NTNU – Trondheim
Norwegian University of
Science and Technology

Department of Computer and Information Science

Examination paper for TDT4215 Web Intelligence

Academic contact during examination: Jon Espen Ingvaldsen

Phone: 98 74 82 89

Examination date: 19th of May, 2015

Examination time (from-to): 09:00 – 13:00

Permitted examination support material: D

Other information:

Language: English

Number of pages (front page excluded):

Number of pages enclosed:

Checked by:

_____

Date Signature

Question 1, Ontologies and SPARQL (20%)

Given the following small ontology:

```
@prefix dc: <http://ntnu.edu/dc/elements/1.1/> .
@prefix stock: <http://ntnu.edu/stock#> .
@prefix inv: <http://ntnu.edu/inventory#> .
stock:book1 dc:title "SPARQL Introduction" .
stock:book1 rdf:type "Computer Science" .
stock:book1 inv:price 10 .
stock:book1 inv:quantity 3 .
stock:book2 dc:title "World War II" .
stock:book2 rdf:type "History" .
stock:book2 inv:price 20 .
stock:book2 inv:quantity 5 .
stock:book3 dc:title "Lost in Space" .
stock:book3 rdf:type "Science Fiction " .
stock:book3 inv:price 5 .
stock:book3 inv:quantity 0 .
stock:book4 dc:title "SpaceX" .
stock:book4 rdf:type "Science Fiction" .
stock:book4 inv:price 20 .
stock:book4 inv:quantity 8 .
stock:book5 dc:title "Novels from Rome" .
stock:book5 rdf:type "Novel collection" .
stock:book5 inv:price 14 .
stock:book5 inv:quantity 2 .
```

a) (10%) What would the following SPARQL-query return?

```
PREFIX dc: <http://ntnu.edu /dc/elements/1.1/>
PREFIX stock: <http://ntnu.edu /stock#>
PREFIX inv: <http://ntnu.edu /inventory#>

SELECT ?book ?title
WHERE {
?book dc:title ?title .
?book inv:price ?price . FILTER ( ?price < 15 )
?book inv:quantity ?num . FILTER ( ?num > 0 ) }
```

Answer, 1a:

SPARQL Introduction

Novels from Rome

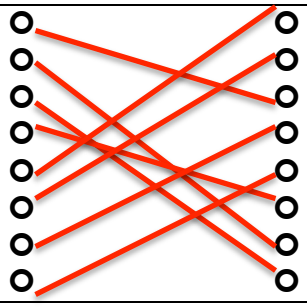
Consider the following OWL-ontology:

```
<!DOCTYPE rdf:RDF [
 <!ENTITY owl "http://www.w3.org/2002/07/owl#">]>
<rdf:RDF xmlns:owl ="http://www.w3.org/2002/07/owl#"
 xmlns:rdf ="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">

<owl:Class rdf:ID="A" />

<owl:Class rdf:ID="B" />
 <rdfs:subClassOf rdf:resource="#A" />
</owl:Class>

<owl:ObjectProperty rdf:ID="C" />

<owl:ObjectProperty rdf:ID="D">
 <rdfs:subPropertyOf rdf:resource="#C" />
</owl:ObjectProperty>

<owl:DatatypeProperty rdf:ID="E" />

<owl:ObjectProperty rdf:ID="F">
 <rdf:type rdf:resource="&owl;TransitiveProperty" />
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="G">
 <rdf:type rdf:resource="&owl;SymmetricProperty" />
</owl:ObjectProperty>

<owl:Class rdf:ID="H">
 <owl:equivalentClass>
 <owl:Restriction>
      <owl:onProperty rdf:resource="#C" />
      <owl:minCardinality
      rdf:datatype="&xsd;nonNegativeInteger">
```

```
      1</owl:minCardinality>
 </owl:Restriction>
 </owl:equivalentClass>
</owl:Class>

</rdf:RDF>
```

b) (10%) The identifiers in the ontology above (ex. "A", "B", etc.) are not very descriptive. Draw lines that connect the identifiers with the correct descriptive identifier.

Identifier	Descriptive Identifier
A	*age*
B	*isTallerThan*
C	*Person*
D	*isFriendOf*
E	*Parent*
F	*hasDaughter*.
G	*Man*
H	*hasChild*



Each correct mapping gives 1.25% points

Question 2, Sentiment Analysis (20%)

a) (5%) What are the inputs and outputs to the unsupervised learning algorithm described in the paper "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews" by Peter D. Turney

Answer, 2a:

INPUT: (2.5%) Review Documents

OUTPUT: (2.5%) Binary classifications (Positive or Negative)

b) (15%) Which steps do this algorithm consists of?

Answer, 2b:

1) (5%) Extract phrases

2) (5%) Estimate Semantic Orientation (SO) of each phrase

3) (5%) Classify the reviews based on average SO

Question 3, Recommender Systems (20%)

Table 1, Dataset of user – item ratings

	Book1	Book2	Book3	Book4	Book5
Alice	1	2	5	?	1
George	5	?	1	?	?
Mary	?	?	4	3	?
Tom	1	1	5	4	?

The table above shows a dataset of user – item ratings. Alice has, for instance, rated "Book3" with a "5" on a 1-to-5 scale, which means that she strongly liked this item. Question marks relate to items that the users have not seen or rated yet.

a) (10%) In a recommender systems setting, -what do we achieve by subtracting the mean from the rating values?

Answer, 3a:

Compensate for personal rating bias.

People use different areas of the rating range in their feedback.

b) (10%) Use the rating-values from Table 1 to calculate the mean-adjusted cosine similarity between by Mary and Tom. Show both the calculation and the result.

Answer, 3b:

Mary has an average rating of **3.5**

Tom has an average rating of **2.75**

Let d1 = 0 0 0.5 -0.5 0

Let d2 = -1.75 -1.75 2.25 1.25 0

Cosine Similarity (d1, d2) = dot(d1, d2) / ||d1|| ||d2||dot(d1, d2) = (0)*(-1.75) + (0)*(-1.75) + (0.5)*(2.25) + (-0.5)*(1.25) + (0)*(0) = 0.5

||d1|| = sqrt((0)^2 + (0)^2 + (0.5)^2 + (-0.5)^2 + (0)^2) = 0.707106781187

||d2|| = sqrt((-1.75)^2 + (-1.75)^2 + (2.25)^2 + (1.25)^2 + (0)^2) = 3.57071421427

Cosine Similarity (d1, d2) = 0.5 / (0.707106781187) * (3.57071421427)

= 0.5 / 2.52487623459

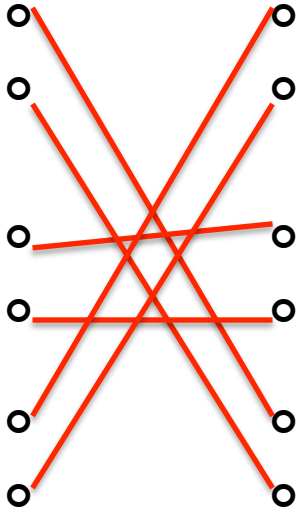
= **0.198029508595**

Correct vector representations give 5%

Correct calculations give 5%.

Question 4, Evaluation of Recommender Systems (20%)

a) (10%) The following table lists a set of evaluation metrics and metric descriptions. Draw lines that connect the metric names with the correct metric descriptions.

Metric		Description
1. Sparsity		*1. Averages Precision and Recall with bias toward the weaker value.*
2. Precision		*2. The average deviation between computed recommendation scores and actual rating values for all evaluated users and all items in their testing sets.*
3. Recall		*3. The proportion of relevant instances that are retrieved.*
4. Accuracy		*4. The proportion of true results (both true positives and true negatives) among the total number of cases examined.*
5. F1		*5. The ratio of empty and total entries in the user–item matrix.*
6. MAE		*6. The proportion of retrieved instances that are relevant.*

Each correct mapping gives 1.67% points

b) (2.5%) Calculate the sparity of dataset in Table 1. Show both the calculation and the result.

Answer, 4b:

$$\text{Sparsity} = 1 - \frac{|R|}{|I| \cdot |U|}$$

- R = ratings
- I= items
- U = users

= 1 – 12 / (5 x 4) = 1 – 12 / 20 = **0.4**

Assume that you built a recommender system recommending our four users with the following items (Alice is recommended with Item1, Item2, and Item3):

Table 2, Recommended items

	Alice	George	Mary	Tom
Item1		X		
Item2	X			
Item3	X		X	X

Item4	X		X	X

When evaluating the recommendations we found out that the following items are considered as truly relevant:

Table 3, Relevant items

	Alice	George	Mary	Tom
Item1				
Item2	X		X	
Item3	X		X	X
Item4			X	X

c) (2.5%) Use the data from Table 2 and 3 to calculate the accuracy the recommendations. Show both the calculation and the result.

Answer, 4c:

$$\text{Accuracy} = \frac{tp+tn}{tp+tn+fp+fn}$$

= 13 / 16 = **0.81**

d) (2.5%) Use the data from Table 2 and 3 to calculate the precision the recommendations. Show both the calculation and the result.

Answer, 4d:

$$\text{Precision} = \frac{tp}{tp+fp}$$

= 6 / (6 + 2) = 6 / 8 = **0.75**

e) (2.5%) Use the data from Table 2 and 3 to calculate the recall the recommendations. Show both the calculation and the result.

Answer, 4e:

$$\text{Recall} = \frac{tp}{tp + fn}$$

= 6 / (6 + 1) = 6 / 7 = **0.86**

Question 5, Linked Data and Recommender Systems (20%)

a) (10%) What is the difference between local and collective entity linking methods?

Answer, 5a:

(5%) Local methods link name mentions in a document by assuming them to be **independent.**

(5%) Collective methods exploit the **interdependence** between the set of entity linking decisions.

b) (10%) What are the strengths and weaknesses of the popularity and semantic relatedness based entity linking approaches?

Answer, 5b:

1. (5%) Popularity based entity linking requires less statistics (only entity frequencies and not entity co-occurrence frequencies)
2. (2.5%) Popularity based entity linking does not require as fresh and updated statistics.
3. (2.5%) Semantic relatedness performs better when co-occurrence data are fresh and updated.