Task 1: Dataset Catalogues

Use the DataHub catalogue on https://old.datahub.io/dataset to find a dataset under the Creative Commons Attribution license about weather that has a SPARQL endpoint and is organized by the Linked Open Data Cloud.
What is the dataset's name, when was it created and how many triples does it currently contain?

> *Resulting dataset after applying the filters and using "weather" as search query: https://old.datahub.io/dataset/knoesis-linked-sensor-data.*
>
> *Name: Linked Sensor Data (Kno.e.sis)*
> *Creation date: 28. Juli 2010, 12:36 (UTC+02:00)*
> *Number of triples: 1730284735*

Task 2: Dataset Description

Given is the following description of the YAGO knowledge graph in natural language[1] [2] [3]:

YAGO is a huge semantic knowledge base, derived from Wikipedia WordNet and GeoNames. Currently, YAGO has knowledge of more than 10 million entities (like persons, organizations, cities, etc.) and contains more than 120 million facts about these entities.

YAGO is special in several ways:

1. The accuracy of YAGO has been manually evaluated, proving a confirmed accuracy of 95%. Every relation is annotated with its confidence value.
2. YAGO combines the clean taxonomy of WordNet with the richness of the Wikipedia category system, assigning the entities to more than 350,000 classes.
3. YAGO is an ontology that is anchored in time and space. YAGO attaches a temporal dimension and a spatial dimension to many of its facts and entities.
4. In addition to a taxonomy, YAGO has thematic domains such as "music" or "science" from WordNet Domains.
5. YAGO extracts and combines entities and facts from 10 Wikipedias in different languages.

YAGO is a joint project of the Max Planck Institute for Informatics and the Telecom ParisTech University.
The YAGO Ontology is licensed under a Creative Commons Attribution 3.0 License by the YAGO team of the Max-Planck Institute for Informatics.
YAGO was first created on October 27, 2009 and last updated on April 28, 2018.

---

[1] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago
[2] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/downloads/
[3] https://old.datahub.io/dataset/yago

2.1 Give three example dataset profile features that you can find in this text. Also name their corresponding categories (general, qualitative, provenance, links, licensing, statistical, dynamics)

- *Qualitative, accuracy/trust: The accuracy of YAGO has been manually evaluated, proving a confirmed accuracy of 95%.*
- *Licensing Features: The YAGO Ontology is licensed under a Creative Commons Attribution 3.0 License by the YAGO team of the Max-Planck Institute for Informatics.*
- *Statistical Features: YAGO has knowledge of more than 10 million entities and contains more than 120 million facts about these entities.*

2.2 Write down a formal description of this dataset using the VoID vocabulary[4].

```
:yago
      a void:Dataset ;
      dcterms:title "YAGO" ;
      dcterms:description "YAGO is a huge semantic knowledge base, derived
      from Wikipedia WordNet and GeoNames." ;
      dcterms:issued "2009-10-27"^^xsd:date ;
      dcterms:modified "2018-04-28"^^xsd:date ;
      dcterms:language <http://id.loc.gov/vocabulary/iso639-1/en>  ;
      dcterms:source :wordnet  ;
      dcterms:license <https://creativecommons.org/licenses/by/3.0/>;
      dcterms:publisher :mpi  ;
      dcterms:publisher :telekom_uni  ;
      .

:telekom_uni a foaf:Organization;
      rdfs:label "Telecom ParisTech University";
      foaf:homepage <http://www.telecom-paristech.fr/eng/>;
      .

:mpi a foaf:Organization;
      rdfs:label "Max Planck Institute for Informatics";
      foaf:homepage <http://www.mpi-inf.mpg.de/>;
      .

:wordnet
      a void:Dataset;
      dcterms:title "WordNet" ;
      .
```

[4] https://www.w3.org/TR/void/

Task 3: Statistical Features of DBpedia

3.1 Write a query to count the total numbers of triples in the English DBpedia (http://dbpedia.org/sparql/).

```
SELECT COUNT(*) WHERE {
      ?a ?b ?c .
}
```

Result: 438336517

3.2 What does the following query return?

```
SELECT COUNT(*) WHERE {
      SELECT DISTINCT(?b) WHERE {
            ?a ?b ?c .
      }
}
```

This query returns the number of different properties used in DBpedia (60649).

3.3 What does the following query return?

```
SELECT ?b ?c {
      dbo: ?b ?c .
}
```

This query returns meta information about the DBpedia ontology itself, e.g. `dbo: dcterms:issued "2008-11-17T12:00Z"` and `http://purl.org/dc/terms/publisher "DBpedia Maintainers"`.

3.4a Write a SPARQL query that, for each type (`?type`), says how many instances it has.

```
SELECT ?type COUNT(*) WHERE {
      ?subject rdf:type ?type .
}
GROUP BY ?type
```

3.4b Change the query from Task 3.4a, such that the types are returned by decreasing frequency.

```
SELECT ?type (COUNT(*) AS ?cnt) WHERE {
      ?subject rdf:type ?type .
}
GROUP BY ?type
ORDER BY DESC(?cnt)
```

Top result: http://xmlns.com/foaf/0.1/Document 12856178

3.5 What happens if we include the line "`?type rdfs:subClassOf* dbo:Person .`" into the query from Task 3.4b?

The types are restricted to dbo:Person and its (transitive) subclasses, e.g. `dbo:TennisPlayer` and `dbo:Athlete`.

Task 4: A Priori Type Probabilities

Infer the type a priori probabilities table from the **French (!)** DBpedia (http://fr.dbpedia.org/sparql).

4.1 Write a SPARQL query to count the total number of instances in DBpedia with at least one type.

```
SELECT COUNT(DISTINCT(?subject)) WHERE {
      ?subject rdf:type ?type .
}
```

Result: 6015374 instances.

4.2 Use the query results from Task 4.1 and Task 3.4b to compute the type distribution in the French DBpedia (a priori probabilities) for the three most frequent types.

Type t	Frequency	P(t) in %
`http://xmlns.com/foaf/0.1/Document`	2944139	$\frac{2944139}{6015374} = 48.94$
`http://www.w3.org/2002/07/owl#Thing`	1527645	$\frac{1527645}{6015374} = 25.40$
`dbo:Image`	1124808	$\frac{1124808}{6015374} = 18.70$

Task 5: Type Distributions for a Given Type

Infer the type distribution table of the property `dbo:location` from the **French** DBpedia.

5.1 Write a SPARQL query that returns the total number of triples that have the property `dbo:location`.

```
PREFIX dbo: <http://dbpedia.org/ontology/>
SELECT COUNT(DISTINCT(?subject)) WHERE {
      ?subject dbo:location ?object.
      ?subject rdf:type dbo:Place .
}
```

Result: 5447 triples. The dbo: namespace is not pre-defined when querying the French DBpedia, so it has to be explicitly defined.

5.2 Write a SPARQL query to find out how many subjects (objects) of the `dbo:location` triples are of a specific type (e.g., 100 subjects in triples with `dbo:location` are having the type `dbo:Person`). Limit the types to those from the `dbo` namespace and order the output by decreasing type count.

```
PREFIX dbo: <http://dbpedia.org/ontology/>
SELECT ?type (COUNT(DISTINCT(?subject/?object)) AS ?frequency) WHERE {
      ?subject dbo:location ?object.
      ?subject/?object rdf:type ?type .
      FILTER(STRSTARTS(STR(?type), "http://dbpedia.org/ontology")) .
} GROUP BY ?type
ORDER BY DESC(?frequency)
```

5.3 For the three most frequent subject and object types: Compute their type distribution using the outputs of Task 5.1 and 5.2.

Type	Subject Count	Subject (%)	Object Count	Object (%)
Place	5447	$\frac{5447}{5447} = 100.00$	1278	$\frac{1278}{5447} = 23.46$
Location	5447	100.00	1278	23.46
Building	5176	95.02	82	1.51
PopulatedPlace	0	0	1117	20.51

Task 6: Type Inference Algorithm

Use the SD-Type Algorithm to infer the type of `dbr:Eiffel_Tower`. Compute the type confidence scores for `dbo:Place`, `dbo:Building` and `dbo:Person`.

Given are the following triples with the entity Eiffel Tower:

```
dbr:Eiffel_Tower dbo:location dbr:Paris
dbr:Eiffel_Tower dbo:architect dbr:Stephen_Sauvestre
dbr:France_Télévisions dbo:location dbr:Eiffel_Tower
```

Additionally, the following property distributions and the a priori type distributions are given:

`dbo:location`

Type	Subject	Object
dbo:Place	0.80	0.75
dbo:Building	0.17	0.01
dbo:Organisation	0.16	0.02
dbo:City	0.03	0.21

`dbo:architect`

Type	Subject	Object
dbo:Place	1.00	0.0
dbo:Building	0.88	0.0
dbo:Person	0.0	0.81
dbo:Architect	0.0	0.36

A priori type distribution:

Type t	P(t)
dbo:Place	0.43
dbo:Building	0.12
dbo:Organisation	0.21

Type t	P(t)
dbo:City	0.08
dbo:Person	0.57
dbo:Architect	0.03

Property Weights:

$$w_{dbo:location} = (0.43 - 0.80)^2 + (0.12 - 0.17)^2 + (0.21 - 0.16)^2$$
$$(0.08 - 0.03)^2 + (0.057 - 0.0)^2 + (0.03 - 0.0)^2 \approx 0.47$$

$$w_{dbo:location}{}^{-1} = (0.43 - 0.75)^2 + (0.12 - 0.01)^2 + (0.21 - 0.02)^2$$
$$(0.08 - 0.21)^2 + (0.57 - 0.0)^2 + (0.03 - 0.0)^2 \approx 0.49$$

$$w_{dbo:architect} = (0.43 - 1.0)^2 + (0.12 - 0.88)^2 + (0.57 - 0.0)^2$$
$$(0.03 - 0.0)^2 + (0.21 - 0.0)^2 + (0.08 - 0.0)^2 \approx 1.28$$

$$conf(dbo : Building(dbr : Eiffel\_Tower))$$
$$= \frac{1}{0.47+0.49+1.28}(0.47 \cdot 0.17 + 0.49 \cdot 0.01 + 1.28 \cdot 0.88) \approx 0.54$$

$$conf(dbo : Place(dbr : Eiffel\_Tower))$$
$$= \frac{1}{0.47+0.49+1.28}(0.47 \cdot 0.80 + 0.49 \cdot 0.75 + 1.28 \cdot 1.0) \approx 0.90$$

$$conf(dbo : Person(dbr : Eiffel\_Tower))$$
$$= \frac{1}{0.47+0.49+1.28}(0.47 \cdot 0.0 + 0.49 \cdot 0.0 + 1.28 \cdot 0.0) = 0$$

Example: With a threshold of 0.5, the Eiffel Tower would be assigned the types dbo:Building and dbo:Place.