

Ατομική Διπλωματική Εργασία

ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΚΑΙ ΕΞΑΓΩΓΗ
ΔΙΚΑΙΩΜΑΤΩΝ ΚΑΙ ΥΠΟΧΡΕΩΣΕΩΝ
ΑΔΕΙΩΝ ΛΟΓΙΣΜΙΚΟΥ ΑΝΟΙΚΤΟΥ ΚΩΔΙΚΑ

Δημήτρης Πασχαλίδης

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ



ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάιος 2016

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μοντελοποίηση Και Εξαγωγή
Δικαιωμάτων Και Υποχρεώσεων
Αδειών Λογισμικού Ανοικτού Κωδικά

Δημήτρης Πασχαλίδης

Επιβλέπουσα Καθηγήτρια

Γεωργία Καπιτσάκη

Η Ατομική Διπλωματική Εργασία υποβλήθηκε προς μερική εκπλήρωση των
απαιτήσεων απόκτησης του πτυχίου Πληροφορικής του Τμήματος Πληροφορικής του
Πανεπιστημίου Κύπρου

Μάιος 2016

Ευχαριστίες

Στο σημείο αυτό θα ήθελα να εκφράσω τις ευχαριστίες μου στην επιβλέπουσα καθηγήτρια μου Δρ. Γεωργία Καπιτσάκη, για την υποστήριξη, την καθοδήγηση και τις συμβουλές που μου προσέφερε καθ' όλη τη διάρκεια αυτής της Διπλωματικής Εργασίας.

Επίσης θα ήθελα να ευχαριστήσω όλους τους καθηγητές του τμήματος και τους συμφοιτητές μου που προσφέρθηκαν να δοκιμάσουν την εφαρμογή. Τέλος, θα ήθελα να εκφράσω ένα μεγάλο ευχαριστώ στην οικογένεια μου και τους φίλους μου, που ήταν δίπλα μου σε όλη τη διάρκεια φοίτησης μου στο Πανεπιστήμιο Κύπρου.

Περίληψη

Η χρήση λογισμικού ανοικτού κώδικα έχει προσελκύσει την προσοχή διαφόρων εταιρειών και οργανισμών, οι οποίοι ενσωματώνουν λογισμικά από τρίτους στις δικές τους υλοποιήσεις. Όταν χρησιμοποιείται μια τέτοια βιβλιοθήκη ή ο πηγαίος κώδικας, έμμεσα ή άμεσα υπάρχει η σύνδεση με μία ή περισσότερες άδειες λογισμικού. Αυτό το γεγονός μπορεί να οδηγήσει σε πιθανές συγκρούσεις μεταξύ των αδειών αυτών, λόγω της δυσκολίας του ανθρώπου να κατανοήσει ή να προσέξει κάποιους σημαντικούς παράγοντες μέσα στο εν λόγω κείμενο. Οπότε απαραίτητη είναι η υποστήριξη όσον αφορά τη χρήση των αδειών λογισμικού με στόχο την αποφυγή τέτοιων συγκρούσεων, καθώς επίσης και την αντιμετώπιση τους εάν υπάρξουν. Αυτό μπορεί να επιτευχθεί μέσω της μοντελοποίησης τους.

Η άδεια λογισμικού έχει μορφή κειμένου και πλησιάζει τη δομή και μορφοποίηση νομικών κειμένων. Μέσα από το κείμενο μπορούμε να διακρίνουμε διάφορους ορισμούς και να τους οργανώσουμε σε τρεις βασικές κατηγορίες, τα δικαιώματα, τις υποχρεώσεις και στις επιπλέον πληροφορίες. Τα δικαιώματα αναφέρονται σε όλες τις παραχωρήσεις που δίνει ο δημιουργός του λογισμικού σε όλους τους χρήστες, οι υποχρεώσεις με τη σειρά τους αναφέρονται σε όλα όσα πρέπει να εφαρμόσει κάποιος χρήστης που επιλέγει να χρησιμοποιήσει οποιοδήποτε λογισμικό που συνοδεύεται από κάποια άδεια και τέλος οι επιπλέον πληροφορίες είναι χαρακτηριστικά που έχουν σχεδόν όλες οι άδειες και δεν εμπίπτουν ούτε στην κατηγορία δικαιώματα αλλά ούτε και στην κατηγορία υποχρεώσεις.

Η συγκεκριμένη διπλωματική εργασία παρουσιάζει το καθολικό μοντέλο άδειας μέσω του οποίου καλύπτονται οι πιθανοί τρόποι υποστήριξης θεμάτων συμβατότητας και διαχείρισης των αδειών λογισμικού ανοικτού κώδικα, εστιάζοντας στη δυνατότητα εξαγωγής του μοντέλου από υφιστάμενα κείμενα αδειών λογισμικού.

Όπως αναφέρθηκε, η χρήση λογισμικού ανοικτού κώδικα έχει αναπτυχθεί και υιοθετηθεί σε μεγάλο βαθμό τα τελευταία χρόνια. Κάθε λογισμικό ανοικτού κώδικα είναι αναγκαίο να μεταφέρει συγκεκριμένη άδεια λογισμικού, την οποία ο αδειοδοτούμενος πρέπει να κατανοεί και στην οποία πρέπει να συμμορφώνεται για να το χρησιμοποιήσει. Το κείμενο της άδειας όμως, το οποίο αποτελεί νομικού τύπου, δεν είναι πάντοτε κατανοητό από τον άνθρωπο, και απαιτεί την εις βάθος ανάλυση του με το ανθρώπινο μάτι. Το μοντέλο αδειών λογισμικού ανοικτού κώδικα το οποίο παρουσιάζεται αυτοματοποιεί την συγκεκριμένη ανάλυση, παρουσιάζοντας τα δικαιώματα, υποχρεώσεις και επιπλέον πληροφορίες της άδειας, μέσω μιας ειδικά διαμορφωμένης οντολογίας. Μέσω του μοντέλου ο άνθρωπος μπορεί ευκολά να διακρίνει τα βασικά δικαιώματα και υποχρεώσεις της άδειας, αλλά ακόμη να εφαρμοστεί μαζί με μια διαδικασία εξαγωγής συμπερασμάτων (μέσω reasoner) για να λάβει περαιτέρω πληροφορίες. Μαζί με το μοντέλο, παρουσιάζεται και ένα σύστημα τεχνητής νοημοσύνης, το οποίο λαμβάνει σαν είσοδο κείμενα αδειών λογισμικού παράγοντας σαν αποτέλεσμα το μοντέλο της συγκεκριμένης άδειας.

Παρόμοιο έργο έχει πραγματοποιηθεί πάνω σε συγκεκριμένο και περιορισμένο αριθμό αδειών λογισμικού ανοικτού κώδικα. Θα γίνει προσπάθεια για γενίκευση της εφαρμογής μοντελοποίησης πάνω σε μεγαλύτερο αν όχι ολόκληρο το σύνολο αυτών των αδειών.

Περιεχόμενα

Κεφάλαιο 1.....	8
Εισαγωγή.....	8
1.1 Πρόλογος.....	8
1.2 Σκοπός διπλωματικής εργασίας.....	9
1.3 Προσέγγιση.....	9
1.4 Δομή διπλωματικής εργασίας.....	10
Κεφάλαιο 2.....	11
Άδειες Λογισμικού Ανοικτού Κώδικα.....	11
2.1 Λογισμικό Ανοικτού Κώδικα.....	11
2.2 Άδειες Λογισμικού Ανοικτού Κώδικα.....	12
2.3 Συμβατότητα Αδειών.....	14
2.4 Προηγούμενες εργασίες.....	14
2.5 Μελέτη Αδειών Λογισμικού Ανοικτού Κώδικα.....	19
Κεφάλαιο 3.....	25
Οντολογία και μοντελοποίηση.....	25
3.1 Ορισμός οντολογίας.....	25
3.2 Μορφή οντολογίας.....	26
3.3 Σύνταξη οντολογίας.....	27
3.4 Επιλογή σύνταξης για το μοντέλο.....	28
3.5 Σχεδίαση και υλοποίηση μοντέλου.....	29
Κεφάλαιο 4.....	40
Σχεδιασμός και υλοποίηση Eclipse IDE Plugin.....	40

4.1 Σχεδιασμός plugin.....	40
4.2 Υλοποίηση plugin.....	46
4.3 Έκδοση plugin.....	54
4.4 Συμπεράσματα.....	54
Κεφάλαιο 5.....	56
Machine Learning και NLP - Natural Language Processing.....	56
5.1 Εισαγωγή.....	56
5.2 Semi-supervised learning.....	57
5.3 Επεξεργασία φυσικής γλώσσας (NLP - Natural language processing).....	58
Κεφάλαιο 6.....	62
Προσέγγιση ανάλυσης και προετοιμασία δεδομένων.....	62
6.1 Προσέγγιση ανάλυσης.....	62
6.2 Μεθόδοι ανάλυσης.....	63
6.3 Προετοιμασία δεδομένων.....	64
Κεφάλαιο 7.....	69
Ανάλυση προσεγγίσεων και Αποτελέσματα.....	69
7.1 Εισαγωγή.....	70
7.2 Μηχανικά πρότυπα μάθησης.....	70
7.3 Αυτόματη εξαγωγή προτύπων.....	86
7.3.2 Topic Modeling.....	86
7.3.3 Αλγόριθμοι για Topic Modeling.....	87
7.3.4 Αλγόριθμος LDA Latent Dirichlet Allocation.....	88
7.3.5 Κατανομή λέξεων κλειδιών.....	89
7.3.6 Topics με μεγαλύτερο score (LDA).....	90
Κεφάλαιο 8.....	95

Ανάλυση Αποτελεσμάτων.....	95
8.1 Σύγκριση Προσεγγίσεων.....	95
8.2 Σύγκριση με προκατασκευασμένα δεδομένα.....	100
Κεφάλαιο 9.....	105
Υλοποίηση REST Υπηρεσίας και Εφαρμογής Διαδικτύου.....	105
9.1 Εισαγωγή.....	106
9.2 Περιγραφή γλωσσών προγραμματισμού και συστατικών στοιχείων.....	106
9.3 Υλοποίηση υπηρεσίας REST.....	108
9.4 Υλοποίηση διαδικτυακής εφαρμογής.....	110
Κεφάλαιο 10.....	118
Συμπεράσματα και μελλοντικές εργασίες.....	118
10.1 Συμπεράσματα.....	118
10.2 Μελλοντικές εργασίες.....	119

Κεφάλαιο 1

Εισαγωγή

1.1 Πρόλογος	10
1.2 Σκοπός διπλωματικής εργασίας	11
1.3 Προσέγγιση	11
1.4 Δομή διπλωματικής εργασίας	11

1.1 Πρόλογος

Το λογισμικό ανοικτού κώδικα (open source software) χρησιμοποιείται ευρέως στις μέρες μας από τις ερευνητικές και αναπτυξιακές κοινότητες, μπορούμε επίσης να παρατηρήσουμε και την υιοθέτηση του από μεγάλες εμπορικές εταιρίες. Αν και πολλές μεγάλες εταιρίες χειρίζονται εσωτερικά ό,τι αφορά τη συμβατότητα των αδειών λογισμικού, σε πολλές περιπτώσεις ο πηγαίος κώδικας τέτοιων λογισμικών επαναχρησιμοποιείται χωρίς να κατανοείται πλήρως το τι πρέπει να πληροί ο χρήστης ή το ποια δικαιώματα έχει, βάσει της άδειας (license) που το συνοδεύει.

Η κατανόηση του περιεχομένου μιας άδειας λογισμικού αποτελεί σημαντικό παράγοντα για την χρήση του λογισμικού στο οποίο αναφέρεται, και κυρίως όσον αφορά το συνδυασμό του με άλλα λογισμικά, έτσι ώστε να αποφύγουμε τις πιθανές παραβιάσεις που μπορούν να οδηγήσουν και σε νομικές διενέξεις.

1.2 Σκοπός διπλωματικής εργασίας

Το νομικό κείμενο μιας άδειας λογισμικού, το οποίο είναι γραμμένο σε φυσική γλώσσα, δεν ακολουθεί κάποια σύμβαση στη συγγραφή του, οπότε αποτελεί δύσκολο έργο η κατανόηση του από τον ανθρώπινο παράγοντα. Υπάρχουν κάποια υφιστάμενα συστήματα μέσω του διαδικτύου, τα οποία μας παρουσιάζουν επεξηγηματικές και σημαντικές πληροφορίες για τις άδειες (π.χ. *tldrlegal* [3]), αλλά λαμβάνουν υπόψη περιορισμένο αριθμό από άδειες οι οποίες αναλύθηκαν χειροκίνητα.

Σκοπός της παρούσης διπλωματικής εργασίας είναι να παρουσιάσει ένα καθολικό μοντέλο που θα μπορεί να εφαρμοστεί σε όλες τις άδειες λογισμικού με σκοπό την εύκολη κατανόηση τους από τον άνθρωπο, καθώς επίσης και την ανάλυσή τους από υπολογιστή. Βασικός στόχος είναι η δυνατότητα εξαγωγής των όρων του μοντέλου αυτού από οποιοδήποτε κείμενο αδειών λογισμικού ανοικτού κώδικα δοθεί ως είσοδος. Για το σκοπό αυτό έχει γίνει χρήση διάφορων τεχνικών και αλγορίθμων. Η κατασκευή του μοντέλου έγινε μέσω οντολογίας σημασιολογικού ιστού (Semantic Web Ontology), με την χρήση των ορισμών που εξήχθησαν στη διπλωματική εργασία της Γεωργίας Χαραλάμπους [1].

1.3 Προσέγγιση

Η προσέγγιση για την κατασκευή του συστήματος μοντελοποίησης αποτέλεσε ένα μεγάλο έργο με πολλή έρευνα και μελέτη. Χωρίζεται σε δύο βασικά μέρη, το πρώτο αφορά την κατασκευή της οντολογίας και της δόμησης του μοντέλου βάσει αυτής. Το δεύτερο αφορά την κατασκευή ενός συστήματος εφαρμογής αλγορίθμων της τεχνητής νοημοσύνης με τη δυνατότητα επεξεργασίας κειμένου φυσικής γλώσσας (*Natural Language Processing, NLP*) για την αυτόματη εξαγωγή των συγκεκριμένων ορισμών από κείμενο που δίνει ως είσοδο ο χρήστης.

1.4 Δομή διπλωματικής εργασίας

Η δομή της συνέχειας έχει ως εξής:

- Στο κεφάλαιο 2 δίνεται ο ορισμός του Λογισμικού Ανοικτού Κώδικα καθώς και ο διαχωρισμός των Αδειών Λογισμικού Ανοικτού Κώδικα στις τρεις βασικές κατηγορίες. Επίσης εξηγείται ο τρόπος που μελετήθηκαν οι άδειες για εντοπισμό των βασικών χαρακτηριστικών τους.
- Στο κεφάλαιο 3 δίνεται ο ορισμός της οντολογίας, καθώς επίσης και άλλων χαρακτηριστικών απαραίτητων για την κατασκευή του μοντέλου αδειών.
- Στο κεφάλαιο 4 περιγράφεται η διαδικασία σχεδιασμού και υλοποίησης του συστήματος μοντελοποίησης μέσω του εργαλείου για σχεδίαση οντολογιών Protégé, καθώς επίσης και η συνένωση του με SPARQL. Περιγράφεται επίσης η διαδικασία σχεδιασμού και υλοποίησης ενός plugin στο Ολοκληρωμένο Περιβάλλον Ανάπτυξης Eclipse IDE για την μοντελοποίηση κειμένων άδειας.
- Στο κεφάλαιο 5 δίνεται ο ορισμός του Machine Learning, καθώς επίσης και τις βασικές κατηγορίες του συμπεριλαμβανομένου του NLP που είναι το πεδίο της τεχνητής νοημοσύνης στο οποίο θα βασιστεί το σύστημα εξαγωγής ορισμών.
- Στο κεφάλαιο 6 περιγράφεται σε συντομία η όλη προσέγγιση για την κατασκευή του συστήματος και η προετοιμασία, όσον αφορά τα δεδομένα που χρησιμοποιήθηκαν.
- Στο κεφάλαιο 7 αναλύονται περισσότερο οι προσεγγίσεις της κατασκευής, δίνοντας τα αποτελέσματα της κάθε μίας, παρουσιάζοντας μια σύγκριση μεταξύ τους και καταλήγοντας σε αυτή που θα χρησιμοποιηθεί στο σύστημα.
- Στο κεφάλαιο 8 γίνεται ανάλυση των αποτελεσμάτων με συγκεκριμένα κείμενα αδειών και η αξιολόγηση του.
- Στο κεφάλαιο 9 περιγράφεται η συνένωση του μοντέλου βάσει οντολογίας μαζί με το σύστημα εξαγωγής των ορισμών από τα κείμενα των αδειών, και η υλοποίηση διαδικτυακής εφαρμογής που χρησιμοποιεί το σύστημα μοντελοποίησης.

- Στο κεφάλαιο 10 δίνονται κάποια συμπεράσματα, καθώς επίσης και πιθανές μελλοντικές εργασίες που μπορεί να γίνουν.

Κεφάλαιο 2

Άδειες Λογισμικού Ανοικτού Κώδικα

2.1 Λογισμικό Ανοικτού Κώδικα	13
2.2 Άδειες Λογισμικού Ανοικτού Κώδικα	14
2.3 Συμβατότητα Αδειών	16
2.4 Προηγούμενες εργασίες	16
2.4.1 Μετα-μοντελοποίηση αδειών λογισμικού	16
2.4.2 Μετα-μοντέλο αδειών λογισμικού του Alsraugh et al	18
2.4.3 Carneades και Qualipso	19
2.4.4 Συμπεράσματα	20
2.5 Μελέτη Αδειών Λογισμικού Ανοικτού Κώδικα	21

2.1 Λογισμικό Ανοικτού Κώδικα

Τα τελευταία χρόνια ο όρος Λογισμικό Ανοικτού Κώδικα γίνεται ευρέως γνωστός αφού όλο και περισσότεροι προγραμματιστές αλλά και εταιρείες παρέχουν τη δυνατότητα χρήσης του λογισμικού τους, αποφασίζοντας παράλληλα να διαθέσουν τον πηγαίο κώδικα του στο ευρύ κοινό. Το Λογισμικό Ανοικτού Κώδικα (Open Source Software - OSS) είναι το λογισμικό το οποίο μπορεί να χρησιμοποιηθεί, τροποποιηθεί και να αναδιανεμηθεί ελεύθερα κάτω από κάποιους περιορισμούς που αφορούν τα

πνευματικά του δικαιώματα (Copyrights) και τη χρήση του σαν Λογισμικό Ανοικτού Κώδικα [9].

Η ραγδαία εξέλιξη της χρήσης Λογισμικού Ανοικτού Κώδικα ίσως να οφείλεται στα σημαντικά πλεονεκτήματα που προσφέρει σε αντίθεση με το Ιδιόκτητο Λογισμικό ή το Λογισμικό Κλειστού Κώδικα (Proprietary Software). Το κυριότερο πλεονέκτημα για τους περισσότερους χρήστες είναι το γεγονός πως η χρήση Λογισμικού Ανοικτού Κώδικα γίνεται χωρίς κόστος, αφού δεν περιορίζει των αριθμών των εγκαταστάσεων και παράλληλα η λήψη κάποια άδειας είναι δωρεάν. Επιπλέον ένα από τα κύρια χαρακτηριστικά του Λογισμικού Ανοικτού Κώδικα που το κάνει τόσο ελκυστικό στο ευρύ κοινό είναι η δυνατότητα προσαρμογής του λογισμικού που παρέχεται στους χρήστες του, ώστε να ικανοποιήσουν τις υφιστάμενες ανάγκες τους, τροποποιώντας και βελτιώνοντας τον πηγαίο κώδικα του λογισμικού. Ακόμη το γεγονός πως υπάρχει υποστήριξη από μεγάλη κοινότητα χρηστών, βοηθά τους χρήστες να επιλύσουν τα όποια προβλήματα παρουσιαστούν στο λογισμικό πολύ γρηγορότερα. Όλα αυτά σε συνδυασμό με ακόμη περισσότερα οφέλη που προσφέρονται από Λογισμικά Ανοικτού Κώδικα, οδηγούν την πλειονότητα των χρηστών και εταιριών, πλέον, να προχωρήσουν σε αντικατάσταση του μέχρι πρότινος παραδοσιακού τρόπου λειτουργίας τους που ήταν η χρήση Ιδιόκτητου Λογισμικού ή Λογισμικού Κλειστού Κώδικα με Λογισμικά Ανοικτού Κώδικα.

Κάθε Λογισμικό Ανοικτού Κώδικα δημοσιεύεται με μια τουλάχιστον άδεια ανοικτού λογισμικού που βρίσκεται είτε στον πηγαίο κώδικα του λογισμικού είτε στην περιγραφή του. Οι άδειες λειτουργούν σα συμβόλαιο μεταξύ του δημιουργού (αδειοδότης) και του χρήστη του λογισμικού (αδειοδοτούμενος) και αφορούν τα δικαιώματα και τις υποχρεώσεις που θα ήθελε να παραχωρήσει ο δημιουργός (Copyright Owner) σε όσους θα χρησιμοποιήσουν το λογισμικό του.

2.2 Άδειες Λογισμικού Ανοικτού Κώδικα

Οι άδειες έχουν νομικό κείμενο για την περιγραφή τους όπου αναφέρουν τα

δικαιώματα αλλά και τις υποχρεώσεις που θα έχει κάποιος χρήστης σε περίπτωση που επιλέξει να χρησιμοποιήσει λογισμικό που συνοδεύεται από τη συγκεκριμένη άδεια. Υπάρχουν διάφορες άδειες που μπορούν να χρησιμοποιηθούν και εντάσσονται σε τρεις βασικές κατηγορίες ανάλογα με το βαθμό ελευθερίας που προσφέρουν στο χρήστη σχετικά με τη χρήση του λογισμικού. Οι τρεις βασικές κατηγορίες είναι οι εξής [10,11]:

- **Επιτρεπτικές Άδειες (Permissive Licenses):** Επιτρέπουν σε οποιοδήποτε να κάνει οτιδήποτε με τον πηγαίο κώδικα του λογισμικού δηλαδή να τον αντιγράψει, τροποποιήσει και αναδιανέμει έχοντας σαν μοναδικό ζητούμενο από τους δικαιούχους (licensees) την αναγνώριση χρήσης κώδικα ανοικτού λογισμικού και την κατάλληλη αναγνώριση στον αρχικό δημιουργό του λογισμικού. Επιπλέον η χρήση μιας τέτοιας άδειας δεν επιβάλλει κάποιο περιορισμό σχετικά με την αδειοδότηση του λογισμικού με κάποια διαφορετική άδεια, είτε την ενσωμάτωση του πηγαίου κώδικα, αυτούσιου ή τροποποιημένου, σε ένα μεγαλύτερο λογισμικό που ίσως είναι εμπορική χρήση. Αυτού του τύπου άδειες μεταξύ άλλων είναι η BSD – Berkeley Software Distribution Licenses και η MIT - Massachusetts Institute of Technology License.
- **Weak Copyleft άδειες:** Επιτρέπουν στους δικαιούχους να αντιγράψουν, να τροποποιήσουν και να αναδιανέμουν τον πηγαίο κώδικα. Σε περιπτώσεις που ο πηγαίος κώδικας χρησιμοποιείται αυτούσιος χωρίς οποιοδήποτε αλλαγές απλά μεταγλωττίζεται ή συνδέεται με ένα άλλο κομμάτι κώδικα, τότε το νέο λογισμικό που προκύπτει μπορεί να δημοσιευτεί με άδεια που χρησιμοποιείται για ιδιόκτητα λογισμικά ή οποιαδήποτε άλλη άδεια ανοικτού κώδικα. Αντιθέτως, σε περιπτώσεις όπου ο δικαιούχος προβεί σε αλλαγές του αρχικού πηγαίου κώδικα τότε αναγκαστικά το νέο λογισμικό που θα προκύψει θα πρέπει να αδειοδοτηθεί είτε με την άδεια του αρχικού πηγαίου κώδικα, είτε με αυστηρότερη. Παραδείγματα τέτοιων αδειών αποτελούν οι άδειες GNU Lesser General Public License (LGPL) σε όλες τους τις εκδόσεις.

- **Strong Copyleft άδειες:** Επιτρέπουν στους δικαιούχους να αντιγράψουν, να τροποποιήσουν και να αναδιανέμουν τον πηγαίο κώδικα του λογισμικού. Όμως ανεξαρτήτως αν θα προβούν σε αλλαγές στον πηγαίο κώδικα ή όχι, το νέο λογισμικό που θα προκύψει θα πρέπει να δημοσιευτεί με την άδεια που έχει ορίσει ο αρχικός δημιουργός του αρχικού πηγαίου κώδικα. Σκοπός τέτοιων αδειών είναι να εξασφαλίσουν πως ο πηγαίος κώδικας θα εξακολουθήσει να είναι κώδικας ανοικτού λογισμικού και προσβάσιμος από όλους. Η πιο γνωστή άδεια τέτοιου τύπου και ίσως η δημοφιλέστερη άδεια γενικότερα είναι όλες οι εκδόσεις της GNU General Public License (GPL).

2.3 Συμβατότητα Αδειών

Πολλές φορές ο δημιουργός του λογισμικού μπορεί να επιλέξει να δημοσιεύσει το λογισμικό με περισσότερες από μια άδεια. Άρα ο χρήστης του συγκεκριμένου λογισμικού έχει τα δικαιώματα και τις υποχρεώσεις που συνοδεύουν κάθε μια από τις άδειες που χρησιμοποιούνται στο λογισμικό. Για να είναι δύο ή περισσότερες άδειες συμβατές μεταξύ τους θα πρέπει τα δικαιώματα που δίνει η μια στο χρήστη να είναι επιτρεπτά και από την άλλη άδεια. Σε περιπτώσεις όπου δυο άδειες δεν είναι συμβατές μεταξύ τους, λογισμικά που χρησιμοποιούν τις συγκεκριμένες άδειες δεν μπορούν να συνδυαστούν για την ανάπτυξη ενός μεγαλύτερου λογισμικού. Είναι σημαντικό λοιπόν να λαμβάνονται υπόψη οι συμβατότητες μεταξύ αδειών από το δημιουργό κάποιου λογισμικού για να αποφεύγονται οι παραβιάσεις μεταξύ αδειών. Όμως στην παρούσα διπλωματική εργασία είναι κάτι που δυστυχώς δεν λαμβάνεται υπόψη εκτός από τις συμβατότητες που αφορούν την άδεια GPL. Ωστόσο, οι συμβατότητες μεταξύ των αδειών αποτελούν σημαντική επέκταση της παρούσας εργασίας που αναλύεται περαιτέρω σε επόμενο κεφάλαιο.

2.4 Προηγούμενες εργασίες

2.4.1 Μέτα-μοντελοποίηση αδειών λογισμικού

Σημαντικό σημείο στην διαδικασία επιλογής της ανάλογης άδειας λογισμικού για το σύστημα μας, είναι η μοντελοποίηση της κάθε πιθανής άδειας, καθώς επίσης και των μεταξύ τους συσχετίσεων. Κάποιες τέτοιες προσεγγίσεις επικεντρώνονται στις πιο διαδεδομένες και ευρέως χρησιμοποιούμενες άδειες λογισμικού, και στην αναπαράσταση αυτών και των συμβατοτήτων τους σε μορφή κατευθυνόμενου γράφου $G(V, E)$. Σε αυτόν τον γράφο, οι κόμβοι αναπαριστούν τις διάφορες άδειες λογισμικού, ενώ οι ακμές την συμβατότητα μεταξύ των παρακείμενων κόμβων [2]. Ένα παράδειγμα τέτοιας προσέγγισης είναι αυτό του γράφου συμβατότητας αδειών του Wheeler [34], ο οποίος αποτελείται από τις πιο διαδεδομένες άδειες λογισμικού, όπως είναι οι MIT, GPL και Apache-2.0. Μια πιο πρόσφατη προσέγγιση που εισάγει μια πιο ολοκληρωμένη μορφή του γράφου συμβατότητας είναι η δουλειά των Γεωργία Μ. Καπιτσάκη, Νικόλαου Δ. Τσελίκα και Ιωάννη Ε. Φουκαράκη, οι οποίοι επέκτειναν το γράφο με περισσότερες και πιο πρόσφατες άδειες. Οι άδειες αυτές βασίζονται σε ανάλυση δια χειρός και διάφορες πρόσθετες πηγές και παρέχουν βασικούς κανόνες για εξαγωγή συμπερασμάτων (reasoning) σε θέματα συμβατότητας αδειών, χωρίς να παρέχουν πληροφορίες σχετικά με συγκεκριμένα δικαιώματα και υποχρεώσεις της κάθε άδειας βάσει του κειμένου της.

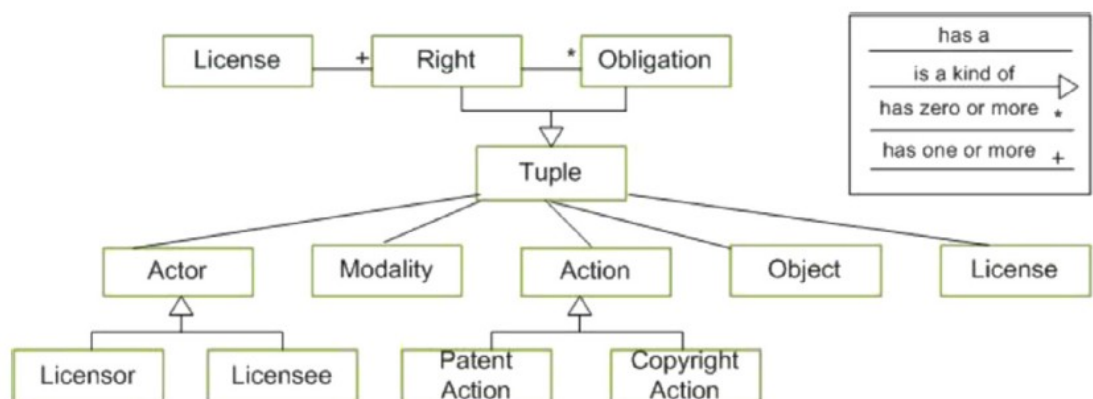
Ακόμη μία προσέγγιση μοντελοποίησης αδειών, όπως περιγράφεται στα περισσότερα από τα άρθρα [2, 4, 7, 8], είναι η σημασιολογική μοντελοποίηση τους μέσω οντολογίας, η οποία γίνεται εις βάθος των κανονισμών της κάθε άδειας και των σχέσεων μεταξύ τους. Έναντι της προηγούμενης προσέγγισης όπου γίνεται η έρευνα για τις πιθανές επιλογές αδειών του συστήματος, σε αυτήν γίνεται η ανάλυση του περιεχομένου της κάθε άδειας για την ανάδειξη των δικαιωμάτων και υποχρεώσεων της, μέσω συστημάτων επιχειρηματολογίας [7]. Η αναπαράσταση της δεν μπορεί να γίνει περιορισμένη σε ένα κόμβο, λόγω της αυξημένης πολυπλοκότητας της ανάλυσης, δηλαδή ένας κόμβος δεν μπορεί να περιέχει επαρκής πληροφορίες που να εξηγούν κάποιο συμπέρασμα σχετικά με τις πληροφορίες της συγκεκριμένης άδειας. Για αυτό το λόγο χρησιμοποιείται η σημασιολογική οντολογία ιστού, μέσω της οποίας μπορούν

να περιγραφούν τέτοιου είδους λεπτομέρειες και σχέσεις. Αυτή η προσέγγιση απαιτεί την λεπτομερή ανάλυση της έννοιας, των γεγονότων και των νομικών κανόνων μιας άδειας, τα οποία εξάγονται από το νομικό κείμενο που περιγράφει την άδεια, όπως γίνεται και στο μετά-μοντέλο αδειών λογισμικού του Alspaugh et al [4].

2.4 .2 Μετα-μοντέλο αδειών λογισμικού του Alspaugh et al

Τα περισσότερα έργα σχετικά με τη μοντελοποίηση αδειών χρησιμοποιούν το μετά-μοντέλο αδειών λογισμικού που προτάθηκε από τον Alspaugh et al [4] το 2009. Το συγκεκριμένο μετά-μοντέλο αποτελείται από στοιχεία που περιγράφουν τα δικαιώματα και τις υποχρεώσεις των αδειών, και είναι βασισμένο στην ανάλυση μερικών από τις πιο διαδεδομένες άδειες λογισμικού. Τα στοιχεία έχουν εξαχθεί, μέσα από τα κείμενα της κάθε άδειας, από τα οποία εντοπίστηκαν όλες οι σημαντικές ετυμολογίες και παρουσιάστηκαν σε μία Δήλωση Φυσικής Γλώσσας (*Natural Language Statement - RNLS*). Τα στοιχεία που απαρτίζουν το σχετικό μετά-μοντέλο, όπως φαίνονται και στο σχήμα 1, περιγράφονται από:

- 1 **Τον εμπλεκόμενο (actor):** δηλαδή αυτόν που παρέχει, ή στον οποίο παρέχεται η άδεια.
- 2 **Την τροπικότητα:** βασισμένη στις νομικές σχέσεις του Hohfeld [35], όπως το τι πρέπει (must) και τι δεν πρέπει.
- 3 **Την ενέργεια:** ενέργεια σε θέματα πνευματικής ιδιοκτησίας, όπως αντιγραφή, μετατροπή και χρήση.
- 4 **Το αντικείμενο της ενέργειας:** το οποίο μπορεί να είναι επιμέρους λογισμικό ή συγκεκριμένη πηγή μαζί με μια πιθανόν μια αναφερόμενη άδεια λογισμικού.



2.4 .3 Carneades & Qualipso

Το Carneades αποτελεί ένα σύστημα επιχειρηματολογίας [7], το οποίο έχει τη δυνατότητα κατασκευής και επικύρωσης επιχειρημάτων, καθώς επίσης την γραφική παρουσίαση τους σε μορφή γράφου. Η λειτουργία του είναι βασισμένη στο μαθηματικό μοντέλο της επιχειρηματικής φιλοσοφίας του Doug Walton [36] και είναι στενά συνδεδεμένο με την Αναιρέσιμη Λογική (*defeasible logic*). Τα επιχειρήματα στο σύστημα του Carneades κατασκευάζονται με τη χρήση επισημοποίησης των γεγονότων, εννοιών, αναιρετικών κανόνων και σχημάτων επιχειρηματολογίας. Γεγονότα και έννοιες παρουσιάζονται μέσω της Γλώσσας Οντολογίας Ιστού (*Web Ontology Language - OWL*), που αποτελεί μια μορφή σημασιολογικής οντολογίας, και οι νομικοί κανόνες παρουσιάζονται μέσω της Μορφής Συναλλαγών Νομικών Γνώσεων (*LKIF – Legal Knowledge Interface Format*).

Το Qualipso [8] με τη σειρά του αποτελεί ένα πρωτότυπο σύστημα με συμβουλευτική, ερευνητική και συγκριτική μορφή όσον αφορά νομικές θεωρίες και την ανάλυση θεμάτων συμβατότητας μεταξύ αδειών λογισμικού ανοικτού κώδικα. Δεν αποτελεί την εκτέλεση ενός στατικού αλγόριθμου, ή αυτόματου θεωρήματος, αλλά ένα εργαλείο διαπροσωπείας. Είναι βασισμένο πάνω στο προαναφερθέν σύστημα επιχειρηματολογίας του Carneades [7], και συγκεκριμένα χρησιμοποιεί τη μηχανή δημιουργίας γραφικής αναπαράστασης των επιχειρημάτων και το εργαλείο συνένωσης τους.

Τα συμπεράσματα σχετικά με την περιγραφική λογική των επιχειρημάτων είναι αυστηρά, αυτό σημαίνει ότι όσο ισχυρή και βοηθητική είναι η Γλώσσα Οντολογίας Ιστού *OWL*, δεν είναι επαρκής για να χρησιμοποιηθεί στην μοντελοποίηση νομικών κανόνων οι οποίοι είναι αναιρέσιμοι. Αυτός είναι ο κύριος λόγος που, ταυτόχρονα και το Carneades, αλλά και το Qualipso, χρησιμοποιούν την *OWL* για να καθορίζουν σύμβολα κατηγορουμένων του πεδίου της εφαρμογής και στον καθορισμό σχέσεων μεταξύ τους, και κάνουν τη χρήση της *LKIF* σαν γλώσσα αναιρέσιμων κανόνων για την μοντελοποίηση νομικών νορμών.

Κανόνες, που πιθανόν να χρησιμοποιηθούν για την παραγωγή συμπερασμάτων, πρέπει να εισαχθούν στο σύστημα σαν έννοιες και όχι σαν φόρμουλες. Αυτό συμβαίνει διότι δύσκολα μπορούμε να καλύψουμε όλες τις πιθανές περιπτώσεις κάτω από τις οποίες μπορεί να εφαρμοστεί ένας κανόνας σε μορφή φόρμουλας, και ειδικά στην περίπτωση νομικού κειμένου σαν αυτό των αδειών λογισμικού ανοικτού κώδικα.

2.4 .4 Συμπεράσματα

Βάσει της ανάλυσης των σχετικών έργων όσον αφορά τη μοντελοποίηση αδειών λογισμικού, εξάγουμε το συμπέρασμα ότι δεν αποτελεί μια εύκολη διεργασία. Για να κινηθεί κάποιος προς την μοντελοποίηση αδειών λογισμικού, δεν πρέπει να έχει απλά νομική γνώση σχετικά με κανόνες πνευματικής ιδιοκτησίας [2], αλλά να κάνει ακόμη και χρήση τεχνικών ανάλυσης φυσικής γλώσσας από το πεδίο της τεχνητής νοημοσύνης.

Όπως περιγράφηκε προηγουμένως, η πρώτη προσέγγιση μοντελοποίησης αδειών μέσω ενός κατευθυνόμενου γράφου $G(V, E)$, χρησιμοποιήθηκε στην ανάπτυξη του Εργαλείου Ελέγχου Συμβατότητας Αδειών Αρχείων (SPDX) [37]. Έγινε η δημιουργία του γράφου συμβατότητας αδειών, από τη δουλειά των Γεωργία Μ. Καπιτσάκη, Νικόλαου Δ. Τσελίκα και Ιωάννη Ε. Φουκαράκη, σαν αντικείμενο στη γλώσσα προγραμματισμού Java μέσω της βιβλιοθήκης *jgraphT*.

Το Software Package Data Exchange (SPDX) αποτελεί αρχείο που περιέχει δεδομένα σχετικά με το συγκεκριμένο λογισμικό. Μαζί με αυτά τα δεδομένα βρίσκονται πληροφορίες σχετικά με την αδειοδότηση του λογισμικού, και πάνω σε αυτά τα δεδομένα βασίστηκε το σύστημα επικύρωσης της συμβατότητας αρχείων SPDX που αναπτύξαμε. Το σύστημα λαμβάνει σαν είσοδο ένα ή περισσότερα αρχεία SPDX και παρουσιάζει τις πιθανές συγκρούσεις μεταξύ των αδειών λογισμικού τους. Εάν αυτές οι συγκρούσεις μπορούν να επιλυθούν, τότε παρουσιάζει τις πιθανές λύσεις βάσει του γράφου που αναφέρθηκε πιο πάνω και παράγει το αρχείο SPDX για το χρήστη. Εκτός

από τις υπηρεσίες ελέγχου, παρέχει και υπηρεσίες ανάλυσης συμβατότητας μεταξύ γνωστών, από το σύστημα, αδειών, καθώς επίσης και υπηρεσία επέκτασης του γράφου συμβατότητας των αδειών.

2.5 Μελέτη Αδειών Λογισμικού Ανοικτού Κώδικα

Για περαιτέρω κατανόηση των Αδειών Λογισμικού Ανοικτού Κώδικα και καλύτερο διαχωρισμό τους, μελετήθηκαν τα κείμενα τους ώστε να εντοπισθούν τα βασικά χαρακτηριστικά που παρέχει η κάθε μια από αυτές στους χρήστες μέσω του λογισμικού που συνοδεύουν [1].

Επιλέχθηκαν οι συγκεκριμένες άδειες για να χρησιμοποιηθούν ως συστάσεις διότι θεωρούνται από τις πιο διαδεδομένες και γνωστές μέσα στο ευρύ κοινό και τα περισσότερα λογισμικά ανοικτού κώδικα χρησιμοποιούν κάποια από αυτές. Τα χαρακτηριστικά αυτά χωρίστηκαν σε τρεις κατηγορίες, στα δικαιώματα, τις υποχρεώσεις αλλά και σε επιπλέον πληροφορίες. Τα δικαιώματα αναφέρονται σε *όλες τις παραχωρήσεις που δίνει ο δημιουργός του λογισμικού σε όλους τους χρήστες*, οι υποχρεώσεις με τη σειρά τους αναφέρονται σε *όλα όσα πρέπει να εφαρμόσει κάποιος χρήστης που επιλέγει να χρησιμοποιήσει οποιοδήποτε λογισμικό που συνοδεύεται από κάποια άδεια* και τέλος οι επιπλέον πληροφορίες είναι *χαρακτηριστικά που έχουν σχεδόν όλες οι άδειες και δεν εμπίπτουν ούτε στη κατηγορία δικαιώματα αλλά ούτε και στην κατηγορία υποχρεώσεις*. Κάποια από τα χαρακτηριστικά που χρησιμοποιήθηκαν υιοθετήθηκαν από παρόμοια προηγούμενη δουλειά του Thomas F. Gordon [8] και αναπτύχθηκαν σε μεγαλύτερη λεπτομέρεια ώστε να καλύψουν όλες τις άδειες και να παρουσιάζουν κάποια μοναδικά χαρακτηριστικά, είτε χαρακτηριστικά που παρέχονται από μόνο κάποιες άδειες. Τα χαρακτηριστικά που έχουν προκύψει στο σύνολο είναι περισσότερα από αυτά της προηγούμενης αυτής δουλειάς. Τα χαρακτηριστικά που υιοθετήθηκαν είναι τα ακόλουθα:

Υποχρεώσεις:

- MustOfferSourceCode: ο πηγαίος κώδικας του λογισμικού θα πρέπει να δημοσιεύεται.
- MustRetainLicenseNotice: θα πρέπει να διατηρηθεί το κείμενο της άδειας.
- NotPermissionToUseTradeMarks: δεν επιτρέπεται η χρήση εμπορικών σημάτων (TradeMarks).
- MustDistributeCopyOfNoticeText: πρέπει να διανέμεται αντίγραφο του κειμένου που δηλώνει τη συγκεκριμένη άδεια (Notice) εάν προϋπάρχει το συγκεκριμένο αρχείο.
- MustRetainNoticeInSourceDistribution: πρέπει να διατηρείται το αρχείο Notice όταν αναδιανέμεται ο πηγαίος κώδικας.

Δικαιώματα:

- MayAddDifferentLicenseTermsToYourModifications: μπορούν να προστεθούν διαφορετικοί όροι άδειας χρήσης στις τροποποιήσεις.
- MayAddYourOwnCopyrightStatementToYourModification: μπορεί να γίνει χρήση διαφορετικού κειμένου που αφορά τα πνευματικά δικαιώματα στις τροποποιήσεις.
- MayDistributeDerivativeWorksInObjectForm: οι παράγωγες δουλειές μπορούν να διανεμηθούν σε μορφή αντικειμένου.
- MayDistributeDerivativeWorksInSourceForm: οι παράγωγες δουλειές μπορούν να διανεμηθούν σε μορφή κώδικα.
- MayDistributeOriginalWorkInObjectForm: η αρχική δουλειά μπορεί να διανεμηθεί σε μορφή αντικειμένου.
- MayDistributeOriginalWorkInSourceForm: η αρχική δουλειά μπορεί να διανεμηθεί σε μορφή κώδικα.
- MaySublicense: το λογισμικό μπορεί να αδειοδοτηθεί από κάποια άλλη άδεια.
- MayPubliclyDisplay: το λογισμικό μπορεί να υπάρχει δημοσίως.

- MayPubliclyPerform: το λογισμικό μπορεί να εκτελεστεί δημοσίως.

Άλλες Πληροφορίες:

- ProvideWithoutWarranty: το λογισμικό παρέχεται χωρίς εγγυήσεις.
- LimitedLiability: ο δημιουργός του λογισμικού φέρει περιορισμένη ευθύνη.

Η συγκεκριμένη μελέτη έγινε από τη Γεωργία Χαραλάμπους κατά την ατομική διπλωματική της εργασία [1], με σκοπό να καθοριστούν τα βασικά χαρακτηριστικά της κάθε άδειας και για να βρεθούν όμοια χαρακτηριστικά μεταξύ αδειών ώστε να καθοριστούν οι ερωτήσεις που θα γίνονται στους χρήστες για εντοπισμό της καταλληλότερης άδειας που θα ικανοποιεί τις ανάγκες και ζητούμενα του χρήστη. Αυτό ήταν απαραίτητο καθώς στόχος της προαναφερόμενης διπλωματικής εργασίας ήταν η κατασκευή ενός συστήματος συστάσεων αδειών λογισμικού ανοικτού κώδικα. Παρακάτω ακολουθούν οι βασικές άδειες που χρησιμοποιήθηκαν κατά τη συγκεκριμένη ανάλυση.

#	License name	Abbreviated name	License category
1	Academic Free License version 3	AFL-3.0	Permissive
2	Apache License 2.0	Apache-2.0	Permissive
3	Artistic License 2.0	Artistic-2.0	Weak copyleft
4	BSD 2-clause "Simplified" or "FreeBSD" License	BSD-2-clause	Permissive
5	BSD 3-clause	BSD-3-clause	Permissive
6	Common Development and Distribution License	CDDL-1.0	Weak copyleft
7	Common Public License	CPL-1.0	Weak copyleft
8	Eclipse Public License	EPL-1.0	Weak copyleft
9	GNU Affero GPL version 3.0	AGPL-3.0	Strong copyleft
10	GNU General Public License version 2.0	GPL-2.0 or GPL-2.0+	Strong copyleft

11	GNU General Public License version 3.0	GPL-3.0 or GPL-3.0+	Strong copyleft
12	GNU Lesser General Public License version 2.1	LGPL-2.1 or LGPL-2.1+	Weak copyleft
13	GNU Lesser General Public License version 3.0	LGPL-3.0 or LGPL-3.0+	Weak copyleft
14	Microsoft Public License	MS-PL	Weak copyleft
15	Microsoft Reciprocal License	MS-RL	Weak copyleft
16	MIT License (X11 is treated similarly)	MIT / X11	Permissive
17	Mozilla Public License version 1.1	MPL-1.1	Weak copyleft
18	Mozilla Public License version 2.0	MPL-2.0	Weak copyleft
19	Open Software License version 3.0	OSL-3.0	Strong copyleft
20	zlib/libpng License	Zlib / Libpng	Permissive
21	Code Project Open License 1.02	CPOL-1.02	Permissive
22	Ce AChrs INRIA Logiciel Libre (<i>French</i>)	CECILL-2.0	Weak copyleft
23	Sun GPL With Classpath Exception v2.0	Sun-GPL-2.0	Strong copyleft
24	World Wide Web Consortium License	W3C	Permissive
25	The Zope Public License 2.0	ZPL-2.0	Permissive
26	International Business Machines – PowerPC and Initialization Booster Software License	IBM-pibs	Weak copyleft
27	LaTeX Project Public License 1.3c	LPPL-1.3c	Permissive
28	Code Project Open License 1.02	CPOL-1.02	Permissive

Για την υλοποίηση του μοντέλου αδειών λογισμικού ανοικτού κώδικα και το σύστημα εξαγωγής ορισμών από τα κείμενα αδειών, χρειάστηκαν περισσότερα δεδομένα. Ο λόγος είναι ότι χρειάστηκε να επικυρωθούν οι ορισμοί που εξηχθησαν, εάν πράγματι μπορούν να χαρακτηρίσουν οποιαδήποτε άδεια λογισμικού ανοικτού κώδικα. Οπότε οι επιπλέον άδειες θα βοηθούσαν σε αυτό, καθώς και στην εύρεση επιπλέον ορισμών.

Ακολουθεί ο πίνακας με τις επιπλέον άδειες που αναλύθηκαν και χρησιμοποιήθηκαν κατά τη μοντελοποίηση τους.

#	License name	Abbreviated name	License category
1	Code Project Open License 1.02	CPOL-1.02	Permissive
2	Ce ACnrs INRIA Logiciel Libre (<i>French</i>)	CECILL-2.0	Weak copyleft
3	Sun GPL With Classpath Exception v2.0	Sun-GPL-2.0	Strong copyleft
4	World Wide Web Consortium License	W3C	Permissive
5	The Zope Public License 2.0	ZPL-2.0	Permissive
6	International Business Machines – PowerPC and Initialization Booster Software License	IBM-pibs	Weak copyleft
7	LaTeX Project Public License 1.3c	LPPL-1.3c	Permissive
8	BSD Zero Clause License	0BSD	Permissive
9	Attribution Assurance License	AAL	Weak copyleft
10	Adobe Postscript AFM License	APAFML	Weak copyleft
11	Caldera Public License	Caldera	Strong copyleft
12	<i>Curl</i> and libcurl Public License	Curl	Permissive
13	The Sleepycat License	Sleepycat	Weak copyleft
14	VOSTROM Public License for OS	VOSTROM	Permissive
15	Xerox License	Xerox	Permissive
16	European Union Public License 1.1	EUPL-1.1	Weak copyleft

Κεφάλαιο 3

Οντολογία και μοντελοποίηση

3.1 Ορισμός οντολογίας	27
3.2 Μορφή οντολογίας	28
3.3 Σύνταξη οντολογίας	29
3.4 Επιλογή σύνταξης για μοντέλο	30
3.5 Σχεδίαση και υλοποίηση μοντέλου	30
3.5.1 Καθορισμός κλάσεων, σχέσεων και αντικειμένων	31
3.5.2 Εύρεση σχέσεων μεταξύ ορισμών	34
3.5.3 Παρατηρήσεις σε οντολογία του <i>findOSLicense</i>	38
3.5.4 Υλοποίηση οντολογίας στο Protégé	38
3.5.5 Συμπεράσματα	41

3.1 Ορισμός οντολογίας

Για τη μοντελοποίηση των αδειών θα χρησιμοποιηθεί η έννοια της οντολογίας, που εφαρμόζεται πάνω στα δεδομένα του Παγκόσμιου Ιστού (Web of Data) [14]. Αποτελεί μια τυπική μέθοδο καθορισμού συγκεκριμένων λεξιλογίων, τα οποία παρέχουν την δυνατότητα περιγραφής και συσχέτισης πολύπλοκων δομών ανάμεσα σε ορισμούς και μέλη που χαρακτηρίζονται από αυτούς.

Είναι μία συλλογή από τριάδες (Triples) που περιγράφουν την σχέση μεταξύ δύο αντικειμένων, κατασκευάζοντας έτσι έναν κατευθυνόμενο γράφο, τον οποίο ονομάζουμε *triplestore* [38]. Το *triplestore* αποτελεί ένα είδος βάσης δεδομένων, πάνω



Σχήμα 2: Δομή triple

στην οποία μπορούμε να εκτελέσουμε σημασιολογικά ερωτήματα. Πιο κάτω παρουσιάζεται η δομή που έχει μια τριάδα (Subject-Predicate-Object), καθώς και ένα παράδειγμα.



Σχήμα 3: Παράδειγμα triple

3.2 Μορφή οντολογίας

Υπάρχουν πολλές μορφές παρουσίασης μιας οντολογίας, με τις πιο διαδεδομένες να είναι οι Resource Description Framework Schema (*RDFS*) [15] και η προαναφερόμενη *OWL* [16]. Όπως αναφέραμε όλες οι μορφές παρουσιάζουν τα δεδομένα τους σε τριάδες. Η διαφορά μεταξύ τους βρίσκεται στην πολυπλοκότητα της περιγραφής των σχέσεων μεταξύ των δεδομένων.

Το RDFS δίνει την δυνατότητα έκφρασης σχέσεων μεταξύ αντικειμένων μέσω των τριάδων, συνοδευόμενο με ένα λεξιλόγιο, το οποίο αποτελεί λέξεις-κλειδιά με τις οποίες αυτές περιγράφονται.

Η γλώσσα OWL είναι παρόμοια αλλά δίνει περισσότερες δυνατότητες στο χρήστη. Ο χρήστης μπορεί να δώσει περισσότερες πληροφορίες και λεπτομέρεια αναφορικά με το μοντέλο των δεδομένων, και η OWL έχει συμβατότητα με συστήματα σημασιολογικών επερωτήσεων και συστημάτων αυτόματης εξαγωγής συμπερασμάτων (reasoners, π.χ. HermiT [40]).

Συνοπτικά η OWL μορφή προσφέρει μεγαλύτερο λεξιλόγιο, το οποίο διευκολύνει την έκφραση των σχέσεων στο μοντέλο των δεδομένων, καθώς και τη δυνατότητα συνδυασμού διαφορετικών οντολογιών. Για την κατασκευή του μοντέλου αδειών λογισμικού ανοικτού κώδικα, χρησιμοποιήθηκε η OWL μορφή για τους λόγους που περιγράφονται πιο πάνω.

3.3 Σύνταξη οντολογίας

Για τη σύνταξη της οντολογίας, υπάρχουν ποικίλες επιλογές για τη διαμοίραση και επεξεργασία της. Οι διάφορες συντάξεις βρίσκονται μεταξύ της επίσημης *RDF/XML* [39], η οποία υποστηρίζεται από οποιοδήποτε OWL εργαλείο, μέχρι και συντάξεις που έχουν σχεδιαστεί για συγκεκριμένους σκοπούς και εφαρμογές. Ο ακόλουθος πίνακας παρουσιάζει τις πιθανές συντάξεις με αντίστοιχο παράδειγμα της τριάδας “*Apache 2.0 license has the right of MayCopy.*”, όπως μπορεί να προκύψει κάνοντας χρήση διαφορετικών αναπαραστάσεων:

Τύπος	Σύνταξη
Φυσική γλώσσα	Apache 2.0 license has the right of MayCopy.

OWL/XML	<pre> <AnnotationAssertion> <AnnotationProperty abbreviatedIRI="ol:hasRight"/> <AbbreviatedIRI>ol:Apache</AbbreviatedIRI> <AbbreviatedIRI>ol:MayCopy</AbbreviatedIRI> </AnnotationAssertion> </pre>
OWL Functional Syntax	<pre> AnnotationAssertion(ol:hasRight ol:Apache ol:MayCopy) </pre>
RDF/XML	<pre> <rdf:Description rdf:about="http://www.cs.ucy.ac.cy/~dpasch01/license-modeling/ontologies/oss-license.owl#Apache"> <ol:hasRight rdf:resource="http://www.cs.ucy.ac.cy/~dpasch01/license-modeling/ontologies/oss-license.owl#MayCopy"/> </rdf:Description> </pre>
Turtle	<pre> ol:Apache ol:hasRight ol:MayCopy . </pre>

3.4 Επιλογή σύνταξης για το μοντέλο

Αν και το World Wide Web Consortium (W3C) προτείνει τη χρήση της RDF/XML σύνταξης, για το μοντέλο που θέλουμε να αναπτύξουμε αναζητούμε μια μορφή που να ικανοποιεί δύο απαιτήσεις, η πρώτη είναι να γίνεται εύκολα κατανοητή από τον άνθρωπο. Αναφέραμε ήδη ότι τα κείμενα των αδειών έχουν μορφή νομικών κειμένων, και αναζητούμε τρόπο μοντελοποίησης τους έτσι ώστε να διευκολύνουμε το έργο της κατανόησης τους, οπότε και η σύνταξη του μοντέλου αυτού πρέπει να είναι όσο πιο κοντά γίνεται στη φυσική γλώσσα. Η δεύτερη απαίτηση είναι να μπορεί να τύχει επεξεργασίας από υπολογιστή, ως επίσης και η συμβατότητα της με συστήματα εξαγωγής συμπερασμάτων (reasoner) και σημασιολογικών επερωτήσεων.

Βλέποντας λοιπόν τον πιο πάνω πίνακα, η πιο εύκολη και κατανοητή σύνταξη είναι η turtle. Αποτελεί μια σύνταξη πολύ πιο συνοπτική και αναγνώσιμη από τον άνθρωπο σε σχέση με τις υπόλοιπες, υποστηρίζεται ευρέως από τα πιο χρησιμοποιούμενα

εργαλεία και προγραμματιστικές διαπρωσωπείες (Application programming interfaces, APIs) όπως και Apache Jena, Fuseki και το OWL API.

3.5 Σχεδίαση και υλοποίηση μοντέλου

3.5.1 Καθορισμός κλάσεων, σχέσεων και αντικειμένων

Καθώς οι προηγούμενες εργασίες πάνω στην κατασκευή μοντέλου για τις άδειες λογισμικού ανοικτού κώδικα, χρησιμοποίησαν ήδη υπάρχοντα και πολύπλοκα λεξιλόγια. Για το μοντέλο που παρουσιάζεται στην παρούσα διπλωματική εργασία, σαν λεξιλόγιο, χρησιμοποιήθηκαν οι ορισμοί που εξήγαγε στην ατομική διπλωματική της εργασία η απόφοιτος του Τμήματος Πληροφορικής Γεωργία Χαραλάμπους [1], όπως προαναφέρθηκε.

Οι ορισμοί θα χωριστούν σε δύο βασικές κλάσεις, την *LicenseTerm*, η οποία δηλώνει τους πιθανούς τύπους των ορισμών, τις *AdditionalCondition*, *Obligation* και *Right*. Η δεύτερη βασική κλάση αποτελεί την *OpenSourceLicense*, που δηλώνει την κατηγορία της άδειας που μοντελοποιείται, αν δηλαδή είναι *Permissive*, *Weak-Copyleft* ή *Strong-Copyleft*. Οι πιθανές σχέσεις που μπορούν να υπάρξουν μεταξύ των αδειών και των ορισμών είναι οι εξής:

- **hasAdditionalCondition:** *OpenSourceLicense* × *AdditionalConditions*
- **hasObligation:** *OpenSourceLicenses* × *Obligations*
- **hasRight:** *OpenSourceLicenses* × *Rights*

Επίσης δηλώθηκαν τα αντικείμενα κάτω από τις κλάσεις *AdditionalCondition*, *Obligation* και *Right*:

- **AdditionalCondition:**
 - ProvideWithoutWarranty: το λογισμικό παρέχεται χωρίς εγγυήσεις.
 - LimitedLiability: ο δημιουργός του λογισμικού φέρει περιορισμένη ευθύνη.

- **Obligation:**
 - MustDistributeCodeUnderThisLicense: ο πηγαίος κώδικας πρέπει να δημοσιεύεται χρησιμοποιώντας την παρούσα άδεια.
 - MustMarkModifications: οι αλλαγές που γίνονται πρέπει να μαρκάρονται/σημειώνονται.
 - MustOfferSourceCode: ο πηγαίος κώδικας του λογισμικού θα πρέπει να δημοσιεύεται.
 - MustPublishSourceCodeWhenDistributedViaNetwork: ο πηγαίος κώδικας του λογισμικού θα πρέπει να δημοσιεύεται στην περίπτωση που διανέμεται μέσω δικτύου.
 - MustRetainCopyrightNotice: θα πρέπει να διατηρηθεί το κείμενο των πνευματικών δικαιωμάτων της άδειας.
 - MustRetainLicenseNotice: θα πρέπει να διατηρηθεί το κείμενο της άδειας.
 - NotSublicense: δεν επιτρέπεται η αδειοδότηση του λογισμικού με κάποια άλλη άδεια.

- **Right:**
 - MayAcceptWarrantyOrAdditionalLiability: επιτρέπεται η αποδοχή εγγύησης ή επιπρόσθετης ευθύνης.
 - MayDistributeDerivativeWorksInObjectForm: επιτρέπεται η διανομή παράγωγης δουλείας σε μορφή αντικειμένου.
 - MayDistributeDerivativeWorksInSourceForm : επιτρέπεται η διανομή της παράγωγης δουλείας σε μορφή κώδικα.
 - MayDistributeOriginalWorkInObjecForm: επιτρέπεται η διανομή της αρχικής δουλείας σε μορφή αντικειμένου (object form).

- `MayDistributeOriginalWorkInSourceForm`: επιτρέπεται η διανομή της αρχικής δουλειάς σε μορφή κώδικα.
- `MayGrantPatentsRghtsFromTheContributorToReceipt`: επιτρέπεται η παραχώρηση δικαιωμάτων πατέντας από το δημιουργό.
- `MayMakeDerivativeWorks`: επιτρέπεται η δημιουργία παράγωγων δουλειών.
- `MayMakeModificarion` : επιτρέπονται τροποποιήσεις στο λογισμικό.
- `MayPubliclyDisplay`: το λογισμικό μπορεί να υπάρχει δημοσίως.
- `MayPubliclyPerfom`: το λογισμικό μπορεί να εκτελεστεί δημοσίως.

Η δήλωση των πιο πάνω κλάσεων, σχέσεων μεταξύ τους και των αντικειμένων βασίζεται σε προηγούμενη δουλειά του *Gordon* πάνω στη μοντελοποίηση αδειών λογισμικού. Υιοθετήθηκαν οι ορισμοί που θεωρήθηκαν σημαντικοί, επεκτείνοντας έτσι το προηγούμενο μοντέλο και την οντολογία του. Να αναφερθεί πως οι όροι που έχουν υιοθετηθεί από το μοντέλο του *Gordon* εξακολουθούν να έχουν τον ίδιο ορισμό και στη δική μας μοντελοποίηση. Στο μοντέλο μας χρησιμοποιήσαμε ορισμούς από αυτό του *Gordon* ως εξής:

- `MustOfferSourceCode`, `MustLicenseDerivativesWorksUnderCombatibleLicense`, `MustMarkModifications`, `MustRetainNoticesInSourceDistributions`, `NoPermission ToUseTrademarks` και `MustDistributeCopyOfNoticeText` προστέθηκαν στην κατηγορία των υποχρεώσεων (obligations).
- `MayProvideAdditionalOrDifferentLicenseTermsToYourModifications` μετονομάστηκε σε `MayAddDifferentLicenseTermsToYourModifications`, και `MayProduceDerivativeWorks` μετονομάστηκε σε `MayMakeDerivativeWorks` και προστέθηκαν στην κατηγορία των δικαιωμάτων (rights).
- Προστέθηκαν περισσότεροι ορισμοί ως δικαιώματα στην αντίστοιχη κατηγορία: `MayDistributeDerivativeWorksInObjectForm`

MayDistributeDerivativeWorksInSourceForm,
MayDistributeOriginalWorkInObjectForm,
MayDistributeOriginalWorkInSourceForm, MaySublicense,
MayAcceptWarrantyOrAdditionalLiability, MayCopy,
MayAddYourOwnCopyrightStatementToYourModifications, MayPubliclyDisplay
και MayPubliclyPerform.

- ProvidedWithoutWarranty (μετονομασία του ProvidedWithoutWarranties) και LimitedLiability αποτελούν τους ορισμούς που προστέθηκαν στην κατηγορία των επιπλέον πληροφοριών (additional conditions).

*Παράγωγη δουλειά (derivative work): Αποτελεί έργο βασισμένο σε ήδη υπάρχουσα δουλειά, η οποία μπορεί να επαναχρησιμοποιηθεί και να υιοθετηθεί από τρίτους. Η διαφορά της παράγωγης δουλειάς από τις απλές αλλαγές (modifications) είναι ότι το πρώτο αναφέρεται στο έργο σαν σύνολο, ενώ το δεύτερο στην κάθε αλλαγή ανεξάρτητα. Σε κάθε παράγωγη δουλειά επαναχρησιμοποιούνται συστατικά από την αρχική, τα πνευματικά δικαιώματα (copyrights) όμως εφαρμόζονται πάνω στο κομμάτι που προσέχθηκε στο συγκεκριμένο έργο.

3.5.2 Εύρεση σχέσεων μεταξύ ορισμών

Μέσω της οντολογίας, μπορούμε να εκφράσουμε και διαφορετικού τύπου σχέσεις από τις πιο πάνω, όπως να δηλώσουμε ότι δύο ορισμοί είναι αντίθετοι (π.χ. κάτι τέτοιο συμβαίνει μεταξύ των όρων NotSublicense και MaySublicense), είτε δηλώνοντας τη συνεπαγωγή ενός ορισμού με έναν άλλον. Αυτό θα βοηθήσει τα διάφορα συστήματα επερωτήσεων στην απόδοσή τους, καθώς επίσης μπορεί να χρησιμοποιηθεί και σε θέματα ελέγχου συμβατότητας.

Μετά από μελέτη των ορισμών, του τί περιγράφουν και στα κομμάτια των κειμένων τα οποία αναφέρονται, προχωρήσαμε στην εξαγωγή σημαντικών σχέσεων μεταξύ τους. Βρέθηκαν πολλοί συμπερασμοί ανάμεσα τους και διασυνδέσεις τις οποίες μπορούμε

να χρησιμοποιήσουμε στην κατασκευή του μοντέλου. Πιο κάτω παρουσιάζονται οι σχέσεις για μερικούς από τους ορισμούς:

- **MustDistributeCodeUnderThisLicense:**

Η ύπαρξη αυτής της υποχρέωσης σε μια άδεια δηλώνει πως πρέπει να προσφέρεται ο πηγαίος κώδικας του έργου. Αυτό σημαίνει ότι η MustOfferSourceCode υποχρέωση πρέπει ακόμα να υπάρχει, αλλά το αντίστροφο δεν ισχύει. Αν ο ορισμός MustOfferSourceCode υπάρχει, αυτό δε σημαίνει ότι ο MustDistributeCodeUnderThisLicense πρέπει επίσης να υπάρχει απαραίτητως. Παραδείγματα αδειών είναι οι: AGPL-3.0, GPL-2.0, GPL-3.0, MS-PL (δεν έχει MustDistributeCodeUnderThisLicense, έτσι ώστε η υποχρέωση MustOfferSourceCode δεν υπάρχει), MS-RL. Ένα άλλο συμπέρασμα από αυτό το χαρακτηριστικό, είναι ότι το έργο πρέπει να αδειοδοτείται κάτω από τη συγκεκριμένη άδεια, η οποία περιορίζει τη διανομή του υπό οποιαδήποτε άλλη. Αυτό σημαίνει ότι ισχύει η υποχρέωση NotSublicense. Η ύπαρξη NotSublicense συνεπάγεται επίσης τον όρο MustDistributeCodeUnderThisLicense, αν υπάρχει δικαίωμα διανομής. Παραδείγματα αδειών είναι οι: AGPL-3.0, GPL-2.0, GPL-3.0.

- **MustRetainLicenseNotice:**

Αν υπάρχει αυτό το χαρακτηριστικό, το οποίο συνηθίζεται στις άδειες των κατηγοριών copyleft, συνεπάγεται την παρουσία της υποχρέωσης MustDistributeCodeUnderThisLicense. Κάποια παραδείγματα είναι οι άδειες CDDL-1.0, GPL-1.0, GPL-2.0, GPL-3.0, EPL-1.0.

- **NotPermissionToUseTradeMarks:**

Αν υπάρχει αυτό το χαρακτηριστικό για τη χρήση εμπορικών σημάτων, συνεπάγεται ότι ο δικαιοδόχος ενδιαφέρεται για τα πνευματικά δικαιώματα που έχει πάνω στο έργο. Αυτό δηλώνει την ύπαρξη της υποχρέωσης MustRetainCopyrightNotice.

- **MustLicenseDerivativesWorksUnderCompatibleLicense:**

Για να υπάρχει η συγκεκριμένη υποχρέωση συνεπάγεται ότι δεν υπάρχει η υποχρέωση του NotSublicense. Ακόμη, με την ύπαρξη αυτού του ορισμού δηλώνεται έμμεσα και το δικαίωμα της δημιουργίας παράγωγης εργασίας από το αρχικό, καθώς επίσης και το δικαίωμα MaySublicense.

- **MustRetainCopyrightNotice:**

Η ύπαρξη της συγκεκριμένης υποχρέωσης δηλώνει ότι ο αδειοδότης θεωρεί σημαντική την κατοχή των πνευματικών δικαιωμάτων στο αρχικό έργο. Έχοντας αυτό υπόψη, και θέλοντας να τα προστατέψει, εάν η άδεια δίνει το δικαίωμα για MayMakeModifications τότε συνεπάγεται και η ύπαρξη του MustMarkModification.

- **MustMarkModifications:**

Η ύπαρξη της συγκεκριμένης υποχρέωσης δηλώνει ότι δίνεται το δικαίωμα στον χρήστη για MayMakeModifications και MayMakeDerivativeWorks, αλλά όχι το αντίστροφο. Έστω και ένα από τα δύο να υπάρχει, δεν σημαίνει ότι υπάρχει και το MustMarkModifications.

- **MustRenameModifiedVersion:**

Από την ύπαρξη της συγκεκριμένης υποχρέωσης φτάνουμε στο συμπέρασμα ότι δίνεται το δικαίωμα του MayMakeModification, που αυτό με τη σειρά του συνεπάγεται με την ύπαρξη του MayMakeDerivativeWork.

- **MaySublicense:**

Η ύπαρξη αυτού του δικαιώματος αντιστοιχεί με την ελευθερία του χρήστη να διανέμει το έργο του υπό τους όρους ή την άδεια που ο ίδιος επιθυμεί. Αυτό αντικρούει τις υποχρεώσεις MustDistributeCodeUnderThisLicense και NotSublicense.

- **MayGrantPatentsRightsFromTheContributorToReceipt:**

Εάν δίνεται το δικαίωμα στον χρήστη να ζητήσει από το δικαιούχο τα πνευματικά δικαιώματα για την πατέντα, αυτό συνεπάγεται ότι δεν τα έχει εξαρχής. Οπότε η

ύπαρξη του συγκεκριμένου δικαιώματος αντιστοιχεί με την ύπαρξη της υποχρέωσης NotPermissionToUseTrademarks και MustRetainCopyrightNotice.

- **MayAddYourOwnCopyrightStatementToYourModifications:**

Το δικαίωμα αυτό δίνει την ελευθερία στο χρήστη να προσθέσει δικούς του όρους όσον αφορά την διανομή του έργου του. Αυτό σημαίνει ότι μπορεί να το θέσει κάτω από μία άδεια παράγωγη της αρχικής. Το πιο πάνω δικαίωμα λοιπόν συνεπάγεται με αυτό του MaySublicense καθώς επίσης και του MayMakeDerivativeWork.

Πιο κάτω ακολουθούν κάποιες από αυτές τις σχέσεις διατυπωμένες σε δεοντική λογική (deontic logic), και συγκεκριμένα για το SPINdle reasoner [42]:

Κανонаς δεοντικής λογικής (deontic logic rule)	Επεξήγηση
MustDistributeCodeUnderThisLicense ->MustOfferSourceCode	Η ύπαρξη του MustDistributeCodeUnderThisLicense συνεπάγεται την ύπαρξη του MustOfferSourceCode.
MustDistributeCodeUnderThisLicense →NotSublicense	Η ύπαρξη του MustDistributeCodeUnderThisLicense συνεπάγεται την ύπαρξη του NotSublicense.
NotPermissionToUseTradeMarks ->MustRetainCopyrightNotice	Η ύπαρξη του NotPermissionToUseTradeMarks συνεπάγεται την ύπαρξη του MustRetainCopyrightNotice.
MustRenameModifiedVersion ->MayMakeModifications	Η ύπαρξη του MustRenameModifiedVersion συνεπάγεται την ύπαρξη του MayMakeModifications.
MaySublicense →-MustDistributeCodeUnderThisLicense	Η ύπαρξη του MaySublicense συνεπάγεται ότι το MustDistributeCodeUnderThisLicense δεν υπάρχει.
MustDistributeCopiesOfOriginalWorkOrInstructionHowToGetThem ->MayCopy	Η ύπαρξη του MustDistributeCopiesOfOriginalWorkOrInstructionHowToGetThem συνεπάγεται την

	ύπαρξη του MayCopy.
MustRenameModifiedVersion ->MayMakeModifications	Η ύπαρξη του MustRenameModifiedVersion συνεπάγεται την ύπαρξη του MayMakeModifications.
MustRenameModifiedVersion ->MayMakeDerivativeWorks	Η ύπαρξη του MustRenameModifiedVersion συνεπάγεται την ύπαρξη του MayMakeDerivativeWorks.

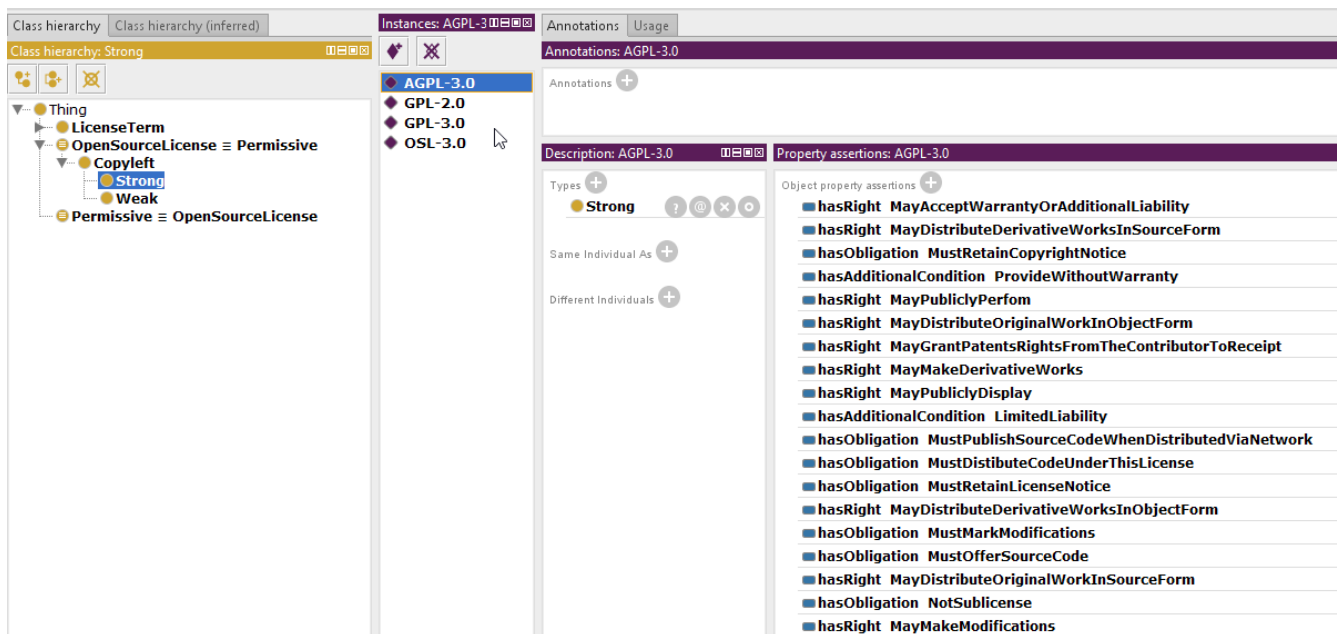
3.5.3 Παρατηρήσεις σε οντολογία του findOSSLicense

Η αρχική οντολογία που δημιουργήθηκε για την παρουσίαση του συστήματος συστάσεων αδειών findOSSLicense [41], είχε λανθασμένη δομή σε σχέση με τα πρότυπα του σημασιολογικού ιστού (semantic web standards). Οι ορισμοί, καθώς επίσης και οι άδειες θα έπρεπε να είναι περασμένα σαν ατομικές δομές (individuals) αντί κλάσεις. Ο κύριος λόγος που αυτό πρέπει να αλλάξει είναι για να μπορούμε να εκτελέσουμε σημασιολογικά επρωτήματα πάνω στην οντολογία μέσω του SPARQL. Το SPARQL (SPARQL Protocol and RDF Query Language) αποτελεί μια γλώσσα επρωτήσεων πάνω σε RDF δεδομένα, δηλαδή μπορεί να εκτελέσει σημασιολογικά επρωτήματα πάνω σε σχεσιακές βάσεις δεδομένων, όπως μια οντολογία, και να εξαγει ή να αλλάξει δεδομένα.

Οπότε υλοποιήθηκε η οντολογία μέσω του εργαλείου Protégé, το οποίο αποτελεί ένα λογισμικό ανοικτού κώδικα για την επεξεργασία οντολογίας και για την απόκτηση γνώσης.

3.5.4 Υλοποίηση οντολογίας στο Protégé

Μέσω λοιπόν του Protégé υλοποιήθηκε η οντολογία για τη μοντελοποίηση αδειών λογισμικού ανοικτού κώδικα, με τις κλάσεις και τα individuals όπως τα αναφέραμε πιο πάνω. Εφαρμόζοντας επίσης τις αλλαγές στην οντολογία του findOSSLicense, πλέον μπορούν να εκτελεστούν σημασιολογικές επρωτήσεις μέσω SPARQL.



Σχήμα 4: Οντολογία στο Protege

Ακολουθεί ένα SPARQL παράδειγμα για την εύρεση αδειών που είναι επιτρεπτικές (permissive):

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX ol: <http://www.cs.ucy.ac.cy/~dpasch01/license-modeling/oss-license.owl#>

SELECT ?license WHERE { ?license rdf:type ol:Permissive }
```

Η ιδιότητα `rdf:type` δηλώνει το ότι ψάχνει αντικείμενα με το συγκεκριμένο τύπο, στην περίπτωση μας `ol:Permissive`. Ακολουθεί το αποτέλεσμα του ερωτήματος πάνω στη βάση δεδομένων της οντολογίας:

```
-----
| license |
=====
| ol:BSD3-clause |
| ol:X11 |
| ol:AFL-3.0 |
| ol:Libpng |
| ol:Zlib |
| ol:MIT |
| ol:Apache-2.0 |
| ol:BSD2-clause |
| ol:ARCH-3.0 |
-----
```

Σχήμα 5: Αποτελέσματα SPARQL ερωτήματος

Όπως φαίνεται το πιο πάνω επερώτημα ζητάει όλες τις πιθανές άδειες λογισμικού που είναι τύπου ol:Permissive.

Έχοντας υλοποιήσει την οντολογία για τη μοντελοποίηση των αδειών λογισμικού ανοικτού κώδικα, έγινε μετατροπή δια χειρός, των αδειών που αναλύθηκαν στη διπλωματική εργασία της Γεωργίας Χαραλάμπους, στην τελική μορφή του μοντέλου σε σύνταξη OWL turtle. Ακολουθεί ένα παράδειγμα μοντελοποίησης της άδειας Apache 2.0:

```
@prefix owl:<http://www.w3.org/2002/07/owl#> .
@prefix rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix ol:<http://www.cs.ucy.ac.cy/~dpasch01/license-modeling/oss-license.owl#>
ol:Apache-2.0
    rdf:type
        ol:Permissive,
        owl:NamedIndividual;
    ol:hasAdditionalCondition
        ol:LimitedLiability,
        ol:ProvideWithoutWarranty ;
    ol:hasRight
        ol:MayAcceptWarrantyOrAdditionalLiability ,
        ol:MayAddDifferentLicenseTermsToYourModifications ,
        ol:MayAddYourOwnCopyrightStatementToYourModifications,
        ol:MayCopy,
        ol:MayDistributeDerivativeWorksInObjectForm,
        ol:MayDistributeDerivativeWorksInSourceForm,
        ol:MayDistributeOriginalWorkInObjectForm,
        ol:MayDistributeOriginalWorkInSourceForm,
        ol:MayGrantPatentsRightsFromTheContributorToReceipt,
        ol:MayMakeDerivativeWorks,
```

ol:MayPubliclyDisplay,

ol:MayPubliclyPerfom,

ol:MaySublicense,

ol:MayUseInHouse ;

ol:hasObligation

ol:MustDistributeCopyOfNoticeText,

ol:MustLicenseDerivativesWorksUnderCombatibleLicense,

ol:MustMarkModifications,

ol:MustRetainInSourceCodeOfDerivativeWorkAllTradeMarksPatentCopyrightsOfOriginalWork,

ol:MustRetainLicenseNotice,

ol:NotPermissionToUseTradeMarks .

3.5.5 Συμπεράσματα

Μέσω αυτής της δομής μοντέλου είναι δυνατή η δημιουργία πλατφόρμας για δημιουργία μοντέλων αδειών. Αυτή η δομή θα χρησιμοποιηθεί σε συνδυασμό με Apache Jena Fuseki, το οποίο αποτελεί υπηρεσία εκτέλεσης SPARQL σημασιολογικών ερωτήσεων για αλλαγή των δεδομένων, για την κατασκευή ενός plugin για το Eclipse IDE – Integrated Development Environment, μέσω του οποίου θα κατασκευάζονται τα μοντέλα των αδειών.

Κεφάλαιο 4

Σχεδιασμός και υλοποίηση Eclipse IDE Plugin

4.1 Σχεδιασμός plugin	42
4.1.1 Τύπος plugin	43
4.1.2 Οπτική παρουσίαση μοντέλου	44
4.1.2.1 Κατηγορίες τεχνικών παρουσίασης	44
4.1.2.2 Plugin παρουσίασης οντολογιών	46
4.1.2.3 Συμπέρασμα	48
4.2 Υλοποίηση plugin	48
4.2.1 Eclipse Plugin	49
4.2.2 Eclipse Editor Plugin	50
4.2.3 Δομή plugin	51
4.2.4 Εξωτερικές βιβλιοθήκες	54
4.3 Έκδοση plugin	56
4.4 Συμπεράσματα	56

4.1 Σχεδιασμός plugin

Στόχος είναι να δημιουργηθεί ένας επεξεργαστής κειμένου ο οποίος να διαχειρίζεται αρχεία με μοντέλα άδειας ανοικτού κώδικα, με όσο πιο απλό και εύκολο τρόπο γίνεται. Για να επιτευχθεί αυτό πρέπει να ληφθούν υπόψη οι πιθανοί χρήστες του συστήματος, και να γίνει προσπάθεια για ικανοποίηση αυτών των ομάδων.

Μετά από έρευνα καταλήξαμε σε τρεις ομάδες χρηστών:

- **Χρήστες που χρησιμοποιούν κώδικα:**

Αναφέρονται στους χρήστες που τους αρέσει να έχουν τον έλεγχο όσον αφορά το μοντέλο, καθώς επίσης να ελέγχουν την έξοδο του αρχείου που περιέχει το μοντέλο. Σε αυτούς τους χρήστες θα πρέπει να δίνεται η δυνατότητα να αλλάζουν και να

μορφοποιήσουν το μοντέλο μέσα από το ίδιο το αρχείο.

- **Χρήστες που προσθέτουν μέσω φόρμας:**

Οι χρήστες αυτοί δεν έχουν ανάγκη από το να δουν τον πηγαίο κώδικα, και προτιμούν να συμπληρώνουν τα στοιχεία του μοντέλου μέσω κάποιας φόρμας (wizard), μέσω της οποίας να δημιουργείται το μοντέλο.

- **Χρήστες που είναι περισσότερο οπτικοί:**

Αναφέρονται στους χρήστες που δεν έχουν ανάγκη από το να δουν τον πηγαίο κώδικα του μοντέλου, αλλά επίσης θεωρούν κουραστικό το να συμπληρώνουν φόρμες με τα στοιχεία του μοντέλου. Για αυτούς τους χρήστες πρέπει να υλοποιηθεί ένα πιο οπτικό και διαδραστικό σύστημα για να ικανοποιηθούν οι ανάγκες τους.

4.1.1 Τύπος plugin

Βάσει των ομάδων χρηστών που αναφέρθηκαν πιο πάνω, καθώς επίσης και για το τί πρέπει να πληροί το plugin, καταλήξαμε ότι ο τύπος του πρέπει να είναι editor plugin. Ένα editor plugin χρησιμοποιείται για την διατήρηση των δεδομένων, μέσω των views. Συνήθως απαιτεί από το χρήστη να επιλέξει ρητά να αποθηκεύσει τα δεδομένα του έτσι ώστε να εφαρμοστούν οι αλλαγές στα αρχεία, και η όψη (view) αλλάζει αμέσως μόλις αλλάξουν τα δεδομένα αυτά.

Θα χρησιμοποιηθούν λοιπόν τρεις διαφορετικές όψεις, ένα για την κάθε ομάδα χρηστών. Θα υλοποιηθεί επίσης ένα είδος αρχείου που θα σχετίζεται με το συγκεκριμένο editor και το μοντέλο άδειας ανοικτού κώδικα που αναπτύξαμε.

4.1.2 Οπτική παρουσίαση μοντέλου

Για την διαδραστική επαφή της οντολογίας, που απευθύνεται στους περισσότερο οπτικούς χρήστες, πρέπει να γίνει επιλογή του τύπου εμφάνισης. Η οπτική

παρουσίαση της οντολογίας αποτελεί σημαντικό παράγοντα στο σχεδιασμό, στη διαχείριση και την περιήγηση της, οπότε για την επιλογή του τύπου παρουσίασης πρέπει να ληφθεί υπόψη η επιθυμητή απόδοση του χρήστη στις πιο πάνω λειτουργίες. Σε αυτό το κομμάτι θα καλυφθούν οι πιθανές επιλογές για παρουσίαση της οντολογίας, καθώς και στην τελική επιλογή που θα συμπεριληφθεί στο plugin.

Υπάρχει μια ποικιλία από εργαλεία για τη διαχείριση και σχεδίαση οντολογιών, όπως τα Protege, KAON2, ontopia κτλ. Υπάρχουν επίσης εργαλεία συγκεκριμένα για την παρουσίαση των οντολογιών, όπως είναι κάποια Protege plugins, το Tulip και το ZGRViewer. Τα περισσότερα εργαλεία οντολογίας, όπως το Protege, προσφέρουν περισσότερα, όσον αφορά την οπτική παρουσίαση των οντολογιών, με τα διάφορα σχετικά plugins.

4.1.2.1 Κατηγορίες τεχνικών παρουσίασης

Υπάρχουν πέντε βασικές κατηγορίες τεχνικών οπτικής παρουσίασης οντολογίας [17]. Η κατηγοριοποίηση αυτών των τεχνικών βασίζεται πάνω στα χαρακτηριστικά της οντολογίας, τα οποία επικεντρώνουν κατά την παρουσίαση, διαδραστικότητα και λειτουργικότητα που υποστηρίζουν. Κάποιες από αυτές περιέχουν ίχνη από άλλες αλλά χωρίστηκαν βάσει τις κυρίαρχες ιδιότητες τους. Ακολουθούν αναλυτικά μερικές από τις κατηγορίες τεχνικών οπτικής παρουσίασης των οντολογιών:

- **Intended lists:**

Στη συγκεκριμένη κατηγορία, η ταξινόμια της οντολογίας παρουσιάζεται σε μορφή δέντρου. Όπως και με το πρόγραμμα περιήγησης κλάσεων στο Protege. Τα θετικά της συγκεκριμένης κατηγορίας είναι η απλότητα όσον αφορά την υλοποίηση και την παρουσίαση της, καθώς επίσης και η οικειότητα των χρηστών, παρόμοιο με το Windows Explorer. Ένα πρόβλημα που αντιμετωπίζει είναι ότι παρουσιάζει την οντολογία σε μορφή δέντρου και όχι γράφου, οπότε παρουσιάζει σχέσεις κληρονομικότητας παρά σχέσεις ρόλων στην οντολογία.

- **Node-link & tree:**

Εδώ οι οντολογίες παρουσιάζονται σαν ένα σύνολο από διασυνδεδεμένους κόμβους με διάταξη από πάνω προς τα κάτω, είτε από δεξιά προς τα αριστερά. Ο χρήστης μπορεί να αλληλεπιδρά με τους κόμβους και τα υποδένδρα της οντολογίας.

- **Zoom-able:**

Η συγκεκριμένη κατηγορία αντιπροσωπεύει τους κόμβους στο χαμηλότερο επίπεδο της ιεραρχίας. Δίνουν τη δυνατότητα στον χρήστη να μεγενθύνει είτε να σμικρύνει τους κόμβους παιδιά, μετατρέποντας τα στο επίπεδο της ιεραρχίας στο οποίο βρίσκονται. Είναι αποδοτικές τεχνικές για την εύρεση συγκεκριμένων κόμβων, εφόσον παρέχουν μια περιεκτική παρουσίαση της ιεραρχίας στην οποία βρίσκεται ο χρήστης.

- **Space-filling:**

Οι συγκεκριμένες τεχνικές βασίζονται στη χρήση ολόκληρης της οθόνης κατά την παρουσίαση. Τέτοιες τεχνικές καταφέρνουν να παρουσιάσουν δέντρα με ιδιότητες (Properties) στους αντίστοιχους κόμβους τους. Δίνουν τη δυνατότητα εφαρμογής χρώματος και αλλαγής μεγέθους των ιδιοτήτων, έτσι ώστε να είναι πιο αποδοτική η αλληλεπίδραση των χρηστών, και η κατανόηση τους για το πού βρίσκονται.

- **Focus and complex or distortion:**

Γίνεται παρουσία της οντολογίας σε μορφή γράφου. Ο κόμβος στον οποίο συγκεντρώνεται ο χρήστης, αποτελεί το κεντρικό κόμβο, ο οποίος μεγαλώνει παρουσιάζοντας τους κόμβους γείτονες του. Ο χρήστης πρέπει να επιλέξει συγκεκριμένο κόμβο έτσι ώστε να τον μεγαλώσει και να εμφανιστούν οι γειτονικοί κόμβοι. Ένα μειονέκτημα είναι ότι τέτοιου είδους τεχνικές δεν διατηρούν σταθερές θέσεις των κόμβων, με αποτέλεσμα ο χρήστης μερικές φορές να αποπροσανατολίζεται.

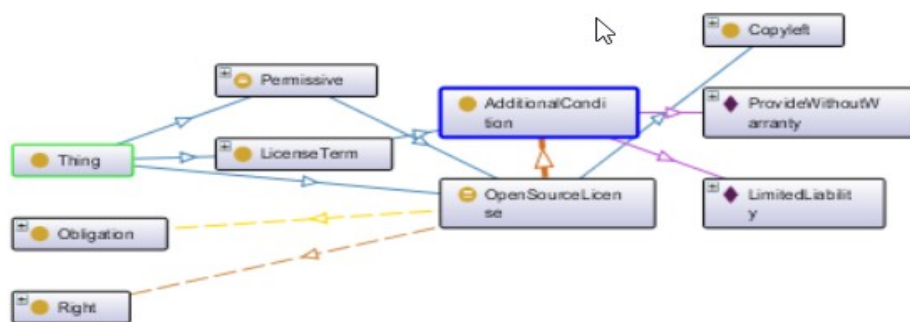
Οι πιο πάνω κατηγορίες αφορούν δυσδιάστατες παρουσιάσεις της οντολογίας. Υπάρχουν ξεχωριστές κατηγορίες τεχνικών για παρουσίαση οντολογιών σε τρισδιάστατη μορφή (π.χ. τοπία), οι οποίες όμως δεν μπορούν να εφαρμοστούν στην περίπτωση μας, οπότε δεν θα τις καλύψουμε.

4.1.2.2 Plugin παρουσίασης οντολογιών

Βάσει των κατηγοριών που αναλύθηκαν πιο πάνω, έγινε σχετική έρευνα για τεχνικές μέσω plugins για το Protege, οι οποίες μπορούν πιθανόν να εφαρμοστούν κατά την υλοποίηση του Eclipse plugin. Ακολουθούν οι αναλύσεις μερικών από αυτών:

- **Ontograf:**

Μπορεί να μετατρέψει την όψη του γράφου σε διαφορετικού τύπου, όπως grid, tree, οριζόντιο ή κατακόρυφο. Ο χρήστης μπορεί επίσης να επιλέξει το είδος των κόμβων και των arc που να παρουσιάζονται. Το συγκεκριμένο plugin δίνει τη δυνατότητα στο χρήστη να εκτελέσει εντολές αναζήτησης καθώς επίσης και την εξαγωγή του γράφου σε εικόνα και αρχείο τύπου DOT.

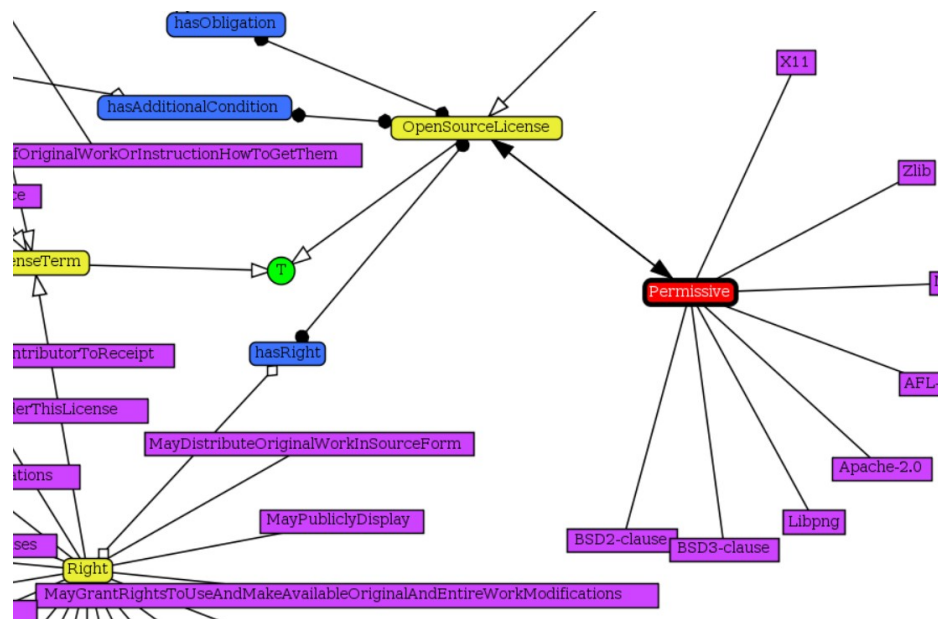


Σχήμα 6: Ontograf visualization of the Ontology.

- **SOVA:**

Αντιστοιχεί σε Simple Ontology Visualization API, και παρουσιάζει όλα τα στοιχεία της οντολογίας, όπως κλάσεις, individuals, ιδιότητες καθώς και τις σχέσεις μεταξύ τους.

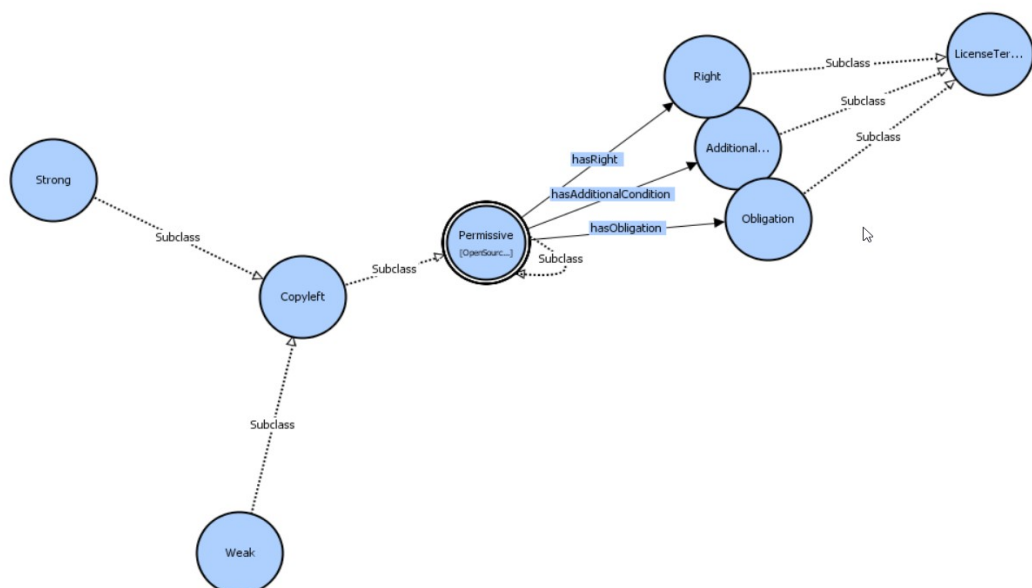
Μπορεί επίσης να εκτελέσει επερωτήσεις αναζήτησης μέσω λέξεων κλειδιών και να ρυθμιστεί έτσι ώστε να φιλτράρει τις πληροφορίες οι οποίες παρουσιάζονται.



Σχήμα 7: SOVA Visualization of the Ontology.

- VOWL:

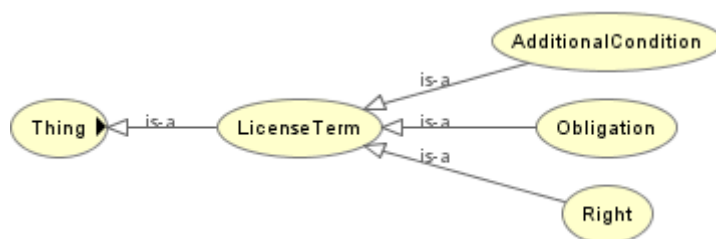
Παρέχει γραφικές απεικονίσεις των στοιχείων που αποτελούν την οντολογία, σε μορφή κατευθυνόμενου γράφου. Παρουσιάζει μόνο τις κλάσεις της οντολογίας και την σχέση κληρονομικότητας μεταξύ τους.



Σχήμα 8: VOWL Visualization of the Ontology.

- OWLViz:

Βασίζεται πάνω στο λογισμικό graphviz dot, το οποίο αποτελεί απαιτούμενο σύστημα για τη λειτουργία του. Δίνει πολλαπλές επιλογές όσον αφορά την παρουσίαση της κληρονομικότητας των διαφόρων κλάσεων της οντολογίας, όπως για παράδειγμα την



Σχήμα 9: OWLViz visualization of the Ontology.

απόκρυψη των κλάσεων παιδιών, είτε την δήλωση της ακτίνας μέσα στην οποία ο χρήστης μπορεί να εξερευνήσει.

4.1.2.3 Συμπέρασμα

Μία οντολογία μπορεί να παρουσιαστεί οπτικά σε μορφή είτε δέντρου είτε γράφου, και μπορεί είτε να παρουσιάζει τις κλάσεις και την κληρονομικότητα τους, είτε τα *individuals* και τις σχέσεις μεταξύ τους. Σχετικά με τη σχεδίαση του Eclipse plugin, καταλήξαμε στη χρήση παρουσίασης σε μορφή γράφου, των *individuals* και τις σχέσεις μεταξύ τους. Η επιλογή αυτή βασίζεται πάνω σε έρευνα που έτυχε για το ποιές μεθόδους οι χρήστες βρίσκουν αποδοτικές και ποιες ενοχλητικές και δύσκολες στη χρήση [18].

4.2 Υλοποίηση plugin

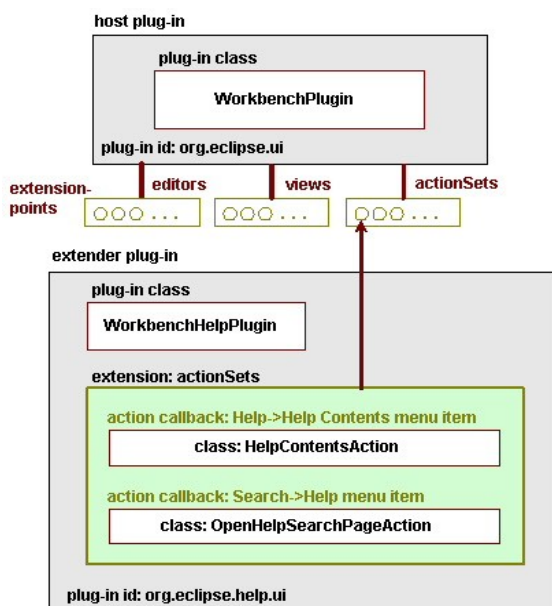
Όσον αφορά την υλοποίηση, καθοδηγήθηκε βάσει τις εντολές από την Eclipse σχετικά με την ανάπτυξη plugin, και σταδιακά μεταφέρθηκε πιο συγκεκριμένα στην ανάπτυξη για editor plugins για το Eclipse IDE. Σε αυτό το κεφάλαιο αναλύεται η κύρια αρχιτεκτονική τέτοιου plugin, οι βασικές κλάσεις για την ανάπτυξη του, καθώς και η

υλοποίηση του Eclipse editor plugin για το μοντέλο αδειών λογισμικού ανοικτού κώδικα.

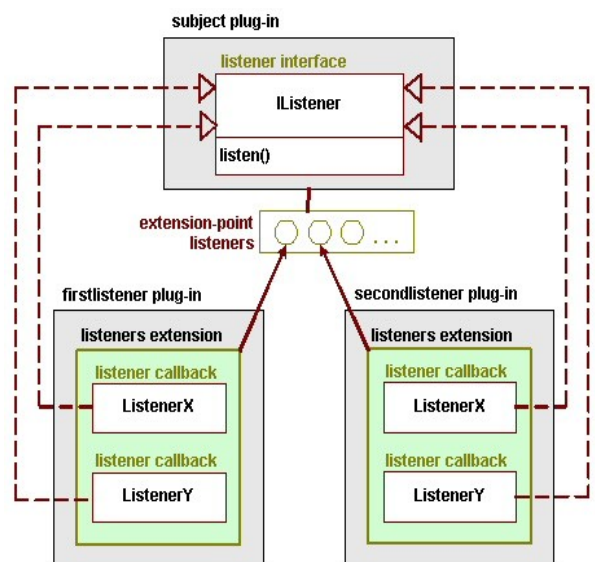
4.2.1 Eclipse Plugin

Ένα plugin μέσα στο Eclipse αποτελεί ένα συστατικό στοιχείο το οποίο παρέχει κάποιου συγκεκριμένου τύπου υπηρεσία μέσα από το Eclipse workbench. Με τον όρο συστατικό στοιχείο εννοούμε ένα αντικείμενο το οποίο μπορεί να ρυθμιστεί σε ένα σύστημα κατά την ανάπτυξη του. Το περιβάλλον εκτέλεσης του Eclipse παρέχει μια υποδομή για την υποστήριξη της λειτουργικότητας ενός συνόλου από plugins μέσα στο Eclipse workbench σαν ένα ενιαίο σύστημα χωρίς ο χρήστης να καταλάβει την ανεξαρτησία τους.

Μια plugin κλάση μέσα στο Eclipse πρέπει να επεκτείνει (extend) την κλάση `org.eclipse.core.runtime.Plugin`, η οποία αποτελεί μία αφηρημένη (abstract) κλάση, που σκοπό έχει την παροχή γενικών λειτουργιών για την διαχείριση τους.



Σχήμα 10: Editor Architectures ontop of the Workbench.



Σχήμα 11: Event Listeners for Editor changes.

4.2.2 Eclipse Editor Plugin

Το Eclipse χρησιμοποιεί τους επεξεργαστές (editors) και τις όψεις (views) για την διατήρηση των δεδομένων του χρήστη. Ένας editor απαιτεί από τον χρήστη να αποθηκεύσει (save) τις αλλαγές που έκανε έτσι ώστε να φυλάξει τα καινούργια δεδομένα στο αρχείο, καθώς μέσω της όψης γίνονται οι συγκεκριμένες αλλαγές και παρουσιάζονται.

Ακολουθούν τα βασικά βήματα για τη δημιουργία ενός editor plugin στο Eclipse:

- Δημιουργία και εμφάνιση του editor area στο perspective του χρήστη.
- Δημιουργία μιας IEditorInput κλάσης.
- Καθορισμός του extension "org.eclipse.ui.editors".
- Υλοποίηση της κλάσης του editor, η οποία πρέπει να υλοποιεί το IEditorPart interface.

Βασικές κλάσεις για την ανάπτυξη του editor plugin στο Eclipse αποτελούν οι παρακάτω:

- **IEditorInput:**

IEditorInput αποτελεί ένα μοντέλο για τον editor και χαρακτηρίζεται από την ανάλαφρη, για το Eclipse, παρουσίαση του. Το Eclipse θα κάνει buffer τα αντικείμενα του τύπου IEditorInput οπότε δεν θα βαρυφορτωθεί για την παρουσίαση του συγκεκριμένου. Για παράδειγμα το Eclipse IDE χρησιμοποιεί αντικείμενα IEditorInput για την αναγνώριση αρχείων, χωρίς όμως να διαχειρίζεται ολόκληρα τα αρχεία.

- **IEditorPart and EditorPart:**

Ο editor καθορίζεται από το extension point "org.eclipse.ui.editors". Η κλάση η οποία υλοποιεί τον editor πρέπει επίσης να υλοποιεί και το interface του "IEditorPart". Στις περισσότερες περιπτώσεις απλά κάνει extend την abstract κλάση του "EditorPart". Ο editor λαμβάνει το IEditorSite και το IEditorInput μέσα στην init() μέθοδο. Η είσοδος

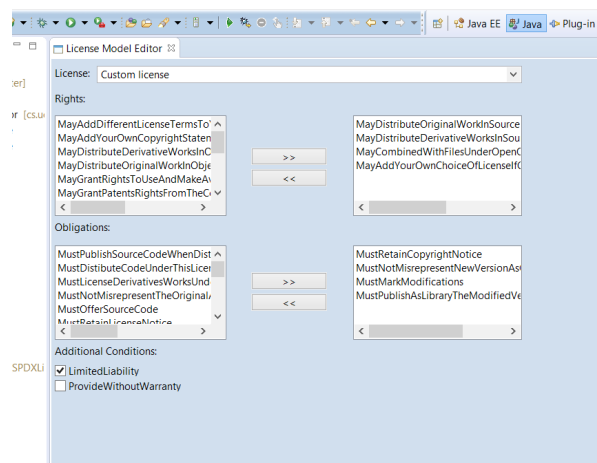
πρέπει να καθοριστεί μέσω της μεθόδου `setInput()` και το site μέσω της μεθόδου `setSite()`.

Για το συγκεκριμένο plugin θα πρέπει να δημιουργηθούν editors αντίστοιχα με τα views που έχουμε, εφόσον οι χρήστες θα μπορούν να αλλάζουν τα δεδομένα μέσα από κάθε μία.

4.2.3 Δομή plugin

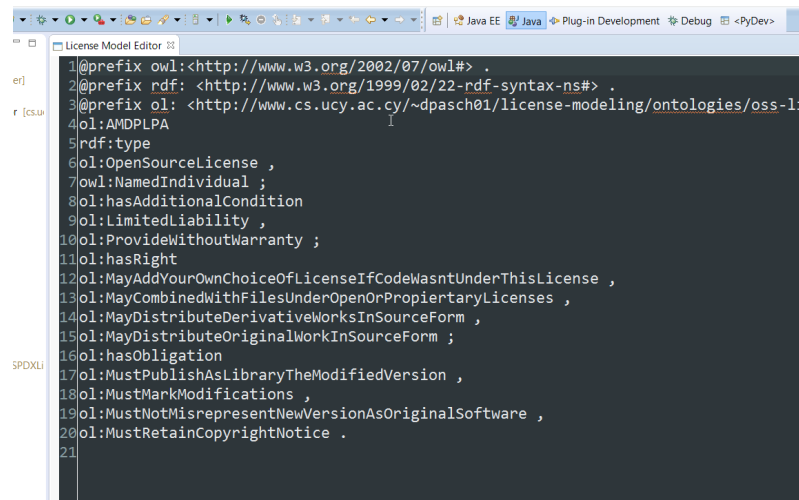
Το plugin αποτελεί, όπως αναφέραμε, ένα editor plugin και με αυτό τον τρόπο θα παρουσιάζεται στον χρήστη. Για την υλοποίηση του όμως, και επειδή θέλουμε να παρέχουμε στο χρήστη τη δυνατότητα να διαχειρίζεται τα δεδομένα του και από τις τρεις διαφορετικές όψεις, αποτελείται από τρεις διαφορετικούς editors.

Ο πρώτος αναφέρεται στις φόρμες συμπλήρωσης σχετικά με τους όρους του συγκεκριμένου μοντέλου. Διαχωρίζεται σε λίστες για δικαιώματα, υποχρεώσεις και επιπρόσθετους ορισμούς, όπου ο χρήστης μπορεί να προσθέσει είτε να αφαιρέσει από αυτούς.



Σχήμα 12: License Model Editor Preferences View.

Ο δεύτερος αναφέρεται στον πηγαίο κώδικα του μοντέλου. Αποτελεί ένα απλό επεξεργαστή κειμένου (text editor), ο οποίος μετατράπηκε έτσι ώστε να καλύπτει δεδομένα τύπου rdf. Υπάρχει δηλαδή περιορισμός στα δεδομένα που εισάγει ο χρήστης, τα οποία ελέγχονται με το που θα εκτελεί την εντολή αποθήκευσης save, έτσι ώστε να είναι στη σωστή μορφή. Για αυτό το λόγο δημιουργήθηκε μια κλάση Validator η οποία διαχειρίζεται την επικύρωση του μοντέλου, όσον αφορά την RDF μορφή του, καθώς επίσης και τα περιεχόμενα του έτσι ώστε να αποτελούν μέρος της οντολογίας. Σε περίπτωση λάθους σε ένα από τα δύο, δημιουργείται LicenseModelSyntaxException, που επίσης είναι μια κλάση για το συγκεκριμένο.

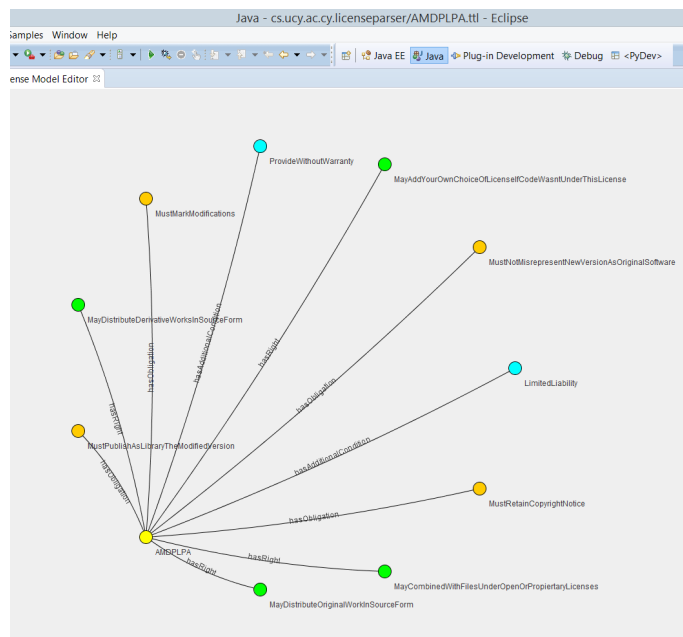


```
1@prefix owl:<http://www.w3.org/2002/07/owl#> .
2@prefix rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
3@prefix ol:<http://www.cs.ucy.ac.cy/~dpasch01/license-modeling/ontologies/oss-li
4ol:AMDPLPA
5rdf:type
6ol:OpenSourceLicense ,
7owl:NamedIndividual ;
8ol:hasAdditionalCondition
9ol:LimitedLiability ,
10ol:ProvideWithoutWarranty ;
11ol:hasRight
12ol:MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense ,
13ol:MayCombinedWithFilesUnderOpenOrProprietaryLicenses ,
14ol:MayDistributeDerivativeWorksInSourceForm ,
15ol:MayDistributeOriginalWorkInSourceForm ;
16ol:hasObligation
17ol:MustPublishAsLibraryTheModifiedVersion ,
18ol:MustMarkModifications ,
19ol:MustNotMisrepresentNewVersionAsOriginalSoftware ,
20ol:MustRetainCopyrightNotice .
21
```

Σχήμα 13: License Model Editor Model View

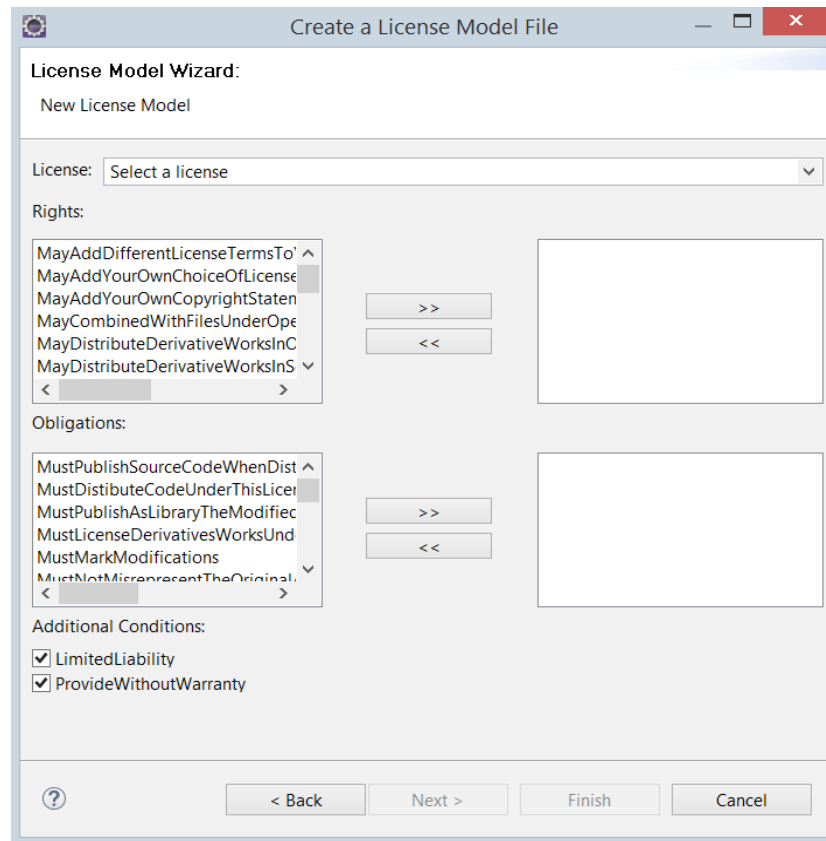
Ο τρίτος και τελευταίος editor αποτέλεσε και τον πιο απαιτητικό και δύσκολο. Αναφέρεται σε αυτού της διαδραστικής διαχείρισης του μοντέλου μέσω διεπαφής γράφου. Για την υλοποίηση του έπρεπε να δημιουργηθεί μια συγχώνευση της βιβλιοθήκης jung μαζί με την κλάση EditorPart, καθώς επίσης και η δημιουργία συγκεκριμένων event listeners με στόχο την αναγνώριση των αλλαγών πάνω στο μοντέλο που παρουσιάζεται στην διεπαφή. Μέσα στο package με το όνομα *cs.ucy.ac.cy.osslicense.model.editor.visualizer* υπάρχουν οι κλάσεις που δημιουργήθηκαν για την αναπαράσταση του γράφου, καθώς επίσης και για την δημιουργία των εργαλείων διαχείρισης του μοντέλου.

Ο χρήστης μπορεί να εισάγει καινούργιο κόμβο είτε ακμή κάνοντας δεξί κλικ και επιλέγοντας “*Insert right*” για την εισαγωγή δικαιώματος, “*Insert obligation*” για την εισαγωγή υποχρέωσης και “*Insert additional condition*” για την εισαγωγή επιπλέον όρου. Επιπλέον πατώντας δεξί κλικ πάνω σε ένα κόμβο ή μια ακμή μπορεί να τα διαγράψει, είτε να δει περισσότερες πληροφορίες για το συγκεκριμένο.



Σχήμα 14: License Model Editor Graph Preview

Εκτός από τους editor, αναφέρθηκε και την δέσμευση συγκεκριμένου τύπου αρχείου με το μοντέλο άδειας λογισμικού ανοικτού κώδικα. Ο συγκεκριμένος τύπος είναι αυτός που αναφέρθηκε κατά την ανάλυση της οντολογίας, δηλαδή του .ttl που αποτελεί το turtle format από το πλαίσιο του RDF. Καθορίζοντας λοιπόν ένα file wizard, ο χρήστης μπορεί να επιλέξει *File* → *New* → *Other* → *License Model Editor* → *License Model* και να δημιουργήσει το δικό του μοντέλο, συμπληρώνοντας την ίδια ακριβώς φόρμα που υπάρχει και στο συγκεκριμένο view του editor, όπως λέχθηκε πιο πάνω.



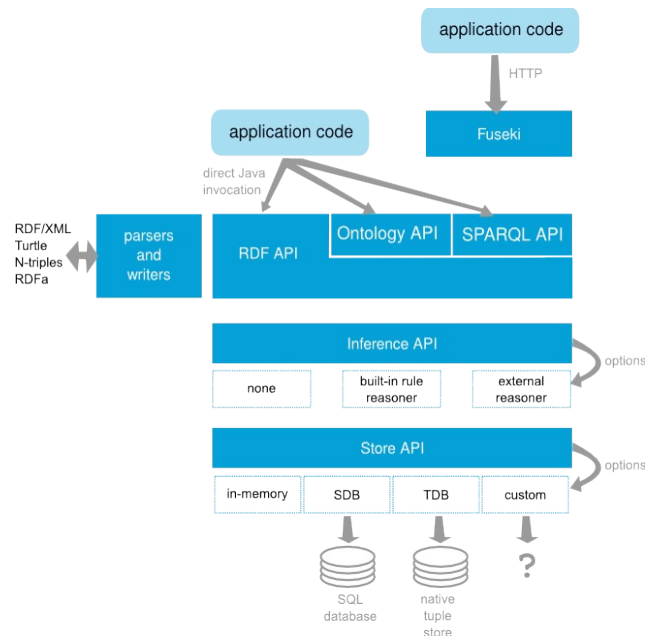
Σχήμα 15: New File Wizard.

4.2.4 Εξωτερικές βιβλιοθήκες

Για τη λειτουργικότητα του editor, που αφορά την διαχείριση του μοντέλου αδειών λογισμικού ανοικτού κώδικα, μέσω της οντολογίας μας, απαιτούνται κάποιες εξωτερικές βιβλιοθήκες. Χρειάζεται την υποστήριξη σημασιολογικών επερωτήσεων, διαχείριση αρχείων .rdf, καθώς και τη βιβλιοθήκη για την γραφική παρουσίαση της οντολογίας όπως αναλύεται σε προηγούμενο κεφάλαιο. Ακολουθούν αναλυτικά οι εξωτερικές βιβλιοθήκες που χρησιμοποιήθηκαν:

- Apache Jena:

Αναφέρεται σε ολόκληρο το πακέτο του Apache Jena [21]. Αποτελεί ένα προγραμματιστικό πλαίσιο ανοικτού κώδικα, για τη δημιουργία λογισμικών με βάσει το σημασιολογικό ιστό και διασυνδεδεμένα δεδομένα (semantic web και linked data). Αποτελείται από διαφορετικά APIs τα οποία μαζί σχηματίζουν ένα ενιαίο σύστημα



Σχήμα 16: *Apache Jena Architecture*

διαχείρισης RDF δεδομένων. Παρέχονται επίσης υπηρεσίες Apache Fuseki, μέσω ενός διαδραστικού συστήματος για την εκτέλεση και επικύρωση των σημασιολογικών ερωτήσεων. Στην εφαρμογή μας χρησιμοποιείται στην εκτέλεση σημασιολογικών ερωτήσεων SPARQL πάνω στην οντολογία μας, με σκοπό την συλλογή των δεδομένων για την παρουσίαση τους στα διάφορα views, καθώς επίσης το validation του μοντέλου κατά τη δημιουργία του, δηλαδή των ελέγχων των ορισμών εάν είναι συμβατοί.

- **rdf2rdf:**

Αποτελεί έναν απλό μετατροπέα του μοντέλου άδειας λογισμικού ανοικτού κώδικα σε διαφορετικά είδη RDF αρχείων από το βασικό, υποστηρίζοντας rdf, rdfs, owl, xml, nt, ttl [22]. Χρησιμοποιείται για την ικανοποίηση των χρηστών που επιθυμούν την εξαγωγή του μοντέλου σε διαφορετικό τύπο για συγκεκριμένο σκοπό.

- **Jung:**

Αντιστοιχεί σε Java Universal Network/Graph Framework και αποτελεί ένα πλαίσιο λογισμικού ανοικτού κώδικα [23] για τη μοντελοποίηση, ανάλυση και γραφική παρουσίαση δεδομένων σε μορφή γράφου. Αποτελεί το κύριο συστατικό στοιχείο που σχετίζεται με την παρουσίαση διαδραστικού περιβάλλοντος για τους χρήστες που είναι περισσότερο οπτικοί, και μέσω αυτού θα υλοποιηθεί η συγκεκριμένη όψη.

4.3 Έκδοση plugin

Για την έκδοση του plugin με στόχο τη δημόσια χρήση του από κοινό, απαραίτητη ήταν η σταδιακή ενσωμάτωση του σε δύο άλλα project ξεχωριστού τύπου. Η τυπική έκδοση ενός Eclipse plugin είναι με τη χρήση features. Ένα feature αποτελεί το αρχικό plugin, μαζί με την ιδιότητα ότι μπορεί να εγκατασταθεί στο Eclipse μέσω του Update Manager.

Οπότε για την έκδοση του plugin χρειάστηκε να δημιουργηθεί ένα feature project, στο οποίο εισάχθηκε το κυρίως plugin, καθώς επίσης και ένα update site project στο οποίο δηλώνεται το update site που περιέχει το feature για να γίνει η εγκατάσταση. Προσωρινά το update site βρίσκεται στην προσωπική μου σελίδα και αναμένεται έγκριση από το *Eclipse Marketplace* [40] για να μπορεί να βρεθεί από τους υπόλοιπους χρήστες και να εγκατασταθεί.

Update site: http://www.cs.ucy.ac.cy/~dpasch01/oss_license/eclipse_editor/update/

4.4 Συμπεράσματα

Εδώ τερματίζεται ό,τι αφορά την υλοποίηση του Eclipse editor plugin για το μοντέλο αδειών λογισμικού ανοικτού κώδικα και θα προχωρήσουμε στις μεθόδους εξαγωγής των ορισμών άδειας από κείμενα. Τα δύο αυτά κομμάτια δεν είναι τελείως ανεξάρτητα, εφόσον διαμέσου της εξαγωγής των ορισμών από το κείμενο θα

δημιουργείται το μοντέλο της συγκεκριμένης άδειας, το οποίο θα μπορεί να εισάγεται μέσα στο plugin για ανάλυση και διαχείριση.

Αυτό θα ανεξαρτητοποιήσει το χρήστη από τη χειροκίνητη μελέτη και εξαγωγή των ορισμών άδειας από το κείμενο για να κατασκευάσει το μοντέλο, εφόσον θα γίνεται αυτόματα από τη μηχανή.

Κεφάλαιο 5

Machine Learning και NLP - Natural Language Processing

5.1 Εισαγωγή	58
5.2 Semi-supervised learning	59
5.3 NLP - Natural language processing	60

5.1 Εισαγωγή

Η μηχανική μάθηση (machine learning [43]) είναι μια περιοχή της τεχνητής νοημοσύνης η οποία αφορά αλγορίθμους και μεθόδους που επιτρέπουν στους υπολογιστές να «μαθαίνουν». Με τη μηχανική μάθηση καθίσταται εφικτή η κατασκευή προσαρμόσιμων (Adaptable) προγραμμάτων υπολογιστών τα οποία λειτουργούν με βάση την αυτοματοποιημένη ανάλυση συνόλων δεδομένων και όχι τη διαίσθηση των μηχανικών που τα προγραμματίσαν. Η μηχανική μάθηση εφαρμόζεται σε μια σειρά από μηχανογραφικές εργασίες, όπου η χρήση αλγορίθμων, τόσο κατά το σχεδιασμό όσο και κατά τον προγραμματισμό τους είναι απαραίτητη. Παραδείγματα εφαρμογών αποτελούν το *spam filtering*, η οπτική αναγνώριση χαρακτήρων (OCR), οι μηχανές αναζήτησης και η υπολογιστική όραση. Η μηχανική μάθηση επικαλύπτεται σημαντικά με τη στατιστική, αφού και τα δύο πεδία μελετούν την ανάλυση δεδομένων, όπως επίσης και με τη εξόρυξη δεδομένων (data mining), ωστόσο η δεύτερη εστιάζει

περισσότερο στη διερευνητική ανάλυση δεδομένων, την οποία εμείς θα χρησιμοποιήσουμε στο πρόβλημα μας.

Οι αλγόριθμοι μηχανικής μάθησης κατηγοριοποιούνται ανάλογα με το επιθυμητό αποτέλεσμα του αλγορίθμου. Οι συνηθέστερες κατηγορίες είναι οι εξής:

- **Supervised learning:**

Όπου ο αλγόριθμος κατασκευάζει μια συνάρτηση που απεικονίζει δεδομένες εισόδους σε γνωστές, επιθυμητές εξόδους (*σύνολο εκπαίδευσης*), με απώτερο στόχο τη γενίκευση της συνάρτησης αυτής και για εισόδους με άγνωστη έξοδο (*σύνολο ελέγχου*).

- **Unsupervised learning:**

Όπου ο αλγόριθμος κατασκευάζει ένα μοντέλο για κάποιο σύνολο εισόδων χωρίς να γνωρίζει επιθυμητές εξόδους για το σύνολο εκπαίδευσης.

- **Reinforcement learning:**

Όπου ο αλγόριθμος μαθαίνει μια στρατηγική ενεργειών για μια δεδομένη παρατήρηση.

Για την εξαγωγή των ορισμών αδειών λογισμικού χρησιμοποιήσαμε ένα συνδυασμό μεταξύ επιβλεπόμενης (supervised) και μη επιβλεπόμενη μάθηση (unsupervised) που ονομάζεται ημι-επιβλεπόμενη μάθηση (semi-supervised learning).

5.2 Semi-supervised learning

Το Semi-supervised learning αποτελεί μια κλάση από supervised learning εργασίες και τεχνικές οι οποίες χρησιμοποιούν μεγάλα ποσοστά δεδομένων χωρίς ετικέτες

(unlabeled data) για εκπαίδευση σε συνδυασμό με μικρά ποσοστά από δεδομένα με ετικέτες (labeled data) για να την υποβοηθήσουν.

Το Semi-supervised learning βρίσκεται μεταξύ του unsupervised learning και του supervised learning. Πολλοί ερευνητές στον τομέα της μηχανικής μάθησης έφτασαν στο συμπέρασμα ότι τα δεδομένα χωρίς ετικέτες (unlabeled data), όταν χρησιμοποιούνται σε συνδυασμό με μικρά ποσοστά από δεδομένα με ετικέτες (labeled data), μπορούν να παράξουν καλύτερα αποτελέσματα όσον αφορά την απόδοση του γνωστικού προβλήματος στο οποίο εφαρμόζονται. Η απόκτηση των δεδομένων με ετικέτες (labeled data) για κάποιο συγκεκριμένο πρόβλημα μπορεί να αποτελέσει πολύ δύσκολο εγχείρημα και απαιτείται κάποιος εξειδικευμένος ανθρώπινος πράκτορας για να το πράξει, όπως συμβαίνει και στην περίπτωση μας.

Όπως και στο supervised learning, έτσι και εδώ, λαμβάνουμε ένα δεδομένο σύνολο

από l ανεξάρτητα παραδείγματα $x_1, x_2, \dots, x_l \in X$ με αντίστοιχα labels $y_1, y_2, \dots, y_l \in Y$.

Επιπρόσθετα λαμβάνουμε U παραδείγματα χωρίς labels $x_{l+1}, x_{l+2}, \dots, x_{l+U} \in X$. Το Semi-supervised learning προσπαθεί να χρησιμοποιήσει αυτές τις δύο πληροφορίες για να προσπεράσει την απόδοση της κατηγοριοποίησης που θα λαμβάναμε χωρίς την χρήση του μικρού ποσοστού από δεδομένα με ετικέτες (labeled data) ακόμη και με την εφαρμογή μόνο του unsupervised learning.

5.3 Επεξεργασία φυσικής γλώσσας (NLP - Natural language processing)

Η **επεξεργασία φυσικής γλώσσας** (Natural language processing) είναι ένας διεπιστημονικός κλάδος της επιστήμης της πληροφορικής, της τεχνητής νοημοσύνης και της γλωσσολογίας και ασχολείται με τις αλληλεπιδράσεις μεταξύ των υπολογιστών και των ανθρώπινων γλωσσών. Ως κατά συνέπεια, η Επεξεργασία του φυσικού λόγου συνδέεται στενά με την αλληλεπίδραση ανθρώπου-υπολογιστή. Πολλές προκλήσεις στην επεξεργασία του φυσικού λόγου περιλαμβάνουν την κατανόηση του, δηλαδή την προσπάθεια να κατασταθούν ικανοί οι υπολογιστές να εξαγάγουν νοήματα από ανθρώπινα ή γλωσσικά δεδομένα.

Η παρακάτω λίστα καταγράφει τα συχνότερα πεδία έρευνας στην Επεξεργασία του φυσικού λόγου. Κάποια από τα πεδία έρευνας έχουν εφαρμογές στην καθημερινή ζωή, ενώ άλλα υφίστανται ως υπο-πεδία που υποβοηθούν την επίλυση μεγαλύτερων ζητημάτων. Το κριτήριο που ξεχωρίζει τα παρακάτω πεδία έρευνας από άλλα πιθανά και πραγματικά πεδία της επεξεργασίας του φυσικού λόγου είναι το γεγονός ότι για το καθένα από αυτά υπάρχει επίσημα ένας καλά ορισμένος χώρος εργασιών και επίλυσης ζητημάτων, ένα καθιερωμένο μετρικό σύστημα για την αξιολόγηση του πεδίου, κάποια δεδομένα σώματα κειμένων πάνω στα οποία κάθε πεδίο αξιολογείται και διαγωνισμοί αφιερωμένοι στο κάθε πεδίο.

- **Ανάλυση λόγου:** Είναι μία ρουμπρίκα που περιλαμβάνει ποικίλες μελέτες. Μία μελέτη αναφέρεται στην αναγνώριση της δομής του λόγου ενός συνδεόμενου κειμένου π.χ. την φύση των σχέσεων του λόγου μεταξύ δύο προτάσεων (π.χ. επεξήγηση, αντίθεση κτλ.). Μία άλλη πιθανή μελέτη είναι η αναγνώριση και η κατηγοριοποίηση των γλωσσικών πράξεων σε ένα κομμάτι κειμένου.
- **Αυτόματη αναγνώριση ομιλίας:** Η αυτόματη μετατροπή του προφερόμενου ανθρώπινου λόγου σε κείμενο από τους υπολογιστές.
- **Αυτόματη ερωταπόκριση:** Η αναζήτηση της σωστής απάντησης σε μία δεδομένη ερώτηση της ανθρώπινης γλώσσας.

- **Αυτόματη μορφολογική τεμαχιοποίηση:** Η κατάτμηση των λέξεων στα μορφήματά τους καθώς και η αναγνώριση και κατηγοριοποίηση αυτών των μορφημάτων. Η δυσκολία του συγκεκριμένου πεδίου μελέτης εξαρτάται σε μεγάλο βαθμό από την περιπλοκότητα της μορφολογίας της εκάστοτε γλώσσας υπό εξέταση.
- **Αυτόματη περίληψη:** Η παραγωγή μίας αναγνώσιμης περίληψης ενός κειμένου. Συχνά χρησιμοποιείται για να παρέχει περιλήψεις σε κείμενα γνωστής διάταξης, όπως άρθρα στο οικονομικό μέρος μίας εφημερίδας.
- **Εξόρυξη πληροφοριών:** Η ανάκτηση πληροφοριών από μη δομημένα ή ημιδομημένα δεδομένα (τυπικά κείμενα γραμμένα σε φυσική γλώσσα, ιστοσελίδες κ.α.).
- **Επίλυση σχέσεων συναναφοράς:** Η αναζήτηση του ποιες λέξεις ("αναφορές") αναφέρονται στα ίδια υποκείμενα ("οντότητες") σε μία δεδομένη πρόταση ή ένα μεγαλύτερο κομμάτι κειμένου. Η επίλυση σχέσεων αναφοράς είναι ένα συγκεκριμένο παράδειγμα αυτού του πεδίου και αναφέρεται συγκεκριμένα στην σύνδεση των αντωνυμιών με τα ουσιαστικά ή τα ονόματα στα οποία αναφέρονται.
- **Επισήμανση των μερών του λόγου:** Ο αυτόματος καθορισμός των μερών του λόγου σε μία δεδομένη πρόταση και η επίλυση της συντακτικής αμφισημίας.
- **Κατανόηση του φυσικού λόγου:** Η μετατροπή κομματιών κειμένου σε πιο τυπικές αναπαραστάσεις όπως σε δομές λογικής πρώτου βαθμού, οι οποίες μπορούν να μεταχειριστούν ευκολότερα από τους υπολογιστές.

- **Μηχανική μετάφραση:** Η αυτόματη μετάφραση ενός κειμένου από μία ανθρώπινη γλώσσα σε μία άλλη.
- **Οπτική αναγνώριση χαρακτήρων:** Ο προσδιορισμός του αντίστοιχου κειμένου από μία δεδομένη εικόνα που αναπαριστά κάποιο τυπωμένο κείμενο.
- **Παραγωγή φυσικού λόγου:** Η μετατροπή των πληροφοριών από υπολογιστικές βάσεις δεδομένων σε αναγνώσιμο φυσικό λόγο.
- **Σύνθεση ομιλίας:** Η αυτόματη, τεχνητή παραγωγή του ανθρώπινου λόγου από τους υπολογιστές.
- **Συντακτική ανάλυση:** Ο αυτόματος καθορισμός του συντακτικού δέντρου μίας δεδομένης πρότασης και η επίλυση των οποιοδήποτε συντακτικών αμφισημιών. Εξαιτίας των πιθανών αμφισημιών που πιθανόν να φέρει μία πρόταση, είναι δυνατόν η εν λόγω πρόταση να αναλυθεί σε παραπάνω από ένα συντακτικά δέντρα.

Για την εξαγωγή των ορισμών αδειών λογισμικού ανοικτού κώδικα θα επικεντρωθούμε στα πεδία της εξόρυξης πληροφοριών, κατανόηση του φυσικού λόγου καθώς επίσης και της συντακτικής ανάλυσης.

Κεφάλαιο 6

Προσέγγιση ανάλυσης και προετοιμασία δεδομένων

6.1 Προσέγγιση ανάλυσης	64
6.2 Μεθόδοι ανάλυσης	65
6.2.1 Part-Of-Speech Tagging	65
6.2.2 Chunking	66
6.2.3 Named Entity Recognition	66
6.2.4 Semantic Role Labeling	66
6.3 Προετοιμασία δεδομένων	66
6.3.1 Συλλογή και είσοδος δεδομένων	66
6.3.2 Προετοιμασία δεδομένων	67
6.3.3 Προεπεξεργασία δεδομένων	68

6.1 Προσέγγιση ανάλυσης

Όπως αναφέραμε προηγουμένως, έγινε μελέτη 32 αδειών λογισμικού από τις οποίες εξάχθηκαν τα δικαιώματα, οι υποχρεώσεις και οι επιπρόσθετες πληροφορίες. Για κάθε μία από αυτές τις άδειες έγινε αντιστοιχία του ορισμού τους με το κομμάτι του κειμένου της άδειας που το δικαιολογεί (καθορίζεται δηλ. το πού ακριβώς στο κείμενο της άδειας εμφανίζεται ο ορισμός). Οπότε το σκεπτικό πίσω από την προσέγγιση είναι η εκπαίδευση ενός νευρωνικού δικτύου, βάσει αυτών των δεδομένων, έτσι ώστε να

μπορεί να αναγνωρίζει μέσα από κείμενα αδειών λογισμικού ανοικτού κώδικα, τους ορισμούς βάσει των κομματιών του κειμένου.

Όπως θα δούμε παρακάτω αυτό μπορεί να γίνει με δύο τρόπους, είτε με τη χρήση δεδομένων που εξήχθησαν δια χειρός όσον αφορά τα κομμάτια του κειμένου, είτε μέσω μάθησης με μερική επίβλεψη (semi-supervised learning [44]). Η πρώτη προσέγγιση αντιστοιχεί στη χρήση προκατασκευασμένων προτύπων μάθησης, που είναι αυτά τα οποία έχουμε σαν δεδομένα, και το δεύτερο είναι η εκπαίδευση ενός συστήματος να αναγνωρίζει από μόνο του κατά πόσον ένα κομμάτι κειμένου μιλάει για συγκεκριμένο ορισμό.

Η κατηγοριοποίηση των κειμένων θα γίνεται βάσει ετικετών (tags) οι οποίες θα αντιπροσωπεύουν το περιεχόμενο τους.

Οι ετικέτες βάσει των οποίων θα γίνεται η κατηγοριοποίηση του κειμένου των αδειών, αντιστοιχούν στους ορισμούς που εξάχθηκαν από την Ατομική Διπλωματική Εργασία της Γεωργίας Χαραλάμπους. Οπότε έχουμε 33 πιθανές ετικέτες που αντιστοιχούν στις κατηγορίες κειμένων, και το κάθε κείμενο θα μπορεί να ανήκει σε μία ή περισσότερες από αυτές.

6.2 Μεθόδους ανάλυσης

6.2.1 Ετικετοποίηση βάσει μέρους του λόγου (Part-Of-Speech Tagging)

Αποτελεί μέθοδο με την οποία λέξεις μέσα στο κείμενο σημαδεύονται σε σχέση με το σχήμα λόγου στο οποίο αντιστοιχούν, βάσει του ορισμού τους, καθώς επίσης και του περιεχομένου στο οποίο βρίσκονται, π.χ. η σχέση με τις γειτονικές λέξεις της πρότασης. Κάποιες βασικές ετικέτες (tags) είναι *VB* που αντιστοιχεί σε ρήμα, *NN* που αντιστοιχεί σε ουσιαστικό, *JJ* που αντιστοιχεί σε επίρρημα κτλ.

6.2.2 Κατατεμάχηση (Chunking)

Αποτελεί ανάλυση της πρότασης μέσω της οποίας αναγνωρίζονται τα συστατικά της χωρίς όμως να καθορίζεται η εσωτερική δομή τους ή ο ρόλος τους στην πρόταση.

6.2.3 Αναγνώριση οντοτήτων (Named Entity Recognition)

Αποτελεί υπο-εργασία της εξαγωγής πληροφορίας, η οποία προσπαθεί να βρει και να κατηγοριοποιήσει στοιχεία μέσα από το κείμενο, σε προκαθορισμένες κατηγορίες όπως είναι τα ονόματα ατόμων, οργανισμών, τοποθεσιών κτλ. Ακολουθεί ένα παράδειγμα:

Παράδειγμα	Αποτέλεσμα
Jim bought 300 shares of Acme Corp. in 2006.	[Jim]Person bought 300 shares of [Acme Corp.]Organization in [2006]Time.

6.2.4 Εξαγωγή σημασιολογικού ρόλου (Semantic Role Labeling)

Αποτελεί μια διαδικασία στο NLP που αποτελείται από την αναγνώριση των σημασιολογικών ορισμών που σχετίζονται με το πρόσωπο ή το ρήμα της πρότασης και την κατηγοριοποίησή τους σε συγκεκριμένους ρόλους. Για παράδειγμα, δοσμένης της πρότασης "Mary sold the book to John", κατά την εξαγωγή σημασιολογικού ρόλου θα πρέπει να αναγνωριστεί το ρήμα "to sell" έτσι ώστε να δηλώνει το στόχο, "Mary" ώστε να δηλώνει τον πράκτορα, "the book" ώστε να δηλώνει το θέμα, και "John" που να δηλώνει τον παραλήπτη. Αυτό αποτελεί ένα σημαντικό βήμα στην κατανόηση της πρότασης.

6.3 Προετοιμασία δεδομένων

6.3.1 Συλλογή και είσοδος δεδομένων

Όσον αφορά τη συλλογή των δεδομένων που χρησιμοποιήθηκαν, υλοποιήθηκε ένας περιηγητής διαδικτύου (Web Crawler) σε Java μέσω της βιβλιοθήκης Jsoup [45]. Ο συγκεκριμένος crawler διέτρεξε κάποιες βασικές ιστοσελίδες που περιέχουν κείμενα αδειών και κατασκεύασε το σύνολο των δεδομένων μας. Οι συγκεκριμένες σελίδες που χρησιμοποιήθηκαν για τη διάσχιση είναι οι:

- <http://spdx.org/licenses/>
- <https://opensource.org/licenses/alphabetical>
- <https://www.openhub.net/licenses/>

Τα αποτελέσματα όμως ήταν πολύ λίγα σε αριθμό με 309 κείμενα αδειών λογισμικού ανοικτού κώδικα συνολικά. Με αυτό το σύνολο δεδομένων, και με το ποσοστό θορύβου που υπάρχει μέσα σε τέτοιου είδους νομικά κείμενα, η απόδοση της εκπαίδευσης ενός συστήματος, θα ήταν πολύ χαμηλή. Οπότε, μέσω του συστήματος FLOSSmole [26] και σχετικού αιτήματος που εστάλη δόθηκε πρόσβαση σε βάση δεδομένων με πληροφορίες από εκατοντάδες έργα λογισμικού ανοικτού κώδικα, συμπεριλαμβανομένου και του κειμένου της άδειας λογισμικού τους. Το FLOSSmole προσφέρει δωρεάν δεδομένα μεγέθους πολλών terabytes, όσον αφορά έργα FLOSS, για σκοπούς έρευνας και ανάλυσης. Κάποια από τα αποθετήρια, από τα οποία το σύστημα λαμβάνει τα δεδομένα τους είναι έργα από Sourceforge, Freecode, Rubyforge, Objectweb, Sourcekibitzer, GitHub κτλ.

Αναμενόμενο ήταν να βρίσκαμε διπλότυπα αδειών, λόγω του περιορισμένου αριθμού τους, οπότε σαν αποτέλεσμα ήταν η αύξηση τους από 309 σε 483, που πάλι αποτελεί μη ικανοποιητικό αριθμό δεδομένων για την εκπαίδευση ενός τέτοιου συστήματος. Επειδή όμως το σύστημα μας αναφέρεται σε περιορισμένο πεδίο αδειών λογισμικών ανοικτού κώδικα, αναγκαστικά θα πρέπει να αρκεστούμε σε αυτά και να βρούμε διαφορετικές μεθόδους για αύξηση της απόδοσης του συστήματος μας.

6.3.2 Προετοιμασία δεδομένων

Όπως αναφέραμε, τα κείμενα αδειών λογισμικού ανοικτού κώδικα, αποτελούν νομικά κείμενα, τα οποία είναι γραμμένα σε φυσική γλώσσα, δεν ακολουθούν κάποια σύμβαση στη συγγραφή τους και η κατανόηση τους αποτελεί δύσκολο έργο. Περιέχουν πολλά αχρείαστα κομμάτια κειμένου, τα οποία είναι θόρυβος σε σχέση με τις κρίσιμες πληροφορίες στις οποίες θέλουμε να επικεντρωθούμε. Τέτοια κομμάτια κειμένου είναι οι διάφορες επεξηγήσεις ορισμών π.χ. ορισμός του “Original Code”, τα διάφορα headers και footers, τα violation notices κ.α.

Για την καλύτερη απόδοση του συστήματος, και πριν από την εκπαίδευση του, απαραίτητη είναι η προετοιμασία των δεδομένων έτσι ώστε να υποβοηθήσουμε το σύστημα μας. Αυτό το στάδιο στοχεύει στο να:

- Μειώσει τον θόρυβο που υπάρχει μέσα στα κείμενα των αδειών.
- Διασπάσει τα κείμενα σε παραγράφους για μείωση της πιθανότητας λανθασμένης κατηγοριοποίησης.
- Διαχωρίσει τα δεδομένα σε δεδομένα εκπαίδευσης (testing) και δεδομένα ελέγχου (testing). Τα δεδομένα εκπαίδευσης αποτελούν αυτά πάνω στα οποία το σύστημα μας θα εκπαιδευτεί και θα προσαρμόσει τις παραμέτρους του, ενώ τα δεδομένα ελέγχου είναι αυτά πάνω στα οποία θα ελέγξει τα αποτελέσματα του και θα εντοπίσει το ποσοστό σφάλματος (error), το οποίο θέλουμε να είναι σε όσο το δυνατόν χαμηλότερα επίπεδα.

6.3.3 Προεπεξεργασία δεδομένων

Η προεπεξεργασία των δεδομένων συμπεριλαμβάνει μια σειρά από γλωσσολογικές τεχνικές με στόχο τη δομική παρουσίαση του κειμένου, όπως για παράδειγμα τις λέξεις, τη γραμματική και τα σημεία στίξης. Κατά την προετοιμασία των κειμένων που συλλέχθηκαν, χρησιμοποιήθηκαν τρεις βασικές τεχνικές:

1) Κατάτμηση (Tokenization): το οποίο αναφέρεται στη διαδικασία διαχωρισμού του κειμένου σε υπομονάδες, αρχικά σε προτάσεις/παραγράφους και μετά σε λέξεις. Αυτόματα μετατρέπει όλα τα κεφαλαία γράμματα σε μικρά για διευκόλυνση της

σύγκρισης των προτύπων.

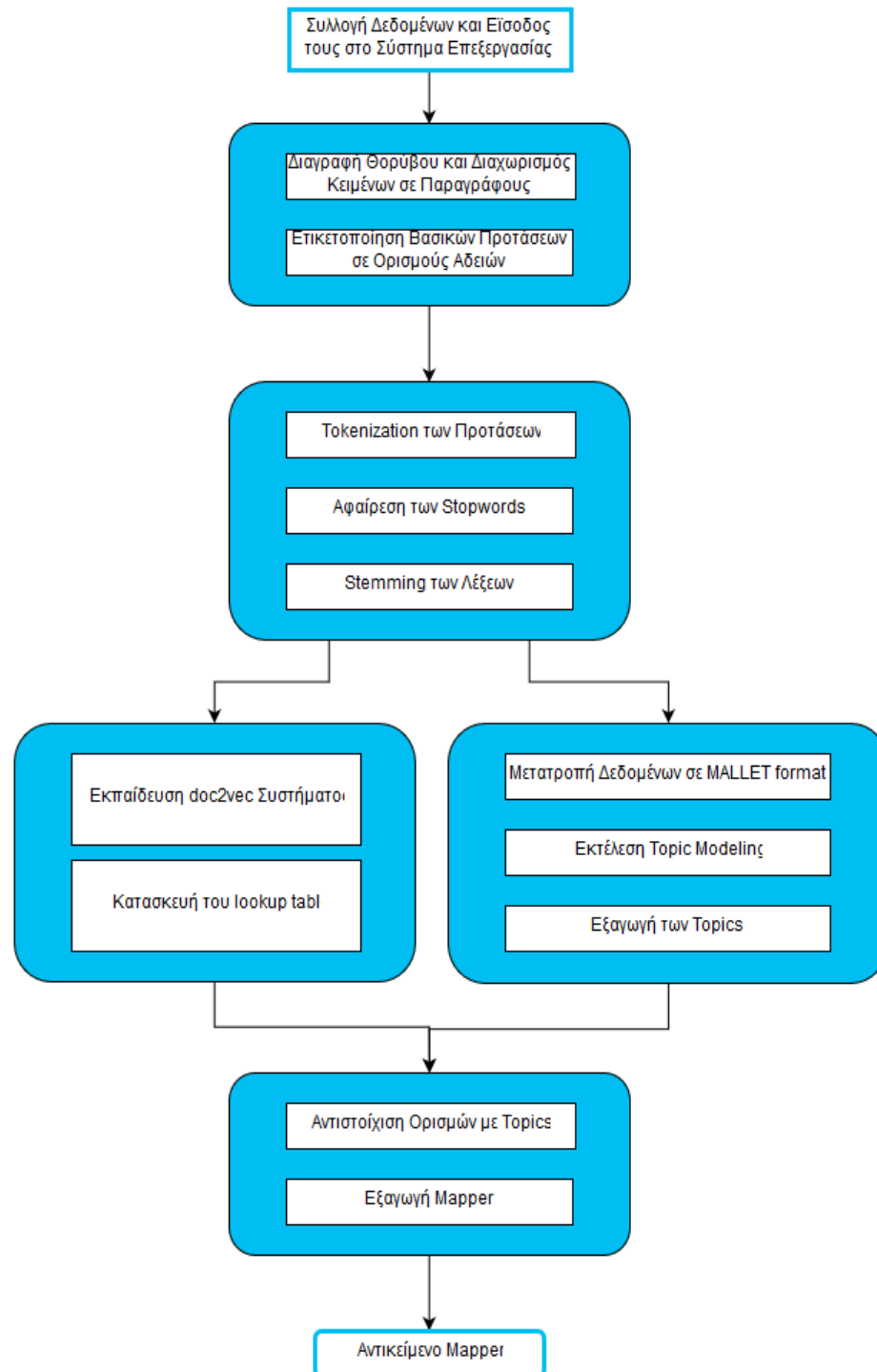
2) Stop-word removal: στοχεύει στην διαγραφή κοινών λέξεων που δεν αποτελούν στοιχειώδες σε κάποια συγκεκριμένη κατηγορία. Με τη διαγραφή τους μειώνεται ο θόρυβος μέσα στο κείμενο και αυξάνεται η απόδοση στο σύστημα. Υλοποιήθηκε με τη χρήση μιας λίστας με λέξεις Stop Word στην Αγγλική γλώσσα.

3) Stemming: είναι η εφαρμογή μορφολογικής ανάλυσης λέξεων μέσα από το κείμενο. Αποτελεί ακόμη μια γλωσσολογική τεχνική που μετατρέπει τις λέξεις στις ρίζες τους, π.χ. η λέξη *environmental* θα γίνει *environ*. Με αυτό τον τρόπο παραλλαγές της ίδιας λέξης θα μετατραπούν στις ρίζες τους και οι σύγκριση μεταξύ τους θα αποδώσει καλύτερα αποτελέσματα. Για το δικό μας σύστημα χρησιμοποιήθηκε το εργαλείο *morpha stemmer* [27].

Για την κατάτμηση των προτάσεων από τα κείμενα των αδειών χρησιμοποιήθηκαν διαφορετικοί *parsers* για έλεγχο αποτελεσμάτων. Ακολουθούν οι διάφοροι *parsers*:

- Break Iterator [28] που αποτελεί κλάση της Java και διαχωρίζει το κείμενο σε προτάσεις βάσει την τοπική σύνταξη τους (πχ. US-Locale).
- Regular expressions για το διαχωρισμό προτάσεων που χρησιμοποιεί το σύστημα GATE [29].
- Stanford NLP Sentence Parser [30] που αποτελεί ένα από τα βασικά (Core) στοιχεία του πακέτου Stanford NLP.

Η πιο αποδοτική ήταν η χρήση του Break Iterator της Java λόγω του καλύτερου διαχωρισμού των προτάσεων από τα υπόλοιπα. Η απόδοση αυτή μετρήθηκε με το μάτι, παρακολουθώντας τα αποτελέσματα, και βλέποντας πως ο συγκεκριμένος έδωσε τα καλύτερα. Οπότε στην περαιτέρω επεξεργασία και εκπαίδευση του νευρωνικού δικτύου, χρησιμοποιούνται τα αποτελέσματα του Break Iterator.



Σχήμα 17: Pipeline για επεξεργασία δεδομένων.

Κεφάλαιο 7

Ανάλυση προσεγγίσεων και Αποτελέσματα

7.1 Εισαγωγή	72
7.2 Μηχανικά πρότυπα μάθησης	72
7.2.1 Εισαγωγή	72
7.2.2 Word2vec	73
7.2.2.1 Cosine similarity	75
7.2.2.2 Υπολογισμός threshold	76
7.2.2.2.1 Threshold για κάθε license term	80
7.2.2.2.2 Ενιαίο threshold	82
7.2.3 doc2vec	83
7.2.4 Συνδυασμός απο license terms	85
7.2.5 Κατηγοριοποίηση βάσει άδειας	86
7.2.6 Δυναμική κατηγοριοποίηση	87
7.2.7 Χρήση Google vectors	87
7.3 Αυτόματη εξαγωγή προτύπων	88
7.3.1 Εισαγωγή	88
7.3.2 Topic Modeling	88
7.3.3 Αλγόριθμοι για Topic Modeling	89
7.3.3.1 LSI Latent Semantic Indexing	89
7.3.3.2 HDP Hierarchical Dirichlet Process	90
7.3.4 Αλγόριθμος LDA Latent Dirichlet Allocation	90
7.3.5 Κατανομή λέξεων κλειδιών	91
7.3.6 Topics με μεγαλύτερο score	92

7.1 Εισαγωγή

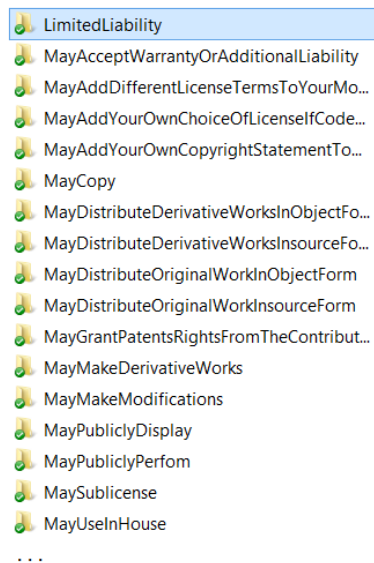
Για την ανάλυση του κειμένου και την υλοποίηση ενός συστήματος κατηγοριοποίησης υπάρχουν δύο κατηγορίες προσεγγίσεων. Η πρώτη αφορά τη χρήση ήδη παρασκευασμένων προτύπων κειμένου (Engineered Features [46]), και η δεύτερη αφορά την αυτόματη εξαγωγή αυτών των προτύπων μέσα από τα ίδια τα κείμενα. Η τελευταία αποτελεί την πιο επιθυμητή προσέγγιση σε προβλήματα NLP.

Για την υλοποίηση του συστήματος μας χρησιμοποιήθηκαν και οι δύο αυτές προσεγγίσεις, κάνοντας συγκρίσεις στα αποτελέσματά τους. Ακολουθούν αναλυτικά οι δύο προσεγγίσεις με τις υποκατηγορίες τους καθώς και τα αποτελέσματα που έδωσε η καθεμία.

7.2 Μηχανικά πρότυπα μάθησης

7.2.1 Εισαγωγή

Μέσω υλοποίησης ενός αναλυτή κειμένου (Parser) στην Java γίνεται κατασκευή της συλλογής δεδομένων που θα χρησιμοποιηθούν, βάσει των πληροφοριών που εξήγαγε η Γεωργία Χαραλάμπους αναφορικά με την αντιστοίχιση όρων αδειών σε σημεία στα κείμενα της κάθε άδειας (data sheet). Η Γεωργία Χαραλάμπους δικαιολόγησε τους όρους των αδειών λογισμικού με κομμάτια προτάσεων από το κείμενο της κάθε άδειας. Για να χρησιμοποιηθούν όμως χρειάζεται να είναι ολοκληρωμένες προτάσεις. Μέσω λοιπόν του Parser γίνεται συλλογή όλων των προτάσεων που αντιστοιχούν στον κάθε όρο άδειας (license term) και δημιουργείται στο σύστημα αρχείων μια διάταξη με φακέλους που περιέχουν τις προτάσεις για το κάθε ένα, και έχουν σαν όνομα το συγκεκριμένο όρο.

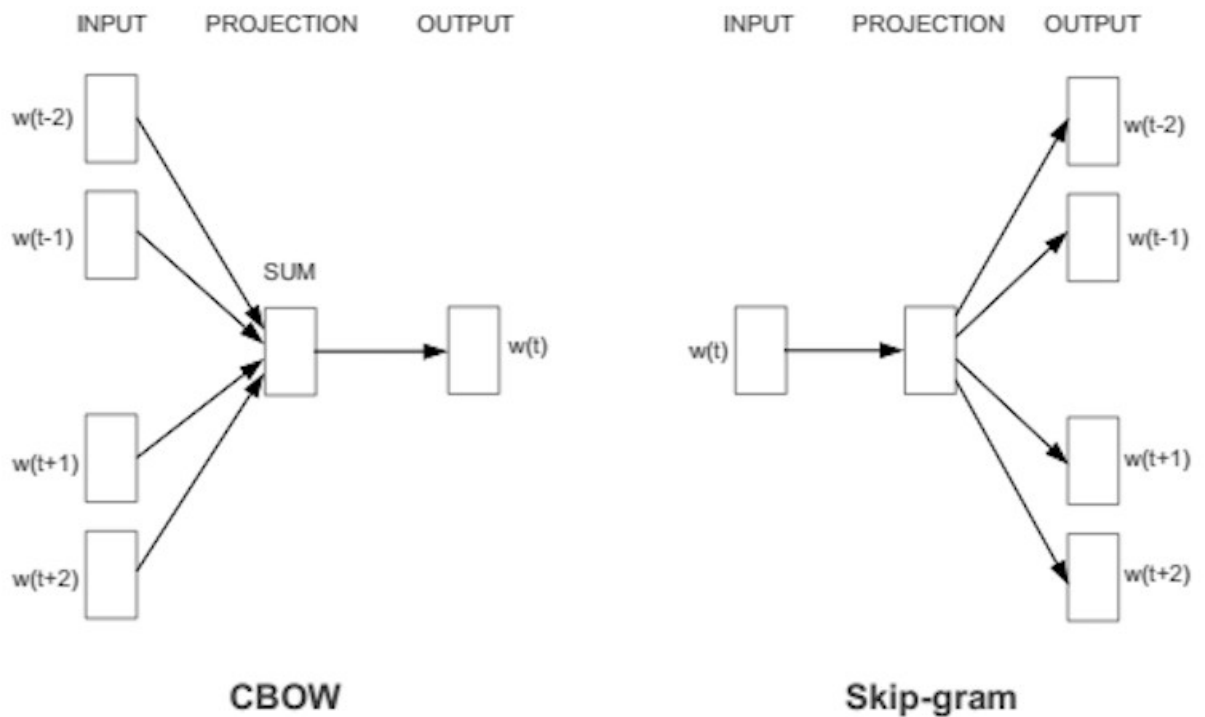


Σχήμα 18: License Term Categorization File System.

Στόχος είναι να γίνει η κατηγοριοποίηση του κειμένου άδειας που θα δώσει ο χρήστης σε ορισμούς αδειών λογισμικού, βάσει του περιεχόμενου του. Δηλαδή η εξαγωγή όλων των license terms από το κείμενο μιας δεδομένης άδειας. Για αυτή την κατηγοριοποίηση, χρησιμοποιώντας το σύνολο δεδομένων (dataset) που δημιουργήθηκε πιο πάνω, εφαρμόστηκαν οι υπόλοιπες προσεγγίσεις όσον αφορά τα παρασκευασμένα πρότυπα.

7.2.2 Word2vec

Αποτελεί ένα είδος νευρωνικού δικτύου δύο επιπέδων, το οποίο επεξεργάζεται κείμενο. Λαμβάνει ως είσοδο μία συλλογή από αρχεία κειμένων και παράγει σαν έξοδο ένα σύνολο από διανύσματα (Vectors), τα οποία περιγράφουν τις λέξεις εκείνων των κειμένων σε σχέση με την παρουσία τους σε αυτά, μέσα στις προτάσεις και γενικά. Με άλλα λόγια περιγράφει τη λέξη σε σχέση με άλλες λέξεις.

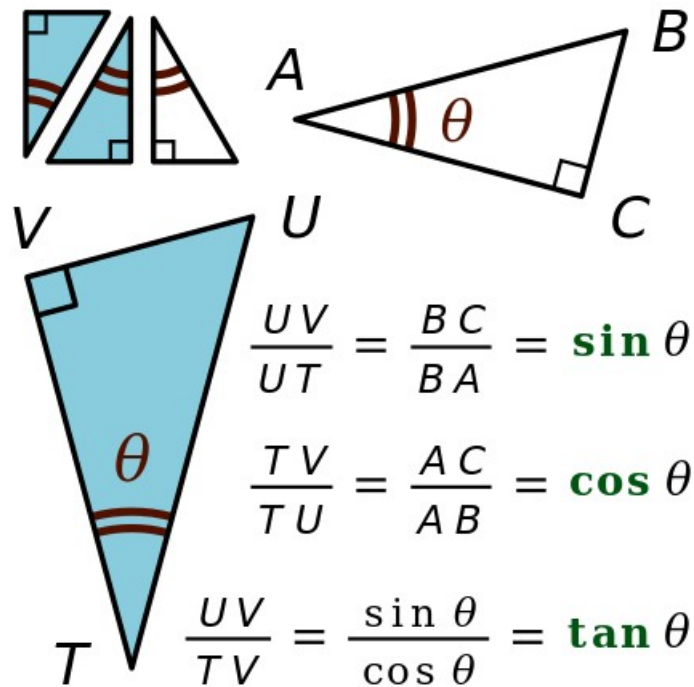


Σχήμα 19: Word2vec CBOW and Skip-gram Algorithms Diagram.

Για να κάνει αυτή την επεξεργασία και τη μετατροπή καθώς και την πρόβλεψη μιας λέξης, χρησιμοποιεί μεθόδους όπως είναι οι CBOW (Continues Bag of Words) και Skip-gram. Το πρώτο χρησιμοποιείται για την πρόβλεψη λέξης βάσει του περιεχομένου του κειμένου και το δεύτερο για την πρόβλεψη του κειμένου βάσει κάποιας λέξης αντίστοιχα. Πιο πάνω βρίσκεται το διάγραμμα για τον τρόπο λειτουργίας τους (σχήμα 18).

Η εφαρμογή του Word2vec στην περίπτωση μας έγινε με το προγραμματιστικό πλαίσιο Deeplearning4j [31]. Τα κείμενα όλων των αδειών λογισμικού δόθηκαν ως είσοδος στο νευρωνικό δίκτυο, και στη συνέχεια δημιουργήθηκε ένας πίνακας αναζήτησης (lookup table) από αντιστοιχία των λέξεων των κειμένων των αδειών με τα διανύσματα (vectors) που εξήχθησαν. Για κάθε πρόταση του κειμένου εισόδου, για το οποίο θέλουμε να κάνουμε την κατηγοριοποίηση, υπολογίζεται ο μέσος όρος του διανύσματος κάθε λέξης και ελέγχεται πόσο κοντά βρίσκεται στο αντίστοιχο διάνυσμα των προτάσεων για κάθε license term.

Ο έλεγχος αυτός γίνεται βάσει του cosine similarity [47] μεταξύ των δύο διανυσμάτων. Τριγωνομετρικές μέθοδοι, όπως αυτές του ημιτόνου, συνημιτόνου και εφαπτόμενης, αποτελούν αναλογίες που χρησιμοποιούν τις αποστάσεις των πλευρών του ορθογώνιου τριγώνου για να υπολογίσουν τις γωνίες του.

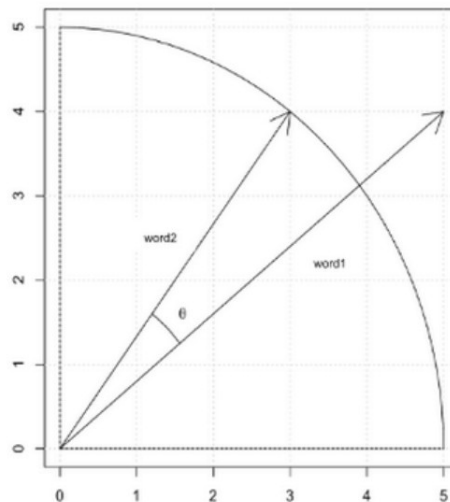


Σχήμα 20: Ορισμοί $\cos$, $\sin$ και $\tan$.

Οι διαφορές μεταξύ δύο διανυσμάτων των λέξεων μπορούν να θεωρηθούν και ως διαφορές σε γωνίες. Αυτή η διαφορά λοιπόν μετρίεται βάσει της cosine similarity μεταξύ των διανυσμάτων. Για να υπολογιστεί η συγκεκριμένη απόσταση μέσω δύο διανυσμάτων χρησιμοποιείται η παρακάτω εξίσωση:

$$\vec{a} \cdot \vec{b} = |\vec{a}| \cdot |\vec{b}| \cdot \cos(\theta)$$

Το οποίο δηλώνει επίσης ότι το συνημίτονο της γωνίας θ είναι ίσο με τη διαίρεση του γινομένου των δύο διανυσμάτων με αυτό των απόλυτων τιμών τους. Οι τιμές που λαμβάνουμε κανονικοποιούνται έτσι ώστε να έχουμε 1 εάν τα διανύσματα είναι όμοια, και 0 εάν δεν έχουν σχέση το ένα με το άλλο.



Σχήμα 21: Σχέση Γωνίας με Λέξεις.

7.2.2.2 Υπολογισμός ορίου

Με το πέρας της διαδικασίας εκπαίδευσης του word2vec, έχουμε τα license terms και κάποια βαθμολογία (Score) όσον αφορά στο πόσο σίγουρα η συγκεκριμένη άδεια περιγράφεται από το συγκεκριμένο license term. Ακολουθεί ανάλυση αυτών των βαθμολογιών για εύρεση συγκεκριμένου ορίου (threshold) το οποίο θα δηλώνει κατά πόσο το ανάλογο κείμενο περιγράφεται με κάποιο license term.

Λήφθηκαν λοιπόν τα αποτελέσματα της κάθε άδειας από τις 22 αρχικές που ανέλυσε η Γεωργία Χαραλάμπους, σε αντιστοιχία με τα license terms που η ίδια είχε εξάγει. Ένα παράδειγμα είναι αυτό της *AFL-3.0* άδειας όπως ακολουθεί πιο κάτω, τα υπόλοιπα μπορούν να βρεθούν στο Παραρτημα Α:

AFL-3.0

Cosine Similarity	License term
0.920651647333 9127	<i>MayMakeDerivativeWorks</i>
0.920651647333 9127	<i>MayPubliclyDisplay</i>
0.920651647333 9127	<i>MayMakeModifications</i>
0.920651647333 9127	<i>MaySublicense</i>
0.920651647333 9127	<i>MayPubliclyPerform</i>
0.920651647333 9127	MayGrantPatentsRightsFromTheContributorToReceipt
0.920651647333 9127	<i>NotPermissionToUseTradeMarks</i>
0.920102506282 5946	<i>LimitedLiability</i>
0.920102506282 5946	<i>ProvideWithoutWarranty</i>
0.713968974837 0922	<i>MayDistributeOriginalWorkInsourceForm</i>
0.713968974837 0922	MustMarkModifications
0.710573648063 527	MayAcceptWarrantyOrAdditionalLiability
0.685201317847 8517	<i>MustProvideInformationHowToGetsourceCode</i>
0.685201317847 8517	NotSublicense
0.685201317847 8517	MustOffersourceCode
0.685201317847	<i>MustDistibuteCodeUnderThisLicense</i>

8517	
0.685201317847 8517	MayDistributeDerivativeWorksInObjectForm
0.685201317847 8517	<i>MayCopy</i>
0.648830989919 9427	<i>MayDistributeDerivativeWorksInSourceForm</i>
0.648830989919 9427	MayDistributeOriginalWorkInObjectForm
0.646173561060 7651	<i>MustRetainCopyrightNotice</i>
0.642859097476 5032	MayUseInHouse
0.637280373580 6974	MayAddDifferentLicenseTermsToYourModifications
0.637280373580 6974	MustLicenseDerivativesWorksUnderCompatibleLicense
0.637280373580 6974	MustDistributeCopyOfNoticeText
0.637280373580 6974	MayAddYourOwnCopyrightStatementToYourModifications
0.598323393803 4308	MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense
0.598323393803 4308	MustRetainLicenseNotice
0.598323393803 4308	<i>MustRetainInSourceCodeOfDerivativeWorkAllTradeMarksPatentCopyri ghtsOfOriginalWork</i>
0.590956600203 7873	MustPublishAsLibraryTheModifiedVersion
0.514833844171 5562	MustNotMisrepresentTheOriginalAuthorOfSoftware

Όσα στοιχεία του πίνακα είναι της μορφής **0.6852013178478517 MayCopy** αντιστοιχούν σε αυτά που βρέθηκαν δια χειρός για κάθε άδεια, και αποτελούν αυτά που θα επιθυμούσαμε να είχαν τη ψηλότερη βαθμολογία, με υψηλότερη θέση πάνω στον πίνακα.

Με παρόμοιο τρόπο βγήκαν και τα αποτελέσματα από τις υπόλοιπες αρχικές άδειες και υπολογίσθηκε ένα κατώφλι για την κατηγοριοποίηση των κειμένων σε όρους αδειών license terms. Για τον υπολογισμό του κατωφλίου ακολουθήθηκαν δύο μέθοδοι, η πρώτη συμπεριλάμβανε τον υπολογισμό κατωφλίου για κάθε μια από τα license terms ξεχωριστά, και η δεύτερη τον υπολογισμό ενός ενιαίου κατωφλίου για όλα. Ακολουθεί παράδειγμα υπολογισμού του κατωφλίου για συγκεκριμένη άδεια.

Term	AFL-3.0
LimitedLiability	0.92
MayAcceptWarrantyOrAdditionalLiability	0.71
MayAddDifferentLicenseTermsToYourModifications	0.64
MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense	0.6
MayAddYourOwnCopyrightStatementToYourModifications	0.64
MayCopy	0.69
MayDistributeDerivativeWorksInObjectForm	0.69
MayDistributeDerivativeWorksInSourceForm	0.65
MayDistributeOriginalWorkInObjectForm	0.65
MayDistributeOriginalWorkInSourceForm	0.71
MayGrantPatentsRightsFromTheContributorToReceipt	0.92
MayMakeDerivativeWorks	0.92
MayMakeModifications	0.92
MayPubliclyDisplay	0.92
MayPubliclyPerfrom	0.92
MaySublicense	0.92
MayUseInHouse	0.64
MustDistributeCodeUnderThisLicense	0.69
MustDistributeCopyOfNoticeText	0.64
MustLicenseDerivativesWorksUnderCompatibleLicense	0.64
MustMarkModifications	0.71
MustNotMisrepresentTheOriginalAuthorOfSoftware	0.51
MustOffersourceCode	0.69
MustProvideInformationHowToGetsourceCode	0.69
MustPublishAsLibraryTheModifiedVersion	0.59
MustRetainCopyrightNotice	0.65
MustRetainInsourceCodeOfDerivativeWorkAllTradeMarksPatentCopyrightsOfOriginalWork	0.6
MustRetainLicenseNotice	0.6
NotPermissionToUseTradeMarks	0.92
NotSublicense	0.69
ProvideWithoutWarranty	0.92
Average	0.73

Σχήμα 22: AFL threshold calculation

Η παραπάνω διαδικασία επαναλαμβάνεται για όλες τις 23 άδειες που αναφέραμε προηγουμένως και παράγονται κατώφλια για κάθε ένα από τους όρους, καθώς επίσης και το ενιαίο κατώφλι.

7.2.2.2.1 Κατώφλι για κάθε license term

Βάσει του μέσου όρου των επιτυχών εξαγωγών των όρων αδειών license term από τα κείμενα, τέθηκαν τα κατώφλια για τις ετικέτες και εκτελέστηκε η κατηγοριοποίηση των license term. Για την κάθε άδεια βρέθηκαν τα license terms που αντιστοιχούν σε αυτή βάσει των κατωφλίων και υπολογίστηκαν τα ποσοστά επιτυχίας (success rate), αυτή τη φορά με την εφαρμογή του threshold.

Άδεια	Success	Error	Total terms
AFL-3.0	0.75	0.31	16
AGPL-3.0	0.75	0	20
Apache-2.0	0.44	0.25	25
Artistic-2.0	0.06	0.12	17
BSD-2-Clause	0.07	0.14	14
BSD-3-Clause	0.07	0.13	15
CDDL-1.0	0.1	0.05	19
CPL-1.0	0.12	0.06	17
EPL-1.0	0.12	0.06	17
GPL-2.0	0.75	0.1	20
GPL-3.0	0.8	0	19
LGPL-2.1	0.74	0.16	19
LGPL-3.0	0	1	3
Libpng	0.12	0.06	17
MIT	0.13	0.07	15
MPL-1.1	0.4	0.11	19
MPL-2.0	0	0.1	18

MS-PL	0.11	0.05	18
MS-RL	0.1	0.05	19
OSL-3.0	0.1	0.05	20
X11	0.13	0.07	15
Zlib	0.12	0.06	17
Average	0.26	0.135	

Όπως φαίνεται τα αποτελέσματα δεν ήταν καθόλου ικανοποιητικά, οπότε προχωρήσαμε στον υπολογισμό ενός ενιαίου κατωφλίου για όλα τα license terms.

7.2.2.2.2 Ενιαίο threshold

Άδεια	Ποσοστό επιτυχίας	Σφάλμα	Σύνολο όρων
AFL-3.0	0.85	0.125	16
AGPL-3.0	0.85	0	20
Apache-2.0	0.49	0.08	25
Artistic-2.0	0.65	0.12	17
BSD-2-Clause	0.86	0.07	14
BSD-3-Clause	0.8	0.07	15
CDDL-1.0	0.58	0.11	19
CPL-1.0	0.6	0.18	17
EPL-1.0	0.6	0.18	17
GPL-2.0	0.85	0.05	20
GPL-3.0	0.9	0	19
LGPL-2.1	0.79	0.16	19
LGPL-3.0	0	4	3
Libpng	0.71	0.06	17
MIT	0.8	0.07	15
MPL-1.1	0.37	0	19
MPL-2.0	0	0	18

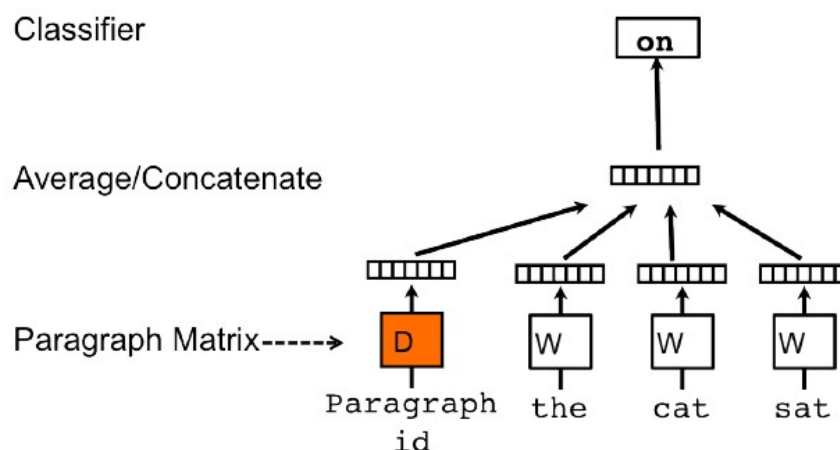
MS-PL	0.56	0.17	18
MS-RL	0.53	0.16	19
OSL-3.0	0.5	0.15	20
X11	0.8	0.07	15
Zlib	0.71	0.06	17
Average	0.61	0.28	

Σαφώς πιο αποτελεσματικό το ενιαίο κατώφλι. Αυτό όμως δεν σημαίνει ότι το 0.61 είναι το μέγιστο ποσοστό επιτυχίας που μπορούμε να φτάσουμε.

7.2.3 doc2vec

Είναι παρόμοιο με το word2vec, με την μόνη διαφορά ότι προσεγγίζει ολόκληρο το κείμενο ως διάνυσμα (vector) και το συσχετίζει με συγκεκριμένες ετικέτες (labels), στην περίπτωση μας τα license terms. Οπότε αντί να συσχετίζει λέξεις με άλλες λέξεις, συσχετίζει ετικέτες με λέξεις.

Κάθε παράγραφος του κειμένου συσχετίζεται με ένα μοναδικό διάνυσμα, και κάθε λέξη της επίσης αναπαρίστανται με ένα μοναδικό διάνυσμα. Το διάνυσμα που αντιστοιχεί στην παράγραφο διαμοιράζεται σε όλα τα περιεχόμενα της αλλά όχι σε άλλες παραγράφους, οι λέξεις όμως διαμοιράζονται σε όλες τις παραγράφους π.χ. το διάνυσμα για τη λέξη “powerful” είναι το ίδιο για όλες τις παραγράφους.



Σχήμα 23: Διάγραμμα Λειτουργίας doc2vec.

Ένα σημαντικό πλεονέκτημα στη χρήση του αλγόριθμου *doc2vec* [32] είναι ότι μαθαίνει μέσα από μη ετικετοποιημένα δεδομένα, οπότε μπορεί να δουλέψει καλά σε περιπτώσεις όπου τα ήδη ετικετοποιημένα δεδομένα είναι λίγα, όπως στη δική μας.

Σαν είσοδο λαμβάνει το *dataset* που δημιουργήσαμε, θεωρώντας ότι οι φάκελοι αποτελούν τις κατηγορίες στις οποίες θέλουμε να διαχωρίσουμε το κείμενο της άδειας μας, και χρησιμοποιεί τα περιεχόμενα του κάθε φακέλου για να υπολογίσει την πιθανότητα το κείμενο που εξετάζει να ανήκει σε αυτό.

Μέσω του *lookup table* που κατασκεύασε ο αλγόριθμος *doc2vec*, ο οποίος έχει καταχωρημένες τις κατηγορίες που θέλουμε να κάνουμε κατηγοριοποίηση, ελέγχεται ο μέσος όρος των λέξεων κάθε *vector* με αυτό των προτάσεων ανά κατηγορία και λαμβάνεται το *cosine similarity* των δύο, όπως γίνεται και στο *word2vec*. Ακολουθεί ένα παράδειγμα:

- Πρόταση από την κατηγορία *ProvidedWithoutWarranty*:

"THE SOFTWARE IS PROVIDED "AS-IS" AND WITHOUT WARRANTY OF ANY KIND, EXPRESS, IMPLIED OR OTHERWISE, INCLUDING WITHOUT LIMITATION, ANY WARRANTY OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE."

- Πρόταση από το κείμενο άδειας λογισμικού:

"THE AUTHORS AND DISTRIBUTORS SPECIFICALLY DISCLAIM ANY WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, AND NON-INFRINGEMENT."

Το σύστημα ελέγχει την κάθε λέξη βάσει του *doc2vec lookup table* και λαμβάνει ένα *vector*. Για κάθε λέξη της πρότασης υπολογίζεται ο μέσος όρος των *vector* και υπολογίζεται το *cosine similarity* των δύο *vector* για το κατά πόσο είναι όμοιες. Τα αποτελέσματα δεν ήταν τόσο καλά, με ποσοστό επιτυχίας 49%.

7.2.4 Συνδυασμός απο license terms

Κατά τη διάρκεια της μελέτης των αποτελεσμάτων της προηγούμενης προσέγγισης παρατηρήθηκε ότι η Γεωργία χρησιμοποίησε σε κάποιες περιπτώσεις την ίδια πρόταση από το κείμενο μιας άδειας για να δικαιολογήσει πολλαπλά license terms. Αυτό οδηγούσε τις προηγούμενες προσεγγίσεις στο να έχουν συγκρούσεις και να παράγουν λανθασμένα αποτελέσματα. Σε αυτή τη προσέγγιση έγινε προσπάθεια να μειωθεί αυτό το σφάλμα με το να βρεθούν όλοι οι συνδυασμοί από license terms που χρησιμοποιούν κοινές προτάσεις και να κατασκευαστούν ξεχωριστές κατηγορίες για αυτές. Ακολουθεί ένα παράδειγμα:

- Πρόταση από την κατηγορία *MaySublicense*.

"To use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software"

- Πρόταση από την κατηγορία *MayCopy*.

"To use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software"

- Πρόταση από την κατηγορία *MayMakeModifications*.

"To use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software"

Το πρόγραμμα βρίσκει αυτά τα τρία παραδείγματα και κατασκευάζει καινούργια κατηγορία με όνομα π.χ. *MayMakeModifications_MayCopy_MaySublicense* που περιείχε αυτή την πρόταση. Αφού τελείωσε την δημιουργία των καινούργιων κατηγοριών και την κατάταξη των προτάσεων σε αυτές, έγινε εξαγωγή του dataset

όπως και προηγουμένως, και εφαρμόστηκε ξανά στις προηγούμενες προσεγγίσεις. Τα αποτελέσματα αυξήθηκαν με το ποσοστό επιτυχίας να φτάνει το 71%.

7.2.5 Κατηγοριοποίηση βάσει άδειας

Για καλύτερη απόδοση της κατηγοριοποίησης έγινε μια προσπάθεια να μειωθεί ο αριθμός των πιθανών κατηγοριών. Αυτό έγινε με το να κατηγοριοποιηθεί το κείμενο της άδειας υπό έλεγχο σε μία από τις 22 αρχικές άδειες που επέλεξε η Γεωργία Χαραλάμπους. Αυτό θα βοηθούσε στο να ακολουθήσει η κατηγοριοποίηση σε μικρότερο όγκο από license terms εφόσον θα χρησιμοποιούσε αυτούς της πιο κοντινής άδειας από τις αρχικές.

Με την ίδια τακτική λοιπόν το σύστημα χωρίστηκε σε δύο μέρη κατηγοριοποίησης. Αρχικά δημιουργήθηκε ένα dataset με τις αρχικές άδειες και τα κείμενα τους, και μετά περνούσε το κείμενο από την κατηγοριοποίηση των license term της άδειας που προβλέφθηκε.

Τα αποτελέσματα αυξήθηκαν ακόμη περισσότερο εφόσον έφτασε το ποσοστό επιτυχίας στα 79%. Το θέμα είναι ότι εάν η άδεια υπό έλεγχο δε βρίσκεται κοντά σε καμία από τις 22 άδειες τότε τα αποτελέσματα μειώνονται σημαντικά.

Όνομα Αδειας Λογισμικού	Ποσοστό Επιτυχίας
Zlib	0.6875
X11	0.75
OSL-3.0	0.9375
MS-RL	0.875
MS-PL	0.875
MPL-2.0	1.0
MPL-1.1	1.0
MIT	0.75
Libpng	0.625
LGPL-3.0	0.0

LGPL-2.1	1.0
GPL-3.0	1.0
GPL-2.0	1.0
EPL-1.0	1.0
CPL-1.0	1.0
CDDL-1.0	0.9375
BSD-3-Clause	1.0
BSD-2-Clause	1.0
Artistic-2.0	0.6875
Apache-2.0	1.0
AGPL-3.0	1.0
AFL-3.0	1.0

7.2.6 Δυναμική κατηγοριοποίηση

Αυτή η προσέγγιση αφορά δυναμική κατηγοριοποίηση των κειμένων. Για παράδειγμα να γίνει έλεγχος για το εάν το κείμενο χαρακτηρίζεται από το MaySublicense, ή από το αντίθετο, που είναι το NotSublicense. Τα αποτελέσματα έδειξαν ότι η εφαρμογή του ήταν αποτυχημένη καθώς τα ποσοστά από τα license terms της Γεωργίας Χαραλάμπους, που δείχνουν ότι ισχύει το MaySublicense είναι κατά μεγάλο βαθμό λιγότερα από το NotSublicense, οπότε η κατηγοριοποίηση που έγινε δεν ήταν σωστή.

7.2.7 Χρήση Google vectors

Τα Google vectors αποτελούν ήδη μετατραπέντα vectors από την συλλογή των Google News, τα οποία έχουν 300 layers και μέγεθος 3.5 gigabytes. Με τη χρήση αυτών πιθανόν να είχαμε μεγαλύτερη σύνδεση στις λέξεις των κειμένων και να συμπληρώναμε την απώλεια όσον αφορά τα δεδομένα εισόδου, αλλά τα αποτελέσματα δεν άλλαξαν σημαντικά εφόσον δεν περιέχουν κείμενο για το συγκεκριμένο πεδίο των αδειών λογισμικού.

7.3 Αυτόματη εξαγωγή προτύπων

Για τους λόγους που αναφέρθηκαν στην αρχή, η καλύτερη λύση είναι να γίνει αυτόματη η εξαγωγή των features από τα κείμενα των αδειών. Θέλουμε λοιπόν να μη βασιζόμαστε πάνω στις προτάσεις της Γεωργίας Χαραλάμπους, εφόσον δεν καλύπτουν όλες τις περιπτώσεις που μπορεί το συγκεκριμένο license term να εκφραστεί.

Χωρίζοντας τα κείμενα των αδειών σε παραγράφους και προτάσεις και εισάγοντας τα σαν είσοδο σε διάφορα είδη νευρωνικών δικτύων, δοκίμασα να επιτύχω την εξαγωγή των features, με μόνο μία μικρή ένδειξη από τα license terms της Γεωργίας Χαραλάμπους. Λόγω όμως της έλλειψης γνώσης για τα πολύπλοκα αυτά συστήματα, ακολούθησα συγκεκριμένα ένα κομμάτι της αυτόματης εξαγωγής προτύπων, αυτό του topic modeling.

7.3.2 Topic Modeling

Το topic modeling λαμβάνει σαν είσοδο ένα κείμενο και εντοπίζει διάφορα μοτίβα που παρουσιάζονται με τη χρήση των λέξεων. Ένα topic αποτελεί, για τον ηλεκτρονικό υπολογιστή, μια λίστα από λέξεις που παρουσιάζονται μέσα στο κείμενο με κάποιο στατιστικό τρόπο. Είναι ένα στατιστικό μοντέλο για εύρεση των αφηρημένων θεμάτων που εμφανίζονται σε μια συλλογή αρχείων. Εργαλεία που εκτελούν αυτή τη δουλειά, αποσυνθέτουν το κείμενο που πέρασε σαν είσοδος σε “καλάθια” πιθανότητας από τα οποία προήλθαν οι λέξεις του κάθε topic. Το εργαλείο επαναλαμβάνει την συγκεκριμένη διαδικασία μέχρι να βρει την πιο πιθανή κατανομή των λέξεων μέσα στα “καλάθια”, τα οποία στο τέλος τα ονομάζουμε topics. Για την εφαρμογή πάνω στην ανάλυση των κειμένων των αδειών, χρησιμοποιήθηκε το εργαλείο MALLET, που αποτελεί εργαλείο για Natural Language Processing στη Java.

Σαν είσοδος δόθηκε ένα dataset που περιέχει αρχεία με την κάθε παράγραφο κάθε άδειας. Το εργαλείο μετατρέπει τα αρχεία σε μορφή (format) που το ίδιο μπορεί να

επεξεργαστεί, και λαμβάνει σαν είσοδο τον αριθμό των πιθανών topic που θέλουμε, καθώς επίσης και τον αριθμό των iterations. Για την επιλογή του συγκεκριμένου αριθμού έγιναν πολλές δοκιμές και καταλήξαμε στα 200 topics και 10 optimize intervals, λόγω εμπειρικών κανόνων και καλύτερων αποτελεσμάτων.

Υπάρχουν πολλοί αλγόριθμοι οι οποίοι χρησιμοποιούνται για topic modeling, όπως οι *lda – latent dirichlet allocation*, *lsi – latent semantic indexing*, *lsa – latent semantic analysis*, *plsa – probabilistic latent semantic analysis*, *hdp – hierarchical dirichlet process* και *ram – pachinko allocation model*. Τρέξαμε το σύστημα μας για τους περισσότερους από αυτούς, τους οποίους θα αναλύσουμε πιο κάτω, και καταλήξαμε ότι με τον LDA-Latent Dirichlet Allocation έχουμε τα καλύτερα αποτελέσματα.

7.3.3 Αλγόριθμοι για Topic Modeling

7.3.3.1 LSI Latent Semantic Indexing

Η ανάλυση Latent semantic analysis αποτελεί μια τεχνική στον τομέα του NLP, σχετικά με την ανάλυση σχέσεων μεταξύ ενός συνόλου απο κείμενα και των ορισμών που περιέχουν. Αυτή η ανάλυση επιτυγχάνεται μέσω της παραγωγής του συνόλου των θεμάτων που περιγράφουν τα συγκεκριμένα κείμενα και τους ορισμούς τους. Θεωρεί ότι λέξεις που είναι κοντά σημασιολογικά, θα επαναληφθούν σε παρόμοιες περιπτώσεις μέσα στα κείμενα. Δημιουργείται ένας πίνακας που περιέχει την κάθε λέξη σε αντιστοιχία με το πόσες φορές επαναλαμβάνεται μέσα σε κάθε παράγραφο, πάνω στον οποίο εφαρμόζεται η μαθηματική τεχνική με το όνομα Singular Value Decomposition (SVD).

Η Singular Value Decomposition χρησιμοποιείται για να μειώσει τον αριθμό των γραμμών καθώς διατηρεί την ομοιότητα μεταξύ των στηλών. Έπειτα οι λέξεις συγκρίνονται με βάσει το συνημίτονο της γωνίας μεταξύ των δύο διανυσμάτων τους, όπως είδαμε και στο word2vec. Τιμές κοντά στο 1 δείχνουν πολύ όμοιες λέξεις, ενώ 0

δείχνουν ότι δεν έχουν καμία σχέση μεταξύ τους.

Για την υλοποίηση του χρησιμοποιήθηκε το gensim [33], μια βιβλιοθήκη ανοικτού κώδικα στη Python για topic modeling.

7.3.3.2 HDP Hierarchical Dirichlet Process

Το HDP αποτελεί επέκταση του αλγόριθμου LDA, που σχεδιάστηκε για να αντιμετωπίσει την περίπτωση όπου ο αριθμός των topics δεν είναι γνωστός προηγουμένως. Χρησιμοποιώντας τον αλγόριθμο LDA για τη μοντελοποίηση του αρχείου (document modeling), το κάθε topic αποτελεί μια κατανομή λέξεων από ένα γνωστό λεξικό. Για κάθε document δημιουργείται ένα μίγμα από topics μέσω του Dirichlet distribution, και κάθε λέξη αποτελεί ξεχωριστή περίπτωση για το κάθε μίγμα. Για το HDP, χρησιμοποιείται το Dirichlet process για την ανάλυση του αριθμού των topics που χρειάζονται. Οπότε επιλέγεται μια βασική στατιστική κατανομή (base distribution), η οποία αντιπροσωπεύει τον άπειρο αριθμό από πιθανά topics, και μετά μια πεπερασμένη κατανομή των topics για κάθε αρχείο με βάση τη βασική κατανομή.

Για την υλοποίηση του χρησιμοποιήθηκε επίσης το gensim [33], μια βιβλιοθήκη ανοικτού κώδικα στη Python για topic modeling.

7.3.4 Αλγόριθμος LDA Latent Dirichlet Allocation

Αναπαριστά τα κείμενα σαν δέσμες από topics τα οποία έχουν, όπως αναφέραμε, “καλάθια” με λέξεις και τις πιθανότητες να βρεθούν αυτές. Θεωρεί ότι το κάθε κείμενο ακολουθεί τα πιο κάτω:

- Έχει προκαθορισμένο αριθμό λέξεων, και αυτό υπολογίζεται από μία κατανομή Poisson.

- Διαλέγει ένα συνδυασμό από topics (topic mixture) για το συγκεκριμένο κείμενο, βάσει της κατανομής dirichlet (dirichlet distribution) πάνω σε ένα σταθερό σύνολο από topics (στην περίπτωση μας 200).
- Δημιουργεί μετά κάθε λέξη από τα κείμενα, λαμβάνοντας ένα από τα topic mixtures που αναφέραμε προηγουμένως, και προσπαθεί να παράξει τη λέξη για το συγκεκριμένο.

Για μία συλλογή από κείμενα, προσπαθεί να κάνει αναζήτηση προς τα πίσω (backtrack) από αυτά για να βρει ένα σύνολο από topics τα οποία να δημιουργήσαν τη συγκεκριμένη συλλογή.

Το εργαλείο λοιπόν παρήγαγε τα topics, και ανάθεσε επίσης τα κείμενα στα αντίστοιχα topics τους. Οπότε έχουμε 200 topics και 33 license terms, χρειαζόμαστε λοιπόν ένα τρόπο αντιστοιχίας των δύο, έτσι ώστε να μπορούμε να πούμε πως εάν το κείμενο κατηγοριοποιείται σε συγκεκριμένο topic, τότε χαρακτηρίζεται από συγκεκριμένα license terms.

Για την συγκεκριμένη αντιστοίχιση ακολούθησα δύο προσεγγίσεις με δεδομένα τα topics που παρήγαγε το εργαλείο MALLET.

7.3.5 Κατανομή λέξεων κλειδιών

Σε αυτή την περίπτωση γίνεται έλεγχος για την ύπαρξη των λέξεων κλειδιών που δημιούργησε το topic modeling για το κάθε topic.

- Με το ίδιο σκεπτικό όπως προηγουμένως, λαμβάνουμε τη μοναδική ύπαρξη αυτών των λέξεων μέσα στα license terms της Γεωργίας, Χαραλάμπους υπολογίζοντας την κατανομή τους μέσα στην πρόταση αυτή.

- Οπότε και πάλι έχουμε αντιστοιχία για κάθε topic τα license terms μαζί με ένα ποσοστό που δείχνει την παρουσία των λέξεων του συγκεκριμένου topic για το συγκεκριμένο license term.
- Μετά λαμβάνουμε τις παραγράφους της κάθε άδειας και βρίσκουμε τα κοινά labels για τα topics τους. Δηλαδή υπολογίζουμε τις λέξεις κλειδιά (keywords) που παρουσιάζονται μέσα στην παράγραφο, και μέρος απο τα labels με το μεγαλύτερο score, χαρακτηρίζουν την παράγραφο. Για παράδειγμα:

Topic	License term	Score
0	ProvideWithoutWarranty	12
45	ProvideWithoutWarranty	9

Υπολογίζουμε για αυτά, το άθροισμα και μετά το μέσο όρο, δηλαδή στο παράδειγμα εφόσον υπάρχουν δύο εμφανίσεις, 10.5 είναι το αποτέλεσμα για το ProvideWithoutWarranty για τη συγκεκριμένη άδεια υπό έλεγχο.

- Εφόσον υπολογιστεί το αποτέλεσμα για το κάθε label της συγκεκριμένης άδειας, τότε υπολογίζεται ο μέσος όρος των αποτελεσμάτων αυτών, και το αποτέλεσμα χρησιμοποιείται σαν threshold για να επιλέξουμε τα license terms που αποτελούν την συγκεκριμένη άδεια.

Το αποτέλεσμα της συγκεκριμένης προσέγγισης ήταν 85% αλλά με πολλά license terms επιπλέον από τα βασικά, τα οποία δεν ισχύουν μετά από ανάλυση του κομματιού του κειμένου που τα δικαιολογεί.

7.3.6 Topics με μεγαλύτερο score (LDA)

Όπως και στην πρώτη προσέγγιση, γίνεται αντιστοίχιση των προτάσεων με τα license terms, στις παραγράφους του κάθε topic.

- Εφόσον μέσω του MALLET έχουμε το topic στο οποίο ανήκει η κάθε παράγραφος, δημιουργούμε ένα σύνολο που αποτελείται από τον αριθμό του topic και τις παραγράφους οι οποίες ανήκουν σε αυτό.
- Διατρέχουμε μετά τα license terms, και εάν υπάρχει αντιστοιχία μεταξύ των προτάσεων με τις παραγράφους, τότε αυξάνουμε τον counter του συγκεκριμένου license term για το συγκεκριμένο topic κατά ένα.
- Αφού διατρέξουμε όλα τα topics και όλα τα license terms, τότε για κάθε topic θα έχουμε μετρήσεις σχετικά με τις προτάσεις που εμφανίζονται στις παραγράφους. Το αποτέλεσμα έχει την εξής μορφή:

Topic 2
<i>MayDistributeOriginalWorkInObjectForm: (1,5)</i>
<i>MayDistributeDerivativeWorksInObjectForm: (1,5)</i>
<i>MustProvideInformationHowToGetSourceCode: (1,4)</i>
<i>MustRetainCopyrightNotice: (1,4)</i>
<i>MayPubliclyDisplay: (1,4)</i>

Θεωρούμε λοιπόν πιο πιθανόν για το topic με τον αριθμό 2 να μιλάει σχετικά με το *MayDistributeDerivativeWorksInObjectForm*, εφόσον είχαν το μεγαλύτερο score.

- Οπότε έτσι γίνεται η αντιστοίχιση των topics με τα license terms, οπότε ελέγχουμε την κάθε άδεια ανάλογα με τα topics που έλαβε κατά την κατηγοριοποίηση του MALLET, και εξάγουμε τα πιθανά license terms.

Με αυτό τον ακολουθούμε την δομή της ημι-επιβλεπόμενης μάθησης (semi-supervised learning) χρησιμοποιώντας αρχικά τα license terms της Γεωργίας, και μέσω της μέθοδου του topic modeling θα γίνει η εισαγωγή γενικευμένων προτάσεων που περιγράφουν το ίδιο topic.

Term	Topic
MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense	191

MayPubliclyPerform	10
MayDistributeDerivativeWorksInsourceForm	80, 84, 164, 69, 22, 10, 191
MayMakeDerivativeWorks	80, 144, 84, 164, 69, 22, 10, 191
MustRetainLicenseNotice	144, 98, 85, 174, 31
MayAcceptWarrantyOrAdditionalLiability	48, 106
MustRetainCopyrightNotice	82, 98, 83, 92, 174, 31
MayDistributeOriginalWorkInsourceForm	80, 51, 164, 69, 5, 8, 10, 31, 191
MayAddDifferentLicenseTermsToYourModifications	72
LimitedLiability	101, 29
NotPermissionToUseTrademarks	126
MayDistributeOriginalWorkInObjectForm	80, 98, 51, 164, 69, 117, 22, 104, 10, 191
MayPubliclyDisplay	164, 69, 5, 165, 8, 10, 44, 142, 80, 144, 51, 20, 126, 31, 191
MayMakeModifications	80, 144, 84, 164, 69, 22, 104, 10, 191
MaySublicense	80, 51, 164, 69, 5, 8, 10, 126, 191
MayPubliclyPerform	164, 69, 5, 165, 8, 10, 44, 142, 80, 144, 51, 20, 31, 191
MustMarkModifications	21, 142, 174
MustProvideInformationHowToGetsourceCode	98, 106, 175
MayGrantPatentsRightsFromTheContributorToReceipt	80, 51, 132, 5, 71, 8, 79
MustPublishAsLibraryTheModifiedVersion	22, 183
MustOffersourceCode	98, 183, 106, 191

NotSublicense	84, 183
MustRetainInSourceCodeOfDerivativeWorkAllTradeMarksPatentCo pyrightsOfOriginalWork	174
MustLicenseDerivativesWorksUnderCompatibleLicense	98, 72
NotPermissionToUseTradeMarks	80, 135, 174, 126
MayUseInHouse	164, 69, 10, 142
MustDistributeCodeUnderThisLicense	98, 84, 183, 55, 142, 111
MayDistributeDerivativeWorksInObjectForm	80, 98, 164, 69, 117, 22, 10, 191
MayCopy	164, 69, 10, 31
MustDistributeCopyOfNoticeText	149, 174
MayAddYourOwnCopyrightStatementToYourModifications	72
MustNotMisrepresentTheOriginalAuthorOfSoftware	119
ProvideWithoutWarranty	32, 112, 101, 74, 141, 61, 46

Τα αποτελέσματα ήταν πολύ καλά, με 91% επιτυχία καθώς επίσης και την εύρεση license terms που η Γεωργία Χαραλάμπους δεν είχε προσέξει τα οποία ήταν ορθά μετά από ανάλυση του κειμένου τους. Εφόσον λοιπόν τα αποτελέσματα όσον αφορά την εξαγωγή των license terms, αλλά και την εύρεση του ορθού κομματιού του κειμένου που τα δικαιολογεί, είναι καλύτερα από τους υπόλοιπους αλγόριθμους, συνεχίζουμε την υλοποίηση του συστήματος μας βάσει τον *LDA*.

Κεφάλαιο 8

Ανάλυση Αποτελεσμάτων

8.1 Σύγκριση Προσεγγίσεων	96
8.1.1 Προσέγγιση προπαρασκευασμένων προτύπων	97
8.1.2 Προσέγγιση Topic Modeling	98
8.1.3 Σύγκριση των δύο προσεγγίσεων	100
8.2 Σύγκριση με προκατασκευασμένα δεδομένα	100

8.1 Σύγκριση Προσεγγίσεων

Ακολουθεί σύγκριση μεταξύ των δύο αρχικών βασικών προσεγγίσεων για την εξαγωγή των ορισμών αδειών λογισμικού ανοικτού κώδικα από τα κείμενα αδειών. Πρώτα με τη χρήση προπαρασκευασμένων προτύπων και μετά με την αυτόματη εξαγωγή τους μέσω μιας ημι-επιβλεπόμενης μάθησης (semi-supervised learning).

Κατά την ανάλυση των αποτελεσμάτων χρησιμοποιούνται ορισμοί όπως το ποσοστό επιτυχίας (success rate) και το σφάλμα (error). Ακολουθούν ορισμοί σε αντιστοιχεία με το πώς εννοούνται στην δική μας περίπτωση.

- **Success rate:**

Αναφέρεται στο ποσοστό των license terms που βρήκε το σύστημα σε σχέση με αυτά της Γεωργίας Χαραλάμπους. Οπότε εάν το σύστημα βρει κάποιο license term το οποίο η Γεωργία Χαραλάμπους δεν είχε εντοπίσει, τότε θα το θεωρήσει λάθος. Για αυτό το λόγο θα ακολουθήσει ως κομμάτι της διαδικασίας αξιολόγησης μια στατική ανάλυση μεταξύ των αποτελεσμάτων του συστήματος και του κειμένου κάποιων πρόσθετων αδειών για να δούμε κατά πόσο μπορεί να γίνει κάποια γενίκευση των αποτελεσμάτων του αυτοματοποιημένου συστήματος.

- **Error:**

Αναφέρεται στο ποσοστό των license terms που βρήκε το σύστημα, τα οποία ήταν εκτός αυτών που βρήκε η Γεωργία Χαραλάμπους, σε σχέση με όλα τα terms που εξήαθησαν. Οπότε εάν, όπως αναφέρθηκε πιο πάνω, το σύστημα βρει κάποια license terms που η Γεωργία Χαραλάμπους δεν είχε εντοπίσει, τότε το σφάλμα θα είναι προφανώς υψηλότερο.

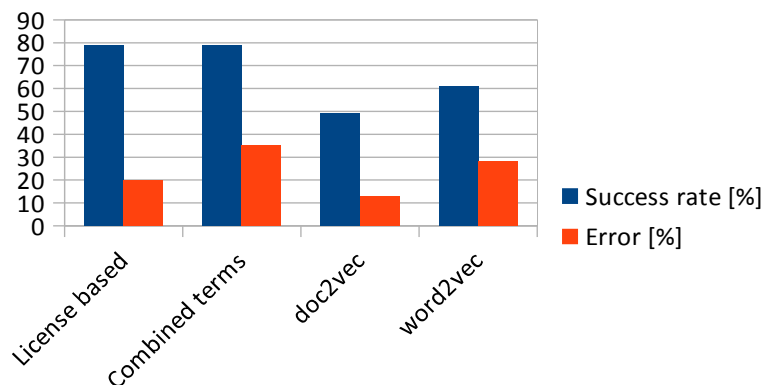
Για τον έλεγχο των αποτελεσμάτων χρησιμοποιήθηκαν, πέραν από τις 22 αρχικές άδειες, ακόμα άλλες 11 που είχε αναλύσει η Γεωργία Χαραλάμπους και τα φυλάξαμε για δεδομένα ελέγχου της ικανότητας του συστήματος μας για γενίκευση.

8.1.1 Προσέγγιση προπαρασκευασμένων πρότυπων

Βάσει των ποσοστών επιτυχίας (success rate) και του σφάλματος (error) για την κάθε μέθοδο, όπως περιγράφηκαν πιο πάνω, η πιο αποδοτική ήταν αυτή κατά την οποία κατηγοριοποιούμε πρώτα με την κοντινότερη άδεια από τις αρχικές 22, και μετά στον μειωμένο πλέον αριθμό license terms. Το success rate ήταν 79% και το error 20%.

Το success rate ήταν το ίδιο με αυτό που πήραμε από την μέθοδο με τον συνδυασμό των κοινών license terms για την αποφυγή διπλότυπων προτάσεων, αλλά το error ήταν αυτό που έκανε τη διαφορά, με την μέθοδο *combined terms* να έχει 35%, ενώ η κατηγοριοποίηση βάσει άδειας να έχει 20%.

Ακολουθούν γραφικές παραστάσεις που δεικνύουν τα συγκεκριμένα αποτελέσματα.

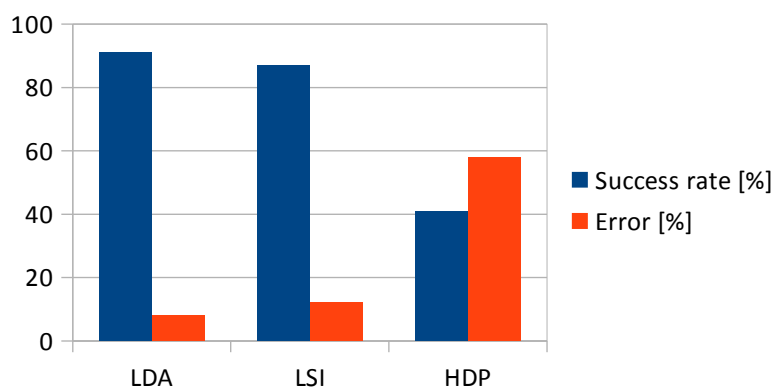


Σχήμα 24: Γραφική Παράσταση Αποτελεσμάτων με τη Χειροκίνητη Προσέγγιση.

8.1.2 Προσέγγιση Topic Modeling

Όπως και στην προηγούμενη προσέγγιση, έτσι και εδώ υπολογίζουμε την μέθοδο με το μεγαλύτερο success rate. Στην περίπτωση του topic modeling, έγινε χρήση των αλγορίθμων lda, hdp, lsi με τον lda να δίνει τα καλύτερα αποτελέσματα. Το success rate του lda βρίσκεται στο 91% με το error στο 8%.

Ακολουθούν γραφικές παραστάσεις που δείχνουν τα συγκεκριμένα αποτελέσματα.



Σχήμα 25: Γραφική Παράσταση με τα Αποτελέσματα Αυτόματης Εξαγωγής.

<u>License</u>	<u>LDA</u>	<u>LSI</u>	<u>HDP</u>
----------------	------------	------------	------------

	Success rate	Error	Success rate	Error	Success rate	Error
<i>MS-PL</i>	94.44	5.5	88.88	11.11	22.22	77.7
<i>Zlib</i>	82.35	17.64	82.35	17.64	17.64	82.35
<i>MS-RL</i>	94.7	5.26	89.47	10.52	26.31	73.7
<i>OSL-3.0</i>	90.0	10.0	90.0	10.0	40.0	60.0
<i>Libpng</i>	76.47	23.52	82.35	17.64	35.29	64.7
<i>AFL-3.0</i>	81.25	18.75	81.25	18.75	37.5	62.5
<i>AGPL-3.0</i>	90.0	10.0	90.0	10.0	70.0	30.0
<i>LGPL-2.1</i>	94.73	5.26	94.73	5.26	57.9	42.1
<i>BSD-2- Clause</i>	92.85	7.14	92.85	7.14	28.57	71.42
<i>BSD-3- Clause</i>	93.33	6.66	86.66	13.33	26.66	73.33
<i>GPL-2.0</i>	95.0	5.0	95.0	5.0	60.0	40.0
<i>GPL-3.0</i>	94.73	5.26	94.73	5.26	73.68	26.31
<i>CDDL-1.0</i>	94.73	5.26	84.21	15.78	52.63	47.36
<i>CPL-1.0</i>	94.11	5.8	88.23	11.76	58.82	41.17
<i>Artistic-2.0</i>	82.35	17.64	76.47	23.52	52.94	47.05
<i>EPL-1.0</i>	94.11	5.8	88.23	11.76	47.05	52.94
<i>MPL-1.1</i>	94.73	5.26	84.21	15.78	57.89	42.10
<i>MPL-2.0</i>	94.44	5.5	88.88	11.11	50.0	50.0
<i>Apache-2.0</i>	92.0	8.0	92.0	8.0	36.0	64.0
<i>X11</i>	93.33	6.66	86.66	13.33	20.0	80.0
<i>MIT</i>	93.33	6.66	86.66	13.33	6.66	93.3

8.1.3 Σύγκριση των δύο προσεγγίσεων

Βάσει των πιο πάνω αποτελεσμάτων έγινε σύγκριση μεταξύ των δύο αποδοτικότερων μεθόδων από τις δύο προσεγγίσεις.

License classification	
Success rate	79%
Error rate	20%
Χαρ/κα	Πολύ καλά αποτελέσματα, αλλά η γενίκευση δεν είναι και τόσο αποδοτική. Με είσοδο αδειας εκτός των αρχικών 26 γίνεται λανθασμένη κατηγοριοποίηση και χάνονται πολλοί ορισμοί.

Top-score topics [LDA]	
Success rate	91%
Error rate	8%
Χαρ/κα	Πολύ καλά αποτελέσματα, πολύ καλή γενίκευση. Συχνά προβλήματα στην δικαιολόγηση βάσει κειμένου, η οποία όμως διορθώθηκε μέσω κώδικα.

Οπότε βλέπουμε ότι η χρήση της δεύτερης προσέγγισης με τη μέθοδο του *Top-score topics* μας δίνει πιο καλά αποτελέσματα και μειωμένο error rate. Όσον αφορά το πρόβλημα με τα κομμάτια του κειμένου, διορθώθηκε με την παρουσίαση όλων των πιθανών κομματιών, έτσι ώστε ο χρήστης να μπορεί να τα συμβουλευτεί όλα σχετικά με την παρουσία του license term.

8.2 Σύγκριση με προκατασκευασμένα δεδομένα

Για τον έλεγχο της γενίκευσης του συστήματος, έγινε ανάλυση δια χειρός των εξής αδειών: 0BSD – Zero Berkley Software Distribution, AAL – Attribution Assurance License, APAFML – Adobe Postscript FM License, Caldera, Curl, Sleepycat, VOSTROM και Xerox. Συμπληρώθηκε μια φόρμα με τους ορισμούς που περιέχει το κείμενο της κάθε άδειας και τη γραμμή του κειμένου στην οποία παρουσιάζεται. Ακολουθεί ένα παράδειγμα για την 0BSD, τα υπολοιπα μπορούν να βρεθούν στο Παραρτημα Β.

<u>License term</u>	<u>Exists</u>	<u>Line</u>
---------------------	---------------	-------------

LimitedLiability		
MayAcceptWarrantyOrAdditionalLiability		
MayAddDifferentLicenseTermsToYourModifications	✓	1
MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense		
MayAddYourOwnCopyrightStatementToYourModifications	✓	1
MayCopy	✓	1
MayDistributeDerivativeWorksInObjectForm	✓	1
MayDistributeDerivativeWorksInSourceForm	✓	1
MayDistributeOriginalWorkInObjectForm	✓	1
MayDistributeOriginalWorkInSourceForm	✓	1
MayGrantPatentsRightsFromTheContributorToReceipt		
MayMakeDerivativeWorks	✓	1
MayMakeModifications	✓	1
MayPubliclyDisplay	✓	1
MayPubliclyPerform	✓	1
MaySublicense	✓	1
MayUseInHouse	✓	1
MustDistributeCodeUnderThisLicense		
MustDistributeCopyOfNoticeText		
MustLicenseDerivativesWorksUnderCompatibleLicense		
MustMarkModifications		
MustNotMisrepresentTheOriginalAuthorOfSoftware		
MustOfferSourceCode		
MustProvideInformationHowToGetSourceCode		
MustPublishAsLibraryTheModifiedVersion		
MustRetainCopyrightNotice		
MustRetainInSourceCodeOfDerivativeWorkAllTradeMarksPatent...		
MustRetainLicenseNotice		
NotPermissionToUseTradeMarks		
NotSublicense		
ProvideWithoutWarranty	✓	2

Μετάπειτα ακολούθησε σύγκριση των αποτελεσμάτων της ανάλυσης μου με αυτά του συστήματος, ενώ έχουν γίνει και κάποιες σχετικές παρατηρήσεις για την εξήγηση της σύγκρισης.

<u>License</u>	<u>Success Rate</u>	<u>Error</u>	<u>Error Χωρίς Συνώνυμα</u>	<u>Παρατηρήσεις</u>
OBSD	11/13	3/13	3/13	Εμφάνισε επιπλέον το <i>MustOfferSourceCode</i> το οποίο δεν αναφέρεται πουθενά μέσα στο κείμενο της άδειας. Επίσης δεν βρήκε το <i>MayUseInHouse</i> .
AAL	12/15	5/15	1/13	Δεν βρέθηκαν τα <i>NotPermissionToUseTradeMarks</i> και το <i>MustNotMissrepresentOriginalAuthorOfSoftware</i> . Επίσης βρέθηκε επιπλέον το <i>MustOfferSourceCode</i> .
APAFML	5/12	8/12	8/12	Εδώ βλέπουμε ότι έχουμε σοβαρό πρόβλημα. Δοκιμάστηκε το συγκεκριμένο κείμενο και με τις υπόλοιπες προσεγγίσεις, αλλά δεν παρουσίαζαν άλλο αποτέλεσμα πέραν του <i>ProvidedWithoutWarranty</i> .
Caldera	12/15	5/15	5/15	Το <i>MustNotMissrepresentOriginalAuthorOfSoftware</i> ήταν επιπλέον αλλά μετά από έλεγχο του κειμένου που το δικαιολογεί ήταν σωστό και μου είχε διαφύγει. Λάθος όμως εμφάνιση του <i>MaySublicense</i> .
Curl	5/13	12/13	10/13	Βρέθηκαν μόνο τα δικαιώματα <i>MayDistributeOriginalWorkInSourceForm</i> και <i>MayDistributeDerivativeWorkInSourceForm</i> από τα οκτώ. Επίσης βρέθηκε επιπλέον το <i>MustProvideInformationHowToGetSourceCode</i> .
Sleepycat	12/12	2/12	2/12	Το <i>MustOfferSourceCode</i> ήταν επιπλέον αλλά μετά από έλεγχο του κειμένου που το

				δικαιολογεί ήταν σωστό και μου είχε διαφύγει. Λάθος όμως εμφάνιση του <i>MaySublicense</i> .
VOSTROM	16/19	6/19	5/19	Επιπλέον εμφάνιση του <i>MustPublishAsLibraryTheModifiedVersion</i> η οποία δεν δικαιολογείται μέσω του κειμένου.
Xerox	5/9	5/9	5/9	Επιπλέον η εμφάνιση του <i>MayDistributeDerivativeWorksInObjectForm</i> καθώς και απώλεια του <i>MustRetainCopyrightNotice</i> .

Βλέπουμε λοιπόν ότι υπάρχουν είδη κειμένων με τα οποία το σύστημα μας δυσκολεύεται να εκτελέσει την ανάλυση και παρουσιάζει λανθασμένα αποτελέσματα. Το πρόβλημα βρίσκεται στο μεγάλο ποσοστό θορύβου μέσα στα κείμενα το οποίο δεν μπόρεσε να εντοπισθεί, καθώς επίσης και στον μειωμένο αριθμό δεδομένων που έχουμε για εκπαίδευση.

Παραδείγματα τέτοιων θορύβων είναι:

- Επεξηγήση ορισμών μέσα στο κείμενο, π.χ. *"Original Author" means the individual or entity who created the Work.*
- Καποιοι τίτλοι π.χ. *There is no standard license header for the license.*
- Ξένες γλώσσες.
- Η έκδοση της συγκεκριμένης άδειας π.χ. *LPPL Version 1.2 1999-09-03*

Για τη διευκόλυνση του χρήστη, θα παρουσιάζουμε τα μέρη του κειμένου με τα οποία δικαιολογείται το συγκεκριμένο license term που εμφανίζει το σύστημα μας, έτσι ώστε και ο ίδιος να αποφασίζει κατά πόσον είναι ορθό. Η δυνατότητα αυτή δόθηκε στο σύστημα που υλοποιήθηκε για τη χρήση της προσέγγισης εξαγωγής ορών αδειών που περιγράφεται στο επόμενο Κεφάλαιο.

Ακόμη μία παρατήρηση κατά τον έλεγχο των αποτελεσμάτων, ήταν ότι μερικοί από τους ορισμούς αδειών λογισμικού ανοικτού κώδικα βρίσκονται πολύ κοντά σημασιολογικά, καθώς επίσης και το περιεχόμενο της πρότασης που τους

χαρακτηρίζει είναι αρκετά όμοιο. Αυτό οδηγεί το σύστημα να παρουσιάζει κάποιο όμοιο ορισμό αντί του αντίστοιχου που δηλώθηκε δια χειρός. Τα αποτελέσματα, λαμβάνοντας υπόψη τους ορισμούς που είναι συνώνυμοι, βρίσκονται στη στήλη του πιο πάνω πίνακα με το όνομα *Error Χωρίς Συνώνυμα*. Για παράδειγμα στην άδεια λογισμικού AAL, αντί του *NotPermissionToUseTradeMarks* και του *MustNotMisrepresentOriginalAuthorOfSoftware*, βρέθηκε το *MustRetainCopyrightNotice*, το οποίο δικαιολογείται από κείμενο που περιγράφει επίσης τους δύο ορισμούς που παραλείφθηκαν. Αυτό αποτελεί ένα κομμάτι που πρέπει να ληφθεί υπόψη κατά την λειτουργία του συστήματος.

Κεφάλαιο 9

Υλοποίηση REST Υπηρεσίας και Εφαρμογής Διαδικτύου

9.1 Εισαγωγή	107
9.2 Περιγραφή γλωσσών προγραμματισμού και συστατικών στοιχείων	107
9.2.1 Jersey	107
9.2.2 JSON	107
9.2.3 Bower	108
9.2.4 Font Awesome	108
9.2.5 jQuery	109
9.2.6 Bootstrap	109
9.3 Υλοποίηση υπηρεσίας REST	109
9.3.1 Εισαγωγή	109
9.3.2 Αρχιτεκτονική υπηρεσίας REST	110
9.4 Υλοποίηση διαδικτυακής εφαρμογής	111
9.4.1 Εισαγωγή	111
9.4.2 Υπηρεσία	112
9.4.3 Εισαγωγή	115
9.4.4 Πηγές	115
9.4.5 Οδηγίες χρήσης	116
9.4.6 API	117
9.4.7 Επικοινωνία	118

9.1 Εισαγωγή

Αφού λάβαμε κάποια ικανοποιητικά αποτελέσματα, θα προχωρήσουμε στην κατασκευή μιας διαδικτυακής εφαρμογής η οποία υποστηρίζεται από μια RESTful υπηρεσία διαδικτύου για την ανάλυση κειμένων αδειών λογισμικού ανοικτού κώδικα από τους χρήστες που το επιθυμούν. Με αυτό τον τρόπο θα μπορεί να αναλυθεί περαιτέρω για τα αποτελέσματα που δίνει το σύστημα μας καθώς να επεκταθούν οι μελέτες που μπορούν να γίνουν πάνω στο συγκεκριμένο θέμα.

Θα ακολουθήσει περιγραφή των συστατικών στοιχείων που θα χρησιμοποιηθούν και πιο κάτω η διαδικασία υλοποίησης και έκδοσης της υπηρεσίας και της διαδικτυακής εφαρμογής.

9.2 Περιγραφή γλωσσών προγραμματισμού και συστατικών στοιχείων

9.2.1 Jersey

Αποτελεί ένα προγραμματιστικό πλαίσιο ανοικτού κώδικα για ανάπτυξη υπηρεσιών REST σε Java. Παρέχει υποστήριξη για JAX-RS, το οποίο με τη σειρά του αποτελεί ένα API για τη κατασκευή των συγκεκριμένων υπηρεσιών με τη χρήση annotations για απλοποίηση της ανάπτυξης και της έκδοσης της υπηρεσίας. Κάποια παραδείγματα από annotations, τα οποία χρησιμοποιήθηκαν και στην ανάπτυξη της δικής μας υπηρεσίας, είναι:

- **@Path**: δηλώνει το μονοπάτι στο οποίο μπορεί να βρεθεί κάποιο resource.
- **@GET, @POST**: καθορίζει τον τύπο του HTTP request για συγκεκριμένο resource.
- **@Produces**: καθορίζει τον τύπο του response βάσει Internet media types.

- **@Consumes:** καθορίζει τον τύπο των request για συγκεκριμένο resource.

9.2.2 JSON

Αντιστοιχεί με *JavaScript Object Annotation* και αποτελεί ένα τύπο δεδομένων ανεξαρτήτου γλώσσας προγραμματισμού. Το Internet Media Type στο οποίο αντιστοιχεί είναι *application/json* και το filename extension είναι *.json*. Αποτελεί την κύρια μέθοδο που χρησιμοποιούμε στο σύστημα μας για λήψη και αποστολή δεδομένων. Ακολουθεί ένα παράδειγμα:

```
{
    "license": "Example",
    "obligation": ["MustOfferSourceCode", "NotSublicense"],
    "additional_condition": ["ProvideWithoutWarranty"],
    "right":
["MayPubliclyPerfrom", "MaySublicense", "MayMakeModifications"]
}
```

9.2.3 Bower

Αποτελεί έναν package manager για το συστατικά στοιχεία τα οποία περιέχουν HTML, CSS και JavaScript. Μέσω ενός αρχείου ρυθμίσεων αναλαμβάνει να βρίσκει και να κατεβάζει τις συγκεκριμένες εκδόσεις βιβλιοθηκών που εξαρτώνται από το σύστημα μας για να λειτουργήσει. Τα τοποθετεί σε καθορισμένο φάκελο για να είναι προσβάσιμα από το πρόγραμμα και διευκολύνει έτσι την ροή εργασίας. Με αυτό τον τρόπο μειώνουμε το φόρτο εργασίας καθώς επίσης το φόρτο πληροφορίας που φυλάγεται μέσα στο αποθετήριο που χρησιμοποιούμε.

9.2.4 Font Awesome

Είναι μια βιβλιοθήκη που σου προσφέρει εικονίδια πολλαπλής χρήσης τύπου `svg-scalable vector graphics`, τα οποία διατηρούν την ποιότητα τους σε όποιου είδους συσκευής παρουσιάζονται. Χρησιμοποιήθηκαν στο σύστημα μας για την καλαισθησία του και την διατήρηση της συνέχειας όσον αφορά την γραφική του παρουσία.

9.2.5 jQuery

Αποτελεί μια γρήγορη, απλή και πολύ ικανή βιβλιοθήκη JavaScript. Μετατρέπει την επεξεργασία ενός HTML document, τη διαχείριση γεγονότων, την κίνηση και τα AJAX request σε κάτι το πολύ απλό, μέσω του πλούσιου API. Χρησιμοποιείται σχεδόν παντού από το σύστημα μας, κυρίως όμως στην εκτέλεση των request προς την υπηρεσία ανάλυσης του κειμένου.

9.2.6 Bootstrap

Είναι ένα από τα διασημότερα έργα στο Github καθώς επίσης και πλαίσιο εργασίας για responsive ανάπτυξη εφαρμογών διαδικτύου. Responsive ανάπτυξη αναφέρεται στην ανάπτυξη εφαρμογών διαδικτύου που στοχεύουν στην χρήση όχι μόνο από υπολογιστές γραφείου και φορητούς υπολογιστές, αλλά και από tablets και κινητές εφαρμογές. Αναλαμβάνει λοιπόν με εύκολο τρόπο την ανάπτυξη μίας μόνο εφαρμογής, η οποία να προσαρμόζεται σε οποιαδήποτε συσκευή.

9.3 Υλοποίηση υπηρεσίας REST

9.3.1 Εισαγωγή

Για να δοθεί η δυνατότητα σε πιθανούς χρήστες να χρησιμοποιήσουν το σύστημα μας, καθώς επίσης να κάνουμε εύκολη τη χρήση του από την εφαρμογή διαδικτύου που θα υλοποιηθεί, έγινε κατασκευή ενός REST API.

Για την σύνταξη του και τον τοπικό έλεγχο χρησιμοποιήθηκε το IntelliJ IDE με τον

Apache Tomcat 7.0 σαν εξυπηρετητή. Επίσης για διαχειριστή των dependencies χρησιμοποιήθηκε το maven.

Μέσω της συγκεκριμένης υπηρεσίας, ο χρήστης θα μπορεί να εκτελέσει την ανάλυση κειμένου μέσω τοπικού αρχείου, μέσω URL, η γράφοντας το με το πληκτρολόγιο. Επίσης θα μπορεί να εφαρμόσει τους διαφορετικούς αλγόριθμους για την ανάλυση των αποτελεσμάτων. Εκτός από θέματα ανάλυσης, θα μπορεί να δημιουργήσει το μοντέλο άδειας λογισμικού ανοικτού κώδικα βάσει των αποτελεσμάτων, όπως το περιγράψαμε σε προηγούμενο κεφάλαιο, θα μπορεί να λάβει τα βασικά topics που χρησιμοποιούμε είτε σε απλό κείμενο, είτε σε μορφή spreadsheet.

9.3.2 Αρχιτεκτονική υπηρεσίας REST

Οι λειτουργίες που περιγράφηκαν πιο πάνω αντιστοιχούν σε URLs πάνω στα οποία θα εκτελεστούν συγκεκριμένα HTTP requests, είτε POST, είτε GET, και θα δώσουν την ανάλογη απάντηση. Για κύριο τρόπο επικοινωνίας στο API χρησιμοποιείται το JSON format, εκτός από τις περιπτώσεις της παραγωγής του μοντέλου άδειας λογισμικού ανοικτού κώδικα, όπου η απάντηση είναι σε .ttl, και στην λήψη των topics είναι σε απλό κείμενο, είτε σε spreadsheet.

Η υπηρεσία λαμβάνει μέσω των αιτημάτων (requests) το κείμενο της άδειας λογισμικού, και το περνάει μέσα από το topic mapper. Το topic mapper αποτελεί ένα

Σχήμα 26: Αρχιτεκτονική REST Υπηρεσίας.

serialized αντικείμενο που κάνει την αντιστοιχία των προτάσεων του κειμένου σε topics, βάσει όπως είπαμε του cosine similarity που υπολογίζεται μέσω του doc2vec μοντέλου. Ανάλογα με το URL στο οποίο εκτελείται το αίτημα, θα παραχθεί το αντίστοιχο αποτέλεσμα.

Ακολουθεί πίνακας με τα βασικά URL, τον τύπο του request, καθώς και με επεξήγηση για το καθένα:

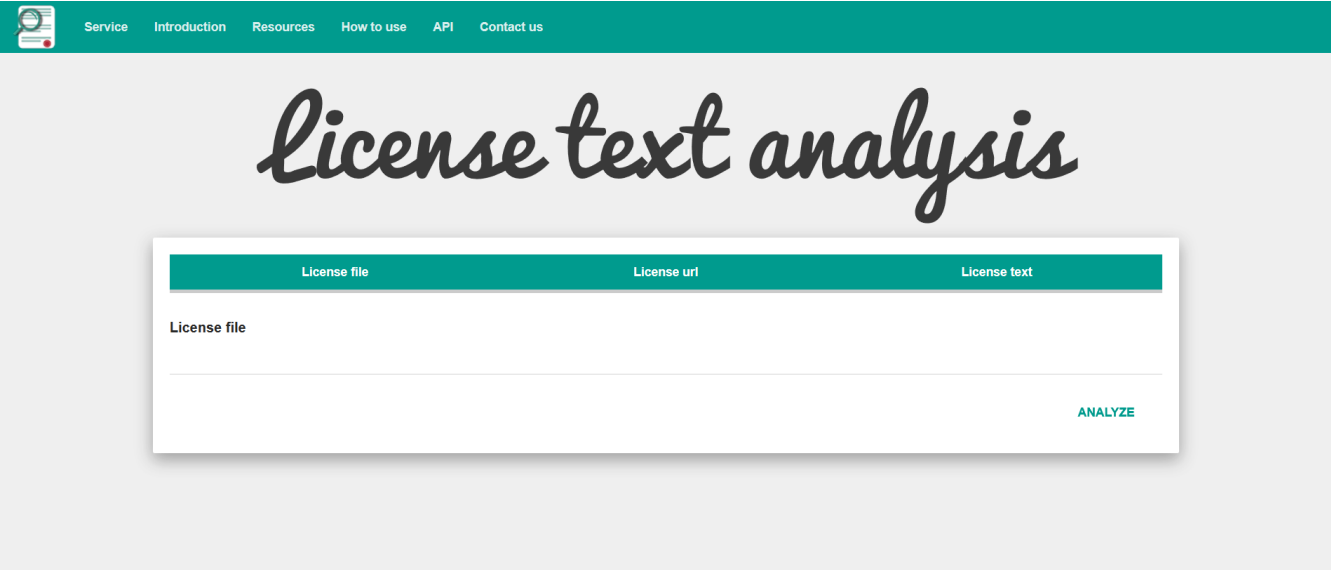
<u>Λειτουργία</u>	<u>Method</u>	<u>URL</u>
Εξαγωγή μόνο των license terms.	@POST	<i>/license/analyze</i>
Εξαγωγή των license terms με το score τους.	@POST	<i>/license/score</i>
Εξαγωγή των license terms με τα αντίστοιχα κομμάτια του κειμένου.	@POST	<i>/license/text/multiple</i>
Ανάλυση άδειας μέσω URL.	@POST	<i>/license/analyze/url</i>
Ανάλυση άδειας μέσω πληκτρολογίου.	@POST	<i>/license/analyze/text</i>
Ανάλυση με τη χρήση του αλγόριθμου LSI.	@POST	<i>/license/analyze/text/lsi</i>
Ανάλυση με τη χρήση του αλγόριθμου HDP.	@POST	<i>/license/analyze/text/hdp</i>
Λήψη κειμένου με τα topics.	@GET	<i>/resources/raw</i>
Λήψη spreadsheet με τα topics.	@GET	<i>/resources/excel</i>
Δημιουργία και λήψη του μοντέλου.	@GET	<i>/resources/generate/model</i>

9.4 Υλοποίηση διαδικτυακής εφαρμογής

9.4.1 Εισαγωγή

Η δημιουργία μιας διαδικτυακής εφαρμογής, και η έκδοση της αποτελεί σημαντικό παράγοντα στην εξάπλωση του συστήματος προς χρήση. Με αυτό τον τρόπο θα μπορεί να χρησιμοποιηθεί από διάφορους χρήστες, καθώς και να αποτελέσει τρόπο εφαρμογής του REST API.

Η συγκεκριμένη εφαρμογή αποτελεί μία μόνο σελίδα χωρισμένη σε έξι μέρη, στα οποία ο χρήστης μπορεί να μεταβεί μέσω του μενού πλοήγησης, είτε με τη χρήση του ποντικιού. Για την κύρια χρήση του συστήματος υπάρχει ένα στοιχείο μέσω του οποίου μπορεί να κάνει ανάλυση κειμένου, αρχείου είτε URL. Τα υπόλοιπα μέρη είναι για σχετικές πληροφορίες. Με αυτό τον τρόπο δείχνουμε ότι το σύστημα είναι εύκολο στη χρήση και επίσης μπορεί να χρησιμοποιηθεί και από κινητές εφαρμογές.



License file	License url	License text
License file ANALYZE		

Σχήμα 27: OSS License Analysis Homepage

9.4.1 Υπηρεσία

Στη συγκεκριμένη σελίδα εμφανίζονται απευθείας οι τρεις τρόποι με τους οποίους μπορεί να εκτελεστεί η ανάλυση. Επιλέγοντας διαφορετικό tab το σύστημα με το πάτημα του κουμπιού “ANALYZE”, θα κάνει την αντίστοιχη αίτηση στο API. Ακολουθεί παράδειγμα εκτέλεσης για το αρχείο της άδειας OBSD.



Σχήμα 28: OSS License Analysis Results

Πατώντας το εικονίδιο που αντιστοιχεί σε μάτι, παρουσιάζεται ένα πλαίσιο με την επεξήγηση του ορισμού που αντιστοιχεί, καθώς και τα κομμάτια του κειμένου τα οποία μπορεί ο χρήστης να συμβουλευθεί για να σιγουρέψει την ύπαρξη του.

Definition of "ProvideWithoutWarranty"



Definition

"The original work provided, comes under no warranty."

Possible parts of text

"THE SOFTWARE IS PROVIDED "AS IS" AND THE AUTHOR DISCLAIMS ALL WARRANTIES WITH REGARD TO THIS SOFTWARE INCLUDING ALL IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS. IN NO EVENT SHALL THE AUTHOR BE LIABLE FOR ANY SPECIAL, DIRECT, INDIRECT, OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF THIS SOFTWARE."

CLOSE

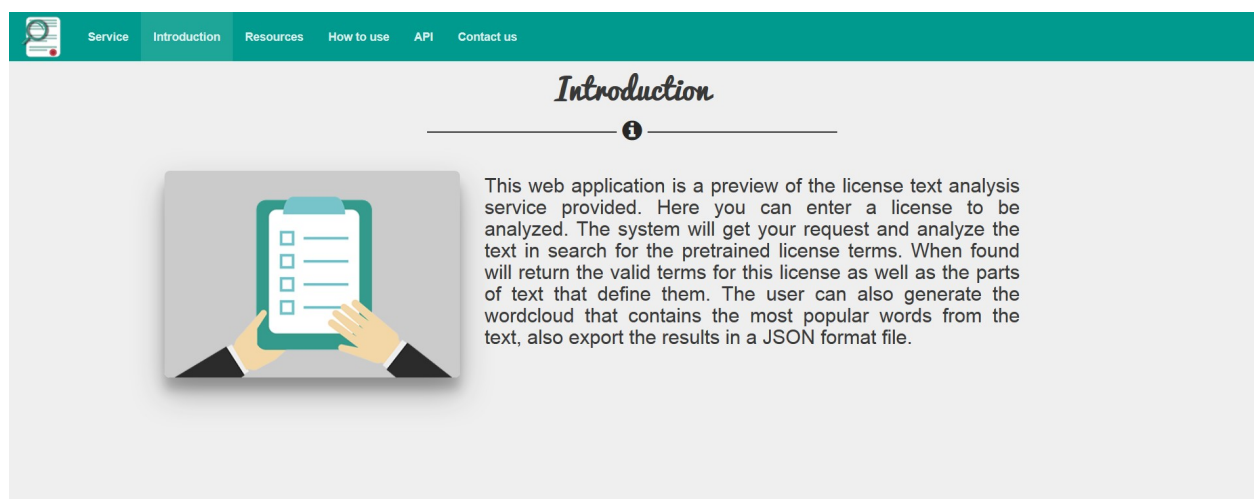
Σχήμα 29: OSS License Analysis License Term and Text Part Presentation

Όσον αφορά τα κουμπιά επιλογής που φαίνονται κάτω δεξιά από το κεντρικό πλαίσιο, αντιστοιχούν στην δημιουργία του μοντέλου άδειας λογισμικού ανοικτού κώδικα, στην αποθήκευση των δεδομένων σε JSON format, στην παρουσίαση του wordcloud για το συγκεκριμένο κείμενο και το κουμπί για επανάληψη της διαδικασίας για διαφορετικό κείμενο.

Το wordcloud αποτελεί μία οπτική παρουσίαση του κειμένου, όσον αφορά την συχνότητα των λέξεων που εμφανίζονται σε αυτό. Λέξεις με μεγαλύτερη συχνότητα παρουσιάζονται μεγαλύτερες σε μέγεθος μέσα στο wordcloud παρά με λέξεις με μικρότερη συχνότητα. Με αυτό τον τρόπο εύκολα παρουσιάζονται οι σημαντικές λέξεις μέσα στο κείμενο.



Σχήμα 30: OSS License Analysis Wordcloud.

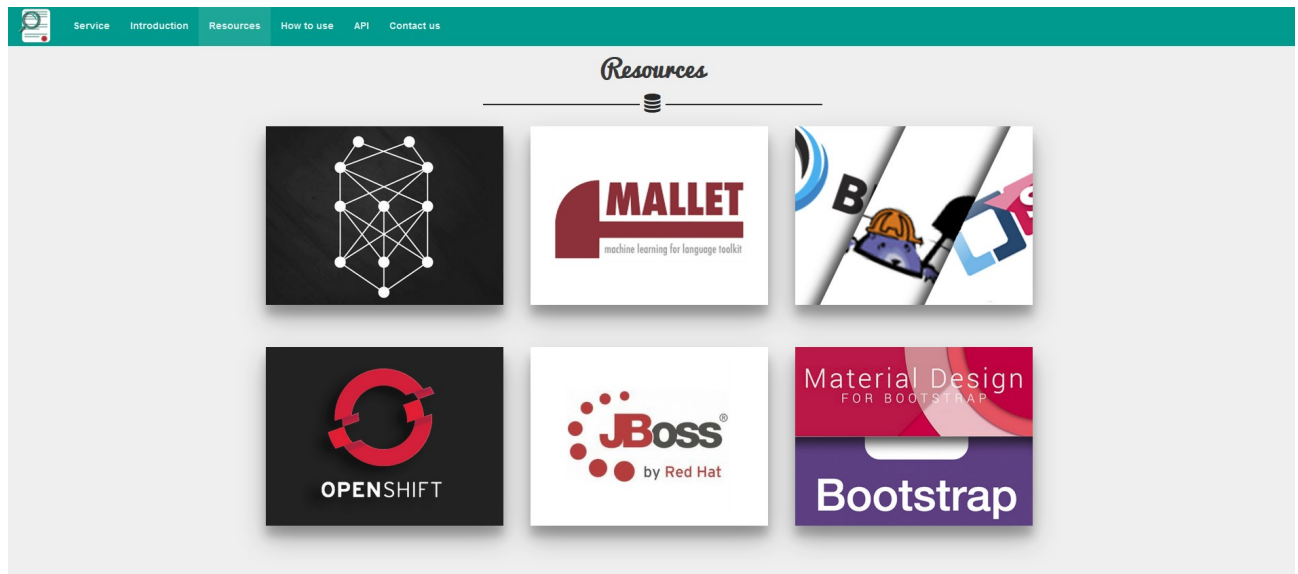


Σχήμα 31: OSS License Analysis Introduction Page.

9.4.2 Εισαγωγή

Στη συγκεκριμένη σελίδα παρουσιάζεται μια απλή εισαγωγή, για τον σκοπό του συστήματος και του τί προσπαθεί να κάνει. Έχει ώση σκοπό να πληροφορήσει καινούργιους χρήστες που έρχονται στο σύστημα χωρίς καμία προδιάθεση.

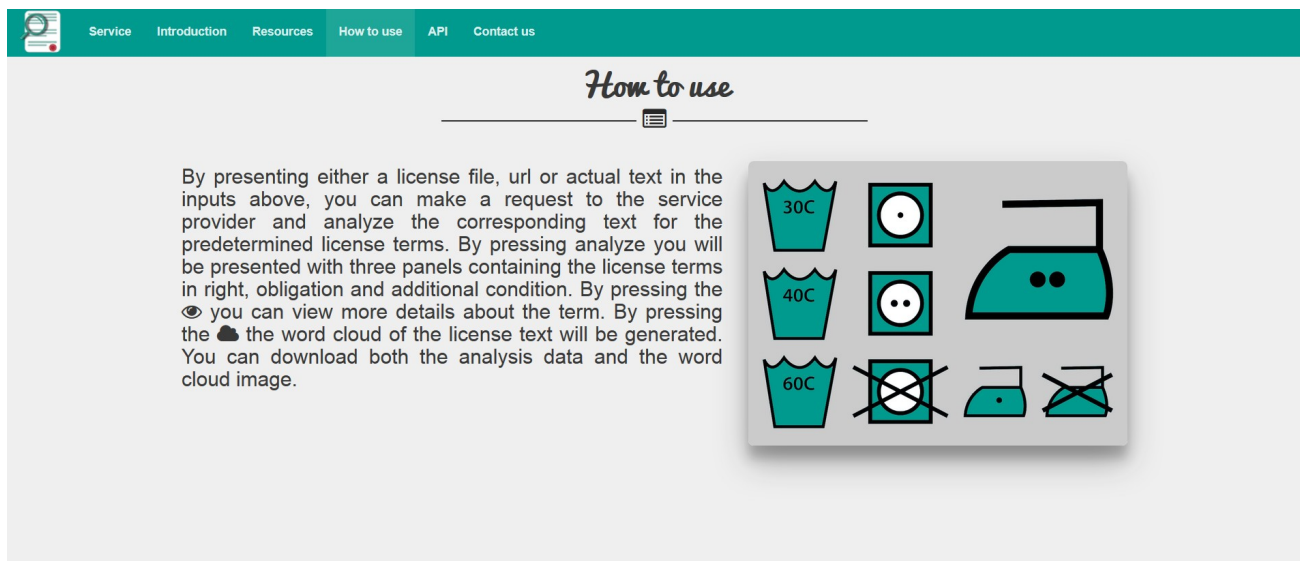
9.4.3 Πηγές



Σχήμα 32: OSS License Analysis Resource Page.

Στη συγκεκριμένη σελίδα παρουσιάζονται οι πηγές πάνω στις οποίες βασίστηκε το σύστημα μας, όσον αφορά τις υπηρεσίες REST αλλά και για την οπτική παρουσία. Περνώντας τον κέρσορα πάνω από ένα, είτε επιλέγοντας το εάν βρίσκεται σε κινητή εφαρμογή, εκτελείτε animation και εμφανίζονται κάποιες βασικές πληροφορίες για την συγκεκριμένη πηγή.

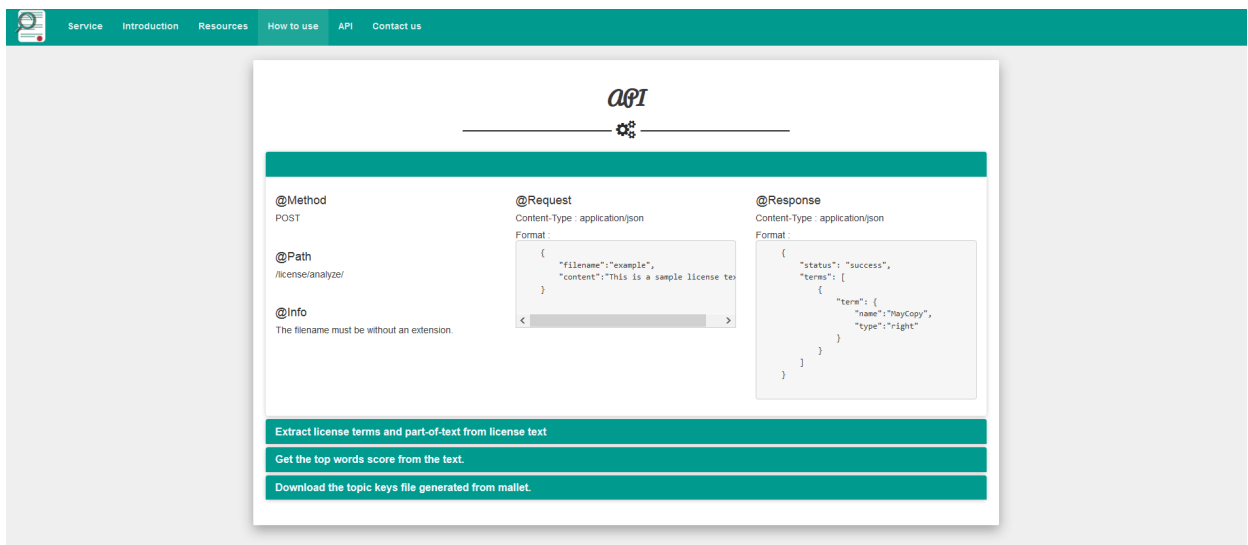
9.4.4 Οδηγίες χρήσης



Σχήμα 33: OSS License Analysis How To Use Page.

Αποτελεί ένα απλό και γρήγορο εγχειρίδιο για την χρήση του συστήματος. Επαναλαμβάνει σε συντομία τα όσα λέμε στην περιγραφή της Υπηρεσίας στο παρόν κεφάλαιο.

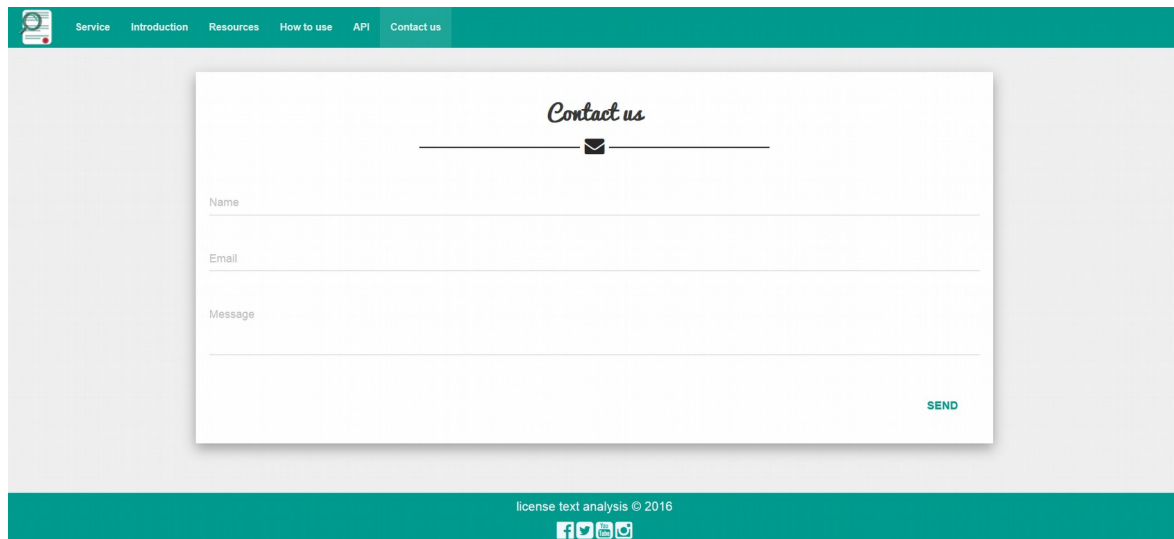
9.4.5 API



Σχήμα 34: OSS License Analysis API Page.

Παρουσιάζει με γραφικό τρόπο τα όσα αναφέραμε στο κομμάτι περιγραφής της

υπηρεσίας REST. Δίνει τις πληροφορίες για τα URL, την μορφή που πρέπει να έχουν τα request και την μορφή των responses. Με αυτό τον τρόπο ο χρήστης μπορεί να ενσωματώσει την υπηρεσία σε κάποιο δικό του σύστημα, και γενικά να τη χρησιμοποιήσει.



The screenshot shows a web page titled "Contact us" with a teal header and footer. The header contains navigation links: Service, Introduction, Resources, How to use, API, and Contact us. The main content area is white and contains a "Contact us" form with fields for Name, Email, and Message, and a "SEND" button. The footer is teal and contains the text "license text analysis © 2016" and social media icons for Facebook, Twitter, and Instagram.

Σχήμα 35: OSS License Analysis Contact Page.

9.4.6 Επικοινωνία

Μέσω αυτής της σελίδας θα γίνεται η επικοινωνία με χρήστες που πιθανόν να έχουν κάποιες απορίες σχετικά με το σύστημα, να θέλουν να δώσουν κάποιες απόψεις είτε να θέλουν να συμβάλουν σε επέκταση του. Με αυτό τον τρόπο θα μπορέσουμε και εμείς να λάβουμε feedback για να βελτιωθεί η υπηρεσία και το όλο σύστημα.

Κεφάλαιο 10

Συμπεράσματα και μελλοντικές εργασίες

10.1 Συμπεράσματα	119
10.2 Μελλοντικές εργασίες	120

10.1 Συμπεράσματα

Η χρήση λογισμικών ανοικτού κώδικα έχει αποκτήσει το κέντρο της προσοχής από διάφορες εταιρείες και οργανισμούς, οι οποίοι ενσωματώνουν λογισμικά από τρίτους στις δικές τους υλοποιήσεις. Όταν χρησιμοποιείται μια τέτοια βιβλιοθήκη είτε ο πηγαίος κώδικας, έμμεσα ή άμεσα υπάρχει η σύνδεση με μία ή περισσότερες άδειες λογισμικού. Αυτό το γεγονός μπορεί να οδηγήσει σε πιθανές συγκρούσεις μεταξύ τους λόγω της δυσκολίας του ανθρώπου να κατανοήσει είτε να προσέξει κάποια σημαντική πρόταση. Οπότε απαραίτητη είναι η υποστήριξη όσον αφορά τη χρήση των αδειών λογισμικού με στόχο την αποφυγή τέτοιων συγκρούσεων, καθώς επίσης και την διαχείριση τους εάν υπάρξουν, μέσω της μοντελοποίησης τους.

Στα πλαίσια της παρούσας Ατομικής Διπλωματικής Εργασίας σχεδιάστηκε ένα καθολικό μοντέλο που μπορεί να εφαρμοστεί πάνω σε όλες τις άδειες λογισμικού με σκοπό την εύκολη κατανόηση τους από άνθρωπο, καθώς επίσης και την ανάλυση τους από υπολογιστή. Με τη χρήση των βασικών 22 ορισμών που εξήγαγε σε προηγούμενη

Διπλωματική Εργασία η Γεωργία Χαραλάμπους μπορεί να περιγραφθεί οποιαδήποτε άδεια λογισμικού ανοικτού κώδικα. Προσαρμόζοντας αυτούς λοιπόν μέσα στο μοντέλο αποτελεί λύση όσον αφορά το πρόβλημα της μοντελοποίησης που αναφέραμε, καθώς επίσης και τη δυσκολία κατανόησης του πολύπλοκου και θορυβώδους νομικού κειμένου τους από τον άνθρωπο.

Σε συνδυασμό με το αυτόματο σύστημα εξαγωγής των ορισμών από τα κείμενα των αδειών αποτελεί μια ολοκληρωμένη λύση και προσέγγιση στο συγκεκριμένο πρόβλημα. Ο χρήστης μέσω του συστήματος μπορεί όχι μόνο να παράξει το μοντέλο της άδειας που θέλει, αλλά να συμβουλευτεί τα κομμάτια του κειμένου που αντιστοιχούν στους ορισμούς της. Με αυτό τον τρόπο διευκολύνεται ακόμη περισσότερο ο χρήστης στην κατανόηση της άδειας, και μπορεί πιο αποδοτικά να τη χρησιμοποιήσει.

Εν κατακλείδι η χρήση του μοντέλου αυτού, καθώς και του συστήματος εξαγωγής των ορισμών, μπορεί να αποτελέσει σημαντικό παράγοντα στη μελέτη των αδειών λογισμικού ανοικτού κώδικα, και βάση για περισσότερες μελέτες και εφαρμογές, όπως για παράδειγμα την ανάλυση συγκρούσεων μεταξύ αδειών, οι οποίες έχουν πολύ μεγάλη σημασία, καθώς μπορούν να οδηγήσουν ακόμα και σε νομικές διενέξεις.

10.2 Μελλοντικές εργασίες

Κατά τη συγκεκριμένη Ατομική Διπλωματική Εργασία, ξεκίνησε μία προσπάθεια όσον αφορά την κατασκευή κανόνων που να παρουσιάζουν τις σχέσεις μεταξύ των ορισμών, όπως περιγράφονται πιο πάνω.

Το σύστημα που σχεδιάστηκε και υλοποιήθηκε μπορεί να χρησιμοποιηθεί σε συνδυασμό με υφιστάμενα εργαλεία εξαγωγής αδειών λογισμικού από κώδικα, όπως είναι το FOSSology, για την παροχή περισσότερων πληροφοριών στο χρήστη. Στην παρούσα μορφή του το FOSSology παρέχει πληροφορίες για την άδεια ή τις άδειες που εμφανίζονται σε κάθε αρχείο ενός συστήματος λογισμικού. Κάνοντας χρήση του

συστήματος που υλοποιήθηκε στη διπλωματική η πληροφορία που παρέχεται στο χρήστη μπορεί να εμπλουτιστεί σημαντικά.

Μέσω του συγκεκριμένου μοντέλου αδειών λογισμικού ανοικτού κώδικα, μπορεί να υποστηριχτεί η εφαρμογή reasoning για την επίτευξη αποφάσεων όπως είναι η επιλογή της κατάλληλης άδειας λογισμικού είτε η κατασκευή άδειας που να είναι συμβατή με τους ξεχωριστούς ορισμούς. Το τελευταίο παρουσιάζεται αναλυτικά και στο σύστημα Carneades argumentation system [7]. Κάποιοι βασικοί κανόνες στα πλαίσια της παρούσας διπλωματικής κατασκευάστηκαν στο SPINdle reasoner, καθώς επίσης και στο Gorgias, μέσω των οποίων εφαρμόζονται θεωρίες της defeasible logic [33]. Παράδειγμα από το SPINdle είναι ο εξής κανόνας *MustDistributeCodeUnderThisLicense* → *NotSublicense, NotSublicense* → *-MaySublicense*.

Οπότε σαν μελλοντική εργασία μπορεί να αποτελέσει η ολοκλήρωση των κανόνων μέ defeasible logic για το SPINdle και Gorgias, τα οποία μπορούν να βοηθήσουν περαιτέρω την ανάλυση των αδειών καθώς επίσης να βοηθήσουν στην αποφυγή συγκρούσεων μεταξύ ορισμών.

Όπως αναφέρθηκε προηγουμένος, κατά τον έλεγχο των αποτελεσμάτων του συστήματος εξαγωγής των ορισμών από το κείμενο, μερικοί από τους ορισμούς αδειών λογισμικού ανοικτού κώδικα βρίσκονται πολύ κοντά σημασιολογικά, καθώς επίσης και το περιεχόμενο της πρότασης που τους χαρακτηρίζει είναι αρκετά όμοιο. Αυτό οδηγεί το σύστημα να παρουσιάζει κάποιο όμοιο ορισμό αντί του αντίστοιχου που δηλώθηκε δια χειρός. Τα αποτελέσματα, λαμβάνοντας υπόψη τους ορισμούς που είναι συνώνυμοι, βρίσκονται στη στήλη του πιο πάνω πίνακα με το όνομα *Error Χωρίς Συνώνυμα*. Για παράδειγμα στην άδεια λογισμικού AAL, αντί του *NotPermissionToUseTradeMarks* και του *MustNotMisrepresentOriginalAuthorOfSoftware*, βρέθηκε το *MustRetainCopyrightNotice*, το οποίο δικαιολογείται από κείμενο που περιγράφει επίσης τους δύο ορισμούς που παραλείφθηκαν. Αυτό αποτελεί ένα κομμάτι που πρέπει να ληφθεί υπόψη κατά την λειτουργία του συστήματος και μπορεί επίσης να αποτελέσει μελλοντική εργασία.

Επιπρόσθετα μπορούν να υλοποιηθούν οι δύο αλγόριθμοι του PAM-Pachinko Allocation Model και PLSA-Probabilistic Latent Semantic Analysis. Μπορεί να γίνει ανάλυση των αποτελεσμάτων τους και να εφαρμοστούν και αυτοί πάνω στο σύστημα εξαγωγής των ορισμών από τα κείμενα των αδειών, μέσω NLP.

Βιβλιογραφία

- [1] ΑΔΕ Γεωργίας Χαραλάμπους 2015 Recommender System for OSS Licenses.
Kapitsaki, G. M., Tselikas, N. D., & Foukarakis, I. E. 2015. An insight into license
- [2] tools for open source software systems. Journal of Systems and Software, 102, 72-87.
- [3] <https://tldrlegal.com/>
Alspaugh, T., Asuncion, H. U., & Scacchi, W. (2009, August). Intellectual property
- [4] rights requirements for heterogeneously-licensed systems. In Requirements Engineering Conference, 2009. RE'09. 17th IEEE International (pp. 24-33). IEEE.
- [5] Cabrio, E., Aproso, A. P., & Villata, S. (2014). These Are Your Rights. In The Semantic Web: Trends and Challenges (pp. 255-269). Springer International Publishing.
- [6] Guido Governatori,, Antonino Rotolo, Serena Villata,, and Fabien Gandon
Gordon TF. Analyzing open source license compatibility issues with Carneades.
- [7] Proc. 13th International Conference on Artificial Intelligence and Law (ICAIL 2011) 2011: 51-55. DOI:10.1145/2018358.2018364.
Gordon TF. Report on a Prototype Decision Support System for OSS License
- [8] Compatibility Issues. Qualipso (IST- FP6-IP-034763), Deliverable A1.D2.1.3 2010. http://www.qualipso.org/sites/default/files/A1.D1.2.1.3_bis.pdf
- [9] Kechagia, M.; Spinellis, D.; Androutsellis-Theotokis, S. Open Source Licensing Across Package Dependencies Publication Year: 2010, Page(s): 27- 32
HOWTO: Pick an open source license (part 1)
- [10] <http://www.zdnet.com/blog/burnette/howto-pick-an-open-source-license-part-1/130>
How to pick an open source license (part 2)
- [11] <http://www.zdnet.com/blog/burnette/how-to-pick-an-open-source-license-part-2/131>
Stephen R. Wall Which open source software license should I use?
- [12] <http://opensource.com/law/13/1/which-open-source-software-license-should-i-use>
- [13] Choosealicense <http://choosealicense.com/>
- [14] A Deontic Logic Semantics for Licenses Composition in the Web of Data, Antonino Rotolo University of Bologna, Italy antonino.rotolo@unibo.it, Serena

Villata INRIA Sophia Antipolis France serena.villata@inria.fr, Fabien Gandon
INRIA Sophia Antipolis France fabien.gandon@inria.fr

[15] <https://www.w3.org/TR/rdf-schema/>

[16] <https://www.w3.org/2001/sw/wiki/OWL>

Ontology Visualization Methods—A Survey, AKRIVI KATIFORI and CONSTANTIN
[17] HALATSIS University of Athens and GEORGE LEPOURAS, COSTAS VASSILAKIS, and
EUGENIA GIANNOPOULOU University of Peloponnese.

Indented Tree or Graph? A Usability Study of Ontology Visualization Techniques
in the Context of Class Mapping Evaluation, B. F., Natalya F. N., Margaret-Anne
[18] Storey, Department of CS, University of Victoria, BC, Canada, Stanford Center for
Biomedical Informatics Research, Stanford University, CA, US, bofu@uvic.ca,
noy@stanford.edu, mstorey@uvic.ca.

Rotolo, A., Villata, S., Gandon, F.: A deontic logic semantics for licenses
[19] composition in the web of data. In: Proceedings of the Fourteenth International
Conference on Artificial Intelligence and Law. pp. 111{120. ACM (2013)

Rotolo, A., Villata, S., Gandon, F.: A deontic logic semantics for licenses
[20] composition in the web of data. In: Proceedings of the Fourteenth International
Conference on Artificial Intelligence and Law. pp. 111{120. ACM (2013)

[21] <https://jena.apache.org/index.html>

[22] <http://www.l3s.de/~minack/rdf2rdf/>

[23] <http://jung.sourceforge.net/index.html>

Semantic deontic modeling and text classification for supporting automated
[24] environmental compliance checking in construction, Dareen Mamdouh Salama
Abdelmoneim, University of Illinois at Urbana-Champaign, 2011

Natural Language Processing (almost) from Scratch, Ronan Collobert
ronan@collobert.com, NEC Labs America, Princeton NJ. Jason Weston
jweston@google.com Google, New York, NY. L'eon Bottou leon@bottou.org
[25] Michael Karlen michael.karlen@gmail.com Koray Kavukcuoglu†
koray@cs.nyu.edu Pavel Kuksa† pkuksa@cs.rutgers.edu NEC Labs America,
Princeton NJ.

[26] <http://flossmole.org/>

[27] <https://github.com/nowitall/morpha>

[28] <https://docs.oracle.com/javase/7/docs/api/java/text/BreakIterator.html>

- [29] <https://gate.ac.uk/>
- [30] <http://nlp.stanford.edu/software/lex-parser.shtml>
- [31] <http://deeplearning4j.org>
 Distributed representations of sentences and documents, quoc le
 qvl@google.com
- [32] Tomas mikolov tmikolov@google.com, google inc, 1600 amphitheatre parkway,
 mountain view, ca 94043
 Rotolo, A., Villata, S., Gandon, F.: A deontic logic semantics for licenses composi-
 tion in the web of data. In: Proceedings of the Fourteenth International
 [33] Conference
 on Artificial Intelligence and Law. pp. 111{120. ACM (2013)
- [34] <http://www.dwheeler.com/essays/floss-license-slide.html>
- [35] <http://www.ics.uci.edu/~alspaugh/cls/shr/hohfeld.html>
- [36] <http://cgi.csc.liv.ac.uk/~floriana/CMNA/WaltonReed.pdf>
- [37] <http://spdxcompatibility-licensetools.rhcloud.com/>
- [38] <https://en.wikipedia.org/wiki/Triplestore>
- [39] <https://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/>
- [40] <http://www.hermit-reasoner.com/>
- [41] <http://findosslicense.cs.ucy.ac.cy/>
- [42] <http://spin.nicta.org.au/spindle/index.html>
- [43] https://en.wikipedia.org/wiki/Machine_learning
- [44] <http://www.acad.bg/ebook/ml/MITPress-%20SemiSupervised%20Learning.pdf>
- [45] <https://jsoup.org/>
- [46] <http://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/>
- [47] <http://deeplearning4j.org/glossary.html#cosine>

Παράρτημα Α

AGPL-3.0	
0.8791667128050654	MayDistributeDerivativeWorksInsourceForm
0.8791667128050654	MayMakeDerivativeWorks
0.8791667128050654	MustRetainCopyrightNotice
0.8791667128050654	MayDistributeOriginalWorkInsourceForm
0.8791667128050654	MayAcceptWarrantyOrAdditionalLiability
0.8791667128050654	LimitedLiability
0.8791667128050654	MayDistributeOriginalWorkInObjectForm
0.8791667128050654	MayPubliclyDisplay
0.8791667128050654	MayMakeModifications
0.8791667128050654	MayPubliclyPerform
0.8791667128050654	MustMarkModifications
0.8791667128050654	MayGrantPatentsRightsFromTheContributorToReceipt
0.8791667128050654	NotSublicense
0.8791667128050654	MustOffersourceCode
0.8791667128050654	MustDistributeCodeUnderThisLicense
0.8791667128050654	MayDistributeDerivativeWorksInObjectForm

0.8791667128050654	ProvideWithoutWarranty
0.5007616759067567	MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense
0.5007616759067567	MustRetainLicenseNotice
0.5007616759067567	MustRetainInsourceCodeOfDerivativeWorkAllTradeMarksPatentCopyrightsOfOriginalWork
0.5007616759067567	MustLicenseDerivativesWorksUnderCompatibleLicense
0.5007616759067567	NotPermissionToUseTradeMarks
0.38162130382036413	MaySublicense
0.37892521596043754	MustProvideInformationHowToGetSourceCode
0.37892521596043754	MustPublishAsLibraryTheModifiedVersion
0.37892521596043754	MayCopy
0.3464063794558486	MayUseInHouse
0.3464063794558486	MustNotMisrepresentTheOriginalAuthorOfSoftware
0.21003007852938768	MayAddDifferentLicenseTermsToYourModifications
0.21003007852938768	MustDistributeCopyOfNoticeText
0.21003007852938768	MayAddYourOwnCopyrightStatementToYourModifications

Apache-2.0

0.999999947770047	MayAcceptWarrantyOrAdditionalLiability
0.999999947770047	MayAddDifferentLicenseTermsToYourModifications

0.999999947770047	LimitedLiability
0.999999947770047	MustMarkModifications
0.999999947770047	MayGrantPatentsRightsFromTheContributorToReceipt
0.999999947770047	MustLicenseDerivativesWorksUnderCombatibleLicense
0.999999947770047	NotPermissionToUseTradeMarks
0.999999947770047	MustDistributeCopyOfNoticeText
0.999999947770047	MayAddYourOwnCopyrightStatementToYourModifications
0.999999947770047	ProvideWithoutWarranty
0.9615167861899608	MayDistributeOriginalWorkInsourceForm
0.9615167861899608	MayPubliclyDisplay
0.9615167861899608	MaySublicense
0.9615167861899608	MayPubliclyPerform
0.635505789454013	MayDistributeDerivativeWorksInsourceForm
0.635505789454013	MayMakeDerivativeWorks
0.635505789454013	MayDistributeOriginalWorkInObjectForm
0.635505789454013	MayMakeModifications
0.635505789454013	MayUseInHouse
0.635505789454013	MayDistributeDerivativeWorksInObjectForm

0.635505789454013	MayCopy
0.6315262501786989	MustRetainCopyrightNotice
0.6315262501786989	MustPublishAsLibraryTheModifiedVersion
0.6315262501786989	MustDistributeCodeUnderThisLicense
0.606060954933757	MustProvideInformationHowToGetsourceCode
0.606060954933757	MustOffersourceCode
0.592514946668153	NotSublicense
0.559670898122906	MayAddYourOwnChoiceOfLicenselfCodeWasntUnderThisLicense
0.559670898122906	MustRetainLicenseNotice
0.559670898122906	MustRetainInsourceCodeOfDerivativeWorkAllTradeMarksPatentCopyrightsOfOriginalWork
0.4899094855352073	MustNotMisrepresentTheOriginalAuthorOfSoftware

Artistic-2.0	
0.8432885792405753	MayAcceptWarrantyOrAdditionalLiability
0.8355029493646704	MayDistributeDerivativeWorksInsourceForm
0.8355029493646704	MayMakeDerivativeWorks
0.8355029493646704	MustRetainCopyrightNotice
0.8355029493646704	MayDistributeOriginalWorkInsourceForm
0.8355029493646704	MayDistributeOriginalWorkInObjectForm

0.8355029493646704	MayPubliclyDisplay
0.8355029493646704	MayMakeModifications
0.8355029493646704	MaySublicense
0.8355029493646704	MayPubliclyPerform
0.8355029493646704	MayUseInHouse
0.8355029493646704	MayDistributeDerivativeWorksInObjectForm
0.8355029493646704	MayCopy
0.7675153084888215	MayAddDifferentLicenseTermsToYourModifications
0.7675153084888215	LimitedLiability
0.7675153084888215	MustMarkModifications
0.7675153084888215	MayGrantPatentsRightsFromTheContributorToReceipt
0.7675153084888215	MustLicenseDerivativesWorksUnderCompatibleLicense
0.7675153084888215	NotPermissionToUseTradeMarks
0.7675153084888215	MustDistributeCopyOfNoticeText
0.7215433449526966	MayAddYourOwnCopyrightStatementToYourModifications
0.67328830505739	MustNotMisrepresentTheOriginalAuthorOfSoftware
0.6607968006728083	ProvideWithoutWarranty
0.6315262501786989	MustProvideInformationHowToGetSourceCode

0.6101537222014184	MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense
0.6101537222014184	MustRetainLicenseNotice
0.6101537222014184	MustOffersourceCode
0.6101537222014184	MustRetainInsourceCodeOfDerivativeWorkAllTradeMarksPatentCopyrightsOfOriginalWork
0.5764958716543769	MustPublishAsLibraryTheModifiedVersion
0.5764958716543769	MustDistributeCodeUnderThisLicense
0.5374964914765656	NotSublicense

BSD-2-Clause	
0.8432885792405753	MayAcceptWarrantyOrAdditionalLiability
0.8355029493646704	MayDistributeDerivativeWorksInsourceForm
0.8355029493646704	MayMakeDerivativeWorks
0.8355029493646704	MustRetainCopyrightNotice
0.8355029493646704	MayDistributeOriginalWorkInsourceForm
0.8355029493646704	MayDistributeOriginalWorkInObjectForm
0.8355029493646704	MayPubliclyDisplay
0.8355029493646704	MayMakeModifications
0.8355029493646704	MaySublicense
0.8355029493646704	MayPubliclyPerform

704	
0.8355029493646 704	MayUseInHouse
0.8355029493646 704	MayDistributeDerivativeWorksInObjectForm
0.8355029493646 704	MayCopy
0.7675153084888 215	MayAddDifferentLicenseTermsToYourModifications
0.7675153084888 215	LimitedLiability
0.7675153084888 215	MustMarkModifications
0.7675153084888 215	MayGrantPatentsRightsFromTheContributorToReceipt
0.7675153084888 215	MustLicenseDerivativesWorksUnderCompatibleLicense
0.7675153084888 215	NotPermissionToUseTradeMarks
0.7675153084888 215	MustDistributeCopyOfNoticeText
0.7215433449526 966	MayAddYourOwnCopyrightStatementToYourModifications
0.6732883050573 9	MustNotMisrepresentTheOriginalAuthorOfSoftware
0.6607968006728 083	ProvideWithoutWarranty
0.6315262501786 989	MustProvideInformationHowToGetSourceCode
0.6101537222014 184	MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense
0.6101537222014 184	MustRetainLicenseNotice
0.6101537222014 184	MustOfferSourceCode
0.6101537222014	MustRetainInSourceCodeOfDerivativeWorkAllTradeMarksPatentCopyrights

184	OfOriginalWork
0.5764958716543769	MustPublishAsLibraryTheModifiedVersion
0.5764958716543769	MustDistributeCodeUnderThisLicense
0.5374964914765656	NotSublicense

BSD-3-Clause	
0.8432885792405753	MayAcceptWarrantyOrAdditionalLiability
0.8355029493646704	MayDistributeDerivativeWorksInsourceForm
0.8355029493646704	MayMakeDerivativeWorks
0.8355029493646704	MustRetainCopyrightNotice
0.8355029493646704	MayDistributeOriginalWorkInsourceForm
0.8355029493646704	MayDistributeOriginalWorkInObjectForm
0.8355029493646704	MayPubliclyDisplay
0.8355029493646704	MayMakeModifications
0.8355029493646704	MaySublicense
0.8355029493646704	MayPubliclyPerform
0.8355029493646704	MayUseInHouse
0.8355029493646704	MayDistributeDerivativeWorksInObjectForm
0.8355029493646704	MayCopy
0.7675153084888	MayAddDifferentLicenseTermsToYourModifications

215	
0.7675153084888 215	LimitedLiability
0.7675153084888 215	MustMarkModifications
0.7675153084888 215	MayGrantPatentsRightsFromTheContributorToReceipt
0.7675153084888 215	MustLicenseDerivativesWorksUnderCombatibleLicense
0.7675153084888 215	NotPermissionToUseTradeMarks
0.7675153084888 215	MustDistributeCopyOfNoticeText
0.7215433449526 966	MayAddYourOwnCopyrightStatementToYourModifications
0.6732883050573 9	MustNotMisrepresentTheOriginalAuthorOfSoftware
0.6607968006728 083	ProvideWithoutWarranty
0.6315262501786 989	MustProvideInformationHowToGetsourceCode
0.6101537222014 184	MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense
0.6101537222014 184	MustRetainLicenseNotice
0.6101537222014 184	MustOffersourceCode
0.6101537222014 184	MustRetainInsourceCodeOfDerivativeWorkAllTradeMarksPatentCopyrights OfOriginalWork
0.5764958716543 769	MustPublishAsLibraryTheModifiedVersion
0.5764958716543 769	MustDistibuteCodeUnderThisLicense
0.5374964914765 656	NotSublicense

CDDL-1.0	
0.8432885792405753	MayAcceptWarrantyOrAdditionalLiability
0.8355029493646704	MayDistributeDerivativeWorksInsourceForm
0.8355029493646704	MayMakeDerivativeWorks
0.8355029493646704	MustRetainCopyrightNotice
0.8355029493646704	MayDistributeOriginalWorkInsourceForm
0.8355029493646704	MayDistributeOriginalWorkInObjectForm
0.8355029493646704	MayPubliclyDisplay
0.8355029493646704	MayMakeModifications
0.8355029493646704	MaySublicense
0.8355029493646704	MayPubliclyPerform
0.8355029493646704	MayUseInHouse
0.8355029493646704	MayDistributeDerivativeWorksInObjectForm
0.8355029493646704	MayCopy
0.7675153084888215	MayAddDifferentLicenseTermsToYourModifications
0.7675153084888215	LimitedLiability
0.7675153084888215	MustMarkModifications
0.7675153084888215	MayGrantPatentsRightsFromTheContributorToReceipt
0.7675153084888	MustLicenseDerivativesWorksUnderCompatibleLicense

215	
0.7675153084888 215	NotPermissionToUseTradeMarks
0.7675153084888 215	MustDistributeCopyOfNoticeText
0.7215433449526 966	MayAddYourOwnCopyrightStatementToYourModifications
0.6732883050573 9	MustNotMisrepresentTheOriginalAuthorOfSoftware
0.6607968006728 083	ProvideWithoutWarranty
0.6315262501786 989	MustProvideInformationHowToGetsourceCode
0.6101537222014 184	MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense
0.6101537222014 184	MustRetainLicenseNotice
0.6101537222014 184	MustOffersourceCode
0.6101537222014 184	MustRetainInsourceCodeOfDerivativeWorkAllTradeMarksPatentCopyrights OfOriginalWork
0.5764958716543 769	MustPublishAsLibraryTheModifiedVersion
0.5764958716543 769	MustDistributeCodeUnderThisLicense
0.5374964914765 656	NotSublicense

EPL-1.0	
0.8432885792405 753	MayAcceptWarrantyOrAdditionalLiability
0.8355029493646 704	MayDistributeDerivativeWorksInsourceForm
0.8355029493646 704	MayMakeDerivativeWorks

0.8355029493646704	MustRetainCopyrightNotice
0.8355029493646704	MayDistributeOriginalWorkInsourceForm
0.8355029493646704	MayDistributeOriginalWorkInObjectForm
0.8355029493646704	MayPubliclyDisplay
0.8355029493646704	MayMakeModifications
0.8355029493646704	MaySublicense
0.8355029493646704	MayPubliclyPerform
0.8355029493646704	MayUseInHouse
0.8355029493646704	MayDistributeDerivativeWorksInObjectForm
0.8355029493646704	MayCopy
0.7675153084888215	MayAddDifferentLicenseTermsToYourModifications
0.7675153084888215	LimitedLiability
0.7675153084888215	MustMarkModifications
0.7675153084888215	MayGrantPatentsRightsFromTheContributorToReceipt
0.7675153084888215	MustLicenseDerivativesWorksUnderCompatibleLicense
0.7675153084888215	NotPermissionToUseTradeMarks
0.7675153084888215	MustDistributeCopyOfNoticeText
0.7215433449526966	MayAddYourOwnCopyrightStatementToYourModifications

0.67328830505739	MustNotMisrepresentTheOriginalAuthorOfSoftware
0.6607968006728083	ProvideWithoutWarranty
0.6315262501786989	MustProvideInformationHowToGetsourceCode
0.6101537222014184	MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense
0.6101537222014184	MustRetainLicenseNotice
0.6101537222014184	MustOffersourceCode
0.6101537222014184	MustRetainInsourceCodeOfDerivativeWorkAllTradeMarksPatentCopyrightsOfOriginalWork
0.5764958716543769	MustPublishAsLibraryTheModifiedVersion
0.5764958716543769	MustDistributeCodeUnderThisLicense
0.5374964914765656	NotSublicense

GPL-2.0	
0.9186759116407093	MustRetainCopyrightNotice
0.9186759116407093	MayDistributeOriginalWorkInsourceForm
0.9186759116407093	MayAcceptWarrantyOrAdditionalLiability
0.9186759116407093	LimitedLiability
0.9186759116407093	MayDistributeOriginalWorkInObjectForm
0.9186759116407093	MayPubliclyDisplay
0.9186759116407093	MayPubliclyPerfom

0.9186759116407093	MustMarkModifications
0.9186759116407093	MustProvideInformationHowToGetsourceCode
0.9186759116407093	NotSublicense
0.9186759116407093	MustOffersourceCode
0.9186759116407093	MustDistributeCodeUnderThisLicense
0.9186759116407093	MayDistributeDerivativeWorksInObjectForm
0.9186759116407093	MayCopy
0.9186759116407093	ProvideWithoutWarranty
0.866557348486769	MayMakeDerivativeWorks
0.866557348486769	MayMakeModifications
0.866557348486769	MustPublishAsLibraryTheModifiedVersion
0.6873708227215999	MayDistributeDerivativeWorksInsourceForm
0.6873708227215999	MayGrantPatentsRightsFromTheContributorToReceipt
0.6047423198585862	MaySublicense
0.4822458990898464	NotPermissionToUseTradeMarks
0.47334925217503276	MayAddDifferentLicenseTermsToYourModifications
0.47334925217503276	MustLicenseDerivativesWorksUnderCombatibleLicense
0.47334925217503276	MustDistributeCopyOfNoticeText

0.47334925217503276	MayAddYourOwnCopyrightStatementToYourModifications
0.43578704012609626	MayUseInHouse
0.42561087448411977	MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense
0.42561087448411977	MustRetainLicenseNotice
0.42561087448411977	MustRetainInSourceCodeOfDerivativeWorkAllTradeMarksPatentCopyrightsOfOriginalWork
0.3991984390740319	MustNotMisrepresentTheOriginalAuthorOfSoftware

GPL-3.0	
0.8791667128050654	MayDistributeDerivativeWorksInSourceForm
0.8791667128050654	MayMakeDerivativeWorks
0.8791667128050654	MustRetainCopyrightNotice
0.8791667128050654	MayDistributeOriginalWorkInSourceForm
0.8791667128050654	MayAcceptWarrantyOrAdditionalLiability
0.8791667128050654	LimitedLiability
0.8791667128050654	MayDistributeOriginalWorkInObjectForm
0.8791667128050654	MayPubliclyDisplay
0.8791667128050654	MayMakeModifications
0.8791667128050654	MayPubliclyPerform
0.8791667128050654	MustMarkModifications

0.8791667128050654	MayGrantPatentsRightsFromTheContributorToReceipt
0.8791667128050654	NotSublicense
0.8791667128050654	MustOffersourceCode
0.8791667128050654	MustDistributeCodeUnderThisLicense
0.8791667128050654	MayDistributeDerivativeWorksInObjectForm
0.8791667128050654	ProvideWithoutWarranty
0.5007616759067567	MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense
0.5007616759067567	MustRetainLicenseNotice
0.5007616759067567	MustRetainInsourceCodeOfDerivativeWorkAllTradeMarksPatentCopyrightsOfOriginalWork
0.5007616759067567	MustLicenseDerivativesWorksUnderCompatibleLicense
0.5007616759067567	NotPermissionToUseTradeMarks
0.38162130382036413	MaySublicense
0.37892521596043754	MustProvideInformationHowToGetsourceCode
0.37892521596043754	MustPublishAsLibraryTheModifiedVersion
0.37892521596043754	MayCopy
0.3464063794558486	MayUseInHouse
0.3464063794558486	MustNotMisrepresentTheOriginalAuthorOfSoftware
0.21003007852938768	MayAddDifferentLicenseTermsToYourModifications

0.2100300785293 8768	MustDistributeCopyOfNoticeText
0.2100300785293 8768	MayAddYourOwnCopyrightStatementToYourModifications

LGPL-2.1	
0.9120236407086 058	MayMakeDerivativeWorks
0.9120236407086 058	MustRetainCopyrightNotice
0.9120236407086 058	MayDistributeOriginalWorkInsourceForm
0.9120236407086 058	MayAcceptWarrantyOrAdditionalLiability
0.9120236407086 058	MayMakeModifications
0.9120236407086 058	MustMarkModifications
0.9120236407086 058	MustPublishAsLibraryTheModifiedVersion
0.9120236407086 058	MustDistributeCodeUnderThisLicense
0.9120236407086 058	MayCopy
0.8700802602026 925	LimitedLiability
0.8700802602026 925	MayDistributeOriginalWorkInObjectForm
0.8700802602026 925	MayPubliclyDisplay
0.8700802602026 925	MayPubliclyPerfom
0.8700802602026 925	MustProvideInformationHowToGetsourceCode
0.8700802602026 925	NotSublicense

0.8700802602026 925	MustOffersourceCode
0.8700802602026 925	MayDistributeDerivativeWorksInObjectForm
0.8700802602026 925	ProvideWithoutWarranty
0.6055493797626 468	MayDistributeDerivativeWorksInsourceForm
0.6055493797626 468	MayGrantPatentsRightsFromTheContributorToReceipt
0.5504371054121 696	MaySublicense
0.4804960413613 448	NotPermissionToUseTradeMarks
0.4528139125168 9795	MayAddDifferentLicenseTermsToYourModifications
0.4528139125168 9795	MustLicenseDerivativesWorksUnderCombatibleLicense
0.4528139125168 9795	MustDistributeCopyOfNoticeText
0.4528139125168 9795	MayAddYourOwnCopyrightStatementToYourModifications
0.4357714660211 8966	MayUseInHouse
0.3936098068517 181	MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense
0.3936098068517 181	MustRetainLicenseNotice
0.3936098068517 181	MustRetainInsourceCodeOfDerivativeWorkAllTradeMarksPatentCopyrights OfOriginalWork
0.3377600422278 177	MustNotMisrepresentTheOriginalAuthorOfSoftware

LGPL-3.0

0.8432885792405753	MayAcceptWarrantyOrAdditionalLiability
0.8355029493646704	MayDistributeDerivativeWorksInsourceForm
0.8355029493646704	MayMakeDerivativeWorks
0.8355029493646704	MustRetainCopyrightNotice
0.8355029493646704	MayDistributeOriginalWorkInsourceForm
0.8355029493646704	MayDistributeOriginalWorkInObjectForm
0.8355029493646704	MayPubliclyDisplay
0.8355029493646704	MayMakeModifications
0.8355029493646704	MaySublicense
0.8355029493646704	MayPubliclyPerform
0.8355029493646704	MayUseInHouse
0.8355029493646704	MayDistributeDerivativeWorksInObjectForm
0.8355029493646704	MayCopy
0.7675153084888215	MayAddDifferentLicenseTermsToYourModifications
0.7675153084888215	LimitedLiability
0.7675153084888215	MustMarkModifications
0.7675153084888215	MayGrantPatentsRightsFromTheContributorToReceipt
0.7675153084888215	MustLicenseDerivativesWorksUnderCompatibleLicense

0.7675153084888215	NotPermissionToUseTradeMarks
0.7675153084888215	MustDistributeCopyOfNoticeText
0.7215433449526966	MayAddYourOwnCopyrightStatementToYourModifications
0.67328830505739	MustNotMisrepresentTheOriginalAuthorOfSoftware
0.6607968006728083	ProvideWithoutWarranty
0.6315262501786989	MustProvideInformationHowToGetsourceCode
0.6101537222014184	MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense
0.6101537222014184	MustRetainLicenseNotice
0.6101537222014184	MustOffersourceCode
0.6101537222014184	MustRetainInsourceCodeOfDerivativeWorkAllTradeMarksPatentCopyrightsOfOriginalWork
0.5764958716543769	MustPublishAsLibraryTheModifiedVersion
0.5764958716543769	MustDistributeCodeUnderThisLicense
0.5374964914765656	NotSublicense

Libpng	
0.8432885792405753	MayAcceptWarrantyOrAdditionalLiability
0.8355029493646704	MayDistributeDerivativeWorksInsourceForm
0.8355029493646704	MayMakeDerivativeWorks
0.8355029493646704	MustRetainCopyrightNotice

0.8355029493646704	MayDistributeOriginalWorkInsourceForm
0.8355029493646704	MayDistributeOriginalWorkInObjectForm
0.8355029493646704	MayPubliclyDisplay
0.8355029493646704	MayMakeModifications
0.8355029493646704	MaySublicense
0.8355029493646704	MayPubliclyPerform
0.8355029493646704	MayUseInHouse
0.8355029493646704	MayDistributeDerivativeWorksInObjectForm
0.8355029493646704	MayCopy
0.7675153084888215	MayAddDifferentLicenseTermsToYourModifications
0.7675153084888215	LimitedLiability
0.7675153084888215	MustMarkModifications
0.7675153084888215	MayGrantPatentsRightsFromTheContributorToReceipt
0.7675153084888215	MustLicenseDerivativesWorksUnderCompatibleLicense
0.7675153084888215	NotPermissionToUseTradeMarks
0.7675153084888215	MustDistributeCopyOfNoticeText
0.7675153084888215	MayAddYourOwnCopyrightStatementToYourModifications
0.7215433449526966	MustNotMisrepresentTheOriginalAuthorOfSoftware

0.67328830505739	ProvideWithoutWarranty
0.6607968006728083	MustProvideInformationHowToGetsourceCode
0.6101537222014184	MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense
0.6101537222014184	MustRetainLicenseNotice
0.6101537222014184	MustOffersourceCode
0.6101537222014184	MustRetainInsourceCodeOfDerivativeWorkAllTradeMarksPatentCopyrightsOfOriginalWork
0.5764958716543769	MustPublishAsLibraryTheModifiedVersion
0.5764958716543769	MustDistributeCodeUnderThisLicense
0.5374964914765656	NotSublicense

MIT	
0.8432885792405753	MayAcceptWarrantyOrAdditionalLiability
0.8355029493646704	MayDistributeDerivativeWorksInsourceForm
0.8355029493646704	MayMakeDerivativeWorks
0.8355029493646704	MustRetainCopyrightNotice
0.8355029493646704	MayDistributeOriginalWorkInsourceForm
0.8355029493646704	MayDistributeOriginalWorkInObjectForm
0.8355029493646704	MayPubliclyDisplay
0.8355029493646704	MayMakeModifications

0.8355029493646704	MaySublicense
0.8355029493646704	MayPubliclyPerform
0.8355029493646704	MayUseInHouse
0.8355029493646704	MayDistributeDerivativeWorksInObjectForm
0.8355029493646704	MayCopy
0.7675153084888215	MayAddDifferentLicenseTermsToYourModifications
0.7675153084888215	LimitedLiability
0.7675153084888215	MustMarkModifications
0.7675153084888215	MayGrantPatentsRightsFromTheContributorToReceipt
0.7675153084888215	MustLicenseDerivativesWorksUnderCompatibleLicense
0.7675153084888215	NotPermissionToUseTradeMarks
0.7675153084888215	MustDistributeCopyOfNoticeText
0.7215433449526966	MayAddYourOwnCopyrightStatementToYourModifications
0.7675153084888215	MustNotMisrepresentTheOriginalAuthorOfSoftware
0.67328830505739	ProvideWithoutWarranty
0.6607968006728083	MustProvideInformationHowToGetSourceCode
0.6101537222014184	MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense
0.6101537222014184	MustRetainLicenseNotice

0.6101537222014184	MustOffersourceCode
0.6101537222014184	MustRetainInsourceCodeOfDerivativeWorkAllTradeMarksPatentCopyrightsOfOriginalWork
0.5764958716543769	MustPublishAsLibraryTheModifiedVersion
0.5764958716543769	MustDistributeCodeUnderThisLicense
0.5374964914765656	NotSublicense

MPL-1.1	
0.9999999333809956	MayDistributeOriginalWorkInsourceForm
0.9999999333809956	LimitedLiability
0.9999999333809956	MayPubliclyDisplay
0.9999999333809956	MaySublicense
0.9999999333809956	MayPubliclyPerform
0.9999999333809956	MustMarkModifications
0.9999999333809956	MayGrantPatentsRightsFromTheContributorToReceipt
0.6639170513541673	MustRetainCopyrightNotice
0.6639170513541673	MayAcceptWarrantyOrAdditionalLiability
0.6639170513541673	MayCopy
0.6639170513541673	ProvideWithoutWarranty
0.659784578996732	MayDistributeDerivativeWorksInsourceForm

0.659784578996732	MayMakeDerivativeWorks
0.659784578996732	MayDistributeOriginalWorkInObjectForm
0.659784578996732	MayMakeModifications
0.659784578996732	NotSublicense
0.659784578996732	MustOffersourceCode
0.659784578996732	MustDistributeCodeUnderThisLicense
0.659784578996732	MayDistributeDerivativeWorksInObjectForm
0.6594273747125473	MustPublishAsLibraryTheModifiedVersion
0.6439063612807063	NotPermissionToUseTradeMarks
0.6439063612807063	MayUseInHouse
0.6434698162962094	MustProvideInformationHowToGetsourceCode
0.6368708098448802	MayAddDifferentLicenseTermsToYourModifications
0.6368708098448802	MustLicenseDerivativesWorksUnderCombatibleLicense
0.6368708098448802	MustDistributeCopyOfNoticeText
0.6368708098448802	MayAddYourOwnCopyrightStatementToYourModifications
0.5201750870719964	MustNotMisrepresentTheOriginalAuthorOfSoftware
0.5137522931996759	MayAddYourOwnChoiceOfLicenselfCodeWasntUnderThisLicense
0.5137522931996759	MustRetainLicenseNotice

0.5137522931996759	MustRetainInsourceCodeOfDerivativeWorkAllTradeMarksPatentCopyrightsOfOriginalWork
--------------------	---

MPL-2.0

0.7129133605647052	MayAcceptWarrantyOrAdditionalLiability
0.7129133605647052	LimitedLiability
0.7129133605647052	MayGrantPatentsRightsFromTheContributorToReceipt
0.4794723036614886	MayDistributeOriginalWorkInsourceForm
0.4794723036614886	MayPubliclyDisplay
0.4794723036614886	MaySublicense
0.4794723036614886	MayPubliclyPerfom
0.44549501324019736	MustMarkModifications
0.43155275050856146	MayDistributeOriginalWorkInObjectForm
0.43155275050856146	NotPermissionToUseTradeMarks
0.420608747679882	ProvideWithoutWarranty
0.414713993963889	NotSublicense
0.4121171701696073	MayDistributeDerivativeWorksInsourceForm
0.4121171701696073	MayMakeDerivativeWorks
0.4121171701696073	MustRetainCopyrightNotice
0.4121171701696073	MayMakeModifications

0.4121171701696073	MayUseInHouse
0.4121171701696073	MayDistributeDerivativeWorksInObjectForm
0.4121171701696073	MayCopy
0.40189882366578633	MustDistributeCodeUnderThisLicense
0.3930310896217433	MustProvideInformationHowToGetsourceCode
0.3930310896217433	MustOffersourceCode
0.3684240911144474	MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense
0.3684240911144474	MustRetainLicenseNotice
0.3684240911144474	MustRetainInsourceCodeOfDerivativeWorkAllTradeMarksPatentCopyrightsOfOriginalWork
0.3684240911144474	MustLicenseDerivativesWorksUnderCompatibleLicense
0.36089453068153093	MustPublishAsLibraryTheModifiedVersion
0.3534850056525189	MayAddDifferentLicenseTermsToYourModifications
0.3534850056525189	MustDistributeCopyOfNoticeText
0.3534850056525189	MayAddYourOwnCopyrightStatementToYourModifications
0.24841341011479312	MustNotMisrepresentTheOriginalAuthorOfSoftware

MS-PL	
0.8432885792405753	MayAcceptWarrantyOrAdditionalLiability
0.8355029493646704	MayDistributeDerivativeWorksInsourceForm

0.8355029493646704	MayMakeDerivativeWorks
0.8355029493646704	MustRetainCopyrightNotice
0.8355029493646704	MayDistributeOriginalWorkInsourceForm
0.8355029493646704	MayDistributeOriginalWorkInObjectForm
0.8355029493646704	MayPubliclyDisplay
0.8355029493646704	MayMakeModifications
0.8355029493646704	MaySublicense
0.8355029493646704	MayPubliclyPerform
0.8355029493646704	MayUseInHouse
0.8355029493646704	MayDistributeDerivativeWorksInObjectForm
0.8355029493646704	MayCopy
0.7675153084888215	MayAddDifferentLicenseTermsToYourModifications
0.7675153084888215	LimitedLiability
0.7675153084888215	MustMarkModifications
0.7675153084888215	MayGrantPatentsRightsFromTheContributorToReceipt
0.7675153084888215	MustLicenseDerivativesWorksUnderCompatibleLicense
0.7675153084888215	NotPermissionToUseTradeMarks
0.7675153084888215	MustDistributeCopyOfNoticeText

0.7675153084888215	MayAddYourOwnCopyrightStatementToYourModifications
0.7215433449526966	MustNotMisrepresentTheOriginalAuthorOfSoftware
0.67328830505739	ProvideWithoutWarranty
0.6607968006728083	MustProvideInformationHowToGetsourceCode
0.6101537222014184	MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense
0.6101537222014184	MustRetainLicenseNotice
0.6101537222014184	MustOffersourceCode
0.6101537222014184	MustRetainInsourceCodeOfDerivativeWorkAllTradeMarksPatentCopyrightsOfOriginalWork
0.5764958716543769	MustPublishAsLibraryTheModifiedVersion
0.5764958716543769	MustDistributeCodeUnderThisLicense
0.5374964914765656	NotSublicense

MS-RL	
0.8432885792405753	MayAcceptWarrantyOrAdditionalLiability
0.8355029493646704	MayDistributeDerivativeWorksInsourceForm
0.8355029493646704	MayMakeDerivativeWorks
0.8355029493646704	MustRetainCopyrightNotice
0.8355029493646704	MayDistributeOriginalWorkInsourceForm
0.8355029493646704	MayDistributeOriginalWorkInObjectForm

0.8355029493646704	MayPubliclyDisplay
0.8355029493646704	MayMakeModifications
0.8355029493646704	MaySublicense
0.8355029493646704	MayPubliclyPerform
0.8355029493646704	MayUseInHouse
0.8355029493646704	MayDistributeDerivativeWorksInObjectForm
0.8355029493646704	MayCopy
0.7675153084888215	MayAddDifferentLicenseTermsToYourModifications
0.7675153084888215	LimitedLiability
0.7675153084888215	MustMarkModifications
0.7675153084888215	MayGrantPatentsRightsFromTheContributorToReceipt
0.7675153084888215	MustLicenseDerivativesWorksUnderCompatibleLicense
0.7675153084888215	NotPermissionToUseTradeMarks
0.7675153084888215	MustDistributeCopyOfNoticeText
0.7675153084888215	MayAddYourOwnCopyrightStatementToYourModifications
0.7215433449526966	MustNotMisrepresentTheOriginalAuthorOfSoftware
0.67328830505739	ProvideWithoutWarranty
0.6607968006728083	MustProvideInformationHowToGetSourceCode

0.6101537222014184	MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense
0.6101537222014184	MustRetainLicenseNotice
0.6101537222014184	MustOffersourceCode
0.6101537222014184	MustRetainInsourceCodeOfDerivativeWorkAllTradeMarksPatentCopyrightsOfOriginalWork
0.5764958716543769	MustPublishAsLibraryTheModifiedVersion
0.5764958716543769	MustDistributeCodeUnderThisLicense
0.5374964914765656	NotSublicense

OSL-3.0

0.8432885792405753	MayAcceptWarrantyOrAdditionalLiability
0.8355029493646704	MayDistributeDerivativeWorksInsourceForm
0.8355029493646704	MayMakeDerivativeWorks
0.8355029493646704	MustRetainCopyrightNotice
0.8355029493646704	MayDistributeOriginalWorkInsourceForm
0.8355029493646704	MayDistributeOriginalWorkInObjectForm
0.8355029493646704	MayPubliclyDisplay
0.8355029493646704	MayMakeModifications
0.8355029493646704	MaySublicense
0.8355029493646704	MayPubliclyPerfom

0.8355029493646704	MayUseInHouse
0.8355029493646704	MayDistributeDerivativeWorksInObjectForm
0.8355029493646704	MayCopy
0.7675153084888215	MayAddDifferentLicenseTermsToYourModifications
0.7675153084888215	LimitedLiability
0.7675153084888215	MustMarkModifications
0.7675153084888215	MayGrantPatentsRightsFromTheContributorToReceipt
0.7675153084888215	MustLicenseDerivativesWorksUnderCombatibleLicense
0.7675153084888215	NotPermissionToUseTradeMarks
0.7675153084888215	MustDistributeCopyOfNoticeText
0.7675153084888215	MayAddYourOwnCopyrightStatementToYourModifications
0.7215433449526966	MustNotMisrepresentTheOriginalAuthorOfSoftware
0.67328830505739	ProvideWithoutWarranty
0.6607968006728083	MustProvideInformationHowToGetsourceCode
0.6101537222014184	MayAddYourOwnChoiceOfLicenselfCodeWasntUnderThisLicense
0.6101537222014184	MustRetainLicenseNotice
0.6101537222014184	MustOffersourceCode
0.6101537222014184	MustRetainInsourceCodeOfDerivativeWorkAllTradeMarksPatentCopyrightsOfOriginalWork

0.5764958716543769	MustPublishAsLibraryTheModifiedVersion
0.5764958716543769	MustDistributeCodeUnderThisLicense
0.5374964914765656	NotSublicense

X11	
0.8432885792405753	MayAcceptWarrantyOrAdditionalLiability
0.8355029493646704	MayDistributeDerivativeWorksInsourceForm
0.8355029493646704	MayMakeDerivativeWorks
0.8355029493646704	MustRetainCopyrightNotice
0.8355029493646704	MayDistributeOriginalWorkInsourceForm
0.8355029493646704	MayDistributeOriginalWorkInObjectForm
0.8355029493646704	MayPubliclyDisplay
0.8355029493646704	MayMakeModifications
0.8355029493646704	MaySublicense
0.8355029493646704	MayPubliclyPerform
0.8355029493646704	MayUseInHouse
0.8355029493646704	MayDistributeDerivativeWorksInObjectForm
0.8355029493646704	MayCopy
0.7675153084888215	MayAddDifferentLicenseTermsToYourModifications

0.7675153084888 215	LimitedLiability
0.7675153084888 215	MustMarkModifications
0.7675153084888 215	MayGrantPatentsRightsFromTheContributorToReceipt
0.7675153084888 215	MustLicenseDerivativesWorksUnderCombatibleLicense
0.7675153084888 215	NotPermissionToUseTradeMarks
0.7675153084888 215	MustDistributeCopyOfNoticeText
0.7675153084888 215	MayAddYourOwnCopyrightStatementToYourModifications
0.7215433449526 966	MustNotMisrepresentTheOriginalAuthorOfSoftware
0.6732883050573 9	ProvideWithoutWarranty
0.6607968006728 083	MustProvideInformationHowToGetsourceCode
0.6101537222014 184	MayAddYourOwnChoiceOfLicenselfCodeWasntUnderThisLicense
0.6101537222014 184	MustRetainLicenseNotice
0.6101537222014 184	MustOffersourceCode
0.6101537222014 184	MustRetainInsourceCodeOfDerivativeWorkAllTradeMarksPatentCopyrights OfOriginalWork
0.5764958716543 769	MustPublishAsLibraryTheModifiedVersion
0.5764958716543 769	MustDistibuteCodeUnderThisLicense
0.5374964914765 656	NotSublicense

Zlib

0.8432885792405753	MayAcceptWarrantyOrAdditionalLiability
0.8355029493646704	MayDistributeDerivativeWorksInsourceForm
0.8355029493646704	MayMakeDerivativeWorks
0.8355029493646704	MustRetainCopyrightNotice
0.8355029493646704	MayDistributeOriginalWorkInsourceForm
0.8355029493646704	MayDistributeOriginalWorkInObjectForm
0.8355029493646704	MayPubliclyDisplay
0.8355029493646704	MayMakeModifications
0.8355029493646704	MaySublicense
0.8355029493646704	MayPubliclyPerform
0.8355029493646704	MayUseInHouse
0.8355029493646704	MayDistributeDerivativeWorksInObjectForm
0.8355029493646704	MayCopy
0.7675153084888215	MayAddDifferentLicenseTermsToYourModifications
0.7675153084888215	LimitedLiability
0.7675153084888215	MustMarkModifications
0.7675153084888215	MayGrantPatentsRightsFromTheContributorToReceipt
0.7675153084888215	MustLicenseDerivativesWorksUnderCompatibleLicense

0.7675153084888 215	NotPermissionToUseTradeMarks
0.7675153084888 215	MustDistributeCopyOfNoticeText
0.7675153084888 215	MayAddYourOwnCopyrightStatementToYourModifications
0.7215433449526 966	MustNotMisrepresentTheOriginalAuthorOfSoftware
0.6732883050573 9	ProvideWithoutWarranty
0.6607968006728 083	MustProvideInformationHowToGetsourceCode
0.6101537222014 184	MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense
0.6101537222014 184	MustRetainLicenseNotice
0.6101537222014 184	MustOffersourceCode
0.6101537222014 184	MustRetainInsourceCodeOfDerivativeWorkAllTradeMarksPatentCopyrights OfOriginalWork
0.5764958716543 769	MustPublishAsLibraryTheModifiedVersion
0.5764958716543 769	MustDistributeCodeUnderThisLicense
0.5374964914765 656	NotSublicense

Παράρτημα Β

AAL		
<u>License term</u>	<u>Exists</u>	<u>Line</u>
LimitedLiability		
MayAcceptWarrantyOrAdditionalLiability		
MayAddDifferentLicenseTermsToYourModifications		
MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense		
MayAddYourOwnCopyrightStatementToYourModifications		
MayCopy	✓	4
MayDistributeDerivativeWorksInObjectForm	✓	4
MayDistributeDerivativeWorksInSourceForm	✓	4
MayDistributeOriginalWorkInObjectForm	✓	4
MayDistributeOriginalWorkInSourceForm	✓	4
MayGrantPatentsRightsFromTheContributorToReceipt	✓	8
MayMakeDerivativeWorks	✓	4
MayMakeModifications	✓	4
MayPubliclyDisplay	✓	4
MayPubliclyPerform	✓	4
MaySublicense		
MayUseInHouse	✓	4
MustDistributeCodeUnderThisLicense		
MustDistributeCopyOfNoticeText		
MustLicenseDerivativesWorksUnderCompatibleLicense		
MustMarkModifications		
MustNotMisrepresentTheOriginalAuthorOfSoftware	✓	4
MustOfferSourceCode		
MustProvideInformationHowToGetSourceCode		
MustPublishAsLibraryTheModifiedVersion		
MustRetainCopyrightNotice	✓	5

MustRetainInsourceCodeOfDerivativeWorkAllTradeMarksPatent...		
MustRetainLicenseNotice		
NotPermissionToUseTradeMarks	√	8
NotSublicense		
ProvideWithoutWarranty	√	10

APAFML		
<u>License term</u>	<u>Exists</u>	<u>Line</u>
LimitedLiability		
MayAcceptWarrantyOrAdditionalLiability		
MayAddDifferentLicenseTermsToYourModifications		
MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense		
MayAddYourOwnCopyrightStatementToYourModifications		
MayCopy	√	1
MayDistributeDerivativeWorksInObjectForm	√	1
MayDistributeDerivativeWorksInsourceForm	√	1
MayDistributeOriginalWorkInObjectForm	√	1
MayDistributeOriginalWorkInsourceForm	√	1
MayGrantPatentsRightsFromTheContributorToReceipt		
MayMakeDerivativeWorks	√	1
MayMakeModifications	√	1
MayPubliclyDisplay		
MayPubliclyPerform		
MaySublicense		
MayUseInHouse	√	1
MustDistributeCodeUnderThisLicense		
MustDistributeCopyOfNoticeText		
MustLicenseDerivativesWorksUnderCompatibleLicense		
MustMarkModifications	√	1
MustNotMisrepresentTheOriginalAuthorOfSoftware		
MustOfferSourceCode		
MustProvideInformationHowToGetSourceCode		

MustPublishAsLibraryTheModifiedVersion		
MustRetainCopyrightNotice	√	1
MustRetainInSourceCodeOfDerivativeWorkAllTradeMarksPatent...	√	1
MustRetainLicenseNotice		
NotPermissionToUseTradeMarks		
NotSublicense	√	1
ProvideWithoutWarranty		

Caldera		
<u>License term</u>	<u>Exists</u>	<u>Line</u>
LimitedLiability		
MayAcceptWarrantyOrAdditionalLiability		
MayAddDifferentLicenseTermsToYourModifications		
MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense		
MayAddYourOwnCopyrightStatementToYourModifications		
MayCopy	√	5
MayDistributeDerivativeWorksInObjectForm	√	5
MayDistributeDerivativeWorksInsourceForm	√	5
MayDistributeOriginalWorkInObjectForm	√	5
MayDistributeOriginalWorkInsourceForm	√	5
MayGrantPatentsRightsFromTheContributorToReceipt		
MayMakeDerivativeWorks	√	5
MayMakeModifications	√	5
MayPubliclyDisplay	√	5
MayPubliclyPerfom	√	5
MaySublicense		
MayUseInHouse	√	5
MustDistibuteCodeUnderThisLicense		
MustDistributeCopyOfNoticeText		
MustLicenseDerivativesWorksUnderCombatibleLicense		
MustMarkModifications		
MustNotMisrepresentTheOriginalAuthorOfSoftware		

MustOffersourceCode		
MustProvideInformationHowToGetsourceCode		
MustPublishAsLibraryTheModifiedVersion		
MustRetainCopyrightNotice	√	6
MustRetainInsourceCodeOfDerivativeWorkAllTradeMarksPatent...	√	9
MustRetainLicenseNotice	√	6
NotPermissionToUseTradeMarks	√	8
NotSublicense	√	6
ProvideWithoutWarranty	√	10

Curl		
<u>License term</u>	<u>Exists</u>	<u>Line</u>
LimitedLiability		
MayAcceptWarrantyOrAdditionalLiability		
MayAddDifferentLicenseTermsToYourModifications		
MayAddYourOwnChoiceOfLicenselfCodeWasntUnderThisLicense		
MayAddYourOwnCopyrightStatementToYourModifications		
MayCopy	√	2
MayDistributeDerivativeWorksInObjectForm	√	2
MayDistributeDerivativeWorksInsourceForm	√	2
MayDistributeOriginalWorkInObjectForm	√	2
MayDistributeOriginalWorkInsourceForm	√	2
MayGrantPatentsRightsFromTheContributorToReceipt		
MayMakeDerivativeWorks	√	2
MayMakeModifications	√	2
MayPubliclyDisplay		
MayPubliclyPerfom		
MaySublicense		
MayUseInHouse	√	2
MustDistributeCodeUnderThisLicense		
MustDistributeCopyOfNoticeText		
MustLicenseDerivativesWorksUnderCombatibleLicense		

MustMarkModifications		
MustNotMisrepresentTheOriginalAuthorOfSoftware		
MustOffersourceCode		
MustProvideInformationHowToGetsourceCode		
MustPublishAsLibraryTheModifiedVersion		
MustRetainCopyrightNotice	√	2
MustRetainInsourceCodeOfDerivativeWorkAllTradeMarksPatent...		
MustRetainLicenseNotice	√	2
NotPermissionToUseTradeMarks	√	4
NotSublicense	√	2
ProvideWithoutWarranty	√	3

Sleepycat		
<u>License term</u>	<u>Exists</u>	<u>Line</u>
LimitedLiability		
MayAcceptWarrantyOrAdditionalLiability		
MayAddDifferentLicenseTermsToYourModifications		
MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense		
MayAddYourOwnCopyrightStatementToYourModifications		
MayCopy	√	1
MayDistributeDerivativeWorksInObjectForm	√	1
MayDistributeDerivativeWorksInsourceForm	√	1
MayDistributeOriginalWorkInObjectForm	√	1
MayDistributeOriginalWorkInsourceForm	√	1
MayGrantPatentsRightsFromTheContributorToReceipt		
MayMakeDerivativeWorks	√	1
MayMakeModifications	√	1
MayPubliclyDisplay		
MayPubliclyPerfom		
MaySublicense		
MayUseInHouse		
MustDistributeCodeUnderThisLicense		

MustDistributeCopyOfNoticeText	√	2
MustLicenseDerivativesWorksUnderCompatibleLicense		
MustMarkModifications		
MustNotMisrepresentTheOriginalAuthorOfSoftware		
MustOffersourceCode		
MustProvideInformationHowToGetsourceCode	√	4
MustPublishAsLibraryTheModifiedVersion		
MustRetainCopyrightNotice	√	2
MustRetainInsourceCodeOfDerivativeWorkAllTradeMarksPatent...		
MustRetainLicenseNotice		
NotPermissionToUseTradeMarks		
NotSublicense	√	2
ProvideWithoutWarranty	√	5

VOSTROM		
<u>License term</u>	<u>Exists</u>	<u>Line</u>
LimitedLiability	√	6
MayAcceptWarrantyOrAdditionalLiability		
MayAddDifferentLicenseTermsToYourModifications		
MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense		
MayAddYourOwnCopyrightStatementToYourModifications		
MayCopy	√	3
MayDistributeDerivativeWorksInObjectForm	√	3
MayDistributeDerivativeWorksInsourceForm	√	3
MayDistributeOriginalWorkInObjectForm	√	3
MayDistributeOriginalWorkInsourceForm	√	3
MayGrantPatentsRightsFromTheContributorToReceipt		
MayMakeDerivativeWorks	√	3
MayMakeModifications	√	3
MayPubliclyDisplay	√	3
MayPubliclyPerform	√	3
MaySublicense	√	9

MayUseInHouse	√	3
MustDistributeCodeUnderThisLicense		
MustDistributeCopyOfNoticeText	√	7
MustLicenseDerivativesWorksUnderCompatibleLicense	√	9
MustMarkModifications	√	5
MustNotMisrepresentTheOriginalAuthorOfSoftware		
MustOffersourceCode	√	4
MustProvideInformationHowToGetsourceCode		
MustPublishAsLibraryTheModifiedVersion		
MustRetainCopyrightNotice	√	7
MustRetainInsourceCodeOfDerivativeWorkAllTradeMarksPatent...		
MustRetainLicenseNotice	√	7
NotPermissionToUseTradeMarks		
NotSublicense		
ProvideWithoutWarranty	√	13

Xerox		
<u>License term</u>	<u>Exists</u>	<u>Line</u>
LimitedLiability		
MayAcceptWarrantyOrAdditionalLiability		
MayAddDifferentLicenseTermsToYourModifications		
MayAddYourOwnChoiceOfLicenseIfCodeWasntUnderThisLicense		
MayAddYourOwnCopyrightStatementToYourModifications		
MayCopy	√	1
MayDistributeDerivativeWorksInObjectForm		
MayDistributeDerivativeWorksInsourceForm		
MayDistributeOriginalWorkInObjectForm		
MayDistributeOriginalWorkInsourceForm		
MayGrantPatentsRightsFromTheContributorToReceipt		
MayMakeDerivativeWorks	√	1
MayMakeModifications	√	1
MayPubliclyDisplay		

MayPubliclyPerform		
MaySublicense		
MayUseInHouse	✓	1
MustDistributeCodeUnderThisLicense	✓	1
MustDistributeCopyOfNoticeText		
MustLicenseDerivativesWorksUnderCompatibleLicense		
MustMarkModifications		
MustNotMisrepresentTheOriginalAuthorOfSoftware		
MustOffersourceCode		
MustProvideInformationHowToGetsourceCode		
MustPublishAsLibraryTheModifiedVersion		
MustRetainCopyrightNotice	✓	1
MustRetainInsourceCodeOfDerivativeWorkAllTradeMarksPatent...		
MustRetainLicenseNotice	✓	1
NotPermissionToUseTradeMarks		
NotSublicense	✓	1
ProvideWithoutWarranty	✓	2