ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΤΜΗΜΑΤΟΣ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΕΞΑΓΩΓΗ ΟΝΤΟΛΟΓΙΩΝ ΑΠΟ ΑΔΟΜΗΤΟ ΚΕΙΜΕΝΟ ΜΕ ΤΗ ΧΡΗΣΗ ΚΑΤΑΛΛΗΛΩΝ ΕΡΓΑΛΕΙΩΝ

Διπλωματική Εργασία

του

Παπατζίκου Νικόλαου

Επιβλέπουσα Καθηγήτρια
Κολωνιάρη Γεωργία

Θεσσαλονίκη, Ιούνιος 2020

Η δομή της παρουσίασης

- Οι στόχοι της εργασίας
- Σημασιολογικός Ιστός – Οντολογίες – RDF/RDFS/OWL
- Το περιβάλλον οντολογιών Protégé
- Η εφαρμογή Text2Onto
- Πειράματα με την εφαρμογή Text2Onto
- Διαδικασίες για την ανάλυση αποτελεσμάτων των πειραμάτων
- Ανάλυση αποτελεσμάτων πειραμάτων – Συμπεράσματα

Οι στόχοι της εργασίας

- Η παρούσα εργασία πραγματεύεται την πειραματική μελέτη της εφαρμογής Text2Onto για την αυτόματη εξαγωγή οντολογικής γνώσης από κείμενο σε φυσική γλώσσα.
- Η εφαρμογή Text2Onto είναι ένα από τα ελάχιστα εργαλεία εκμάθησης οντολογιών το οποίο έχει αναπτυχθεί για να υποστηρίξει τη δημιουργία οντολογιών από αδόμητο κείμενο.
- Στόχος της εργασίας είναι η εξαγωγή συμπερασμάτων σχετικά με την έκταση των δυνατοτήτων και τους τρόπους εφαρμογής του εργαλείου στον τομέα των οντολογιών.

Σημασιολογικός Ιστός - Οντολογίες

- Ο Σημασιολογικός Ιστός αναφέρεται σε ένα σύνολο τεχνολογιών και μεθόδων μέσω των οποίων οι υπολογιστές είναι σε θέση να αντιλαμβάνονται τη σημασία της πληροφορίας που διαχειρίζονται.
- Στον Σημασιολογικό Ιστό, οι οντολογίες ορίζουν τις έννοιες και τις σχέσεις που χρησιμοποιούνται για να περιγράψουν και να αντιπροσωπεύσουν έναν τομέα ενδιαφέροντος.
- Οι κυριότερες γλώσσες/πρότυπα οντολογιών είναι οι RDF, RDFS και OWL.

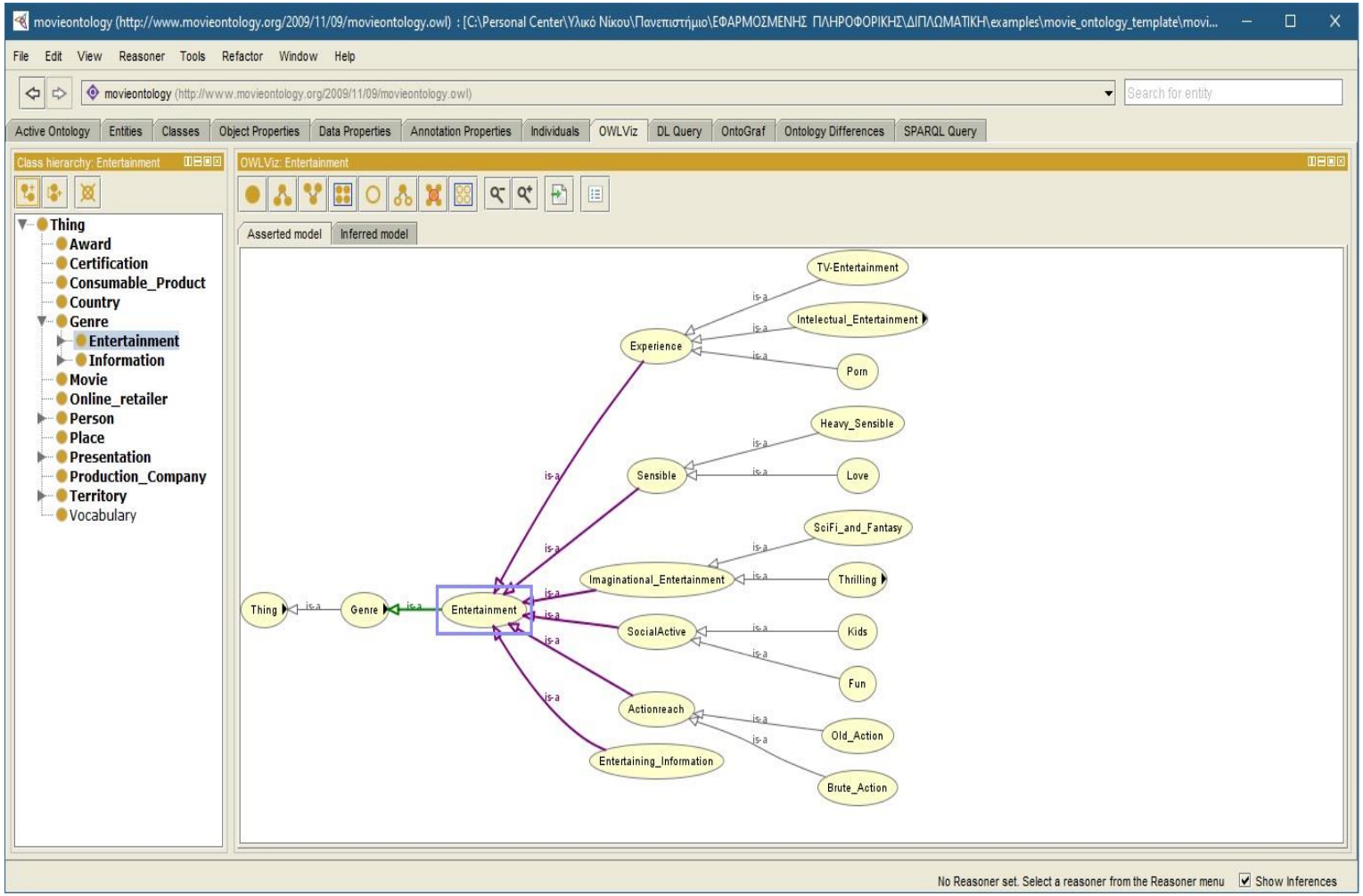
RDF/RDFS/OWL

- Το μοντέλο δεδομένων που η υποδομή του Σημασιολογικού Ιστού χρησιμοποιεί για την απεικόνιση αυτού του κατανεμημένου ιστού δεδομένων ονομάζεται RDF (Resource Description Framework). Βασική δομικό του στοιχείου είναι η τριπλέτα Υποκείμενο – Κατηγορούμενο – Αντικείμενο.
  - Παράδειγμα: mov:TomCruise mov:PlayedIn mov:TopGun
- Η RDFS (Resource Description Framework Schema) είναι η γλώσσα σχήματος για το RDF. Περιγράφει δομικά στοιχεία για έννοιες αντικειμένων (κλάσεις), συσχετίσεις μεταξύ κλάσεων (υποκλάσεις), ιδιότητες που περιγράφουν αντικείμενα και συσχετίσεις μεταξύ ιδιοτήτων (υπο-ιδιότητες).
  - Παράδειγμα: mov:Comedy rdfs:subClassOf mov:Movie
- Η OWL (Web Ontology Language) βασίστηκε στις RDF και RDFS, έχει εκτεταμένες δυνατότητες σε σχέση με αυτές και καλύπτει περισσότερες απαιτήσεις. Είναι προσανατολισμένη στην επεξεργασία του περιεχόμενου της πληροφορίας. Βασικά στοιχεία της γλώσσας είναι οι κλάσεις, τα άτομα και οι ιδιότητες.
  - Παράδειγμα: mov:Film owl:equivalentClass mov:Movie

Το περιβάλλον οντολογιών Protégé

- Το Protégé είναι ένα πλούσιο σε λειτουργίες περιβάλλον οντολογιών με πλήρη υποστήριξη της OWL.
- Υποστηρίζει τη δημιουργία και την επεξεργασία μιας ή περισσοτέρων οντολογιών σε έναν ενιαίο χώρο εργασίας μέσω ενός πλήρως προσαρμόσιμου περιβάλλοντος χρήστη.
- Τα εργαλεία απεικόνισης επιτρέπουν τη διαλογική πλοήγηση των σχέσεων οντολογίας.

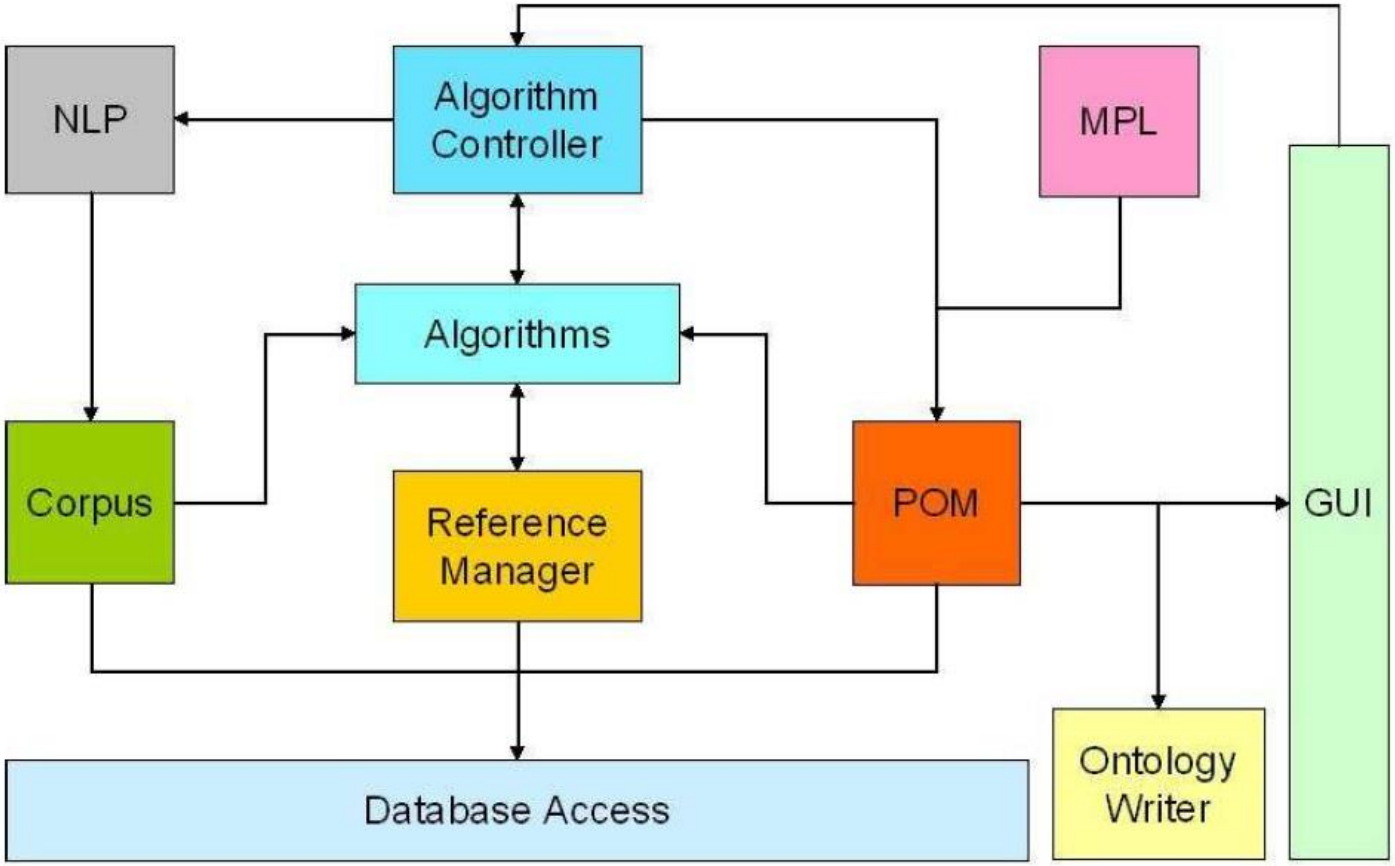
Το περιβάλλον οντολογιών Protégé



Η εφαρμογή Text2Onto

- Corpus. Πρόκειται για το σώμα των δεδομένων κειμένου που χρησιμοποιούνται ως είσοδος στο Text2Onto.
- NLP (Natural Language Processing). Το Text2Onto συνδυάζει προσεγγίσεις μάθησης μηχανής με γλωσσική επεξεργασία, όπως διαχωρισμός λέξεων, προτάσεων και γραμματική ανάλυση.
- MPL (Modeling Primitive Library). Το Text2Onto υποστηρίζει τα θεμελιακά στοιχεία μοντελοποίησης Concept, Instance, Similarity, SubclassOf, InstanceOf, Relation, SubtopicOf.
- Αλγόριθμοι-Ελεγκτής Αλγορίθμων. Το Text2Onto χρησιμοποιεί διαφόρους αλγορίθμους εκμάθησης οντολογιών ανά θεμελιακό στοιχείο οι οποίοι ελέγχονται και συνδυάζονται από τον ελεγκτή αλγορίθμων.
- POM (Probabilistic Ontology Model). Η αρχιτεκτονική του Text2Onto επικεντρώνεται γύρω από το Πιθανοτικό Οντολογικό Μοντέλο (POM) που αποθηκεύει τα αποτελέσματα των διαφορετικών αλγορίθμων εκμάθησης οντολογιών προσδίδοντάς τους μια τιμή πιθανότητας.
- Ontology Writers. Χρησιμοποιούνται για την μετάφραση των θεμελιακών στοιχείων σε διάφορες γλώσσες οντολογικής αναπαράστασης όπως RDFS και OWL.

Η εφαρμογή Text2Onto-Αρχιτεκτονική



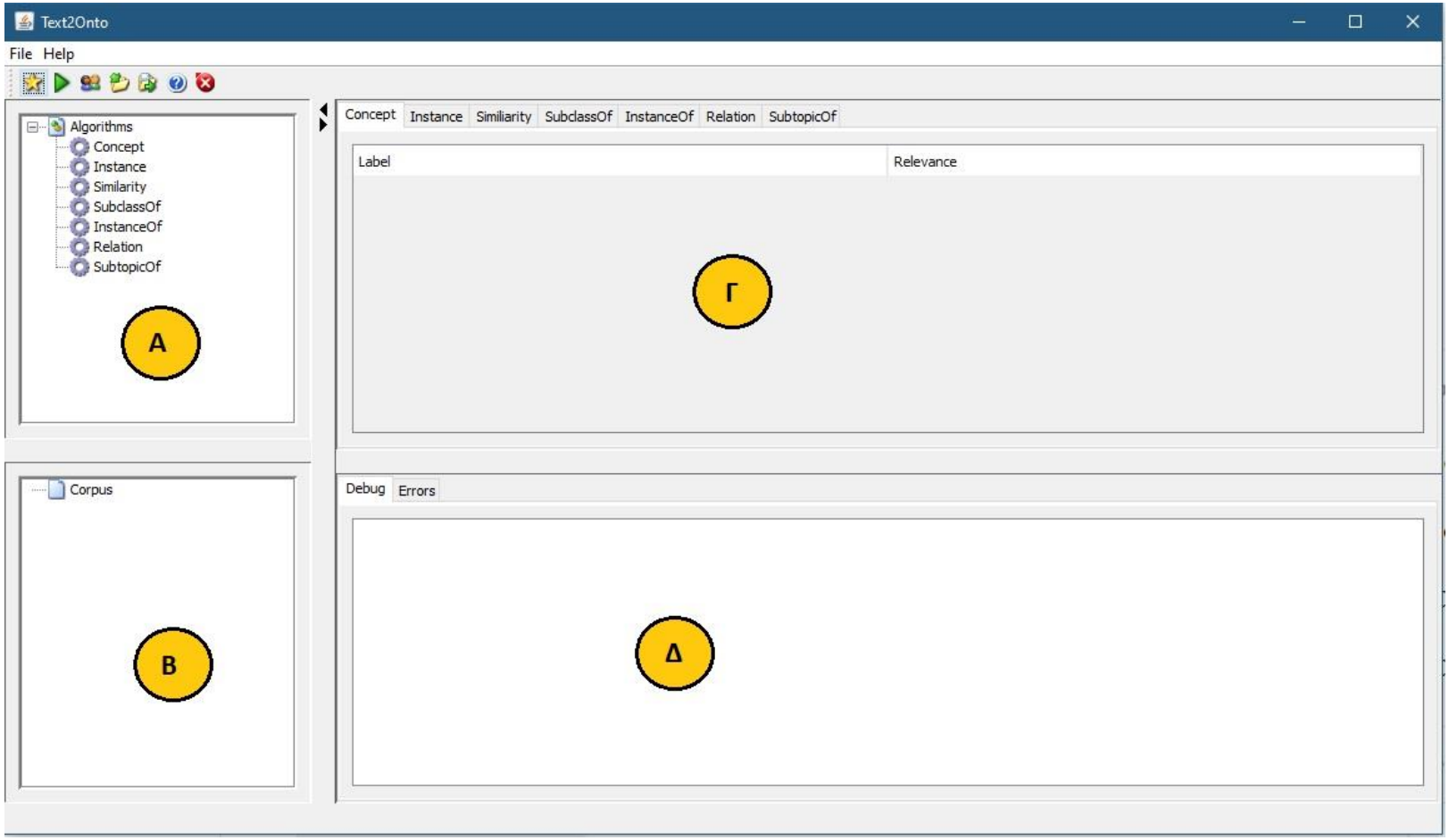
Η εφαρμογή Text2Onto-Θεμελιακά στοιχεία μοντελοποίησης

Θεμελιακό στοιχείο	**Επεξήγηση**	**OWL**
Concept	Μια έννοια (κλάση) C. Παράδειγμα: movie, actor	C
Instance	Ένα στιγμιότυπο a. Παράδειγμα: tom cruise, hannibal	a
Similarity	Ισοδυναμία μεταξύ των εννοιών C1 και C2. Παράδειγμα: similarity (film, movie)	$C1 \equiv C2$
SubclassOf	Κληρονομικότητα κλάσεων C1, C2. Παράδειγμα: subclassof (comedy, movie)	$C1 \subset C2$
InstanceOf	Στιγμιότυπο έννοιας a, C. Παράδειγμα: instanceof(tom cruise, actor)	C(a)
Relation	Μια σχέση R μεταξύ των εννοιών C1 και C2. Παράδειγμα: hasDirector(movie, direrto r)	$C1 \subset \forall R.C2$
SubtopicOf	Επιμέρους θέμα C1, C2 Παράδειγμα: subtopicof(horror_movie, movie)	$C1 \subset C2$

Η εφαρμογή Text2Onto-Αλγόριθμοι

Θεμελιακό στοιχείο	**Αλγόριθμοι**
Concept	EntropyConceptExtraction ExampleConceptExtraction RTFConceptExtraction TFIDFConceptExtraction
Instance	ExampleInstanceExtraction TFIDFInstanceExtraction
Similarity	ContextSimilarityExtraction
SubclassOf	PatternConceptClassification VerticalRelationsConceptClassification WordNetConceptClassification
InstanceOf	ContextInstanceClassification PatternInstanceClassification
Relation	SubcatRelationExtraction
SubtopicOf	SubtopicOfRelationConversion SubtopicOfRelationExtraction

Η εφαρμογή Text2Onto-Περιβάλλον εργασίας



Πειράματα με την εφαρμογή Text2Onto

- Δεδομένα εισόδου από κριτικές και στοιχεία κινηματογραφικών ταινιών.
- Τέσσερις κατηγορίες δεδομένων εισόδου.
  - Κριτικές ταινιών από την Amazon. Επιλογή κριτικών από τυχαίες ταινίες.
  - Κριτικές ταινιών από την Amazon. Επιλογή κριτικών από 4 συγκεκριμένες ταινίες.
  - Κριτικές ταινιών από την IMDb. Επιλογή κριτικών από τυχαίες ταινίες.
  - Στοιχεία ταινιών από την IMDb.
- Ποσοτική μεταβολή των δεδομένων για την 1η και 3η κατηγορία δεδομένων εισόδου της τάξης των 20, 100 και 1000 κριτικών. Στην 2η και 4η κατηγορία τα δεδομένα εισόδου είναι της τάξης των 1000 κριτικών και στοιχείων ταινιών.
- Πέντε είδη πειραμάτων. Για τα πρώτα τέσσερα είδη πειραμάτων χρησιμοποιήθηκε από ένας αλγόριθμος για κάθε θεμελιακό στοιχείο. Στο πέμπτο είδος πειραμάτων έγινε ένας συνδυασμός όλων των διαθέσιμων αλγορίθμων ανά θεμελιακό στοιχείο.
- Κωδικοποίηση ονομασίας των πειραμάτων της μορφή rX_expY_Z.

Πειράματα με την εφαρμογή Text2Onto

Πείραμα r1000_exp05_amazon_same

Text2Onto

File Help

- Algorithms
  - Concept
    - EntropyConceptExtraction
    - ExampleConceptExtraction
    - RTFConceptExtraction
    - TFIDFConceptExtraction
  - Instance
    - ExampleInstanceExtraction
    - TFIDFInstanceExtraction
  - Similarity
    - ContextSimilarityExtraction
  - SubclassOf
    - PatternConceptClassification
    - VerticalRelationsConceptClassification
    - WordNetConceptClassification
  - InstanceOf
    - ContextInstanceClassification
    - PatternInstanceClassification
  - Relation
    - SubcatRelationExtraction
  - SubtopicOf

Concept | Instance | Similiarity | SubclassOf | InstanceOf | Relation | SubtopicOf

Label	Domain	Range	Confidence
unleash	director mary harron	disappointment	1.0
think_of	investigator	bateman	1.0
complain_about	bartender	way	1.0
portray_of	bale	patrick bateman	1.0
urge_into	society	logic	1.0
urge	society	people	1.0
evoke	movie	time period	1.0
believe	bateman	none	1.0
kick	bateman	man	1.0
put_on	christian bale	show	1.0
greet_with	bateman	voice	1.0
greet	bateman	world	1.0
pull	bale	mask	1.0
transport	director mary harron	viewer	1.0
dissolve	bateman	friendship	1.0
listen	nobody	period	1.0
poke	movie	fun	1.0
miss	confines	point	1.0
feature	track	guinevere turner	1.0
point	apologium speech	film	1.0
know	one	difference	1.0
enhance	harron	film	1.0
justify	work	lifestyle	1.0
cover_as	cd	substitute	1.0
coincide_with	approach	ultra	1.0
serve_as	movie	time capsule	1.0

Corpus

C:\Users\User\Downloads\same_movies_reviews\r10

C:\Users\User\Downloads\same_movies_reviews\r10

Debug | Errors

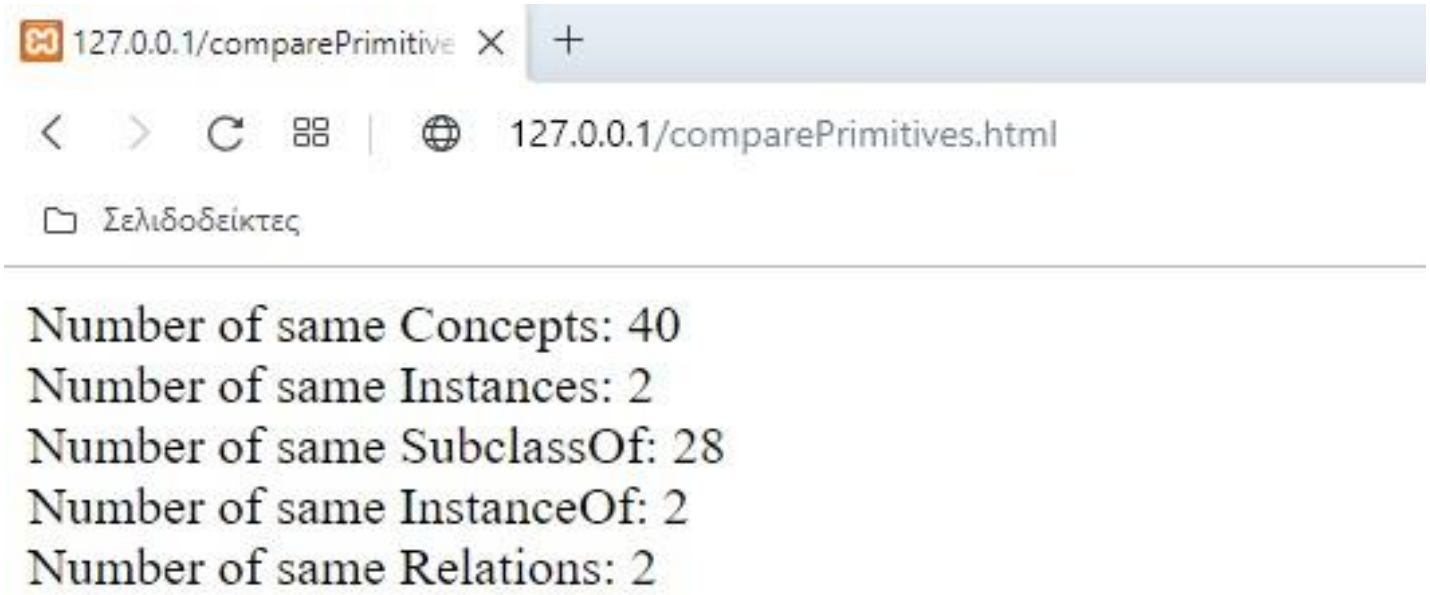
teman, character), instance-of(bateman, psychopath), instance-of(american psycho, film), instance-of(satire, something), instance-of(christian bale, key), instance-of(humanity, annoyance), instance-

nstance-of(patrick bateman, broker), instance-of(american psycho, satire), instance-of(bateman, sort), instance-of(patrick, man), instance-of(bateman, serial killer), instance-of(bateman, victim),

Διαδικασίες για την ανάλυση αποτελεσμάτων των πειραμάτων

- Λόγω του μεγάλου όγκου των αποτελεσμάτων των πειραμάτων κρίθηκε αναγκαίο να αναπτυχθούν ορισμένες διαδικασίες ώστε να διευκολυνθεί η ποσοτική και ποιοτική ανάλυση των αποτελεσμάτων των πειραμάτων.
- Για την ανάπτυξη των διαδικασιών χρησιμοποιήθηκε η γλώσσα προγραμματισμού Javascript. Οι διαδικασίες αυτές είναι οι εξής:
  - Η διαδικασία μέτρησης θεμελιακών στοιχείων.
  - Η διαδικασία διαγραφής θεμελιακών στοιχείων.
  - Η διαδικασία σύγκρισης θεμελιακών στοιχείων.
  - Η διαδικασία πιθανοτικής ομαδοποίησης θεμελιακών στοιχείων.
  - Η διαδικασία μετατροπής θεμελιακών στοιχείων.
  - Η διαδικασία σύγκρισης με πρότυπο οντολογίας.

Διαδικασίες για την ανάλυση αποτελεσμάτων των πειραμάτων



Διαδικασίες για την ανάλυση αποτελεσμάτων των πειραμάτων

Elements Console Sources Network Performance Memory »

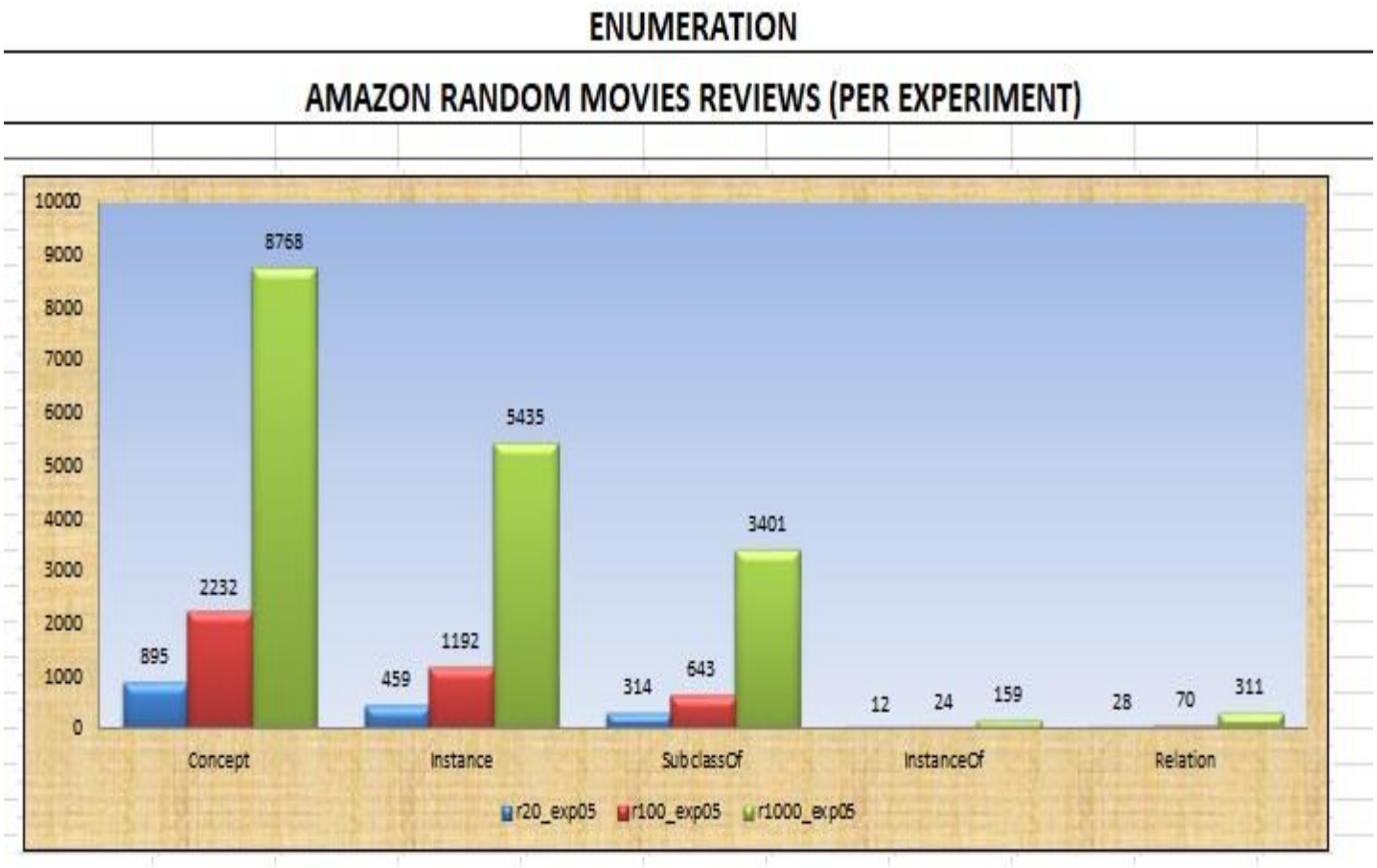
top Filter Default levels 2 hidden

```
</owl:Class>
<owl:Class rdf:about="http://www.text2onto.org/ontology#wrong">
    <rdfs:subClassOf rdf:resource=""/>
</owl:Class>
<owl:Class rdf:about="http://www.text2onto.org/ontology#year">
    <rdfs:subClassOf rdf:resource="http://www.text2onto.org/ontology#per
</owl:Class>
<owl:Class rdf:about="http://www.text2onto.org/ontology#yesteryear">
    <rdfs:subClassOf rdf:resource=""/>
</owl:Class>
<owl:Class rdf:about="http://www.text2onto.org/ontology#yore">
    <rdfs:subClassOf rdf:resource=""/>
</owl:Class>
<owl:Class rdf:about="http://www.text2onto.org/ontology#zany_charatcer">
    <rdfs:subClassOf rdf:resource="http://www.text2onto.org/ontology#charatcer"/>
</owl:Class>
<owl:NamedIndividual rdf:about="http://www.text2onto.org/ontology#actor_brian_peck">
    <rdf:type rdf:resource=""/>
</owl:NamedIndividual>
<owl:NamedIndividual rdf:about="http://www.text2onto.org/ontology#actor_don_calfa">
    <rdf:type rdf:resource=""/>
</owl:NamedIndividual>
<owl:NamedIndividual rdf:about="http://www.text2onto.org/ontology#actor_thom_mathew">
    <rdf:type rdf:resource=""/>
</owl:NamedIndividual>
<owl:NamedIndividual rdf:about="http://www.text2onto.org/ontology#adam_scott">
    <rdf:type rdf:resource=""/>
</owl:NamedIndividual>
<owl:NamedIndividual rdf:about="http://www.text2onto.org/ontology#added">
    <rdf:type rdf:resource=""/>
</owl:NamedIndividual>
<owl:NamedIndividual rdf:about="http://www.text2onto.org/ontology#afghan_mountain">
    <rdf:type rdf:resource=""/>
</owl:NamedIndividual>
<owl:NamedIndividual rdf:about="http://www.text2onto.org/ontology#afghanistan">
    <rdf:type rdf:resource=""/>
```

Clear console
Clear console hist...
Save as...

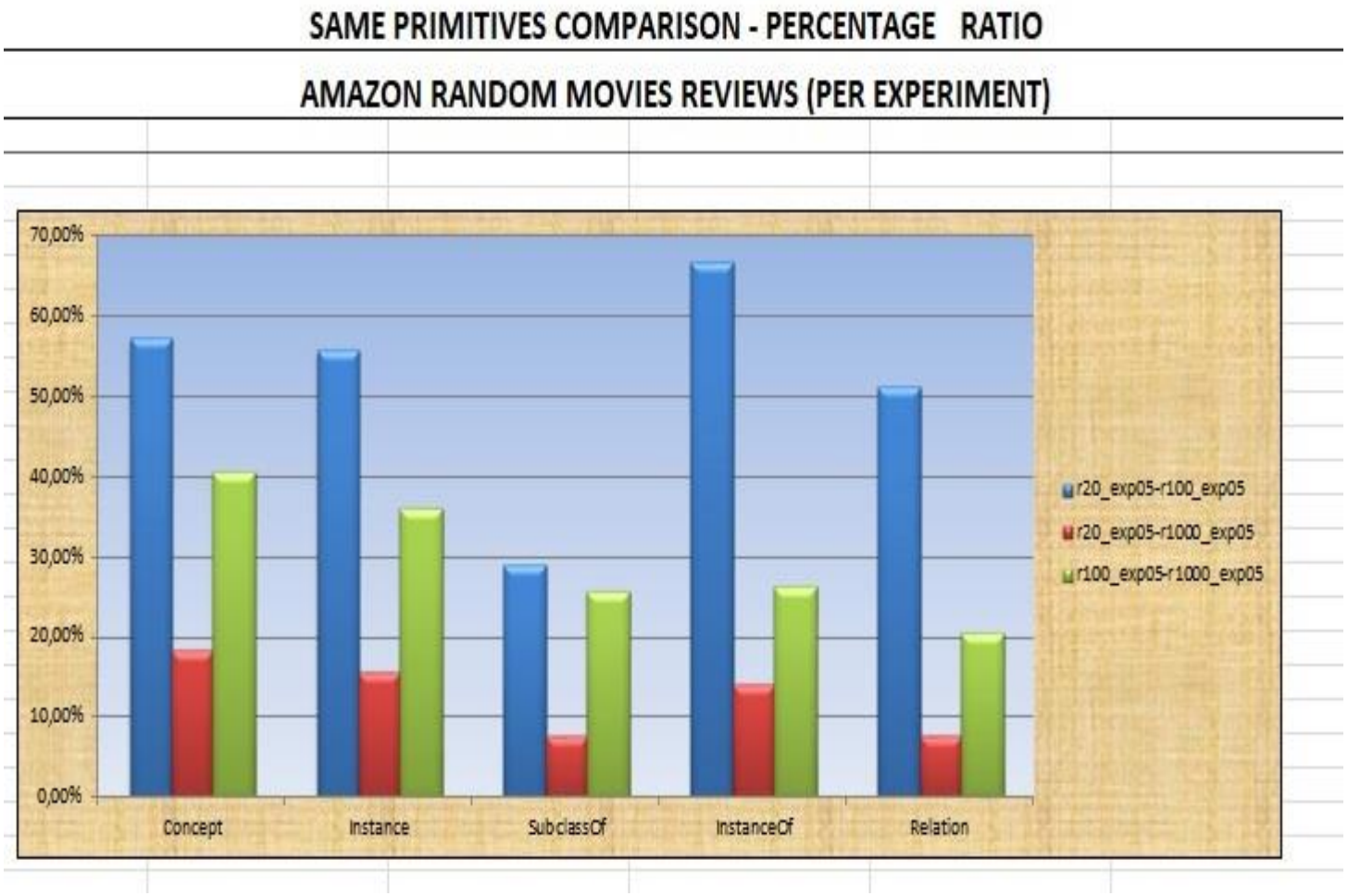
Ανάλυση αποτελεσμάτων πειραμάτων

Μέτρηση αριθμού στοιχείων οντολογίας



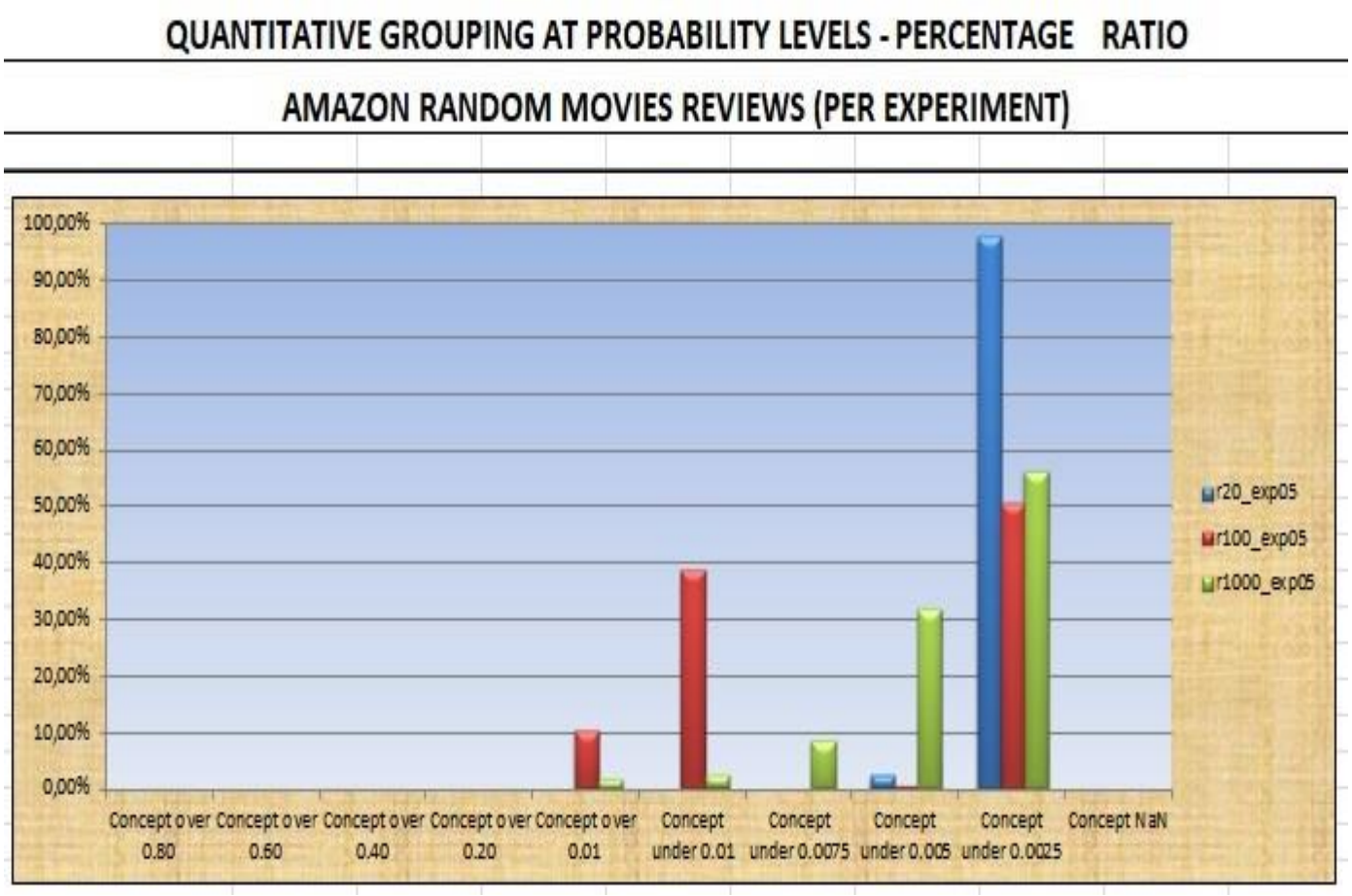
Ανάλυση αποτελεσμάτων πειραμάτων

Σύγκριση ως προς τον αριθμό κοινών στοιχείων οντολογιών



Ανάλυση αποτελεσμάτων πειραμάτων

Ποσοτική ομαδοποίηση στοιχείων οντολογίας με βάση την τιμή πιθανότητας

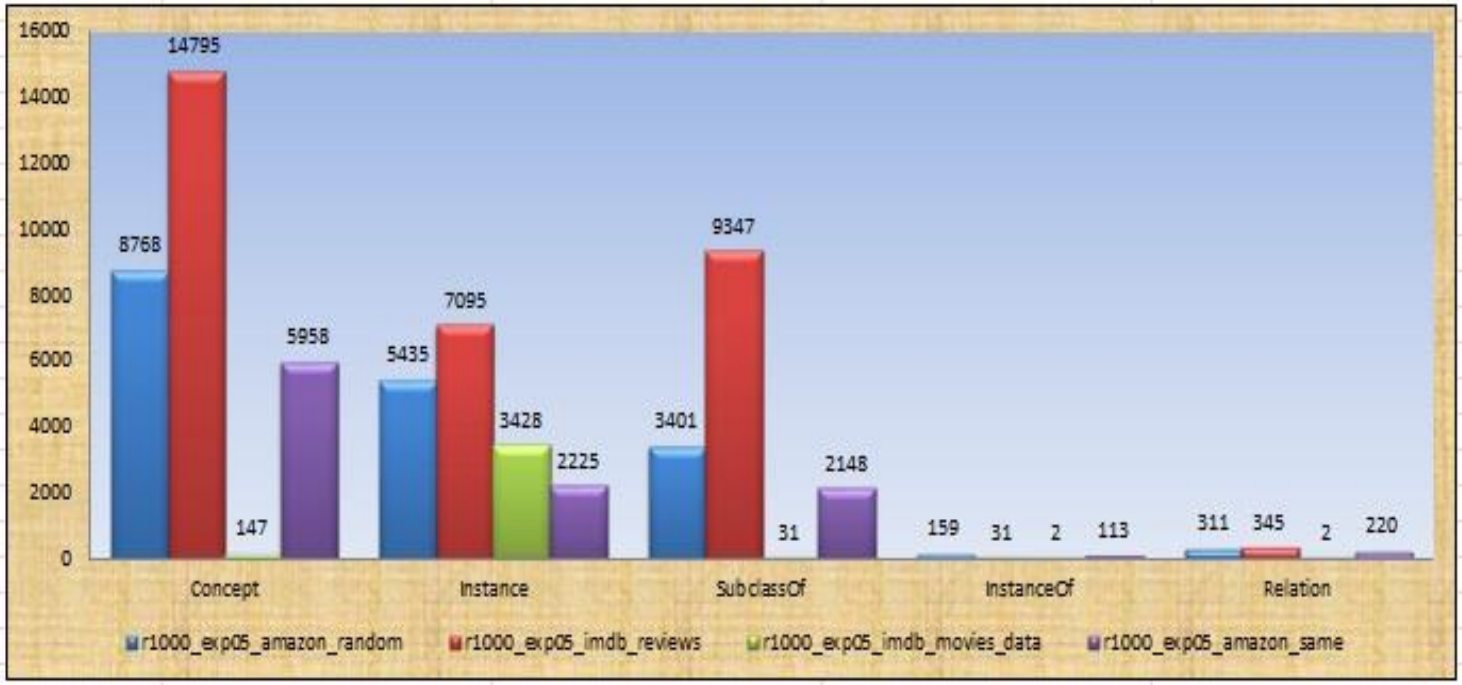


Ανάλυση αποτελεσμάτων πειραμάτων

Σύγκριση μετρήσεων αριθμού στοιχείων οντολογιών μεταξύ διαφορετικών κατηγοριών δεδομένων εισόδου

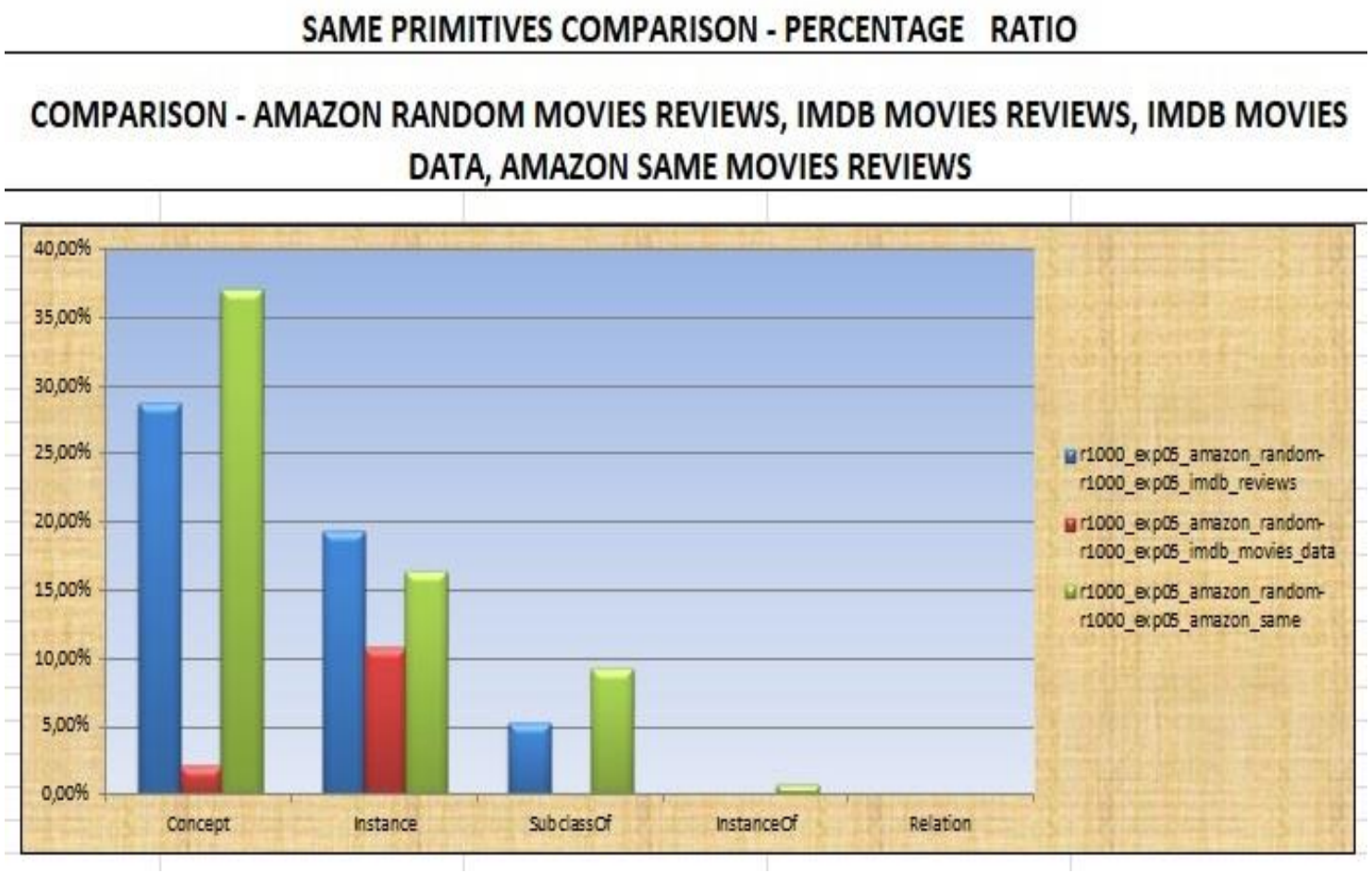
ENUMERATION

COMPARISON - AMAZON RANDOM MOVIES REVIEWS, IMDB MOVIES REVIEWS, IMDB MOVIES DATA, AMAZON SAME MOVIES REVIEWS



Ανάλυση αποτελεσμάτων πειραμάτων

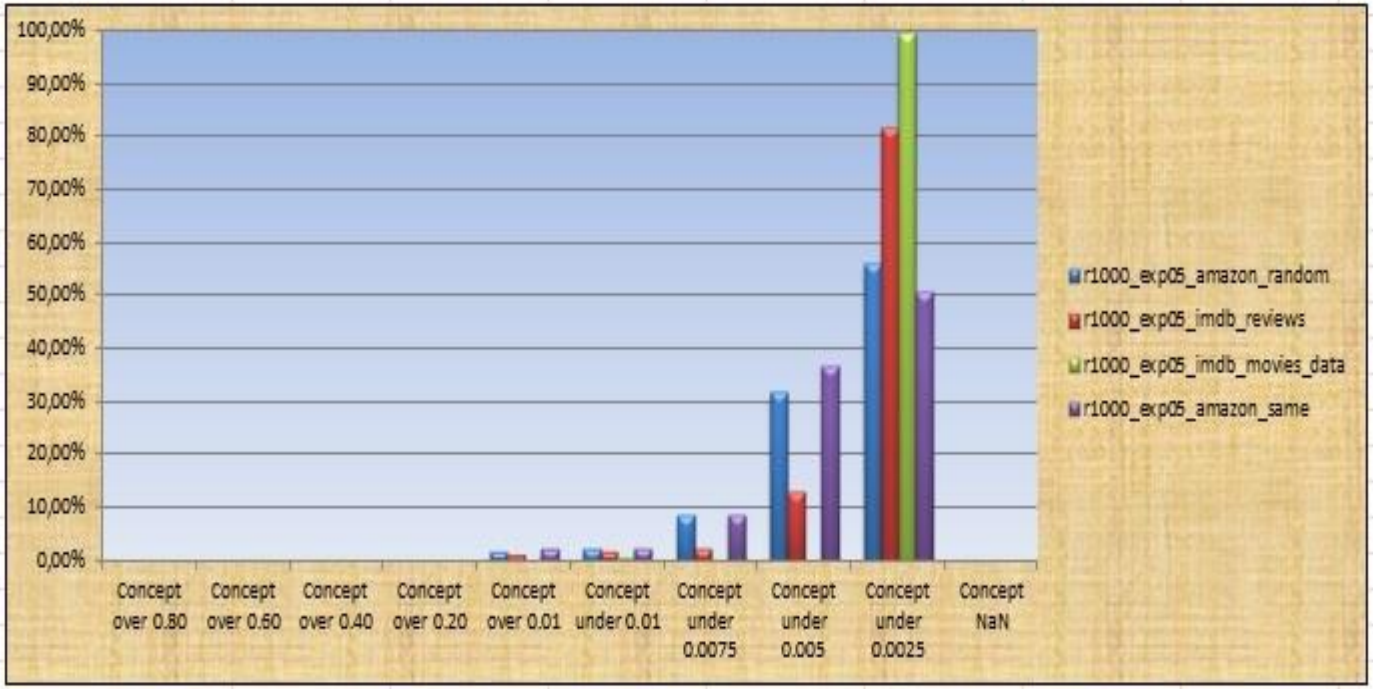
Σύγκριση ως προς τον αριθμό κοινών στοιχείων οντολογιών μεταξύ διαφορετικών κατηγοριών δεδομένων εισόδου



Ανάλυση αποτελεσμάτων πειραμάτων

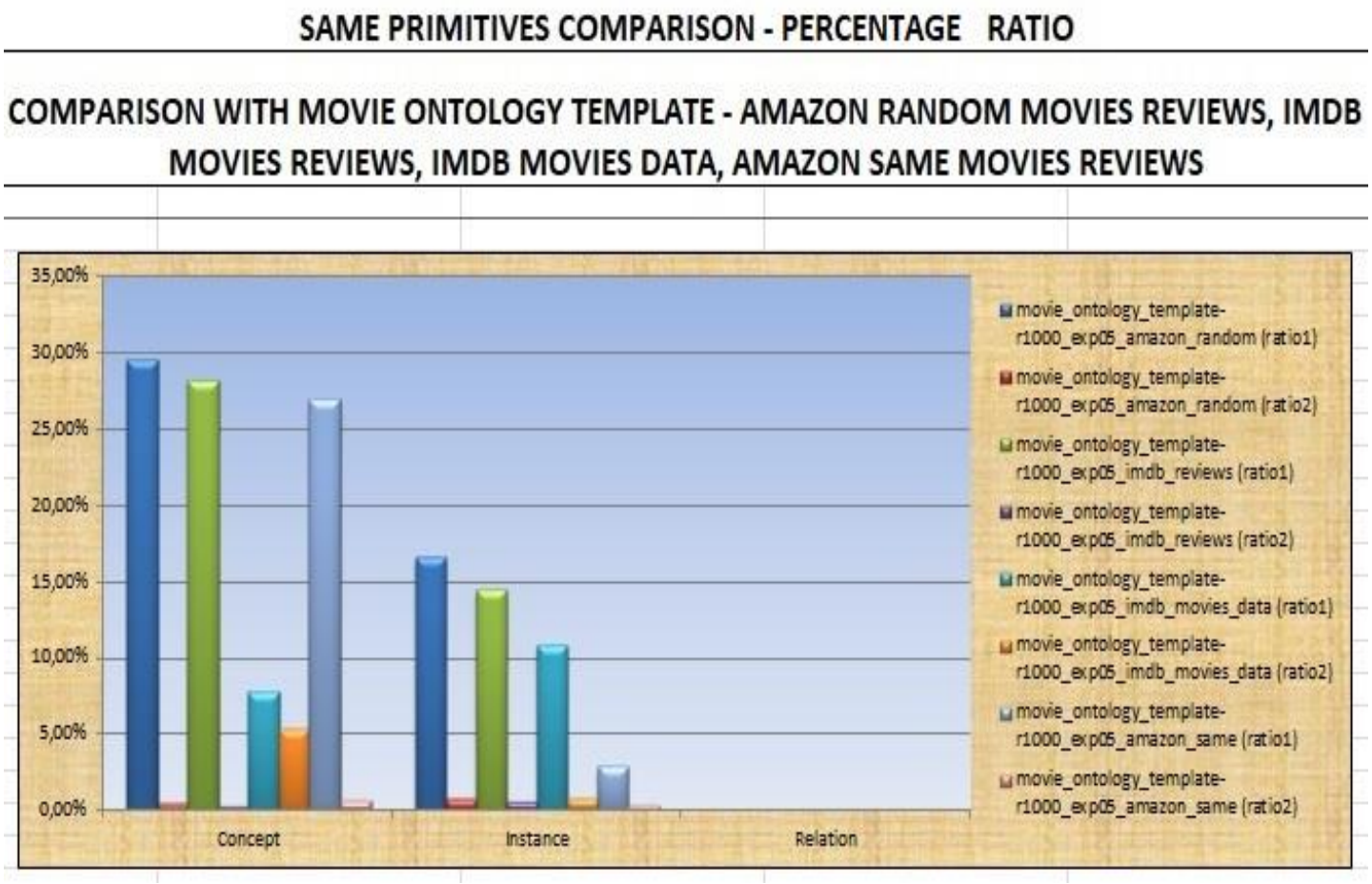
Σύγκριση ως προς την ποσοτική ομαδοποίηση στοιχείων οντολογίας με βάση την τιμή πιθανότητας μεταξύ διαφορετικών κατηγοριών δεδομένων εισόδου

QUANTITATIVE GROUPING AT PROBABILITY LEVELS - PERCENTAGE RATIO COMPARISON - AMAZON RANDOM MOVIES REVIEWS, IMDB MOVIES REVIEWS, IMDB MOVIES DATA, AMAZON SAME MOVIES REVIEWS



Ανάλυση αποτελεσμάτων πειραμάτων

Σύγκριση ως προς τον αριθμό κοινών στοιχείων οντολογιών με πρότυπο οντολογίας κινηματογραφικών ταινιών



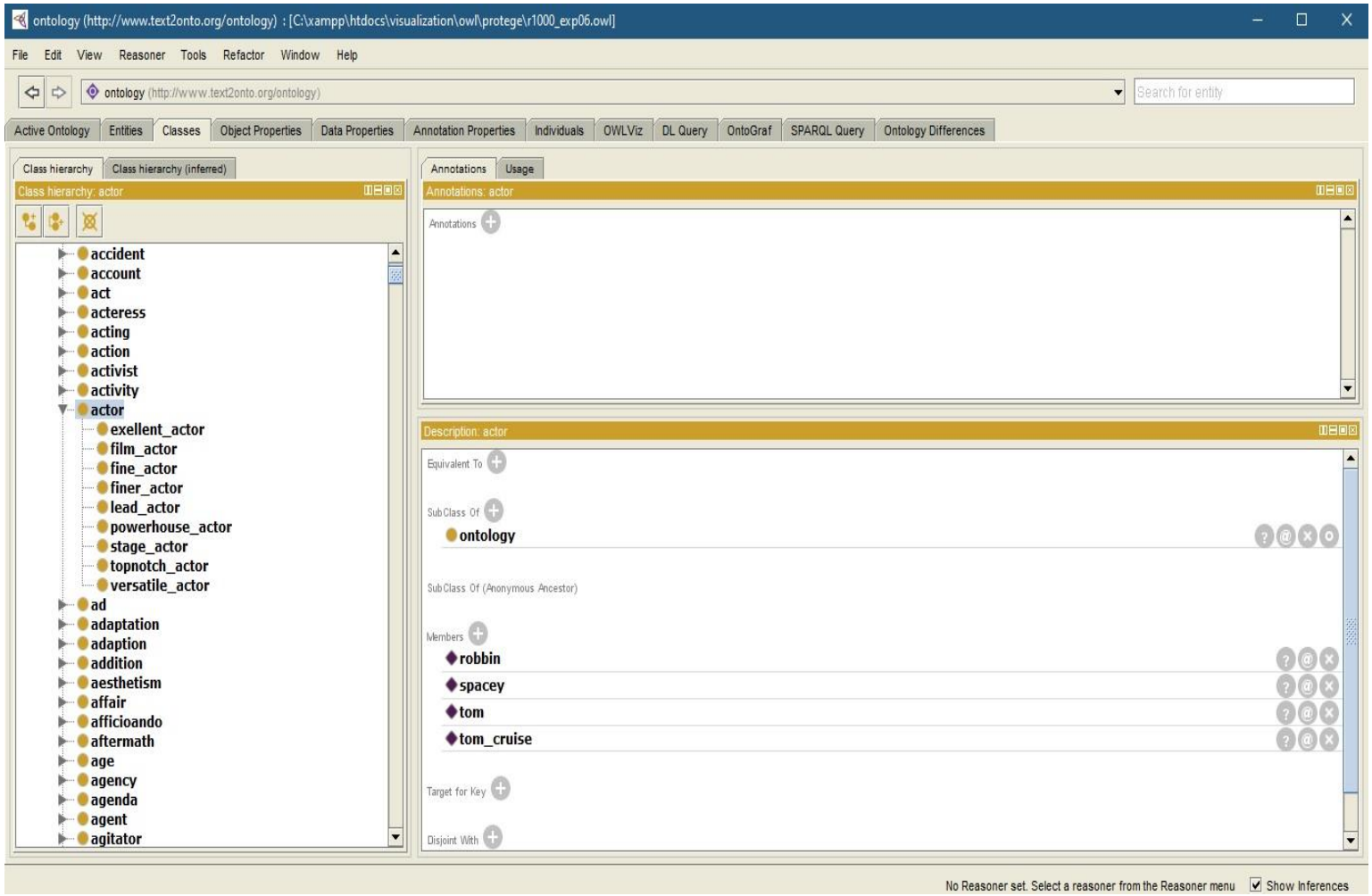
Ανάλυση αποτελεσμάτων πειραμάτων

Είσοδος εξαγόμενης οντολογίας από την εφαρμογή Text2Onto στο περιβάλλον οντολογιών Protégé

- Νέο είδος πειράματος. Επιλέχθηκε η κατηγορία δεδομένων εισόδου που αφορά κριτικές της Amazon από ίδιες ταινίες. Η ποσότητα δεδομένων εισόδου είναι 1000 κριτικές.
- Επιλέχθηκαν οι αλγόριθμοι που μας έδωσαν τα καλύτερα αποτελέσματα ανά θεμελιακό στοιχείο.
  - Concept-TFIDFConceptExtraction
  - Instance-TFIDFInstanceExtraction
  - SubclassOf-VerticalRelationsConceptClassification
  - InstanceOf-PatternInstanceClassification
  - Relation-SubcatRelationExtraction
- Διαγραφή θεμελιακών στοιχείων κάτω από ένα κατώφλι πιθανότητας μέσω της διαδικασίας διαγραφής θεμελιακών στοιχείων.
- Μέσω της διαδικασίας μετατροπής θεμελιακών στοιχείων μετατροπή του owl αρχείου σε μορφή κατάλληλη για είσοδο στο περιβάλλον οντολογιών Protégé.

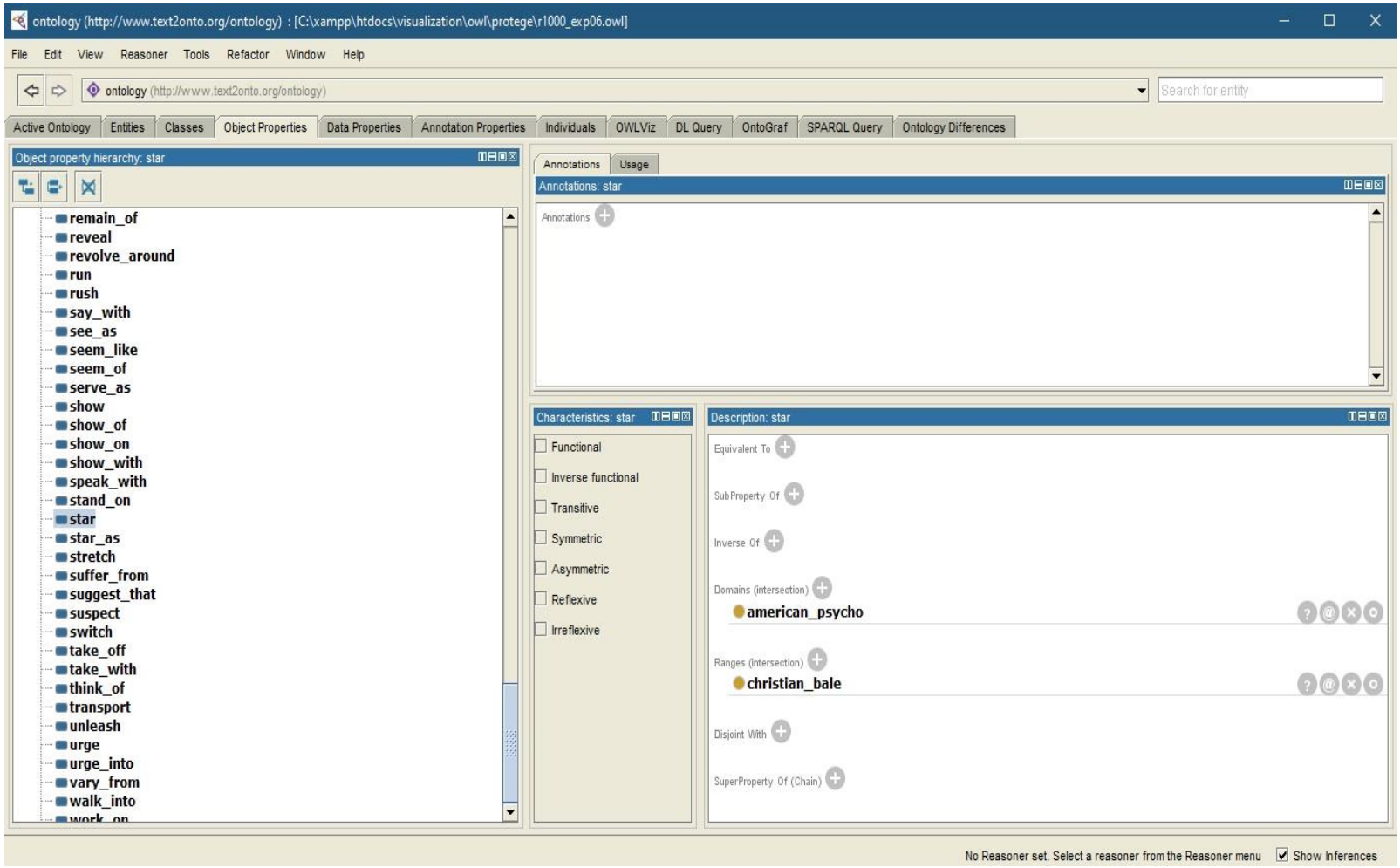
Ανάλυση αποτελεσμάτων πειραμάτων

Είσοδος εξαγόμενης οντολογίας από την εφαρμογή Text2Onto στο περιβάλλον οντολογιών Protégé



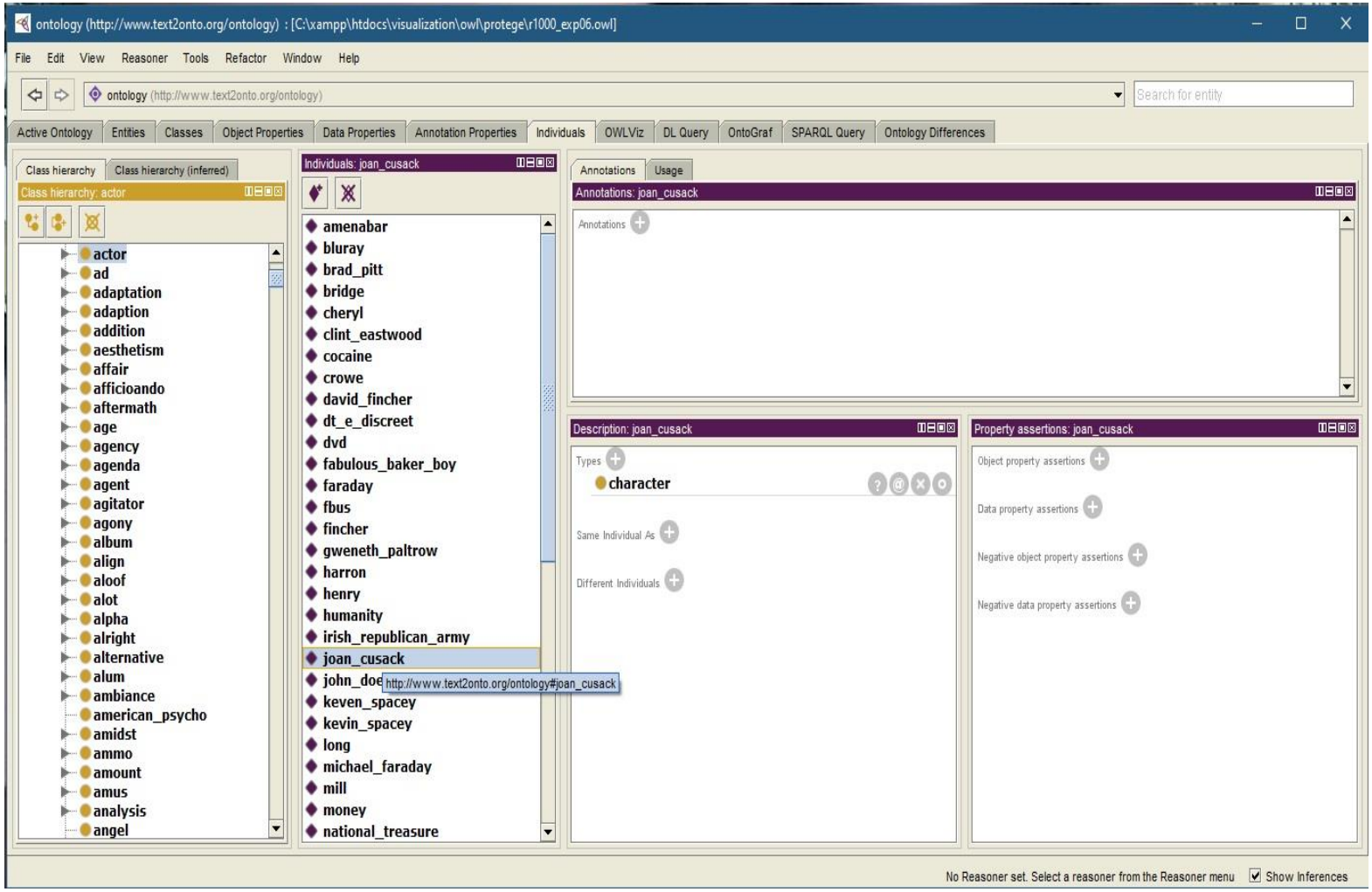
Ανάλυση αποτελεσμάτων πειραμάτων

Είσοδος εξαγόμενης οντολογίας από την εφαρμογή Text2Onto στο περιβάλλον οντολογιών Protégé



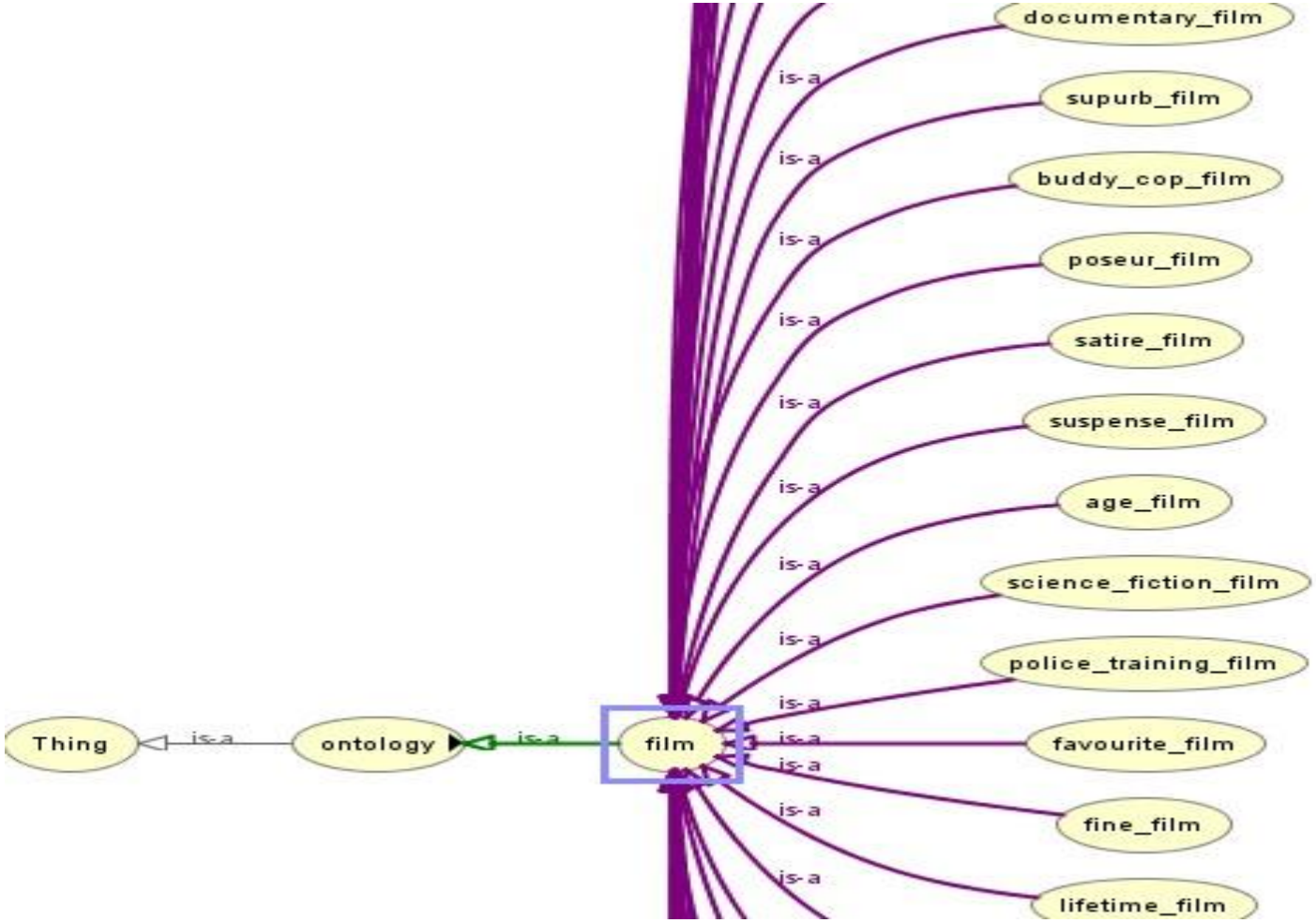
Ανάλυση αποτελεσμάτων πειραμάτων

Είσοδος εξαγόμενης οντολογίας από την εφαρμογή Text2Onto στο περιβάλλον οντολογιών Protégé



Ανάλυση αποτελεσμάτων πειραμάτων

Είσοδος εξαγόμενης οντολογίας από την εφαρμογή Text2Onto στο περιβάλλον οντολογιών Protégé



Συμπεράσματα

- Η εφαρμογή Text2Onto δεν μπορεί να κατασκευάσει σε πλήρη έκταση αυτόματα μια οντολογία μέσω εκμάθησης επάνω σε ένα υποκείμενο σώμα δεδομένων αδόμητου κειμένου.
- Η εφαρμογή Text2Onto μπορεί να λειτουργήσει ως βοηθητικό εργαλείο του χρήστη στην κατασκευή μιας οντολογίας αλλά χρειάζεται βελτίωση.
- Οι κριτικές κινηματογραφικών ταινιών μπορούν να χρησιμοποιηθούν για τη δημιουργίας οντολογίας. Οι περιορισμοί και ο βαθμός αξιοποίησης τους έχουν να κάνουν με τις δυνατότητες του εργαλείου που δεν δίνουν τη δυνατότητα δημιουργίας της οντολογίας αυτόματα και σε πλήρη έκταση αλλά μπορεί να δημιουργηθεί ένας σκελετός που θα βοηθήσει στην περαιτέρω κατασκευή και βελτίωση της οντολογίας.

Συμπεράσματα

- Τα αποτελέσματα για τα θεμελιακά στοιχεία Concept και Instance σε όλα τα πειράματα που πραγματοποιήθηκαν έδιναν στοιχεία οντολογίας μεγάλου πλήθους με μικρές τιμές πιθανότητας. Ειδικότερα το πλήθος τους αυξανόταν πολύ καθώς αυξανόταν και η ποσότητα των δεδομένων εισόδου χωρίς να αυξάνονται ιδιαίτερα οι τιμές πιθανότητάς τους.
- Για το θεμελιακό στοιχείο Similarity η εφαρμογή δεν έδωσε αποτελέσματα σε κανένα από τα πειράματα που πραγματοποιήθηκαν.
- Τα αποτελέσματα για τα θεμελιακά στοιχεία SubclassOf, InstanceOf και Relation έδιναν στοιχεία οντολογίας με αρκετά μεγάλες τιμές πιθανότητας.
- Η σύγκριση των εξαγόμενων οντολογιών των πειραμάτων ως προς τον αριθμό των κοινών στοιχείων με ένα πρότυπο οντολογίας σχετικά με κινηματογραφικές ταινίες έδωσε σε κάποιες περιπτώσεις πειραμάτων θετικά αποτελέσματα.

Σας ευχαριστώ για την προσοχή σας

Ερωτήσεις