

DSCI 558: Building Knowledge Graphs

Final Exam Sample Questions – Fall 2020

** please note that the total number of points does not sum up to 100. Each number of points assigned to each question is to give an indicator for how many points such question would be worth (and how much time you're expected to spend on it).*

Q1. RDF (10 points)

Convert the following sentences into a single RDF graph using the schema.org ontology:

- Amazon is a store
- Barnes and Noble is a store
- Hamlet is a book
- Hamlet was written by Shakespeare
- Amazon sells a Hamlet book for \$5
- Barnes and Noble sells a Hamlet book for \$7

Classes to use: [the schema.org classes pic on the right]

Properties to use: [all RDF/RDFs properties + author, itemOffered, priceCurrency, price, seller from schema.org]

Classes:

CreativeWork
Book
Intangible
Offer
Demand
Organization
LocalBusiness
Store
Library
Person
Place
Product

Q2. Structured Data (15 points)

Boardgamegeek is a website for board game enthusiasts that collects information about board games. The following shows a screenshot of the BGG webpage with three tables.

The Hotness

Games | People | Company

A

Wingspan

Nemesis

KeyForge: Call of the Archons

Gloomhaven

The Rise of Queensdale

Tiny Epic Zombies Deluxe Edition

Architects of the West Kingdom

Root

Terraforming Mars

Tainted Grail: Fall of Avalon

War Chest

Teotihuacan: City of Gods

Blood Rage

Spirit Island

Arkham Horror: The Card Game

Brettspiel Adventskalender 2018

Scythe

B

1, 2, 3, 4, 5


Next »

[1033]

Board Game Rank ▲

C

1



Gloomhaven (2017)


8.618

8.93

20449


List: \$140.00

Lowest Amazon: \$114.99

New Amazon: \$133.26 

[Shop]

2




Pandemic Legacy: Season 1 (2015)

8.496

8.65


28601

List: \$69.99

New Amazon: \$52.96 

[Shop]

3



Through the Ages: A New Story of Civilization (2015)

8.280

8.57

14065


List: \$69.95

New Amazon: \$67.40

iOS App: \$9.99

[Shop]

4



Terraforming Mars (2016)


8.224

8.39

32333


List: \$69.95

Lowest Amazon: \$40.00

New Amazon: \$44.77 

[Shop]

5



Twilight Struggle (2005)


8.184

8.33

32852

List: \$64.99

Lowest Amazon: \$24.99

New Amazon: \$54.99 

iOS App: \$6.99

[Shop]

1. What are the table types of the three tables shown in the image above?

A. _____

B. _____

C. _____

2. Boardgamegeek provides reports of how many times a game is played in a time range (Qty). You want to represent these values using the RDF Data Cube ontology.

(a) What are the Measure, Attributes, Dimensions, and Observations in this table? Provide a general description of each and examples for the attributes, dimensions, and observations.

Measure: _____

Attributes: _____

Dimensions: _____

Observations: _____

Date Range: 2018-11-01 to 2018-11-30

Format: By Game

Type: Board Game

Go

Games Played in November 2018

Game	Qty	Unique users
KeyForge: Call of the Archons	12852	2384
Azul	8407	3902
Terraforming Mars	6452	3192
Gloomhaven	6053	1926
Magic: The Gathering	5022	438
Ganz schön clever	4762	1151
The Mind	4056	1413
Codenames	3914	1375

(b) Using the classes qb:observation, qb:MeasureProperty, qb:DimensionProperty, and qb:AttributeProperty, Write triples for the number of times Azul was played in November. You do not need to ontologize the attribute and dimension values (use strings).

Q3. Information Extraction (15 points)



REIMPLEMENTS: PANDEMIC
REIMPLEMENTED BY: PANDEMIC LEGACY:....
RANK: OVERALL 2 THEMATIC 2 STRATEGY 2

8.7 **Pandemic Legacy: Season 1** (2015)
29K Ratings & 4.6K Comments · GeekBuddy Analysis

2-4 Players Community: 2-4 — Best: 4	60 Min Playing Time	Age: 13+ Community: 12+	Weight: 2.83 / 5 'Complexity' Rating
--	-------------------------------	-----------------------------------	--

Designer: Rob Daviau, Matt Leacock
Artist: Chris Quilliams
Publisher: Z-Man Games + 10 more
[See Full Credits](#)

1. Boardgamegeek curates information about board games. How would you perform information extraction on the snippet of the board game profile shown above? Remember to describe how you would perform the three steps of information extraction: define the domain, learn extractors, and score facts.



Note: This is a spoiler free review of *Pandemic: Legacy*. Any information discussed here is found in the rulebook or pages loose in the box when opened.

When it comes to original board games on the market today, you'd be hard pressed to find an experience more unique than a Legacy game.

Published in 2011, *Risk: Legacy* turned tabletop gaming on its head by bringing a game to the market that players physically altered over time. Players wrote on, ripped up, and otherwise defaced their game during play, and they loved it. Who would have thought!

Since then, gamers have been chomping at the bit for the next legacy game to tickle our imaginations. Well fret no more because *Risk: Legacy* designer Rob Daviau has teamed up with Matt Leacock, the designer of the perennial best seller *Pandemic* ([review here](#)). What these two industry veterans have brought us is *Pandemic: Legacy*. A game played over twelve months where players must fight off the infections plaguing humanity. Can you save a world that will literally never be the same. Lets find out!

Pandemic: Legacy is a cooperative, campaign style game for 2-4 players that takes about 45-60 minutes to play. *Pandemic: Legacy* plays well with any number of players.

2. Board game reviews can also contain many important attributes about a board game. Using the review given above on the left as an example, describe the three steps of information extraction again. Describe how your approach to information extraction would differ. Be specific! How would you use the Boardgamegeek extractions to help with information extraction?

Q4. Entity Resolution (10 points)

Entity resolution for board games is a challenge, since games have multiple editions, international printings, expansions, and sequels. There are also unrelated games with similar titles. For example, the board game Pandemic has been so popular, there have been many printings of the same game, which each have slightly different names for the same game, such as “Pandemic: Tenth Anniversary Edition,” or “Pandemic - English First Edition (2008)” or “Pandemic - English Edition (2015).” Board games are also loved internationally, so there are many different language variants -- in Spain, Portugal, and Denmark “Pandemic” is called “Pandemia” and the French call it “Pandemié.” Since Pandemic was so popular, there have been many separate games published as sequels (not the same entity!) including “Pandemic Legacy: Season 1,” “Pandemic Legacy: Season 2,” “Pandemic: Iberia,” “Pandemic: On the Brink,” “Pandemic: In the Lab,” and “Pandemic: Fall of Rome.” There are, of course, board games totally unrelated to Pandemic which have similar titles, like “Risk: Legacy” or “The Fall of Rome.” And there are many games with unrelated titles, “Gloomhaven,” “Through the Ages,” “Terraforming Mars,” and “Twilight Struggle” are examples from an earlier question.

For each of the string similarity functions below, explain whether it would be a good choice for entity linking board game names and explain its strengths and weaknesses using examples from this description to explain when it would succeed and fail.

- a. Levenshtein
- b. Jaccard
- c. Jaro Winkler
- d. Smith Waterman
- e. TF-IDF

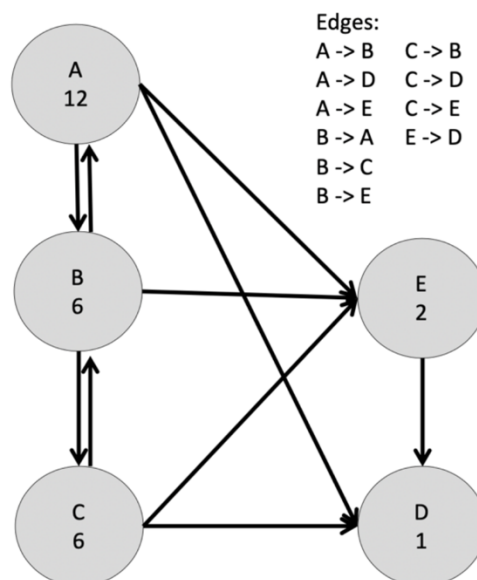
Q5. KG Embeddings (5 points)

Which of the following statements about knowledge graph embeddings is/are true, as taught in class?

- They encode concepts from a knowledge graph into low-dimensional continuous vectors
- They can be generated only if an ontology or taxonomy of classes is already defined
- They embed the entities in the graph but do not embed the relations
- They cannot support graphs with asymmetrical relations
- They cannot support graphs with 1-to-many relations
- They have poor performance on sparse and noisy graphs extracted from text
- Their performance is harmed when the original graph is extended with noisy data
- Their performance degrades with sparsity
- They can have complex values

Q6. Graph Analytics (5 points)

Each node in the following graph is labeled with an initial PageRank. Provide the updated values for each node PageRank after one iteration of the PageRank algorithm (with damping factor 0). At convergence, which node will have the highest PageRank?



(Please refer to the node number when writing each answer: A, B, ...)

Q7. SPARQL Queries (10 points)

Suppose you have a graph that includes thousands of `rdf:Statements` of the following form:

```
:s_i      a rdf:Statement ;  
          rdf:subject <subject_i> ;  
          rdf:object <object_i> ;  
          rdf:predicate <predicate_i> ;  
          ex:confidence <confidence_i> .
```

`<confidence_i>` is a real number between 0 and 1.

- a) Write a SPARQL CONSTRUCT query that generates triples of the form shown below for the `rdf:Statements` that have confidence ≥ 0.5 .

```
<subject_i> <predicate_i> <object_i> .
```

- b) Write a SPARQL query that counts the number of `rdf:Statement` with `ex:confidence` < 0.5

Q8. Real-World KGs (6 points)

You decide to link your board game knowledge graph to some of the large, public knowledge graphs currently published. Explain which real-world KG you would use for each scenario and why:

- a. Most board games in your KG have associated Wikipedia pages
- b. You want to add your triples to an existing KG
- c. You want to translate rules into a different language

Q9. Graph Embeddings (6 points)

You decide to embed your board game knowledge graph to predict missing relations and score the existing facts. You are considering the tensor-factorization approach RESCAL or the embedding approach Trans-H. One issue you're worried about is that games have multiple designers and designers design multiple games over their career. Which graph embedding method will work better? Why?

Q10. Probabilistic Models (15 points)

One of the properties of board games on Boardgamegeek is their “weight” or complexity. “Heavy” games take a long time to play and require a great deal of strategy, while “Light” games are quick and easy to play and often good for families. Board game weights are currently determined by having boardgamers vote, but you want to build a PSL model to infer the weight of all games simultaneously.

1. Write a PSL model to infer the target $\text{Heavy}(\text{Game})$ using the predicates Similar (two games are similar), DesignedBy (the designer of a game), ShortPlayingTime (if the game takes less than an hour), Strategy (whether the game is in the strategy category), and VotedHeavy (whether gamers have voted this game as heavy).
2. Given the logical atoms from a KG:
 $\text{Similar}(\text{Pandemic}, \text{Pandemic Legacy})=0.9$, $\text{VotedHeavy}(\text{Pandemic})=0.5$, $\text{Strategy}(\text{Pandemic})=1$,
and $\text{ShortPlayingTime}(\text{Pandemic})=1$
write out the groundings that your model will produce.
3. Write out the mathematical loss function for the ground rules for Strategy and ShortPlayingTime . You should have two simple inequalities for $\text{Heavy}(\text{Pandemic})$.