιον όρο του λεξιλογίου, όπως φαίνεται παρακάτω, προκειμένου να εμφανίζεται ένα κομμάτι των επισημειώσεων.

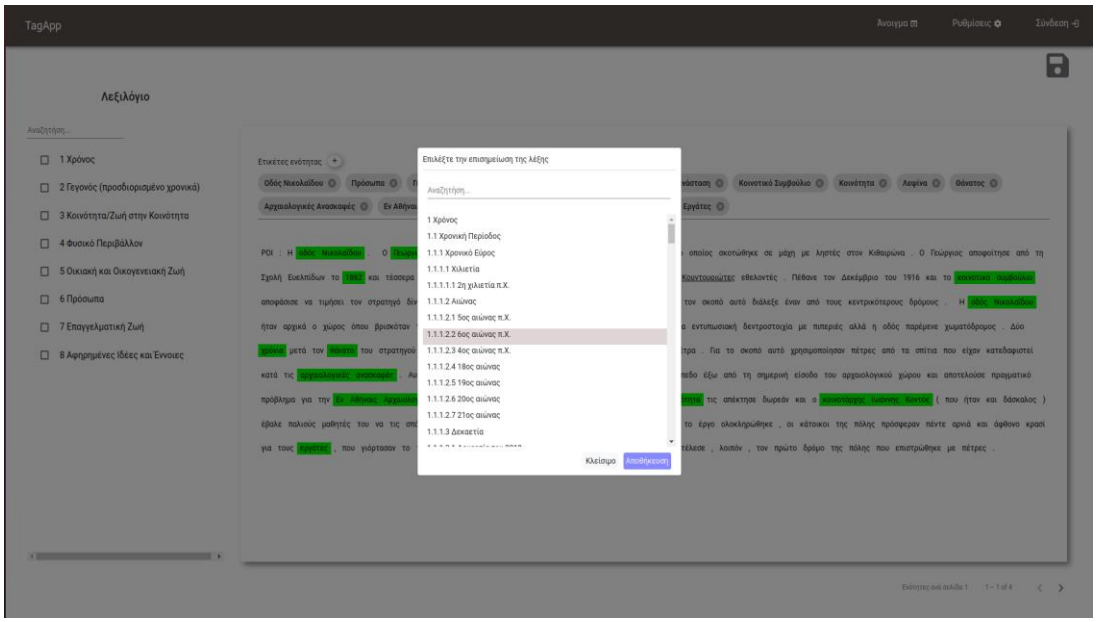
Εικόνα 23: Αρχική σελίδα με φιλτραρισμένο κείμενο

Επιπλέον, αν έχει επιλέξει διαφορετικό χρώμα επισημείωσης για κάποιον από τους όρους, για παράδειγμα για τον χρόνο, το κείμενο θα εμφανίζεται, όπως φαίνεται παρακάτω.

Εικόνα 24: Αρχική σελίδα με πολύχρωμο κείμενο

Η κύρια ενέργεια των χρηστών σε αυτήν την εφαρμογή είναι η επεξεργασία των επισημειώσεων ενός κειμένου. Στις παρακάτω εικόνες φαίνονται οι επιλογές του χρήστη κατά την προσθήκη καινούργιας επισημείωσης, αλλά και κατά την επιλογή μίας προτεινόμενης επισημείωσης.

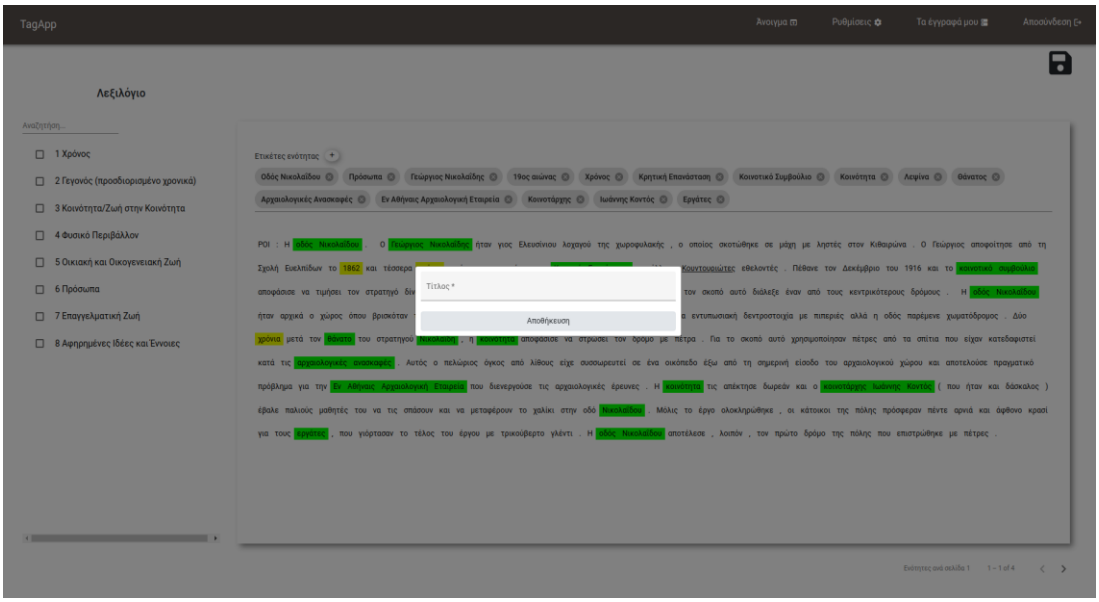
Εικόνα 25: Προσθήκη νέας επισημείωσης



Εικόνα 26: Επιλογή ετικέτας για την επισημείωση

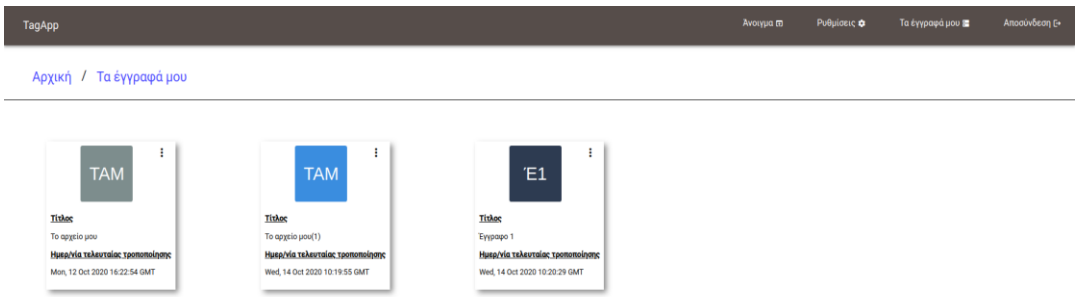
Εικόνα 27: Επιλογή προτεινόμενης επισημείωσης

Αν ο χρήστης έχει κάνει είσοδο στην εφαρμογή, προσφέρεται η επιπλέον δυνατότητα αποθήκευσης των κειμένων του, πατώντας το κουμπί αποθήκευσης, και όπως είναι λογικό και η επιλογή για την μετάβαση στα έγγραφά του.

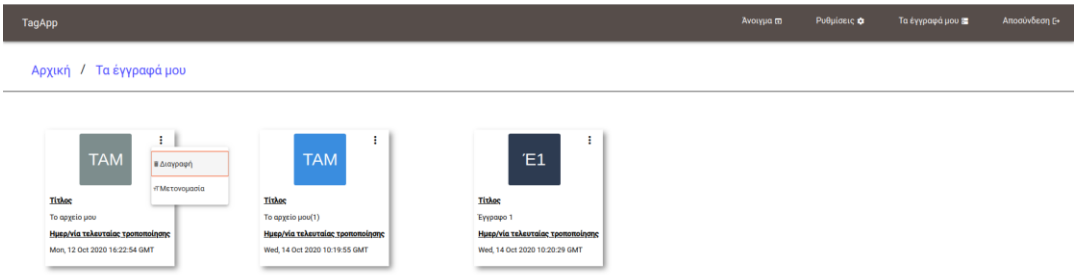


Εικόνα 28: Επιλογή αποθήκευσης κειμένου

Στις παρακάτω εικόνες απεικονίζονται οι σελίδες με τα έγγραφα ενός χρήστη. Όπως φαίνεται εμφανίζονται κάποιες βασικές πληροφορίες για τα έγγραφα και με το πάτημα ενός από αυτά, ο χρήστης μπορεί να ανοίξει κάποιο από τα έγγραφά του και να συνεχίσει την επεξεργασία. Επιπλέον δίνεται η δυνατότητα διαγραφής και μετονομασίας των εγγράφων.

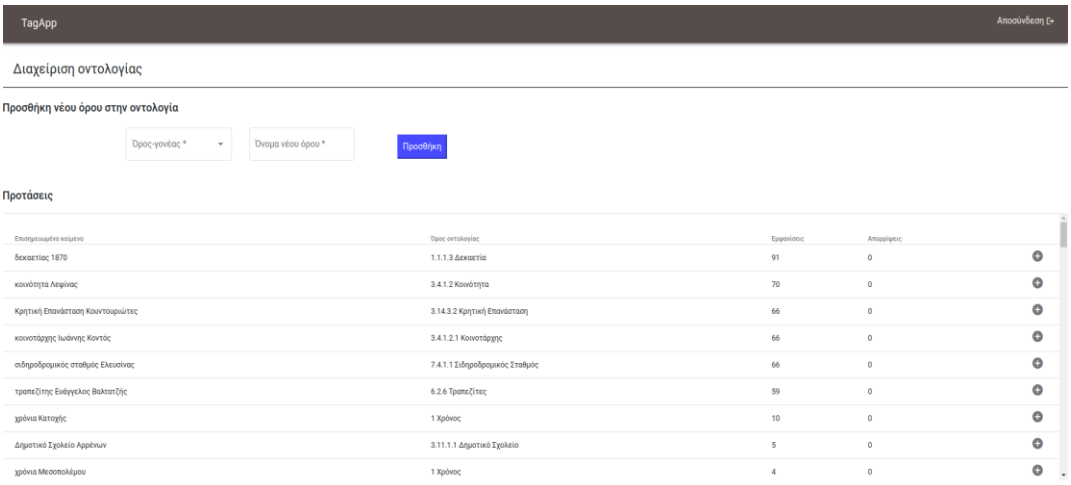


Εικόνα 29: Σελίδα εγγράφων

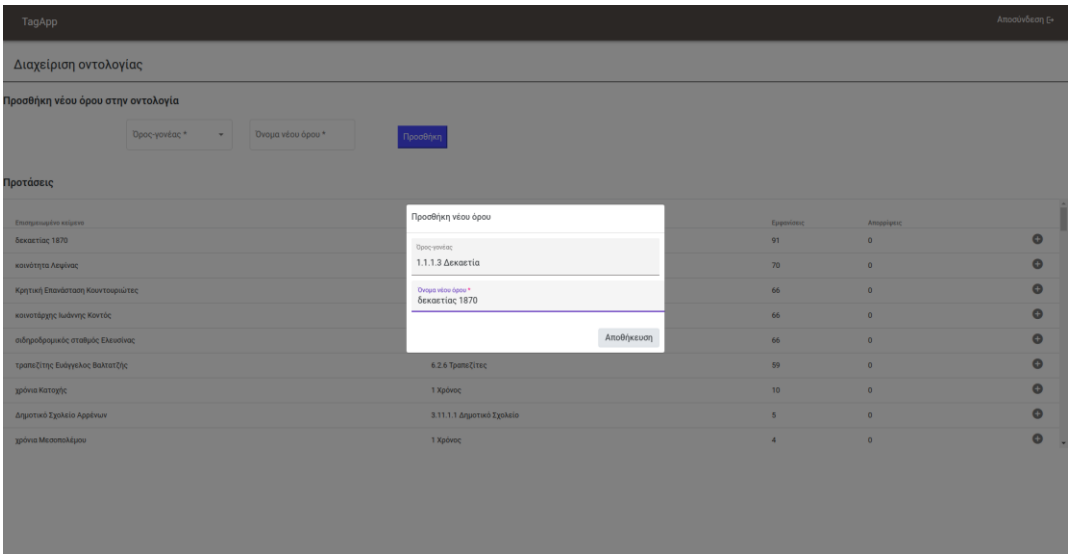


Εικόνα 30: Επιλογές εγγράφου του χρήστη

Η τελευταία σελίδα της εφαρμογής είναι η σελίδα του διαχειριστή. Σε αυτήν την σελίδα ο διαχειριστής της εφαρμογής μπορεί να προσθέσει έναν όρο στην οντολογία είτε μέσω της φόρμας, είτε με την επιλογή κάποιου προτεινόμενου όρου, που έχει εμφανιστεί στα κείμενα των χρηστών.

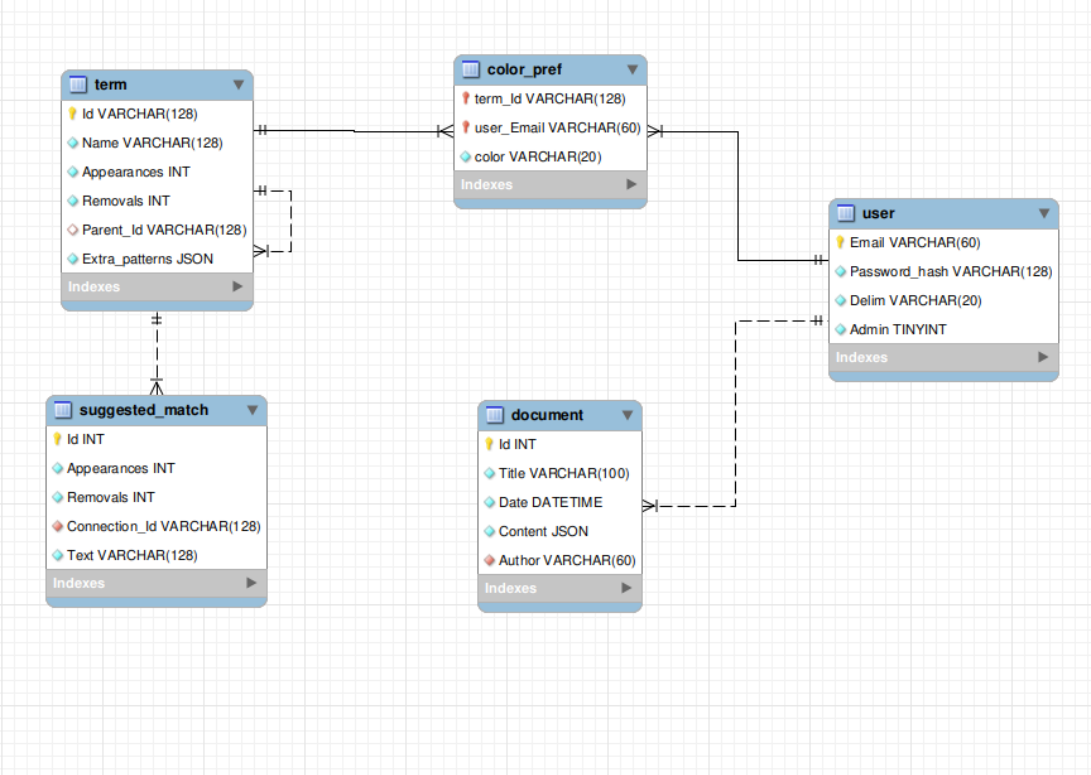


Εικόνα 31: Σελίδα διαχειριστή



Εικόνα 32: Σελίδα διαχειριστή επιλογές

5.3.2 Βάση δεδομένων



Εικόνα 33: Βάση δεδομένων της εφαρμογής

Στην παραπάνω εικόνα φαίνεται το σχήμα της βάσης δεδομένων της εφαρμογής.

Όπως μπορεί να διαπιστωθεί αποτελείται από 5 πίνακες, οι οποίοι περιγράφονται αναλυτικά στην συνέχεια.

1) Πίνακας user : Είναι ο πίνακας για την αναπαράσταση των χρηστών της εφαρμογής. Ο κάθε χρήστης έχει ένα μοναδικό email, έναν κωδικό, ο οποίος για λόγους ασφάλειας δεν αποθηκεύεται αυτούσιος στη βάση, μία προτίμηση στο τρόπο με τον οποίο επιθυμεί να διαχωρίζονται οι ενότητες στα κείμενα, τα οποία ανοίγει στην εφαρμογή, και, τέλος, ένα πεδίο που καθορίζει αν ο χρήστης είναι διαχειριστής της εφαρμογής.
2) Πίνακας document : Αποτελεί τον πίνακα για την αναπαράσταση ενός αποθηκευμένου επισημειωμένου αρχείου. Το κάθε αρχείο αποτελείται από τον τίτλο του, την ημερομηνία τροποποίησης και το περιεχόμενό του. Επιπλέον συνδέεται με τον χρήστη, ο οποίος το επεξεργάστηκε και το αποθήκευσε και διαθέτει ένα μοναδικό αναγνωριστικό.
3) Πίνακας term : Ο πίνακας term αναπαριστά έναν όρο της οντολογίας. Στα πεδία του περιέχονται ένα μοναδικό αναγνωριστικό, το όνομα του όρου, ο αριθμός εμφανίσεων και απορρίψεων του όρου στα κείμενα τα οποία έχουν εμφανιστεί στην εφαρμογή, το αναγνωριστικό του όρου-γονέα με βάση την ιεραρχία της οντολογίας και τα επιπλέον μοτίβα που έχουν του όρου.
4) Πίνακας color_pref : Ο πίνακας αυτός αναπαριστά την χρωματική προτίμηση του χρήστη για την επισημείωση ενός όρου της οντολογίας, καθώς και των απογόνων του όρου αυτού. Αποτελείται από το αναγνωριστικό του όρου της

οντολογίας, το email του χρήστη, τον οποίο αφορά η προτίμηση, και το χρώμα προτίμησης. Τα πρώτα δύο πεδία είναι το πρωτεύον κλειδί του πίνακα.

5) Πίνακας suggested_match : Αποτελεί τον πίνακα για την αναπαράσταση των προτεινόμενων επισημειώσεων που έχουν εμφανιστεί στα κείμενα της εφαρμογής. Περιέχει ένα μοναδικό αναγνωριστικό της προτεινόμενης επισημείωσης και το αναγνωριστικό του όρου της οντολογίας, με τον οποίο συνδέεται. Επιπροσθέτως, περιέχει το πεδίο text, το οποίο αντιπροσωπεύει το κείμενο το οποίο έχει επισημειωθεί, και τις εμφανίσεις και τις απορρίψεις αυτής της προτεινόμενης επισημείωσης στα κείμενα που έχουν εμφανιστεί στην εφαρμογή.

5.3.3 Back-end

Στην ενότητα αυτή γίνεται μία σύντομη περιγραφή των συναρτήσεων που αναπτύχθηκαν στο back-end, καθώς και μία ανάλυση κάποιων επιλογών υλοποίησης.

5.3.2.1 Περιγραφή συναρτήσεων

Το σύνολο του κώδικα για το back-end αναπτύσσεται στα παρακάτω αρχεία:

1) **routes.py :**

Στο αρχείο αυτό περιλαμβάνονται τα endpoints του API που δημιουργήθηκε. Στον παρακάτω πίνακα αναφέρεται η κάθε συνάρτηση του αρχείου και η λειτουργικότητά της.

Πίνακας 2: Συναρτήσεις του αρχείου routes.py

Όνομα συνάρτησης	**Λειτουργία**
get_suggested_matches	Επιστρέφει το σύνολο των αντικειμένων του πίνακα suggested_match, δηλαδή το σύνολο των προτεινόμενων συνδέσεων που έχουν εμφανιστεί στα κείμενα των χρηστών μέχρι στιγμής.
add_term_to_ontology	Αποθηκεύει έναν καινούργιο όρο της οντολογίας στη βάση δεδομένων και επιστρέφει το id του νέου όρου.
login	Ελέγχει το e-mail και τον κωδικό που δόθηκαν και επιστρέφει αν τα στοιχεία ήταν έγκυρα ή όχι, καθώς και αν ο χρήστης είναι διαχειριστής.
unique_email	Επιστρέφει αν το e-mail που δόθηκε είναι μοναδικό ή αν υπάρχει ήδη στην βάση δεδομένων.
register	Προσθέτει έναν καινούργιο χρήστη στην βάση δεδομένων.

upload_file	Δέχεται το κείμενο ενός αρχείου, το χωρίζει σε ενότητες ανάλογα με το διαχωριστικό που δίνεται και επιστρέφει τις ετικέτες και τις επισημειώσεις της κάθε ενότητας.
get_user_document	Επιστρέφει το περιεχόμενο ενός κειμένου που έχει αποθηκεύσει ένας χρήστης.
get_ontology	Επιστρέφει τους όρους της οντολογίας, που είναι αποθηκευμένη στη βάση δεδομένων.
get_user_document_info	Επιστρέφει πληροφορίες (πχ τον τίτλο) για τα κείμενα που έχει αποθηκεύσει ένας χρήστης.
save_user_document	Αποθηκεύει το κείμενο ενός χρήστη στη βάση δεδομένων.
update_user_document	Ενημερώνει τον τίτλο και τα περιεχόμενα ενός κειμένου του χρήστη.
rename_user_document	Αλλάζει τον τίτλο ενός κειμένου στη βάση δεδομένων.
delete_user_document	Διαγράφει ένα κείμενο από την βάση δεδομένων.
change_user_delim	Αλλάζει την προτίμηση που έχει ο χρήστης για το διαχωρισμό των ενοτήτων.
update_color_pref	Αλλάζει τις προτιμήσεις ενός χρήστη για τον χρωματισμό των κύριων όρων της οντολογίας.
update_term_removals	Αυξάνει τις απορρίψεις του όρου της οντολογίας στην βάση δεδομένων κατά 1.
update_term_appearances	Αυξάνει τις εμφανίσεις ενός όρου της οντολογίας στην βάση δεδομένων κατά 1.
get_main_terms_colors	Επιστρέφει τις προτιμήσεις ενός χρήστη για τον χρωματισμό των κύριων όρων.

update_suggested_match_removals	Αυξάνει τις απορρίψεις μιας προτεινόμενης σύνδεσης κατά 1.
update_suggested_match_appearances	Αυξάνει τις εμφανίσεις μιας προτεινόμενης σύνδεσης κατά 1.
update_term_patterns	Ενημερώνει τα επιπλέον μοτίβα που αντιστοιχούν σε έναν όρο της οντολογίας.

2) annotate_doc.py:

Το αρχείο αυτό αποτελείται από την συνάρτηση annotate_doc, η οποία είναι υπεύθυνη για την επισημείωση των ενοτήτων ενός κειμένου. Ειδικότερα, αφού χωρίσει το κείμενο σε ενότητες χρησιμοποιεί τον matcher του spacy (βλ. παρακάτω ενότητα 5.3.2.2) για να μπορέσει να βρει όλες τις άμεσες συνδέσεις με τους όρους της οντολογίας. Μετά την εύρεση αυτών των συνδέσεων ελέγχει για συγκρούσεις, καθώς κάθε λέξη μπορεί να έχει μία επισημείωση, κρατώντας την σύνδεση που περιέχει τον ειδικότερο όρο της οντολογίας.

Στη συνέχεια βρίσκει τις προτεινόμενες συνδέσεις με τους όρους της οντολογίας, όπως επίσης και τα επιπλέον μοτίβα που έχουν προστεθεί από τους χρήστες, ελέγχει και αυτές για συγκρούσεις και τελικά επιστρέφει τις λέξεις του κειμένου, τις ετικέτες της κάθε ενότητας και τις επισημειώσεις που βρέθηκαν για την ενότητα αυτή.

3) matches.py:

Στο αρχείο αυτό περιλαμβάνονται όλες οι συναρτήσεις που ασχολούνται με τα μοτίβα, από τα οποία προκύπτουν οι επισημειώσεις. Το όνομα και η λειτουργικότητα αυτών των συναρτήσεων παρουσιάζεται στον παρακάτω πίνακα.

Πίνακας 3: Συναρτήσεις του αρχείου matches.py

Όνομα συνάρτησης	**Λειτουργικότητα**
create_pattern	Δημιουργεί ένα μοτίβο, που αποτελείται από το λήμμα του κειμένου που δέχεται ως όρισμα.
create_patterns	Δημιουργεί μοτίβα, καλώντας την create_pattern για όλους τους όρους της οντολογίας και για κάθε extra_pattern του όρου.

create_suggested_pattern	Δημιουργεί κάποια μοτίβα, με βάση τα dependency labels του spacy, για το κείμενο που δέχεται ως όρισμα.
create_suggested_patterns	Δημιουργεί μοτίβα, καλώντας την create_suggested_pattern για κάθε όρο της οντολογίας και για κάθε extra_pattern του κάθε όρου.
create_words_indexes	Δέχεται την λέξη-βάση ενός μοτίβου (την λέξη που υπάρχει ήδη στο ελεγχόμενο λεξιλόγιο) και την λέξη-σύνδεση (την καινούργια λέξη που εντοπίστηκε στο κείμενο) και επιστρέφει έναν πίνακα με το σύνολο των θέσεων των λέξεων που αποτελούν μέρος αυτής της επισημείωσης. Ουσιαστικά προσθέτει στις λέξεις της επισημείωσης συντακτικούς απογόνους της λέξης-σύνδεσης, που θεωρούνται σημαντικοί.
create_nmod_annotation	Επιστρέφει έναν πίνακα με το id και τις θέσεις των λέξεων μιας προτεινόμενης επισημείωσης αφού πρώτα ελέγξει για την εγκυρότητά της.
create_flat_annotation	Επιστρέφει έναν πίνακα με το id και τις θέσεις των λέξεων μιας προτεινόμενης επισημείωσης αφού πρώτα ελέγξει για την εγκυρότητά της.
create_annotation	Δημιουργεί μία επισημείωση ανάλογα με το είδος της σύνδεσης που δέχεται σαν όρισμα.

4) helpers.py:

Το αρχείο helpers.py περιλαμβάνει κάποιες βοηθητικές συναρτήσεις, η λειτουργικότητα των οποίων περιγράφεται στον παρακάτω πίνακα.

Πίνακας 4: Συναρτήσεις του αρχείου helpers.py

Όνομα	**Λειτουργικότητα**
compare_text	Δημιουργεί strings με τα λήμματα των λέξεων των δύο κειμένων που δέχεται σαν όρισμα και επιστρέφει την Levenshtein απόστασή τους.
find_root	Επιστρέφει την συντακτική ρίζα του κειμένου που δέχεται σαν όρισμα, αν αυτή υπάρχει.
hierarchy_level	Επιστρέφει το επίπεδο της ιεραρχίας στο οποίο βρίσκεται ένας όρος της οντολογίας.
lemmatize_text	Επιστρέφει την πρόταση που της δόθηκε ως όρισμα αντικαθιστώντας την κάθε λέξη με το λήμμα της.

5) **app.py :**

   Σε αυτό το αρχείο γίνονται οι αρχικοποιήσεις που είναι απαραίτητες κατά την εκκίνηση του server.

6) **config.py :**

   Το config.py περιέχει όλες τις ρυθμίσεις που χρειάζονται (για παράδειγμα τις πληροφορίες για την σύνδεση με την βάση δεδομένων) που είναι απαραίτητες για την λειτουργία του back-end.

5.3.2.2 Αιτιολόγηση επιλογών υλοποίησης

Σε αυτήν την υποενότητα αναλύεται το σκεπτικό πίσω από κάποιες βασικές επιλογές υλοποίησης για το back-end.

Χρήση του matcher

Για την επισημείωση του κειμένου χρησιμοποιείται ο Matcher του spacy, ο οποίος βρίσκει ακολουθίες από tokens σε ένα κείμενο, με βάση τα μοτίβα που του έχουν προστεθεί. Το spacy διαθέτει και τον PhraseMatcher, ο οποίος λειτουργεί με παρόμοιο τρόπο. Η κύρια διαφορά των δύο, αλλά και ο λόγος για τον οποίο επιλέχθηκε ο πρώτος, είναι η ευελιξία που προσφέρει ο matcher αναφορικά με τα patterns τα οποία δημιουργούνται. Αναλυτικότερα, ο PhraseMatcher επιτρέπει την εισαγωγή μοτίβων, που αποτελούνται από Doc του spacy, ενώ ο Matcher προσφέρει μεγαλύτερη ελευθερία, δίνοντας την δυνατότητα να προστεθούν πιο δυναμικά μοτίβα. Για παράδειγμα, μπορεί να προστεθεί το μοτίβο [{“LOWER”: “hello”}, {“LOWER”: “world”}], το οποίο μπορεί να εντοπίζει σε ένα κείμενο όλες τις εμφανίσεις της φράσης “hello world” είτε είναι

γραμμένη με μικρά είτε με κεφαλαία, καθώς το μοτίβο που έχει προστεθεί ψάχνει για λέξεις των οποίων η μετατροπή από κεφαλαία σε πεζά ταιριάζει με την λέξη που δίνεται.

Επέκταση patterns

Μία από τις βασικότερες επιλογές ήταν ο τρόπος με τον οποίο θα μπορεί το σύστημα να αναγνωρίζει επισημειώσεις, εκτός από αυτές που αποτελούν απλή αντιστοίχηση με τους όρους του ελεγχόμενου λεξιλογίου. Για την επίτευξη του σκοπού αυτού αξιοποιήθηκαν τα dependency labels του spacy, με την βοήθεια των οποίων μπορεί να επιτευχθεί η σύνδεση λέξεων του ελεγχόμενου λεξιλογίου με καινούργιες λέξεις, λόγω της συντακτικής τους συσχέτισης. Μετά από πειράματα διαπιστώθηκε πως οι περισσότερες έγκυρες συνδέσεις προκύπτουν από τις ετικέτες nmod και flat, που διαθέτει ο dependency parser του spacy.

Το πλήρες όνομα της πρώτης ετικέτας είναι nominal modifier και η λέξη στην οποία αναφέρεται αποτελεί ένα ουσιαστικό, που αντιστοιχεί λειτουργικά σε ένα επίρρημα όταν συνδέεται με ένα ρήμα, επίθετο ή άλλο επίρρημα, ή σε ένα χαρακτηριστικό ή ένα γενετικό συμπλήρωμα όταν συνδέεται με ένα ουσιαστικό. Για παράδειγμα στην πρόταση “οι άθλοι του Ηρακλή”, η λέξη Ηρακλής αποτελεί nmod της λέξης άθλοι.

Η δεύτερη ετικέτα, ετικέτα flat, ή αλλιώς η επίπεδη σχέση μεταξύ λέξεων, χρησιμοποιείται για να δηλώσει την εξάρτηση μεταξύ ουσιαστικών. Για παράδειγμα αν είχαμε τις λέξεις “Χαρίλαος Τρικούπης” σε μία πρόταση, στην λέξη Τρικούπης θα αποδιδόταν η ετικέτα flat του dependency parser του spacy.

Τα επιπρόσθετα μοτίβα που δημιουργούνται, χρησιμοποιούνται για τον εντοπισμό όρων της οντολογίας που ακολουθούνται από κάποια λέξη με ένα από τα παραπάνω dependency labels, στην ίδια πρόταση.

Επιλογή επικρατέστερης επισημείωσης

Στην παρούσα εφαρμογή η κάθε λέξη μπορεί να έχει μία μόνο επισημείωση. Για αυτό το λόγο, κατά τη διαδικασία επισημείωσης ενός κειμένου, αφαιρούνται οι συγκρούσεις μεταξύ των επισημειώσεων. Ως επικρατέστερη επισημείωση ανάμεσα στις συγκρουόμενες επισημειώσεις επιλέγεται η επισημείωση με την χαμηλότερη ιεραρχική θέση στην οντολογία. Η επιλογή αυτή έγινε, καθώς για την συγκεκριμένη αξιοποίηση της εφαρμογής θεωρήθηκε η πιο αποδοτική.

Ομοιότητα μεταξύ φράσεων

Επιπλέον, μία ακόμα απόφαση που έπρεπε να παρθεί ήταν ο τρόπος με τον οποίο θα κρινόταν αν μία προτεινόμενη σύνδεση θα έπρεπε να προστεθεί στην βάση δεδομένων ως νέα ή αν θα έπρεπε να θεωρηθεί παρόμοια με μία υπάρχουσα. Για την επίλυση του ζητήματος αυτού, χρειάστηκε μία συνάρτηση που να υπολογίζει την ομοιότητα μεταξύ δύο φράσεων. Μετά από μία σύντομη έρευνα καταλήξαμε στην απόσταση Levenshtein, η οποία υπολογίζει τον αριθμό των μετατροπών που χρειάζονται για να μπορέσει η πρώτη φράση να μετασχηματιστεί στην δεύτερη. Ως φράγμα για την ομοιότητα των δύο φράσεων, ύστερα από δοκιμές και με δεδομένο το γεγονός ότι συγκρίνουμε φράσεις που αποτελούνται από λήμματα, θεωρήθηκε το 3 επί το πλήθος των λέξεων της φράσης. Ουσιαστικά, λόγω των λαθών κυρίως στις καταλήξεις των λημμάτων των λέξεων, που υπάρχουν στο spacy, θεωρήθηκε επιτρεπτός ένας μέσος όρος στα τρία διαφορετικά γράμματα ανά λέξη, για τις επισημειώσεις με βάση την συγκεκριμένη οντολογία.

6 ΑΞΙΟΛΟΓΗΣΗ

Απαραίτητο στάδιο κατά την διάρκεια ανάπτυξης της εφαρμογής αλλά και μετά την ολοκλήρωσή της είναι η αξιολόγηση, προκειμένου να διαπιστωθούν και να αντιμετωπιστούν πιθανά προβλήματα. Στην ενότητα αυτή περιγράφεται η διαδικασία της τελικής αξιολόγησης της εφαρμογής, καθώς και τα αποτελέσματα που προέκυψαν.

6.1 Διαδικασία αξιολόγησης χρηστών

Για την αξιολόγηση της εφαρμογής χρησιμοποιήθηκε ένας συνδυασμός μεθόδων. Αρχικά ζητήθηκε από 5 χρήστες να εκτελέσουν το σενάριο χρήσης που περιγράφεται στη συνέχεια, ενώ παρατηρούσαμε τις κινήσεις τους και ζητήσαμε να αναφέρουν τις σκέψεις και τα προβλήματα τους κατά τη διάρκεια της χρήσης του εργαλείου (Think aloud protocol). Μετά τη χρήση του εργαλείου ζητήθηκε από τους χρήστες να απαντήσουν σε ένα ερωτηματολόγιο και πραγματοποιήθηκε μία σύντομη συνέντευξη. Το ερωτηματολόγιο περιέχει το user experience questionaire (https://www.ueq-online.org/) για τα ελληνικά και κάποιες ερωτήσεις ανοικτού τύπου. Το user experience questionaire είναι ένα γρήγορο, αξιόπιστο ερωτηματολόγιο, που μετράει την εμπειρία χρήστη σε διαδραστικά προϊόντα και είναι διαθέσιμο για πάνω από 30 γλώσσες.

Σενάριο χρήσης

1. Αρχικά θα πρέπει να ανοίξετε το αρχείο example.txt.
2. Στην εφαρμογή μπορείτε να αφαιρέσετε και να προσθέσετε επισημειώσεις στις λέξεις. Για παράδειγμα, προσπαθήστε να αφαιρέσετε την επισημείωση της λέξης “χρόνια” από την πρώτη ενότητα και να προσθέσετε στην λέξη “1879” την επισημείωση “19ος αιώνας”.
3. Εκτός από τις επισημειώσεις μπορείτε να αλλάξετε και τις ετικέτες της ενότητας. Για παράδειγμα, αφαιρέστε την ετικέτα “Χρόνος” από την πρώτη ενότητα του κειμένου.
4. Όπως μπορεί να έχετε ήδη προσέξει εκτός από τις χρωματισμένες λέξεις, που αναπαριστούν τις λέξεις οι οποίες είναι επισημειωμένες, υπάρχουν και κάποιες υπογραμμισμένες λέξεις. Οι λέξεις αυτές αποτελούν προτεινόμενες επισημειώσεις. Μεταβείτε στην 4η παράγραφο του κειμένου και αποδεχτείτε την προτεινόμενη επισημείωση για το “1870”.
5. Αφού εξοικειωθήκατε λίγο με τις ετικέτες και τις επισημειώσεις, μπορείτε να δείτε τις επιλογές που υπάρχουν για να διευκολύνουν την αναζήτηση κάποιας συγκεκριμένης κατηγορίας επισημειώσεων ή τον διαχωρισμό των κύριων κατηγοριών. Αρχικά, μπορείτε να εφαρμόσετε ένα φίλτρο στο κείμενο επιλέγοντας, οποιοδήποτε όρο του λεξιλογίου. Για παράδειγμα, επιλέξτε τον όρο “Χρονική περίοδος”. Αφού παρατηρήσετε τα αποτελέσματα μπορείτε να αφαιρέσετε το φίλτρο.
6. Στη συνέχεια, μεταβείτε στις ρυθμίσεις και αλλάξτε τα χρώματα των κύριων κατηγοριών του λεξιλογίου για να μπορείτε να τις διακρίνετε με ευκολία στο κείμενο.
7. Τέλος, περιηγηθείτε στις ενότητες του κειμένου και αλλάξτε 3-4 επισημειώσεις ή ετικέτες που θεωρείτε ότι χρειάζονται αλλαγή.

6.2 Αποτελέσματα αξιολόγησης χρηστών

Από την παραπάνω διαδικασία αξιολόγησης προέκυψαν αρκετές ενδιαφέρουσες παρατηρήσεις από τον κάθε χρήστη. Κάποιες από αυτές τις παρατηρήσεις των χρηστών και τα προβλήματα που αντιμετώπισαν ήταν κοινά και θεωρήθηκε σημαντικό να αναφερθούν.

Αρχικά η θετική παρατήρηση που έγινε από όλους τους χρήστες ήταν η ευκολία στην εκμάθηση της εφαρμογής και η απλότητά της. Το κυριότερο πρόβλημα που αντιμετώπισαν οι χρήστες ήταν μετά το τέλος του βήματος 6, του σεναρίου αξιολόγησης, όταν έπρεπε να επιστρέψουν στην αρχική οθόνη της εφαρμογής, καθώς η πλειοψηφία των χρηστών δεν μπόρεσε να βρει το σύνδεσμο για την επιστροφή στην αρχική. Ένα ακόμα πρόβλημα που αντιμετώπισαν οι χρήστες αφορά τα φίλτρα στα αριστερά της αρχικής οθόνης. Αναλυτικότερα, η εμφάνιση και η λειτουργία των φίλτρων δεν ήταν απολύτως κατανοητή στους χρήστες, οι οποίοι καθυστέρησαν μέχρι να καταφέρουν να τα χρησιμοποιήσουν, ενώ πολλοί δεν χρησιμοποίησαν την αναζήτηση που υπάρχει στο πάνω μέρος ή προσπάθησαν να τα χρησιμοποιήσουν όταν τους ζητούνταν να αλλάξουν ενότητα. Τέλος, ένα σοβαρό πρόβλημα της εφαρμογής είναι ο τρόπος με τον οποίο επιλέγονται οι λέξεις του κειμένου. Πολλοί χρήστες προσπάθησαν να προσθέσουν επισημείωση ενώ δεν είχαν επιλέξει την λέξη ή τις λέξεις που πίστευαν.

6.3 Διαδικασία αξιολόγησης αυτόματης επισημείωσης

Για την αξιολόγηση της αυτόματης επισημείωσης ελέγχθηκαν τρία καινούργια κείμενα από το έργο Pros-eleusis. Τα κείμενα αυτά φορτώθηκαν στην εφαρμογή και μελετήθηκαν ποιοτικά και ποσοτικά οι προτεινόμενες επισημειώσεις που εμφανίστηκαν.

6.4 Αποτελέσματα αξιολόγησης αυτόματης επισημείωσης

Μετά το πέρας του παραπάνω πειράματος, προέκυψαν τα παρακάτω στοιχεία για το κάθε κείμενο.

Πίνακας 5: Αποτελέσματα αξιολόγησης αυτόματης επισημείωσης

	Πλήθος επισημειώσεων	Πλήθος προτεινόμενων επισημειώσεων
Κείμενο 1	253	30
Κείμενο 2	245	34
Κείμενο 3	164	17

Από τον παραπάνω πίνακα προκύπτει ότι περίπου για κάθε 10 επισημειώσεις δημιουργείται και μία προτεινόμενη επισημείωση. Παρά το γεγονός ότι το ποσοστό αυτό θα θέλαμε να ήταν μεγαλύτερο, αν σκεφτούμε πως οι προτεινόμενες επισημειώσεις προκύπτουν από την επέκταση των όρων του ελεγχόμενου λεξιλογίου, δηλαδή αποτελούν επέκταση των επισημειωμένων λέξεων, για τις οποίες αρκετές φορές δεν υπάρχουν παραπάνω λέξεις με τις οποίες συνδέονται συντακτικά και θα ήταν επιθυμητό να επισημειωθούν, καταλαβαίνουμε ότι το ποσοστό αυτό δεν είναι και τόσο αποθαρρυντικό.

Επιπλέον, ένα θετικό στατιστικό που προέκυψε από το πείραμα ήταν πως το πλήθος των προτεινόμενων επισημειώσεων, για τις οποίες θα μπορούσαμε με σιγουριά να πούμε ότι δεν θα έπρεπε να υπάρχουν ήταν μόνο 2.

Από τα παραπάνω μπορούμε να υποθέσουμε ότι με την πλήρη αξιοποίηση των συντακτικών συνδέσεων που μπορούν να φανούν χρήσιμες για την εύρεση νέων επισημειώσεων αλλά και από την ρύθμιση της ανοχής στα λάθη θα μπορούσε να επέλθει αρκετά μεγάλη βελτίωση.

7 ΣΥΜΠΕΡΑΣΜΑΤΑ - ΜΕΛΛΟΝΤΙΚΕΣ ΕΠΕΚΤΑΣΕΙΣ

7.1 Συμπεράσματα

Μετά την ολοκλήρωση της εφαρμογής αλλά και την εξοικείωση με τον τομέα της επισημείωσης, λόγω της παρούσας πτυχιακής εργασίας, έγινε κατανοητό ότι η ημι-αυτόματη επισημείωση κειμένων είναι μια πολύπλοκη διαδικασία. Η επιτυχία της εξαρτάται σε μεγάλο βαθμό από τις μεθόδους που θα αναπτυχθούν για την αποδοτικότερη εύρεση των λέξεων που χρήζουν επισημείωσης αλλά και από την όσο τον δυνατόν αναλυτικότερη και ορθότερη επεξεργασία της φυσικής γλώσσας. Επιπλέον, η αποδοτικότητα ενός εργαλείου ημι-αυτόματης επισημείωσης, εξαιτίας των διαφορετικών αναγκών επισημείωσης του εκάστοτε κειμένου, δεν είναι καθολική. Όμως, παρά τις παραπάνω δυσκολίες είναι σαφής η ανάγκη έστω και για μια ημιτελής ημι-αυτόματη επισημείωση κειμένων, καθώς η διαδικασία της χειροκίνητης επισημείωσης είναι χρονοβόρα και δύσκολη. Εκτός αυτού αν ο σκοπός της επισημείωσης είναι παρόμοιος με αυτόν του έργου Pros-eleusis, που αποτέλεσε και τον λόγο για την δημιουργία αυτής της πτυχιακής, η χειροκίνητη επισημείωση ελλοχεύει και τον κίνδυνο σε παρόμοιες λέξεις να προστίθενται διαφορετικές επισημειώσεις, με αποτέλεσμα να αλλοιώνονται τα αποτελέσματα της σύγκρισης μεταξύ των παραγράφων.

7.2 Μελλοντικές επεκτάσεις

Αναμενόμενο είναι πως κατά την διάρκεια της υλοποίησης της εφαρμογής υπήρξαν ιδέες για πιθανές επεκτάσεις, αλλά και, μετά την ολοκλήρωσή της, διαπιστώσεις για μελλοντικές βελτιώσεις. Στο κεφάλαιο, αυτό, παρατίθενται κάποιες ιδέες, που στοχεύουν στην βελτίωση της αποτελεσματικότητας της εφαρμογής.

1) **Βάρος επισημείωσης για την κάθε ενότητα**

   Το σημαντικότερο κομμάτι της εφαρμογής για τον χρήστη είναι οι ετικέτες που προκύπτουν για την κάθε ενότητα. Συνεπώς, είναι σημαντική και οποιαδήποτε βελτίωση είναι σχετική με τις ετικέτες αυτές. Μία τέτοια βελτίωση, θα μπορούσε να είναι η απεικόνιση με τρόπο ξεκάθαρο για τον χρήστη, του ποσοστού συσχέτισης της κάθε ετικέτας με την ενότητα.

2) **Δυνατότητα κοινής χρήσης του κειμένου με άλλους χρήστες**

   Όσο ικανοποιητική και να είναι η επισημείωση του κειμένου είναι σχεδόν σίγουρη η ανάγκη του χρήστη για αλλαγές, λόγω λαθών της εφαρμογής αλλά και λόγω των διαφορετικών απαιτήσεων του χρήστη από την επισημείωση του κάθε κειμένου. Συνεπώς, θα αποτελούσε μεγάλη βοήθεια στο χρήστη η δυνατότητα διόρθωσης του κειμένου σε συνεργασία με άλλους χρήστες, όταν η επισημείωση αποτελεί τον σκοπό μιας ομάδας ατόμων.

3) **Δυνατότητα αποθήκευσης του επισημειωμένου κειμένου εκτός εφαρμογής**

   Ο σκοπός χρήσης της εφαρμογής είναι προφανώς η επισημείωση του κειμένου. Μετά όμως από την επιτυχημένη επισημείωση είναι δεδομένο ότι ο χρήστης θα θέλει να εκμεταλλευτεί το επισημειωμένο κείμενο. Στην παρούσα εφαρμογή δεν υπάρχει η δυνατότητα αποθήκευσης του κειμένου εκτός του πλαισίου της εφαρμογής. Επομένως, γίνεται φανερή η ανάγκη της προσθήκης μιας επιλογής αποθήκευσης εκτός της εφαρμογής, που να μπορεί να χρησιμοποιηθεί και να διαβαστεί από τους χρήστες.

4) **Δυνατότητα προσθήκης νέου ελεγχόμενου λεξιλογίου**

Η επιτυχής επισημείωση των κειμένων της εφαρμογής βασίζεται κατά ένα μεγάλο βαθμό στο ελεγχόμενο λεξιλόγιο που χρησιμοποιείται. Επομένως, για την επιτυχημένη επισημείωση κειμένων διαφορετικού περιεχομένου, τα οποία δεν καλύπτονται από το υπάρχον λεξιλόγιο, θα πρέπει να υπάρχει η δυνατότητα χρήσης διαφορετικού λεξιλογίου, το οποίο να επιλέγεται από τον χρήστη ανάλογα με τις ανάγκες του.

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Ξενόγλωσσος όρος	**Ελληνικός όρος**
E-mail	Ηλεκτρονικό ταχυδρομείο
Username	Όνομα χρήστη
Password	Κωδικός
Checkboxes	Κουτιά επιλογής
Front-end	Ο κώδικας που γράφεται και αφορά το εμφανισιακό κομμάτι μιας εφαρμογής
Web application framework	Πλαίσιο εφαρμογής ιστού
Back-end	Ο κώδικας που γράφεται για την λειτουργικότητα της εφαρμογής και τη επικοινωνία του front-end με την βάση δεδομένων
Microframework	Μικροπλαίσιο
Endpoints	Τελικά σημεία
Dependency labels	Ετικέτες εξάρτησης
Dependency parser	Αναλυτής εξάρτησης
Nominal modifier	Ονομαστικός τροποποιητής
Flat	Επίπεδο

ΣΥΝΤΜΗΣΕΙΣ - ΑΡΚΤΙΚΟΛΕΞΑ - ΑΚΡΩΝΥΜΙΑ

AAT	Art & Architecture Thesaurus
SHIC	Social History and Industrial Classification
CRM	Conceptual Reference Model
DOM	Document Object Model
HTML	Hypertext Markup Language
CSS	Cascading Style Sheets
IDE	Integrated Drive Electronics
API	Application Programming Interface
id	identifier

ΑΝΑΦΟΡΕΣ

[1] Art and Architecture Thesaurus [Ιστοσελίδα], Διαθέσιμο: https://www.getty.edu/research/tools/vocabularies/aat/index.html
[2] firstBASE [Ιστοσελίδα], Διαθέσιμο: http://www.shcg.org.uk/firstBASE-home
[3] SHIC [Ιστοσελίδα], Διαθέσιμο: http://www.shcg.org.uk/About-SHIC
[4] Nomenclature [Ιστοσελίδα], Διαθέσιμο: https://www.nomenclature.info/apropos-about.app?lang=en
[5] CIDOC-CRM [Ιστοσελίδα], Διαθέσιμο: http://www.cidoc-crm.org/
[6] Natural language proccessing [Ιστοσελίδα], Διαθέσιμο: https://en.wikipedia.org/wiki/Natural_language_processing
[7] spaCy [Ιστοσελίδα], Διαθέσιμο: https://spacy.io/
[8] Prodigy [Ιστοσελίδα], Διαθέσιμο: https://prodi.gy/
[9] Tagtog [Ιστοσελίδα], Διαθέσιμο: https://www.tagtog.net/
[10] INCEpTION [Ιστοσελίδα], Διαθέσιμο: https://inception-project.github.io/
[11] Marky [Ιστοσελίδα], Διαθέσιμο: http://www.sing-group.org/marky/index.html
[12] CHESS Roussou, M., & Katifori, A. (2018). Flow, Staging, Wayfinding, Personalization: Evaluating User Experience with Mobile Museum Narratives. Multimodal Technologies and Interaction, *2*(2), 32. http://www.mdpi.com/2414-4088/2/2/32/pdf
[13] Maria Roussou, Akrivi Katifori, Laia Pujol, Maria Vayanou, Stefan Rennick Egglestone: A life of their own: museum visitor personas penetrating the design lifecycle of a mobile experience. CHI Extended Abstracts 2013: 547-552
[14] Wikipedia «Angular (web framework)» [Ιστοσελίδα], Διαθέσιμο: https://en.wikipedia.org/wiki/Angular_(web_framework)
[15] Wikipedia, «React (web framework)» [Ιστοσελίδα], Διαθέσιμο: https://en.wikipedia.org/wiki/React_(web_framework)
[16] Wikipedia, «Django (web framework)» [Ιστοσελίδα], Διαθέσιμο: https://en.wikipedia.org/wiki/Django_(web_framework)
[17] Wikipedia, «Flask (web framework)» [Ιστοσελίδα], Διαθέσιμο: https://en.wikipedia.org/wiki/Flask_(web_framework)
[18] Wikipedia, «HTML» [Ιστοσελίδα], Διαθέσιμο: https://en.wikipedia.org/wiki/HTML
[19] Wikipedia, «CSS» [Ιστοσελίδα], Διαθέσιμο: https://en.wikipedia.org/wiki/CSS
[20] Wikipedia, «Typescript» [Ιστοσελίδα], Διαθέσιμο: https://en.wikipedia.org/wiki/TypeScript
[21] Wikipedia, «Python(programming language)» [Ιστοσελίδα], Διαθέσιμο: https://en.wikipedia.org/wiki/Python_(programming_language)
[22] Angular Material [Ιστοσελίδα], Διαθέσιμο: https://material.angular.io/
[23] Wikipedia, «Visual Studio Code» [Ιστοσελίδα], Διαθέσιμο: https://en.wikipedia.org/wiki/Visual_Studio_Code
[24] Wikipedia, «Pycharm» [Ιστοσελίδα], Διαθέσιμο: https://en.wikipedia.org/wiki/PyCharm
[25] Wikipedia, «MySQL Workbench» [Ιστοσελίδα], Διαθέσιμο: https://en.wikipedia.org/wiki/MySQL_Workbench