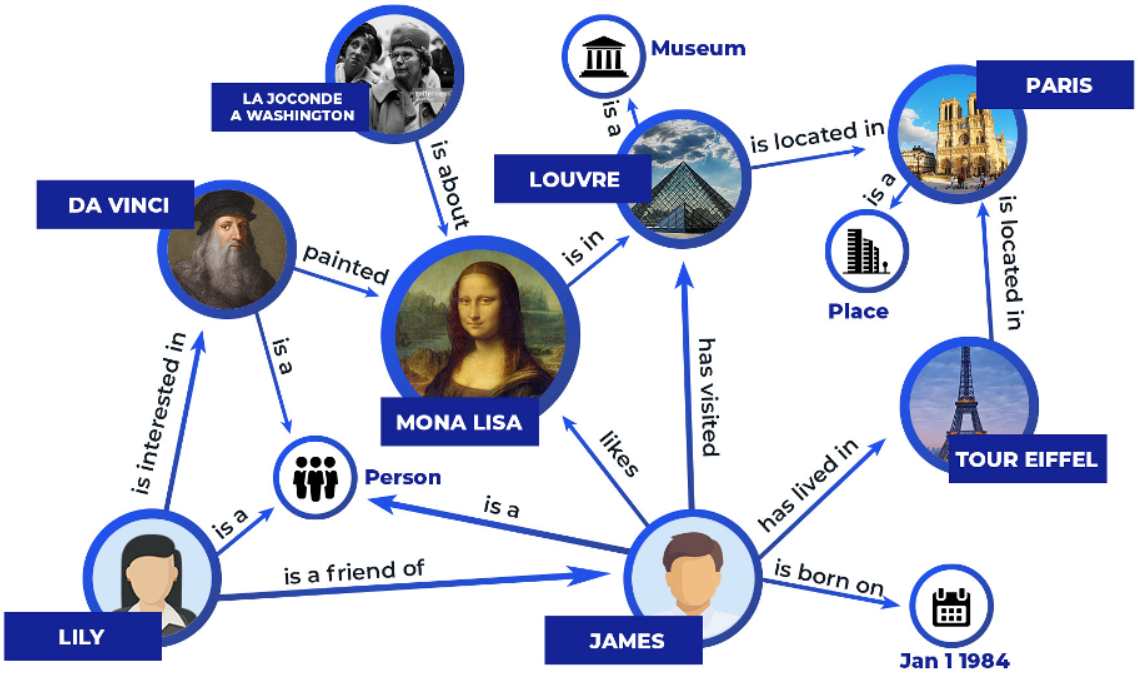
Κ08 Δομές Δεδομένων και Τεχνικές Προγραμματισμού (Τμήμα Φοιτητών/τριών με Άρτιο ΑΜ)
Διδάσκων: Μανόλης Κουμπαράκης
Εαρινό Εξάμηνο 2024-2025
Bonus Εργασία
Ανακοίνωση: 21 Μαΐου 2025
Παράδοση: 6 Ιουλίου 2025, ώρα 23:59
0.5 μονάδες στις 10 του μαθήματος

Προσοχή: Τη συγκεκριμένη εργασία θα πρέπει να την παραδώσετε μέσω του eclass του μαθήματος, όχι μέσω GitHub.

Απορίες/Ερωτήσεις: Στο Piazza ή στο skefalidis@di.uoa.gr

Σημαντικό: **Η συμμετοχή στην εργασία είναι προαιρετική. Αν συμμετάσχετε, θα βοηθήσετε την έρευνα της ομάδας μας (https://ai.di.uoa.gr/).** Ο ερευνητικός στόχος της εργασίας είναι η δημιουργία ενός συνόλου δεδομένων μέσω πληθοπορισμού (crowdsourcing). Για να έχει αξία αυτή η διαδικασία θα πρέπει η χρήση Chatbots/LLMs (ChatGPT, Gemini, DeepSeek κ.α.) να περιοριστεί μέσα σε συγκεκριμένα πλαίσια. Σας παρακαλούμε θερμά να ακολουθήσετε αυτά τα πλαίσια και να επικοινωνήσετε μαζί μας τυχόν αμφιβολίες ή δυσκολίες.

Εισαγωγή
Ένας *Γράφος Γνώσης* (*Knowledge Graph*) είναι μία μέθοδος αναπαράστασης δεδομένων που χρησιμοποιεί ένα μοντέλο γράφου για την αποθήκευση και πρόσβαση στην πληροφορία που εμπεριέχει. Σε ένα τέτοιο μοντέλο οι κορυφές αναπαριστούν οντότητες, και οι ακμές αναπαριστούν σχέσεις μεταξύ οντοτήτων.



Τέτοιου τύπου μοντέλα χρησιμοποιούνται ευρέως, με ίσως πιο γνωστή την περίπτωση της Google που χρησιμοποιεί έναν γράφο γνώσης για να προσφέρει εμπλουτισμένα αποτελέσματα στους χρήστες (περισσότερα).

Στις μέρες μας, οι γράφοι γνώσης έχουν φτάσει στο σημείο να εμπεριέχουν εκατομμύρια κορυφές και δισεκατομμύρια ακμές με αποτέλεσμα να αποτελούν εύρωστες πηγές πληροφοριών. Δύο από τους πιο γνωστούς γράφους γνώσης είναι οι *DBpedia* και *Wikidata*, τους οποίους και θα χρησιμοποιήσουμε στη συγκεκριμένη εργασία. Ο DBpedia είναι ένας γράφος που αντλεί τις πληροφορίες του από τη Wikipedia. Κατα κύριο λόγο από τα infoboxes που βρίσκονται σε κάθε σελίδα. Από την αντίθετη τροχιά, ο Wikidata περιέχει πληροφορίες οι οποίες χρησιμοποιούνται συμπληρωματικά στη Wikipedia.

Παράλληλα με την αύξηση του μεγέθους των γράφων γνώσης, δημιουργήθηκε και η ανάγκη πρόσβασης στις πληροφορίες τους από μη ειδικούς. Κανονικά, για την ανάκτηση πληροφοριών από έναν γράφο γνώσης χρειάζεται η συγγραφή ερωτημάτων στην γλώσσα SPARQL, που εκφράζουν το αίτημα μας προς τον γράφο. Αυτή η μέθοδος πρόσβασης απαιτεί εξοικείωση με τη γλώσσα SPARQL αλλά και με τον εκάστοτε γράφο γνώσης.

Για να αντιμετωπιστεί αυτό το πρόβλημα αναπτύσσονται συστήματα απάντησης ερωτημάτων (Question Answering Engines) που λαμβάνουν ένα αίτημα σε φυσική γλώσσα και επιστρέφουν ως απάντηση πληροφορίες που έχουν εξάγει από τον γράφο.

Εργασία

Περιληπτικά

Σε αυτή την εργασία, σας ζητείται να σκεφτείτε και να γράψετε **δέκα (10) πολύπλοκα ερωτήματα σε φυσική γλώσσα**, τα οποία θα μπορούσαν να δοθούν σε ένα σύστημα απάντησης ερωτημάτων για γράφους γνώσης για να λάβετε μία απάντηση. Οι γράφοι γνώσης που μας ενδιαφέρουν είναι οι DBpedia και Wikidata.

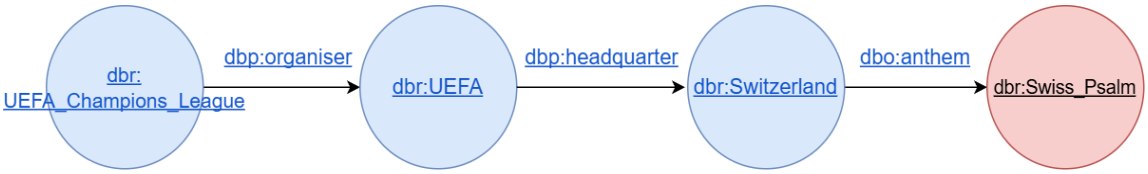
Ορισμοί

Ερώτημα σε φυσική γλώσσα: μια ερώτηση, για παράδειγμα: “Πότε γεννήθηκε ο Albert Einstein;”. Για να απαντηθεί αυτό το ερώτημα στον γράφο DBpedia, θα πρέπει αρχικά να πάμε στη σελίδα (κορυφή του γράφου) που αναφέρεται στον Άλμπερτ Αϊνστάιν. Δηλαδή στη διεύθυνση https://dbpedia.org/page/Albert_Einstein. Ύστερα θα πρέπει σε αυτή τη σελίδα να βρούμε τη σχέση που μας δίνει την ημερομηνία γέννησης του (δηλαδή την ακμή που συνδέει τον κόμβο dbr:Albert_Einstein με την απάντηση). Στη συγκεκριμένη περίπτωση η σχέση είναι η http://dbpedia.org/ontology/birthDate.

Πολυπλοκότητα ερωτημάτων: υπάρχουν δύο βασικοί άξονες που κατηγοριοποιούν ένα ερώτημα ως απλό ή πολύπλοκο.

- *Πλήθος οντοτήτων και σχέσεων*: Το ερώτημα του παραδείγματος απαιτεί μόνο μία κορυφή (dbr:Kyriakos_Mitsotakis) και μία ακμή (dbp:birthDate) για να απαντηθεί. Τέτοια ερωτήματα αναφέρονται στη βιβλιογραφία ως *one-hop*, μιας και χρειάζεται μόνο ένα “άλμα” από μία οντότητα (dbr:Albert_Einstein) προς την απάντηση (ημερομηνία γέννησης). Κάθε ακμή που μας πηγαίνει από μία οντότητα σε μία άλλη, και τέλος στην απάντηση μετράει ως ένα άλμα.

Όσο αυξάνεται το πλήθος των οντοτήτων και των “αλμάτων” που χρειάζεται για να απαντηθεί ένα ερώτημα, τόσο πιο πολύπλοκο είναι. Για παράδειγμα: “Ποιός είναι ο εθνικός ύμνος της χώρας όπου εδρεύει ο οργανισμός που διοργανώνει το Champions League;”. Σε αυτή την περίπτωση η απάντηση θα δοθεί από το μονοπάτι



dbr:UEFA_Champions_League → dbp:organiser → dbr:UEFA → dbp:headquarter → dbr:Switzerland → dbo:anthem → dbr:Swiss_Psalm. Το μονοπάτι αυτό περιέχει 3 άλματα, και 4 οντότητες.

- *Χρήση πολύπλοκης συλλογιστικής και πράξεων*: Ένας άλλος άξονας πολυπλοκότητας, είναι η χρήση πολύπλοκης συλλογιστικής και πράξεων για την απάντηση του ερωτήματος. Αντί δηλαδή να πρέπει απλά να ανακτήσουμε πληροφορία, πρέπει να γίνουν κάποιες πράξεις για να λάβουμε τη σωστή απάντηση.
  Για παράδειγμα: “Ποιές είναι οι γειτονικές χώρες της 5ης μεγαλύτερη οικονομίας στον κόσμο;”. Στη συγκεκριμένη ερώτηση, πρέπει να ταξινομηθούν οι οικονομίες των χωρών, να εξάγουμε την πέμπτη μεγαλύτερη, και στη συνέχεια να αναζητήσουμε τις γειτονικές της χώρες μέσω γεωχωρικών υπολογισμών.
  Ένα άλλο παράδειγμα είναι η χρήση πράξεων σε αριθμούς ή αλφαριθμητικά: “Ποιό ποσοστό των πρωθυπουργών της Ελλάδας είχαν επίθετο που τελειώνει σε -άκης;”. Εδώ θα πρέπει να βρεθούν οι πρωθυπουργοί της Ελλάδας, να βρεθεί το πλήθος τους, αλλά και το πλήθος αυτών που το όνομα τους έχει τη συγκεκριμένα κατάληξη, και τέλος να υπολογιστεί ο λόγος των δύο αυτών τιμών.

Ζητούμενα

Όπως αναφέρθηκε και προηγουμένως, σας ζητείται να γράψετε δέκα (10) πολύπλοκα ερωτήματα σε φυσική γλώσσα τα οποία μπορούν να απαντηθούν με τη χρήση του DBpedia ή του Wikidata (ανάλογα με το ποιός γράφος περιέχει τις απαραίτητες πληροφορίες). Τα ερωτήματα σας δηλαδή θα πρέπει να έχουν είτε μεγάλο πλήθος οντοτήτων και αλμάτων (τουλάχιστον 3 άλματα), είτε να περιέχουν κάποιες πράξεις (υπολογισμός μέσου όρου/μεγίστου/ελαχίστου, string manipulation, ταξινόμηση κλπ), ή και τα δύο. Επιπροσθέτως, καλείστε να γράψετε τις οντότητες, κλάσεις και σχέσεις που χρησιμοποιούνται για την απάντηση του ερωτήματος σας.

Μπορείτε να δείτε μερικά **παραδείγματα ερωτημάτων** που έχουν την επιθυμητή πολυπλοκότητα. Μπορείτε να λάβετε έμπνευση από αυτά, αλλά θα θέλαμε να προσπαθήσετε να σκεφτείτε ιδιαίτερες και ενδιαφέρουσες ερωτήσεις. Για να το πετύχετε αυτό σας προτείνουμε να φανταστείτε ότι έχετε κάποιον συγκεκριμένο ρόλο. Για παράδειγμα:

1. Είστε πολιτικοί αναλυτές στις αμερικανικές εκλογές και αναζητάτε μοτίβα στις προτιμήσεις των ψηφοφόρων βάσει των δημογραφικών στοιχείων κάθε πολιτείας.
2. Έχετε αρρωστήσει και αναζητάτε πληροφορίες για την κοινή γρίπη.
3. Είστε σπουδαστές του ιστορικού τμήματος και κάνετε έρευνα για μία παρουσίαση.

Επειδή οι ερωτήσεις θα πρέπει να έχουν απάντηση στους γράφους γνώσης, σας προτείνουμε να σκέφτεστε το θέμα σας και μερικές πιθανές ερωτήσεις, ύστερα να αναζητάτε την ιστοσελίδα του θέματος και από εκεί να βλέπετε αν μπορούν να απαντηθούν οι ερωτήσεις σας ή κάποια παρόμοια ερώτηση.

Προσπαθήστε να μην είναι όλες οι ερωτήσεις σας για την ίδια θεματική. Όπως επίσης και να μην έχουν όλες την ίδια μορφή. Προσπαθήστε να έχετε διαφορετικά είδη πολυπλοκότητας.

Χρήση LLMs: Πιθανόν να είναι μεγάλος ο πειρασμός να χρησιμοποιήσετε κάποιο AI Chat για να σας γράψει τις ερωτήσεις. Σας παρακαλούμε όμως να μην το κάνετε. Αρχικά, δεν θα είναι κατάλληλες, μιας και πρέπει να αναφέρονται σε συγκεκριμένα reasoning chains στις βάσεις που σας δίνονται κάτι το οποίο δύσκολα θα γίνει με επιτυχία. Ύστερα, κάτι τέτοιο θα χαλάσει την ποιότητα των δεδομένων, με αποτέλεσμα να μειώνεται η αξία της συγκεκριμένης προσπάθειας. Μπορείτε να χρησιμοποιήσετε LLMs για να σας βοηθήσουν να μπείτε σε έναν ρόλο, μπορείτε να τους ζητήσετε επιπλέον παραδείγματα. **Μπορείτε να γράψετε τις ερωτήσεις σας είτε στα Ελληνικά είτε στα Αγγλικά**. Μην χρησιμοποιήσετε LLMs για να σας μεταφράσουν την ερώτηση από τα Ελληνικά στα Αγγλικά.

Σε οποιαδήποτε περίπτωση, αν κάνετε χρήση υπηρεσιών όπως ChatGPT, Gemini, Grok, Claude σας παρακαλούμε να το αναφέρετε στο παραδοτέο, καθώς και πως τα χρησιμοποιήσατε.
Είναι προτιμότερο να γράψετε λιγότερες από 10 ερωτήσεις που να μην είναι γραμμένες από LLM, παρά να γράψετε 10 που να μην τις έχετε σκεφτεί εσείς.
Δεν θα είναι αυστηρή η βαθμολόγηση, θέλουμε μόνο να κάνετε μία τίμια προσπάθεια ώστε η ποιότητα των ερωτημάτων να είναι καλή και να πληρεί τις παραμέτρους που αναφέρουμε.

Παρουσίαση των Γράφων Γνώσης

Οι γράφοι που σας προσφέρονται περιέχουν ένα μεγάλο εύρος πληροφοριών, σαν μεγάλες εγκυκλοπαίδειες γενικού ενδιαφέροντος. Ο DBpedia είναι κάπως πιο εύκολος στην πλοήγηση, αλλά ο κάθε γράφος έχει διαφορετικές πληροφορίες και θα θέλαμε να προσπαθήσετε να τους χρησιμοποιήσετε και τους δύο. Επιπλέον των πληροφοριών που υπάρχουν στους γράφους και θα βρείτε στις ιστοσελίδες τους (https://www.dbpedia.org/ και https://www.wikidata.org/wiki/Wikidata:Main_Page) μπορείτε να θεωρήσετε ότι υπάρχουν διαθέσιμες και οι ακόλουθες πληροφορίες για να χρησιμοποιήσετε:

- Αναλυτικά δημογραφικά χαρακτηριστικά για τις αμερικανικές πολιτείες και κομητείες. Συγκεκριμένα για κάθε πολιτεία και κομητεία υπάρχουν οι σχέσεις:
    - pnyqa:hasPopulation
    - pnyqa:has[White|Black|NativeAmerican|Asian|Hispanic|Other]Population (επιλέγετε μία από τις επιλογές
    - pnyqa:has[Adult|Children|Teen|Senior]Population
    - pnyqa:has[Male|Female]Population
    - pnyqa:populationWith[PrimarySchool|HighSchool|Bsc|Msc|Phd]Education (πληθυσμός με δίπλωμα από δημοτικό, γυμνάσιο-λύκειο, πτυχίο, δίπλωμα/μεταπτυχιακό, διδακτορικό)

    - pnyqa:hasMedianIncome
- Αναλυτικά εκλογικά αποτελέσματα για τις αμερικανικές κομητείες και πολιτείες για τις προεδρικές εκλογές των ετών 2020 και 2024:
    - pnyqa:has[Democrat|Republican|Libertarian|Green]Vote
- Αναλυτικές γεωχωρικές πληροφορίες για οικισμούς, δήμους, νομούς, πολιτείες, κομητείες, χώρες και γενικά οποιοδήποτε επίπεδο αυτοδιοίκησης, καθώς και για ποτάμια, λίμνες, βουνά, δάση, θάλασσες. Αυτές οι γεωχωρικές πληροφορίες μπορούν να χρησιμοποιηθούν μέσω γεωχωρικών πράξεων. Π.χ.
    - “Ποιά είναι η μεγαλύτερη λίμνη σε νομούς που συνορεύουν με την Αρκαδία και πόσο μεγαλύτερη είναι από τη δεύτερη μεγαλύτερη;”. Εδώ, ο υπολογισμός των συνορευοντων νομών, καθώς και τα μεγέθη των λιμνών μπορούν να υπολογιστούν από τις αναλυτικές γεωμετρίες που υπάρχουν.
    - Η σχέση που δίνει πρόσβαση στη γεωχωρική πληροφορία μιας οντότητας με γεωχωρική πληροφορία είναι η:
        - geo:hasGeometry

Παράδοση

Πρέπει να παραδώσετε ένα αρχείο όπως αυτό τον παραδειγμάτων (χωρίς τις επιπλέον εξηγήσεις). Η παράδοση γίνεται μέσω eclass. Για τυχόν απορίες σας παρακαλούμε να επικοινωνήσετε μαζί μας μέσω Piazza, ή στέλνοντας ένα email στο skefalidis@di.uoa.gr