ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΤΕΧΝΟΛΟΓΙΕΣ ΓΝΩΣΗΣ

ΕΑΡΙΝΟ ΕΞΑΜΗΝΟ 2019 **ΘΕΜΑ**

Σε αυτή τη άσκηση καλείστε να κατασκευάσετε μια σημασιολογική βάση γνώσης που θα έχει ως στόχο να εξυπηρετεί ερωτήματα χρηστών σχετικά με οπτικοακουστικές παραγωγές (κινηματογραφικές ταινίες, τηλεοπτικές σειρές, ταινίες μικρού μήκους, κλπ). Συγκεκριμένα, θα πρέπει να κατασκευάσετε μια οντολογία η οποία να επιτρέπει την επαρκή μοντελοποίηση των σχετικών δεδομένων, να συγκεντρώσετε κατάλληλα δεδομένα και να τα αναπαραστήσετε ως στιγμιότυπα των εννοιών και ρόλων και ιδιοτήτων τύπων δεδομένων της οντολογίας ώστε να φτιάξετε μια ολοκληρωμένη βάση γνώσης, και τέλος να διασυνδέσετε τη βάση γνώσης με τρίτες πηγές του Σημασιολογικού Ιστού.

Ως βασική πηγή δεδομένων θα χρησιμοποιήσετε το IMDB. Τα δεδομένα διατίθενται στη διεύθυνση https://www.imdb.com/interfaces/. Όπως θα διαπιστώσετε, το IMDB διαθέτει ελεύθερα ένα μόνο μέρος των δεδομένων του σε μορφή αρχείων TSV. Τα αρχεία αυτά αφορούν βασικές πληροφορίες για τους διάφορους τίτλους (δηλαδή τις οπτικοακουστικές παραγωγές) (title.basics.tsv.gz) και τα διάφορα πρόσωπα που συμμετέχουν (name.basics.tsv.gz), καθώς και πληροφορίες για τους σκηνοθέτες και παραγωγούς κάθε τίτλου (title.crew.tsv.gz), τους βασικούς συντελεστές του (title.principals.tsv.gz) και την αξιολόγησή του από τους χρήστες (title.ratings.tsv.gz).

Αφού μελετήσετε τα περιεχόμενα αυτών των αρχείων, θα πρέπει να σχεδιάσετε αρχικά μια οντολογία OWL2 η οποία να επιτρέπει την ορθή σημασιολογική μοντελοποίηση των περιεχομένων τους, με τελικό στόχο την εξυπηρέτηση σημασιολογικών ερωτημάτων. Ο οντολογία σας θα πρέπει κατ΄ ελάχιστο να καλύπτει τα δεδομένα που διατίθενται από το IMDB, αλλά δεδομένου ότι αυτά αποτελούν ένα μικρό μόνο μέρος των συνολικών δεδομένων, μπορεί να μοντελοποιεί πληρέστερα το συγκεκριμένο πεδίο. Για την ανάπτυξη της οντολογίας προτείνεται η χρήση του εργαλείου ανάπτυξης και επεξεργασίας οντολογιών Protege, που διατίθεται στη διεύθυνση https://protege.stanford.edu/.

Στη συνέχεια, θα πρέπει να μετατρέψετε τα περιεχόμενα των αρχείων TSV του IMDB σε ένα σύνολο δεδομένων RDF, το οποίο και θα εισαγάγετε σε μία αποθήκη RDF. Η μετατροπή θα συνίσταται ουσιαστικά στη δημιουργία των κατάλληλων URI για κάθε πόρο που υπάρχει στα δεδομένα και των συνοδευτικών δηλώσεων RDF που του αποδίδουν ιδιότητες. Για να επιτύχετε τη μετατροπή θα χρειαστεί να γράψετε κώδικα, σε Java ή Python, για την κατάλληλη επεξεργασία των δεδομένων αρχείων. Ο κώδικας που θα γράψετε θα πρέπει να παράγει αρχεία δηλώσεων RDF είτε σε μορφή N-Triples/N-Quads είτε σε μορφή Turtle/Trig. Για τις ανάγκες παραγωγής των δηλώσεων RDF, στην Java μπορείτε, προαιρετικά, να χρησιμοποιήσετε κάποια από τις βιβλιοθήκες Apache Jena (https://jena.apache.org/) ή RDF4J (http://rdf4j.org/) Προσέξτε την κλιμακωσιμότητα των υλοποιήσεών σας, δεδομένου του μεγάλου όγκου των αρχείων. Ως αποθήκη RDF για την αποθήκευση των δεδομένων που θα παραγάγετε προτείνεται το OpenLink Virtuoso (http://vos.openlinksw.com/owiki/wiki/VOS), καθώς μπορεί να υποστηρίξει χωρίς προβλήματα μεγάλους όγκους δεδομένων.

Ως τελευταίο στάδιο της δημιουργίας της βάσης γνώσης, καλείστε να συνδέσετε τα δεδομένα που κατασκευάσατε με βάσεις δεδομένων του νέφους των Συνδεδεμένων Δεδομένων (Linked Open Data). Θα πρέπει να τα συνδέσετε κατ' ελάχιστο με τις DBPedia και Wikidata. Η ταυτοποίηση των πόρων σας με τους αντίστοιχους πόρους των τρίτων βάσεων δεδομένων θα γίνει μέσω των αναγνωριστικών IMDB id που υπάρχουν στα δεδομένα που διαθέτει το IMDB. Συγκεκριμένα θα μελετήσετε σύντομα το λεξιλόγιο αναπαράστασης γνώσης των DBPedia και WikiData, ώστε να μπορέσετε να εντοπίσετε τους πόρους που διαθέτουν αναγνωριστικό IMDB, να λάβετε τα URI τους, να κατασκευάσετε τις κατάλληλες δηλώσεις RDF που αποτυπώνουν τη σύνδεση, και να τις εντάξετε το σύνολο δεδομένων RDF που έχετε ήδη κατασκευάσει. Γενικές πληροφορίες για την DBPedia υπάρχουν στη διεύθυνση https://wiki.dbpedia.org/ και η υποβολή ερωτημάτων SPARQL και η λήψη των αποτελεσμάτων μπορεί να γίνει μέσω του SPARQL endpoint https://dbpedia.org/sparql (η ίδια διεύθυνση λειτουργεί και ως διεπαφή χρήστη). Πληροφορίες για την υποβολή ερωτημάτων SPARQL στα Wikidata υπάρχουν στη διεύθυνση https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service και https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/queries/examples. Το SPARQL endpoint είναι https://query.wikidata.org/bigdata/namespace/wdq/sparql και μια διεπαφή χρήστη δίνεται στη διεύθυνση https://query.wikidata.org/. (Προσέξτε ότι κάποια SPARQL endpoint μπορεί να θέτουν (σιωπηρά) ένα άνω όριο στο μέγιστο πλήθος των απαντήσεων που επιστρέφουν, οπότε για να λάβετε το σύνολο των αποτελεσμάτων θα χρειαστεί να χρησιμοποιήσετε τις λειτουργίες LIMIT-OFFSET της SPARQL.)

Αφού κατασκευάσετε τη βάση γνώσης και την αποθηκεύσετε στην αποθήκη RDF, μπορείτε να αρχίσετε να υποβάλετε ερωτήματα SPARQL. Για να επιδείξετε τη σωστή κατασκευή της βάσης γνώσης σας, σας ζητείται να κατασκευάστε, διατυπώνοντας ένα μοναδικό ερώτημα SPARQL για κάθε περίπτωση, τα εξής:

1. Έναν πίνακα που να περιέχει για κάθε είδος (genre) κινηματογραφικής ταινίας, το πλήθος των ταινιών αυτού του είδους που περιέχει η βάση του IMDB. Ο πίνακας θα πρέπει να είναι ταξινομημένος κατά φθίνουσα σειρά πλήθους ταινιών.
2. Έναν πίνακα που να περιέχει τις εταιρείες διανομής (dbo:distributor) των τίτλων της IMDB που έχουν βαθμολογηθεί από

περισσότερους από 100000 χρήστες με μέση βαθμολογία τουλάχιστον 7 και για τους οποίους υπάρχει η σχετική πληροφορία στην DBPedia. Κάθε εταιρεία διανομής θα πρέπει να συνοδεύεται από το πλήθος των ταινιών που διένειμε και ο πίνακας θα πρέπει να είναι ταξινομημένος κατά φθίνουσα σειρά πλήθους ταινιών.

3. Έναν πίνακα που να περιέχει τα ονόματα των χωρών γέννησης των σκηνοθετών των τίτλων της IMDB που έχουν βαθμολογηθεί από περισσότερους από 200000 χρήστες με μέση βαθμολογία τουλάχιστον 8,5 και για τους οποίους υπάρχει η σχετική πληροφορία στην Wikidata. Κάθε όνομα χώρας θα πρέπει να συνοδεύεται από το πλήθος των αντίστοιχων ταινιών και ο πίνακας θα πρέπει να είναι ταξινομημένος κατά φθίνουσα σειρά πλήθους ταινιών.

4. Έναν πίνακα όπου να φαίνονται ποιοι τίτλοι της IMDB που έχουν βαθμολογηθεί από περισσότερους από 10000 χρήστες δεν περιλαμβάνονται στη βάση ούτε της DBPedia ούτε της Wikidata.

Τέλος, σας ζητείται να εμπλουτίσετε σημασιολογικά τη βάση γνώσης που κατασκευάσατε (την οντολογία και τα στιγμιότυπα των δεδομένων) ορίζοντας και εντάσσοντας σε αυτήν επιπλέον χρήσιμες κατά την άποψή σας έννοιες οι οποίες θα μπορούσαν να χρησιμοποιηθούν από κάποια διεπαφή αναζητήσεων (π.χ. ασπρόμαυρες ταινίες, καθαρές κωμωδίες (δηλαδή ταινίες που είναι κωμωδίες και το μοναδικό είδος στο οποίο ανήκουν είναι η κωμωδία), συγγραφείς κωμωδιών, δραματικοί ηθοποιοί, ευρωπαϊκές ταινίες, αποκλειστικά αγγλόφωνες ταινίες, ταινίες κάποιας συγκεκριμένης δεκαετίας, κλπ.). Για κάθε μία από αυτές τις έννοιες θα πρέπει να δώσετε τον ορισμό της είτε μέσω αξιωμάτων της οντολογίας, είτε διατυπώνοντας για αυτή ένα SPARQL ερώτημα μέσω του οποίου να μπορούν να ληφθούν τα στιγμιότυπά της από τις βάσεις δεδομένων που έχετε χρησιμοποιήσει (IMDB, DBPedia, Wikidata).

Για την αξιολόγηση της δουλειάς σας θα παραδώσετε α) μια αναφορά όπου θα περιγράφετε αναλυτικά τον σχεδιασμό και την υλοποίηση της βάσης γνώσης, β) το αρχείο της οντολογίας που κατασκευάσατε και γ) τυχόν κώδικα που αναπτύξατε για τις ανάγκες της υλοποίησης. Πιο συγκεκριμένα, στην αναφορά να πρέπει

1. Να παρουσιάστε την οντολογία που αναπτύξατε και να εξηγήστε τη λογική που ακολουθήσατε. Θα πρέπει να προσέξετε η οντολογία να επιτρέπει την «ορθή» σημασιολογική μοντελοποίηση του συγκεκριμένου πεδίου γνώσης.

2. Να περιγράψετε τη διαδικασία μετατροπής των δεδομένων της IMDB σε τριάδες RDF, και να αιτιολογήσετε τις διάφορες σχεδιαστικές αποφάσεις που λάβατε.

3. Να παρουσιάστε ενδεικτικά αποσπάσματα των μετασχηματισμένων δεδομένων (δηλαδή μικρά τμήματα του συνόλου δεδομένων RDF είτε σε N-Triples/N-Quads είτε σε Turtle/Trig), στα οποία να αποτυπώνονται οι επιλογές που κάνατε για το σύνολο των ειδών δεδομένων που επεξεργαστήκατε.

4. Να παρουσιάστε και να περιγράψετε τα ερωτήματα SPARQL που σας ζητούνται μαζί με τους πίνακες αποτελεσμάτων τους.

5. Να περιγράψετε τον σημασιολογικό εμπλουτισμό της βάσης γνώσης, παρουσιάζοντας αναλυτικά τις νέες έννοιες που ορίσατε και τους τρόπους λήψης των στιγμιοτύπων τους. Να σχολιάσετε επίσης τον τρόπο με τον οποίο θα μπορούσε να υλοποιηθεί μια διαδικασία απάντησης ερωτημάτων επί της βάσης γνώσης με βάση την οντολογία.

6. Να σχολιάσετε τυχόν προβλήματα που αντιμετωπίσατε.

Παρατηρήσεις για το OpenLink Virtuoso

1. Αναλυτικές πληροφορίες για την εγκατάσταση και τη χρήση του OpenLink Virtuoso υπάρχουν στη διεύθυνση http://vos.openlinksw.com/owiki/wiki/VOS.

2. Μετά την εγκατάσταση μπορείτε να εκκινήσετε το OpenLink Virtuoso πηγαίνοντας στον κατάλογο bin της εγκατάστασης και εκτελώντας την εντολή virtuoso-t +foreground +configfile ../database/virtuoso.ini. Με τον τρόπο αυτό θα βλέπετε απευθείας στην γραμμή εντολών μηνύματα σχετικά με τη λειτουργία και τυχόν σφάλματα. Αν η εγκατάσταση έχει δημιουργήσει κάποιο service του λειτουργικού συστήματος που εκτελείται αυτόματα θα πρέπει πρώτα να το απενεργοποιήσετε αν θέλετε να ελέγχετε χειροκίνητα την εκκίνηση και τον τερματισμό. (Στα Windows αυτό γίνεται από το Control Panel (Πίνακας Ελέγχου)/Administrative Tools (Εργαλεία Διαχείρισης)/Services (Υπηρεσίες): επιλέγετε την υπηρεσία OpenLink Virtuoso Server [vos], διακόπτετε τη λειτουργία [με την επιλογή που δίνεται επάνω αριστερά], και με δεξί κλικ/Properties (Ιδιότητες)/General (Γενικά) στο μενού Startup type επιλέγετε Manual (Μη αυτόματα)). Τερματίζετε το OpenLink Virtuoso πατώντας Ctrl+C στο παράθυρο από τον οποίο το εκκινήσατε.

3. Το OpenLink Virtuoso είναι μεταξύ άλλων μια αποθήκη τριάδων RDF και περιλαμβάνει δύο server, έναν database server και έναν web server. Ο πρώτος επιτρέπει την απευθείας διαχείριση της βάσης δεδομένων μέσω μιας γραμμής εντολών και ο δεύτερος παρέχει μια γραφική διεπαφή διαχείρισης και υποβολής ερωτημάτων SPARQL. Εισέρχεστε στο περιβάλλον του database server εκτελώντας την εντολή isql από τη γραμμή εντολών στον κατάλογο bin της εγκατάστασης. Ο web server ακούει (με βάση τις προκαθορισμένες ρυθμίσεις στη διεύθυνση http://localhost:8890/. Η διεπαφή υποβολής ερωτημάτων SPARQL είναι η http://localhost:8890/sparql στην οποία μπορείτε να εκτελέσετε απευθείας οποιοδήποτε ερώτημα SPARQL. Στη γραμμή εντολών του database server, για αν εκτελέστε κάποιο ερώτημα SPARQL πρέπει να γράψετε πρώτα SPARQL και να συνεχίστε με το ερώτημα.

4. Στο virtuoso.ini ορίζονται διάφορες ρυθμίσεις για τη λειτουργία του OpenLink Virtuoso. Ενδιαφέρουν ιδιαίτερα:

   (α΄) Το πεδίο Parameters/DirsAllowed που έχει ως τιμές (χωρισμένες μεταξύ τους με κόμματα) τους καταλόγους από τους οποίους επιτρέπεται να διαβαστούν αρχεία δεδομένων προκειμένου να εισαχθούν στη βάση γνώσης.

   (β΄) Το πεδίο Paremeters/ServerPort που έχει ως τιμή την πόρτα στον οποία θα ακούει ο database server.

   (γ΄) Το πεδίο HTTPServer/ServerPort που έχει ως τιμή την πόρτα στον οποία θα ακούει ο web server και το sparql endpoint.

   (δ΄) Το πεδίο SPARQL/ResultSetMaxRows που έχει ως τιμή το μέγιστο πλήθος αποτελεσμάτων που επιτρέπεται να επιστρέψει ένα ερώτημα SPARQL.

5. Σε περίπτωση κακού τερματισμού του server και αδυναμίας επανεκκίνησής του, δοκιμάστε να σβήσετε το αρχείο virtuoso.lck στον κατάλογο database της εγκατάστασης.

6. Λόγου του μεγάλου όγκου των δεδομένων που καλείστε να χειριστείτε προτείνεται η χρήση του database server για την εισαγωγή και διαχείρισή τους. Από τη γραμμή εντολών του database server μπορείτε να δίνετε εντολές εισαγωγής και διαγραφής δεδομένων, αλλά και να κάνετε ερωτήματα. Επειδή το OpenLink Virtuoso είναι υλοποιημένο πάνω σε σχεσιακή βάση, η διαχείριση των δεδομένων γίνεται με τη μεσολάβηση σχεσιακών πινάκων.

7. Για οδηγίες για τη μαζική εισαγωγή αρχείων δεδομένων RDF, συμβουλευτείτε τη σελίδα http://vos.openlinksw.com/owiki/wiki/VOS/VirtBulkRDFLoader που εξηγεί αναλυτικά τη διαδικασία.

8. Για ευκολότερη διαχείριση των δεδομένων σας προτείνεται να τα εντάξετε σε έναν ή περισσότερους ονοματισμένους γράφους, ώστε να μπορείτε π.χ. να τα διαγράψετε εύκολα μαζικά αλλά και επιλεκτικά αν χρειαστεί. Η διαγραφή των περιεχομένων ενός ονοματισμένου γράφου γίνεται μέσω της εντολής SPARQL CLEAR GRAPH <graph-name>; στη γραμμή εντολών του database server. Βλ. και http://vos.openlinksw.com/owiki/wiki/VOS/VirtTipsAndTricksGuideDeleteLargeGraphs

9. Για μπορείτε να εκτελέσετε ομόσπονδα ερωτήματα SPARQL θα πρέπει να δώσετε στη γραμμή εντολών του database server τις δύο εντολές
   grant execute on ”DB.DBA.SPARQL_SINV_IMP” to ”SPARQL”; και
   grant select on ”DB.DBA.SPARQL_SINV_2” to ”SPARQL”;