netcompany

intrasoft

«Αναβάθμιση και επέκταση λειτουργικότητας του Ολοκληρωμένου Συστήματος Διαχείρισης Δικαστικών Υποθέσεων Πολιτικής & Ποινικής Δικαιοσύνης (ΟΣΔΔΥ-ΠΠ) - ΟΣΔΔΥ-ΠΠ Β’ Φάση»

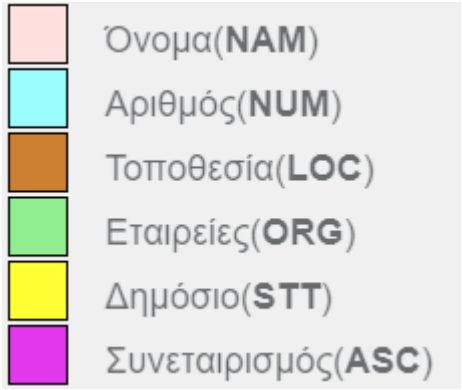
Χρήση ΤΝ στην ανωνυμοποίηση δικαστικών αποφάσεων

Σετ Δεδομένων

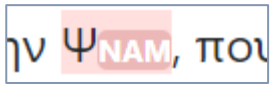
- Υλοποίηση ενός web crawler για την συλλογή δημοσιευμένων αποφάσεων στο διαδίκτυο.
- Συλλογή 7.500 ανωνυμοποιημένων αποφάσεων από την ιστοσελίδα του Αρείου Πάγου.
  - https://www.areiospagos.gr/nomologia/apofaseis.asp
- Επισημείωση των δεδομένων.
- Χρήση των επισημειωμένων δεδομένων για εκπαίδευση νευρωνικών δικτύων.

Επισημείωση Δεδομένων

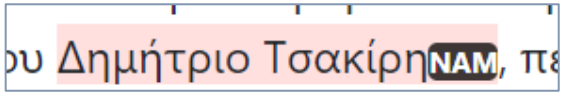
- Σε κάθε μία από τις ανωνυμοποιημένες αποφάσεις επισημειώνονται οι εξής οντότητες:



- (οι παραπάνω κατηγορίες έχουν προκύψει με βάση τους κανόνες ανωνυμοποίησης)
- Κάθε οντότητα χαρακτηρίζεται επίσης σαν:
  - Ανωνυμοποιημένη (ΜΟ 22.83/Απόφαση):



  - Μη ανωνυμοποιημένη (ΜΟ 98.18/Απόφαση)



netcompany
intrasoft

Κανόνες Ανωνυμοποίησης

ΑΝΩΝΥΜΟΠΟΙΟΥΜΕ

- Ονόματα διαδίκων (ιδιώτες, πατριάρχες, μητροπολίτες, δικηγόροι)
- Ονόματα Ενόρκων
- Ονόματα Μαρτύρων
- Ονόματα Δικαστικών συμπαραστατών
- Ονόματα άλλων εμπλεκομένων με την υπόθεση που παρέχουν γνωματεύσεις (πχ. Πραγματογνώμονες, γιατροί, μηχανικοί, ορκωτοί λογιστές, κλπ) - optional
- Ονόματα Δικαστικών επιμελητών - optional
- Στα ονόματα πλοίων αφήνουμε μόνο το πρώτο γράμμα, πχ. “ΑΡΓΩ Α…” . Αν το αρχικό γράμμα είναι το ίδιο για δυο ή περισσότερα πλοία τότε συμπληρώνουμε δίπλα στο αρχικό γράμμα και έναν αριθμό πχ Α1, Α2, …
- Εταιρίες (ΟΕ, ΕΕ, ΑΕ, ΕΠΕ κλπ, ή αγνώστου τύπου), το όνομα, την περιοχή και το νομό που εδρεύουν
- Κοινοπραξίες που έχουν ονόματα προσώπων.
- Οδούς, περιοχές που συγκεκριμενοποιούν τον εμπλεκόμενο (όχι τις οδούς που αναφέρονται σε ατύχημα)
- Αριθμούς Τηλεφώνου

netcompany

intrasoft

Κανόνες Ανωνυμοποίησης

ΑΝΩΝΥΜΟΠΟΙΟΥΜΕ

- Αριθμούς κυκλοφορίας οχημάτων
- Αριθμούς δελτίου ταυτότητας, διαβατηρίου, ληξιαρχικές πράξεις θανάτου, αριθμούς δημοτολογίου, τραπεζικών λογαριασμών, τραπεζικών συμβάσεων
- Αριθμούς χρηματιστηριακών λογαριασμών, τραπεζικών επιταγών
- Αριθμούς συμβολαίων, κτηματολογίου, κληροτεμαχίου, Ο.Τ. (Οικοδομικού Τετραγώνου)
- Αριθμούς μητρώου ΙΚΑ ή άλλου ασφαλιστικού Ταμείου,
- Αριθμούς που συνδέονται απόλυτα με κάποιον εμπλεκόμενο στην υπόθεση.
- Συνεταιρισμούς, μόνο εάν αναφέρεται όνομα φυσικού προσώπου, το οποίο σβήνουμε
- Ν.Π.Ι.Δ, μόνο αν αναφέρεται όνομα φυσικού προσώπου, το οποίο σβήνουμε

netcompany
intrasoft

Κανόνες Ανωνυμοποίησης

ΔΕΝ ΑΝΩΝΥΜΟΠΟΙΟΥΜΕ

- Δικαστές, παρέδρους, δικηγόρους, πληρεξούσιους δικηγόρους, συμβολαιογράφους, εκτός εάν είναι διάδικοι
- Δικαστικές αποφάσεις και δικαστήρια
- Συνεταιρισμούς - αλλά αν αναφέρεται όνομα φυσικού προσώπου, το σβήνουμε
- Ν.Π.Ι.Δ, - αλλά αν αναφέρεται όνομα φυσικού προσώπου, το σβήνουμε
- Ο.Τ.Α, δηλαδή δήμους και κοινότητες (ανωνυμοποιείται ο δήμος/η κοινότητα που ζει κάποιος)
- Συνδικαλιστικές οργανώσεις
- Ελληνικό Δημόσιο, Ν.Π.Δ.Δ, Δ.Ε.Κ.Ο

Παράδειγμα Επισημειωμένης Απόφασης

Συγκροτήθηκε από τους Δικαστές: Ασπασία Καρέλλου NAM, Αντιπρόεδρο του Αρείου Πάγου, Νικόλαο Πάσσο NAM, Παναγιώτη Κατσιρούμπα NAM, Δήμητρα Κοκοτίνη NAM και Γεώργιο Μιχολιά NAM, Αρεοπαγίτες.
Συνεδρίασε δημόσια στο Κατάστημά του, στις 11 Οκτωβρίου 2016 NUM, με την παρουσία και της γραμματέως Αγγελικής Ανυφαντή NAM, για να δικάσει την εξής υπόθεση μεταξύ:
Της αναιρεσείουσας: Α. Α. NAM - Β. NAM του Α. NAM, κατοίκου ... LOC, η οποία εκπροσωπήθηκε από τον πληρεξούσιο δικηγόρο της Γεώργιο Κουφογιάννη NAM, με δήλωση κατ' άρθρο 242 παρ. 2 NUM του Κ.Πολ.Δ., που κατέθεσε προτάσεις.
Της αναιρεσίβλητης: ανώνυμης εταιρείας με την επωνυμία "... Α.Ε." ... ORG, που εδρεύει στο ... LOC, η οποία εκπροσωπήθηκε από την πληρεξούσια δικηγόρο της Ευγενία Σούμπαση, με δήλωση κατ' άρθρο 242 παρ. 2 NUM του Κ.Πολ.Δ., που κατέθεσε προτάσεις
Η ένδικη διαφορά άρχισε με την από 22/8/2005 NUM αγωγή της ήδη αναιρεσείουσας, που κατατέθηκε στο Ειρηνοδικείο Αμαρουσίου STT.
Εκδόθηκαν οι αποφάσεις: 510/2007 NUM του ίδιου Δικαστηρίου και 3808/2012 NUM του Πολυμελούς Πρωτοδικείου Αθηνών STT.
Την αναίρεση της τελευταίας απόφασης ζητεί η αναιρεσείουσα με την από 2/4/2015 NUM αίτησή της.
Κατά τη συζήτηση της αίτησης αυτής, που εκφωνήθηκε από το πινάκιο, οι διάδικοι παραστάθηκαν όπως σημειώνεται πιο πάνω.

netcompany

intrasoft

Μη-Ανωνυμοποιημένες Αποφάσεις

Alignment

- Αφού ολοκληρώθηκε η επισημείωση των ανωνυμοποιημένων αποφάσεων και μας δόθηκαν οι μη-ανωνυμοποιημένες (master) αποφάσεις:
  - Έγινε ένα alignment του κειμένου μεταξύ των αντίστοιχων ανωνυμοποιημένων/μη-ανωνυμοποιημένων αποφάσεων.
  - Η επισημειωμένη πληροφορία αντιστοιχίστηκε στις μη-ανωνυμοποιημένες αποφάσεις.
  - Χρησιμοποιήθηκαν οι (master) μη-ανωνυμοποιημένες αποφάσεις με τις αντίστοιχες (aligned) επισημειώσεις για την εκπαίδευση των μοντέλων (νευρωνικών δικτύων).

RoBERTa

netcompany
intrasoft

- Το RoBERTa (Robustly Optimized BERT Approach) είναι μια παραλλαγή του μοντέλου BERT (Bidirectional Encoder Representations from Transformers), το οποίο αναπτύχθηκε από ερευνητές της Facebook AI (open source).
- Όπως το BERT, το RoBERTa είναι ένα μοντέλο γλώσσας βασισμένο σε transformer που χρησιμοποιεί self-attention για να επεξεργάζεται ακολουθίες εισόδου και να παράγει εξειδικευμένες αναπαραστάσεις λέξεων σε μια πρόταση.
- Το RoBERTa εκπαιδεύτηκε σε ένα σετ δεδομένων 160GB, περισσότερο από δέκα φορές μεγαλύτερο από αυτό που χρησιμοποιήθηκε για το BERT.
- Έχει αποδειχθεί ότι το RoBERTa υπερτερεί του BERT και άλλων κορυφαίων μοντέλων σε διάφορες εργασίες επεξεργασίας φυσικής γλώσσας, όπως η μετάφραση γλώσσας, η ταξινόμηση κειμένου, και η απάντηση ερωτήσεων, γίνοντας δημοφιλής επιλογή για έρευνα και βιομηχανικές εφαρμογές.

Τροποποιήσεις σε σχέση με το BERT

- Κατάργηση της Πρόβλεψης Επόμενης Πρότασης (NSP):
  - Στην πρόβλεψη επόμενης πρότασης, το μοντέλο προσπαθεί να ανιχνεύσει αν τα τμήματα κειμένου προέρχονται από το ίδιο ή διαφορετικά έγγραφα με χρήση της απώλειας NSP. Οι συγγραφείς διαπίστωσαν ότι η απαλοιφή της NSP βελτιώνει ή τουλάχιστον διατηρεί την απόδοση σε μεταγενέστερες εργασίες.
- Μεγαλύτερα Μεγέθη Batches & Μεγαλύτερες Ακολουθίες:
- Το RoBERTa εκπαιδεύεται με μεγαλύτερα batches και περισσότερα δεδομένα ανά batch, βελτιώνοντας τη γενική ακρίβεια και ευκολότερη παραλληλοποίηση μέσω διανεμημένης εκπαίδευσης.
- Δυναμική Αλλαγή του Μοτίβου Μασκαρίσματος:
  - Σε αντίθεση με το BERT, το RoBERTa χρησιμοποιεί δυναμικό μασκάρισμα, αλλάζοντας το μοτίβο μάσκας κάθε φορά που τα δεδομένα εισάγονται στο μοντέλο, αυξάνοντας την ικανότητα γενίκευσης του μοντέλου.

Εκπαίδευση

- Το custom dataset μας χωρίστηκε σε train, validation και test set με τα εξής ποσοστά:
  - Train set: 70%
  - Validation set: 15%
  - Test set: 15%
- Το μοντέλο εκπαιδεύτηκε για 15 epochs.
- Αναγνωρίζει τις οντότητες και τις κατηγοριοποιεί στις εξής **2 κλάσεις**:
  - **Ανωνυμοποίηση** (οντότητες προς ανωνυμοποίηση)
  - **Μη ανωνυμοποίηση** (οντότητες που δεν πρέπει να ανωνυμοποιηθούν π.χ. ονόματα δικαστών)