γλωσσAPI (https://github.com/eellak/glossAPI)

Project Plenary Meeting, 11/7/2024

- **Αποστολή:** ***Ελληνικό γλωσσικό μοντέλο ανοιχτού λογισμικού «Greek OSS LLM», σε συνεργασία με όλα τα μέλη της ΕΕΛΛΑΚ.***

- Διαφανής και ανθρωποκεντρική ενσωμάτωση της ΤΝ και των Μεγάλων Γλωσσικών Μοντέλων στην Αλληλεπίδραση των Πολιτών με τη Δημόσια Διοίκηση.
- Υπάρχουν πολύγλωσσα LLMs διαθέσιμα όπως το meltemi, αλλά τα περισσότερα χρησιμοποιούν μόνο ένα μικρό μέρος των ελληνικών δεδομένων (BLOOM, Pythia) ή ακόμη και καθόλου ελληνικά δεδομένα (LLaMA).
- Μοντέλα όπως το GPT-4 & DEMINI φαίνεται να αποδίδουν καλά σε ελληνικά δεδομένα, αλλά δεν είναι ανοικτά, ενώ άλλα ελληνικά μοντέλα, όπως το meltemi, έχουν εκπαιδευτεί σε σύνολο ελληνικών κειμένων υψηλής ποιότητας, τα οποία ωστόσο δεν είναι διαθέσιμα στο κοινό λόγω ζητημάτων αδειοδότησης

- **Οφέλη του GlossAPI**: (α) Απόρρητο και ασφάλεια δεδομένων, (β) Εξάρτηση και προσαρμογή, (γ) Κόστος και επεκτασιμότητα, (δ) Πρόσβαση και διαθεσιμότητα.

digiGOV innoHUB

ΕΕΛΛΑΚ ΟΡΓΑΝΙΣΜΟΣ ΑΝΟΙΧΤΩΝ ΤΕΧΝΟΛΟΓΙΩΝ

- Το GlossApi είναι ένα έργο εντελώς διαφορετικής κλίμακας και εμβέλειας. Η διακυβέρνηση και ο ποιοτικός έλεγχος του συνόλου δεδομένων και η δημιουργία ενός ανοιχτού, δεοντολογικά αποκτημένου, μηχανικά προσπελάσιμου, και αντιπροσωπευτικού της ελληνικής γλώσσας σώματος παραδειγμάτων εκπαίδευσης είναι πρωταρχικής σημασίας.

Πιο συγκεκριμένα το GlossAPI εκτείνεται σε τομείς όπως:

- Καταγραφή, ιεράρχηση, απόκτηση και καθαρισμός δεδομένων κειμένου.
- Ενοποίηση με δεδομένα άλλων φορέων για ολοκληρωμένα σύνολα δεδομένων.
- Ποιοτικός έλεγχος και ανάπτυξη ενδιάμεσων εργαλείων επισημείωσης και καθαρισμού.
- Ανάπτυξη υλικών αξιολόγησης σε διαφορετικούς τομείς: γλωσσική ορθότητα, πραγματολογική συνοχή, κοινωνική καταλληλότητα και συμπερίληψη.

- Στατιστική επεξεργασία σε όλα τα στάδια από το μετασχηματισμό δεδομένων και τη συμφωνία μεταξύ επισημειωτών με την τελική επίδοση ενδιάμεσων βοηθητικών μοντέλων και μεγάλων γλωσσικών μοντέλων.
- Διαμόρφωση λύσεων βασισμένων στα ανοιχτά δεδομένα με έμφαση στην ανοιχτή επιστήμη και την ανακαλυψιμότητα συνόλων δεδομένων από και για την έρευνα.
- Σχεδιασμός λύσεων διεπίδρασης ανάμεσα σε γλωσσικά μοντέλα και γνωσιακές βάσεις, συνδυασμός ανάκτησης πληροφοριών με λογοπαραγωγική τεχνολογία (Retrieval Augmented Generation).

digiGOV innoHUB

Απογραφή, Διαλογή, και Απόκτηση Υλικού

Κατέχει κεντρική θέση στο έργο και αποτελεί σημείο αφετηρίας του η καθοδηγούμενη από κριτήρια κατάρτιση του υλικού εκπαίδευσης, μέσα από την ανθρώπινη επισημείωση και τον συνεχή ποιοτικό έλεγχο.

Στις διαδικασίες απογραφής, διαλογής, και απόκτησης περιλαμβάνονται:

- καταγραφή διαθέσιμων πηγών,
- προτεραιοποίηση πηγών στη βάση κριτηρίων
- αυτόματη απόκτηση κειμενικού περιεχομένου,
- εξαγωγή μηχανικά επεξεργάσιμης πληροφορίας,
- καθαρισμός και επιλογή δεδομένων στη βάση κριτηρίων
- κατάρτιση συνόλων δεδομένων εκπαίδευσης
- κατάρτιση συνόλων δεδομένων αξιολόγησης

Στατιστική Επεξεργασία

Η στατιστική επεξεργασία κατέχει κεντρικό ρόλο σε όλα τα στάδια

- Μετασχηματισμός και καθαρισμός πρωτογενών δεδομένων
- Επιτήρηση της επισημείωσης, ως προς ποιοτικά χαρακτηριστικά και χαρακτηριστικά περιεχομένου
- Δειγματοληψία από διαφορετικές πηγές, με έμφαση στις επιθυμητές κατανομές χαρακτηριστικών
- Ποσοτική αξιολόγηση της επίδοσης βοηθητικών εργαλείων
- Ποσοτική αξιολόγηση της συμφωνίας μεταξύ κριτών
- Κατάρτιση συνόλων δεδομένων εκπαίδευσης με αυτόματο τρόπο
- Μείωση διαστάσεων για απλοποίηση της εκπαίδευσης βοηθητικών εργαλείων
- Ανάλυση της σημασιολογικής σύστασης του σώματος κειμένων
- Εκπαίδευση παραδοσιακών μοντέλων γλωσσικής τεχνολογίες και μηχανικής μάθησης
- Αξιολόγηση της επίδοσης μεγαλύτερων γλωσσικών μοντέλων
- Παρουσίαση των αποτελεσμάτων, διανομή πηγαίου κώδικα, εξασφάλιση αναπαραγωγιμότητας

Τι έχει συγκεντρωθεί μέχρι τώρα (Ο συνολικός αριθμός λεξικών τύπων είναι 126,740,807)

Συλλογή	Τεκμ	Προτ	ΜΜ
anodos	71	5150	75.55
cyprus	250	62387	47.13
dimodis	228	181260	40.64
ebooks	137	18662	33.25
glc	27957	27957	4233.07
greek-language	30	1159	544.03
gutenberg	351	266820	262.39
kallipos	57	3298	33303.46
kentra	1	58	22.34
kodiko	106	11131	1.12
pergamos	289	83289	2158.89
themata-lyseis	57	3306	22.34

Αριθμός τεκμηρίων: 29,537
Αριθμός προτάσεων: 664,477

Αριθμός λεξικών τύπων: 82,348,668

Συλλογές από το OPUS :

Συλλογή	Τεκμ.	MM
Bible	33260	126.74
Europarl	1597955	166.83
GlobalVoices	26621	123.39
HNC	4614	133.91
Wikipedia	116298	108.13

Αριθμός αποσπασμάτων: 1,778,748

Αριθμός λεξικών τύπων: 44,392,139

Επόμενα Βήματα

Κατάρτιση αυτόματων εργαλείων ικανά να αναγνωρίζουν αυτόματα το κειμενικό είδος ή τη γλωσσική ποικιλία, διευκολύνοντας την επισημείωση.

Χρήση ανάλυσης μεταδεδομένων από ψηφιακές συλλογές, όπως ο Κάλλιπος, για την περαιτέρω βελτίωση της απογραφής πηγών.

Σύγκλιση των δεδομένων εκπαίδευσης με υλικά από διάφορους φορείς συνδυάζοντας δεδομένα από διάφορους τομείς γνώσης.

Ταξινόμηση και περίληψη των τομέων γνώσης δημιουργία συμβατών ταξινομήσεων με την ανοιχτή επιστήμη και τα ανοιχτά δεδομένα.

Αγορά εξοπλισμού για τις ανάγκες του έργου (8 A100 σε server 4U)

Ευχαριστούμε,

eellak.gr & opengov.ellak.gr/

Κεντρικός στόχος της ΕΕΛΛΑΚ είναι να συμβάλλει στη δημιουργία συνόλων **ανοιχτών δεδομένων** για εκπαίδευση μοντέλων ΤΝ με **ανοιχτούς αλγορίθμους** και **ανοιχτό υλισμικό**.

digiGOV innoHUB

ΕΕΛΛΑΚ
ΟΡΓΑΝΙΣΜΟΣ ΑΝΟΙΧΤΩΝ ΤΕΧΝΟΛΟΓΙΩΝ