Government Gazette Text Mining, Cross-Linking and Codification - 3gm

Google Summer of Code

FOSSCOMM 2018
Συνέδριο Κοινοτήτων Ελεύθερου / Ανοικτού Λογισμικού

Marios Papachristou
GFOSS - Open Technologies Alliance

github.com/eellak/gsoc2018-3gm

Mentors: D. Spinellis, A. Zavras, S. Kapidakis

Problem & Project Statement

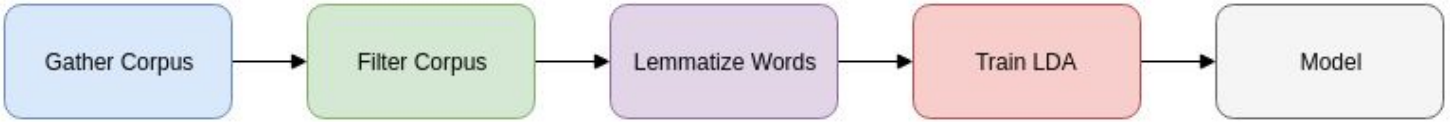
- Codification (Wikipedia): In law, codification is the process of collecting and restating the law of a jurisdiction in certain areas, usually by subject, forming a legal code, i.e. a codex (book) of law.
- Done by hand!
- Automate it!

Features

1. Document parser
2. Named Entities for Legal Acts (e.g. Laws, Legislative Decrees etc.) encoded in regular expressions.
3. Similarity analyzer using topic models for finding Government Gazette Issues that have the same topics.
4. MongoDB Integration
5. Fetching Tool for automated fetching of documents from ET
6. Parallelized tool for batch conversion of documents with pdf2txt
7. Digitized archive of Government Gazette Issues from 1976 - today in PDF and plaintext format
8. Web application hosted at 3gm.ellak.gr
9. RESTful API
10. Versioning system
11. Ranking System.
12. Summarization Module using TextRank for providing summaries at the search results. (alpha)
13. Amendment Detection Procedure

Topic Modelling Pipeline

1. Export Corpus from DB
2. Filter Corpus (remove stopwords, numbers, symbols etc.)
3. Lemmatize words using spaCy's lemmatizer
4. Train LDA model with scikit-learn
5. Produce LDA Model



Amendment Detection

Heuristic methods / Hybrid for detecting amendments. For example (taken from Greek Government Gazette):

Amendment

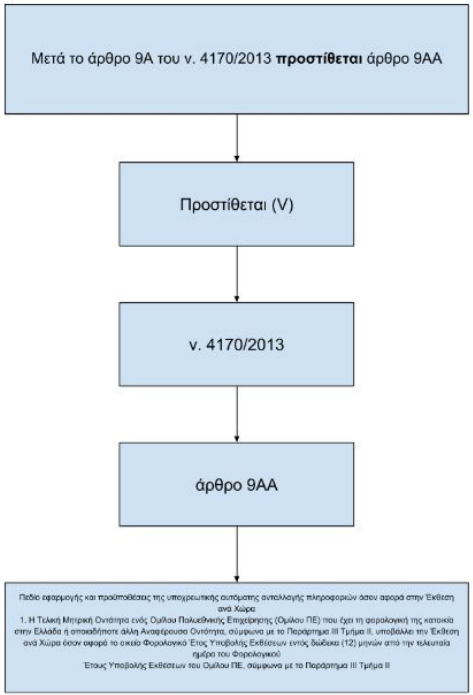
> Μετά το άρθρο 9Α του ν. 4170/2013 **προστίθεται** *άρθρο 9ΑΑ*, ως εξής:

Main Body / Extract

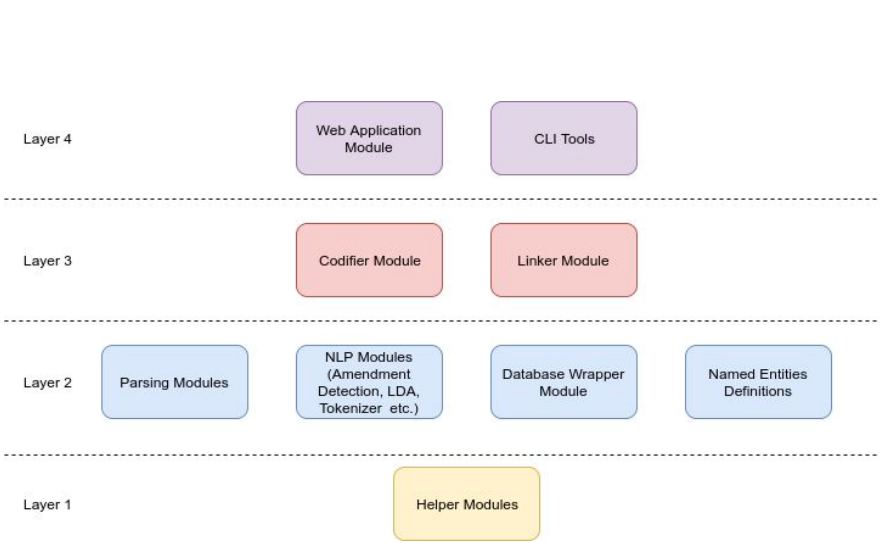
> Άρθρο 9ΑΑ
>
> Lorem Ipsum.....

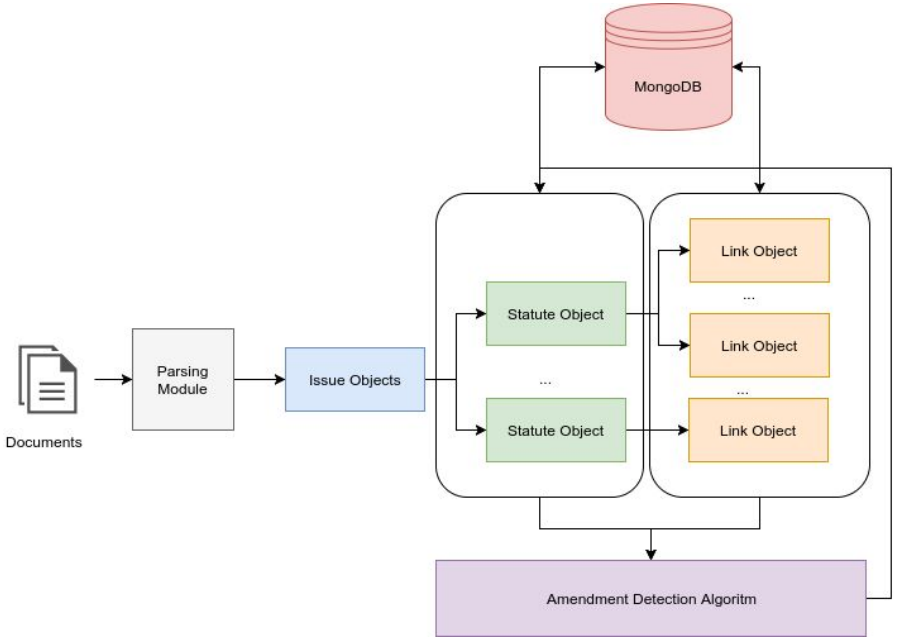
Results in a database query that adds an article on a MongoDB

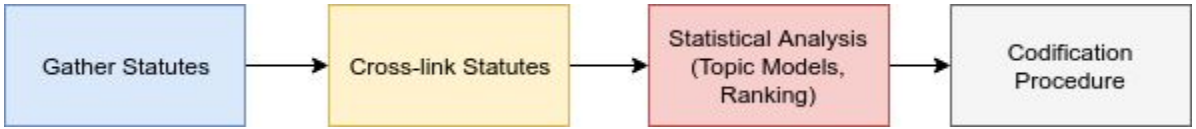
database



Project Architecture







Examples & Results (law 4009/2011) - diff page

https://goo.gl/Wvsf3e

καλύπτει μια ενότητα συγγενών επιστημονικών κλάδων και εξασφαλίζει τη διεπιστημονική προσέγγιση, τη μεταξύ τους επικοινωνία και τον αναγκαίο για τη διδασκαλία και την έρευνα συντονισμό τους. Η σχολή συντονίζει και εποπτεύει τη λειτουργία των προγραμμάτων σπουδών, αναθέτει την υλοποίησή τους σε τμήματα κατά την έννοια της επόμενης παραγράφου και απονέμει τους αντίστοιχους τίτλους σπουδών, κατά τα οριζόμενα στον Οργανισμό και τον Εσωτερικό Κανονισμό του ιδρύματος. Τα προγράμματα σπουδών οργανώνονται ή καταργούνται με απόφαση του πρύτανη, που εκδίδεται ύστερα από εισήγηση της κοσμητείας και γνώμη της Συγκλήτου, εγκρίνεται από το Συμβούλιο και δημοσιεύεται στην Εφημερίδα της Κυβερνήσεως.

- 2. Το τμήμα αποτελεί τη βασική εκπαιδευτική μονάδα, του ιδρύματος, προάγει την επιστήμη, την τεχνολογία ή τις τέχνες στο αντίστοιχο επιστημονικό πεδίο, οργανώνει τη διδασκαλία στο πλαίσιο ενός προγράμματος σπουδών και εξασφαλίζει τη συνεχή βελτίωση της μάθησης σε αυτό. Το τμήμα αποτελείται από σύνολο των καθηγητών της σχολής που διδάσκουν σε ένα πρόγραμμα σπουδών. .

+ 2. Το Τμήμα αποτελεί τη βασική εκπαιδευτική και ακαδημαϊκή μονάδα του Ιδρύματος, προάγει την επιστήμη, την τεχνολογία ή τις τέχνες στο αντίστοιχο επιστημονικό πεδίο, οργανώνει τη διδασκαλία στο πλαίσιο του προγράμματος σπουδών και εξασφαλίζει τη συνεχή βελτίωση της μάθησης σε αυτό. Το Τμήμα αποτελείται από το σύνολο των Καθηγητών, των λεκτόρων, των μελών του Ειδικού Εκπαιδευτικού Προσωπικού (ΕΕΠ), των μελών του Εργαστηριακού Διδακτικού Προσωπικού (ΕΔΙΠ) και των μελών του Ειδικού Τεχνικού Εργαστηριακού Προσωπικού (ΕΤΕΠ), που υπηρετούν σε αυτό.

3. Τα εργαστήρια, οι κλινικές και τα μουσεία υπάγονται στις σχολές, όπως ορίζεται στον Οργανισμό του ιδρύματος.

- 4. Η σχολή μεταπτυχιακών σπουδών αποτελεί τη βασική διοικητική μονάδα που εξασφαλίζει τη διεπιστημονική συνεργασία και επικοινωνία, συντονίζει και οργανώνει τα προγράμματα μεταπτυχιακών και διδακτορικών σπουδών του ιδρύματος και αναθέτει την υλοποίησή τους σε τμήματα ή ομάδες διδασκόντων. Η σχολή αυτή ιδρύεται με τον Οργανισμό του ιδρύματος.

+ 4. Η Σχολή συντονίζει και οργανώνει τα προγράμματα μεταπτυχιακών και διδακτορικών σπουδών που υπάγονται σε αυτή και αναθέτει την υλοποίησή τους στα Τμήματα ή σε ομάδες διδασκόντων. Τα προγράμματα μεταπτυχιακών και διδακτορικών σπουδών λειτουργούν σύμφωνα με τον Οργανισμό του Ιδρύματος.

5. Η σχολή δια βίου μάθησης αποτελεί τη βασική διοικητική μονάδα που εξασφαλίζει το συντονισμό και τη διεπιστημονική συνεργασία στην ανάπτυξη προγραμμάτων δια βίου μάθησης. Η σχολή αυτή ιδρύεται με τον Οργανισμό του ιδρύματος.

- 6. Με την επιφύλαξη των οριζομένων στις παραγράφους 4 και 5, με προεδρικό διάταγμα, που εκδίδεται με πρόταση των Υπουργών Διοικητικής Μεταρρύθμισης και Ηλεκτρονικής Διακυβέρνησης, Οικονομικών και Παιδείας, Δια Βίου Μάθησης και Θρησκευμάτων, ύστερα από γνώμη του Συμβουλίου των οικείων ιδρυμάτων και της Αρχής Διασφάλισης και Πιστοποίησης της Ποιότητας στην Ανώτατη Εκπαίδευση (ΑΔΙΠ), μπορούν να συγχωνεύονται, κατατέμνονται, μετονομάζονται και

Challenges

1. Government Gazette Issues may not always follow guidelines
2. Improving detection methods
3. No substantial NLP progress in Greek

Contributing to the project

We follow the **fork & pull-request** workflow for contributions.

The issue page contains things that you can work on.

Thank you!

3gm.ellak.gr | github.com/eellak/gsoc2018-3gm

papachristoumarios [at] gmail [dot] com

5862 31C0 05F4 8C71 A5B6 A1CE D6BC 45BD E0DC 0EDA