Práctica 4 (opcional). Relación entre entidades en Wikidata

Fecha de entrega: 14 de junio de 2020

Uno de los usos más interesantes de las redes semánticas es calcular la relación entre dos entidades recorriendo las aristas del grafo hasta encontrar una entidad común. Por ejemplo, utilizando sólo la relación *subclass of* podemos calcular la relación entre las entidades piano (Q5994) y electronic keyboard (Q1343007) que es la siguiente:

```
piano (Q5994) -> subclass of (P279) -> keyboard instrument (Q52954)
```

```
electronic keyboard (Q1343007) -> subclass of (P279) -> keyboard
instrument (Q52954)
```

En este caso las dos entidades representan instrumentos que son de teclado. Además, podemos calcular una distancia semántica entre ellas sumando las longitudes de los caminos que llevan desde cada una de ellas a la entidad común. En este caso sólo hemos tenido que recorrer una arista desde cada una de ellas para llegar a *keyboard instrument* así que podemos decir que se encuentran a distancia 1+1=2 y que la entidad común está a distancia 1 de cada una de ellas.

El algoritmo que calcula la entidad común más cercana lo hemos visto en clase. Se basa en crear una onda a partir de cada una de las entidades que estamos comparando y expandirlas alternativamente hasta que intersecan.

Parte 1

Implementa una función que, dados dos identificadores de ítems y una distancia máxima, devuelva los ítems comunes que están como máximo a esa distancia de los dos ítems iniciales.

Como Wikidata es muy grande vamos a limitarnos a buscar entidades comunes a distancia máxima 3 usando sólo el siguiente subconjunto de propiedades:

```
instance of (P31), subclass of (P279), has part (P527),
instrument (P1303), genre (P136)
```

Si no hay entidades comunes a esa distancia diremos que los ítems no están relacionados.

Calcula las relaciones entre las siguientes entidades dos a dos y explica en lenguaje natural el tipo de relación encontrada:

```
piano (Q5994), electronic keyboard (Q1343007),
String synthesizer (Q2355465), Hans Zimmer (Q76364),
Pirates of the Caribbean: The Curse of the Black Pearl (Q46717),
The Lion King (Q36479), Toy Story (Q171048), Iron Man (Q192724)
```

Consejos:

- Obtener la información de una entidad Wikidata con *get_entity_dict_from_api* es una operación lenta. Te recomendamos que crees una cache de entidades (map entidad_id -> entidad) que evite pedir una y otra vez las mismas entidades a Wikidata.
- Te recomendamos crear una función *get_claims(item_id)* que devuelva las relaciones que salen a partir de una entidad en formato (id_propiedad, id_valor). Sólo debes considerar los *truthy_claims* que tengan como valor un item de Wikidata (entidades que empiezan por 'Q'). Recuerda que también hemos limitado las propiedades que debes considerar.
- Es importante tener en cuenta que, cuando expandimos una onda, la intersección se puede producir con cualquiera de las entidades de la otra onda (y no sólo con las de la frontera).

Parte 2

Implementa una versión avanzada del algoritmo visto en clase que no sólo devuelva las entidades comunes sino también los caminos que conectan cada una de las entidades originales con esas entidades comunes. En este caso, la función sólo debe devolver los caminos de longitud mínima. Si no hay caminos de longitud menor o igual a 5 diremos que las entidades no están relacionadas.

Vuelve a comparar las entidades del apartado anterior dos a dos mostrando los caminos de longitud mínima. Sólo es necesario considerar las propiedades indicadas previamente.

Consejos:

- La intersección de las dos ondas puede contener varias entidades. Debes considerar todas ellas a la hora de calcular las posibles soluciones. Además, cada una de las entidades comunes puede dar lugar a varias soluciones si hay distintos caminos que conectan las entidades iniciales con esa entidad común.
- De todas las soluciones posibles sólo debes devolver las de longitud mínima, entendiendo que la longitud de la solución es la suma de las longitudes de los caminos desde las entidades iniciales hasta la entidad común.

- Hay muchas formas de implementar este algoritmo. Una de ellas consiste en almacenar todos los caminos de cada una de las ondas e ir expandiéndolos alternativamente hasta que las ondas intersequen. Para calcular la intersección deberás almacenar también los ítems de cada una de las ondas.
- Ten en cuenta que para calcular todas las posibles soluciones de longitud <= n debes expandir cada una de las ondas n veces (puede que la relación vaya del primer ítem hasta el segundo).

Nota: se puede obtener la máxima calificación entregando sólo la parte 2 si se explica en lenguaje natural las relaciones encontradas.