Assignment 2: Ontology management

Rok Zidarn, Miha Pešič

December 2016

1 Uvod

To nalogo smo izdelovali v parih. Najprej smo si izbrali zbirko podatkov in iz njene vsebine razbrali povezave med podatki. Te povezave je bilo potrebno predstaviti v ontologiji z orodjem Protégé[1], nato pa implementirati avtomatsko dodajanje podatkov v ontologijo s pomočjo Apache Jena Fuseki[2] strežnika in SPARQL poizvedb. Spisali smo tudi 5 poizvedb, relevantnih za izbrano zbirko podatkov in ročni vnos podatkov v ontologijo.

2 Podatki

Za zbirko podatkov sva si izbrala *20 Newsgroups*, ki je dosegljiva na http://qwone.com/ jason/20Newsgroups/. Zbirka vsebuje približno 20000 dokumentov novic, ki so večinoma enakomerno porazdeljeni v 20 skupin glede na vsebino.

Vsaka novica ima svojega avtorja, ki ima ime, priimek in email. Avtor je zaposlen pri organizaciji s svojim nazivom in distribucijo. Novica ima tudi podatek o času nastanka, ki ga razbijemo na čas, časovno območje in datum. Iz datuma razberemo še dan nastanka. Novica spada v eno izmed 20 vsebinskih skupin, vsaka od teh ima svoje ime. Podatki o vsebini novice so razbiti v povzetek, zadevo in število vrstic v novici.

3 Ontologija

Slika 1 prikazuje ontologijo, zgrajeno v orodju Protégé.

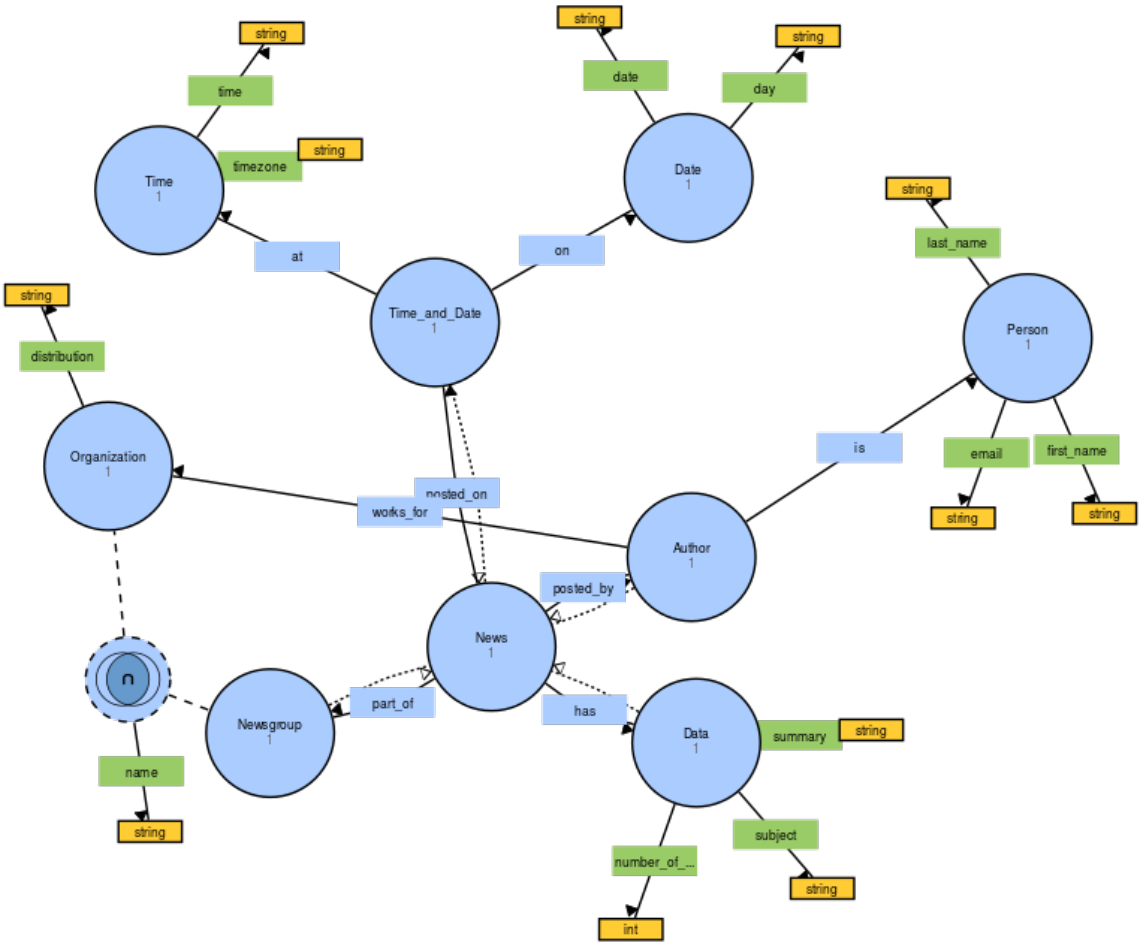

Figure 1: Zgradba ontologije

4 Python implementacija

V programskem jeziku Python je implementirano iskanje podatkov iz dokumentov s pomočjo regularnih izrazov. Zaradi lažjega testiranja je število procesiranih dokumentov odvisno od tega, koliko se jih nahaja v direktoriju news/. Vsi pridobljeni podatki so s pomočjo SPARQL INSERT stavkov vnešeni v ontologijo, ki se nahaja na aktivnem Fuseki strežniku. Ko so podatki vnešeni v ontologijo, lahko nad njimi izvajamo ročne poizvedbe in vnašanja novih podatkov.

5 SPARQL

Vse poizvedbe v naslednjih razdelkih na začetku vsebujejo še naslednjo kodo:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rm: <http://www.semanticweb.org/2016/ontology/rm#>
```

5.1 INSERT

Tu je primer ročnega vnosa podatkov v ontologijo. Prikazani podatki so izmišljeni in se lahko zamenjajo za najdene povezave iz dokumentov, ki jih Python skripta ni zaznala. Vnosi so razbiti na več kosov zaradi lažje preglednosti.

```
INSERT DATA {
  rm:Elon_Musk rdf:type rm:Person .
  rm:Elon_Musk rm:first_name "Elon" .
  rm:Elon_Musk rm:last_name "Musk" .
  rm:Elon_Musk rm:email "elon@musk.com"  .}


INSERT DATA {
  rm:Tesla rdf:type rm:Organization .
  rm:Tesla rm:name "Tesla Motors" .
  rm:Tesla rm:distribution "USA" .}


INSERT DATA {
  rm:Top_Secret rdf:type rm:Data .
  rm:Top_Secret rm:subject "New technology" .
  rm:Top_Secret rm:summary "A new car that runs on oxygen" .
  rm:Top_Secret rm:number_of_lines 200 .}
```

```
INSERT DATA {
  rm:Rec_Autos rdf:type rm:Newsgroup .
  rm:Rec_Autos rm:name "rec.autos" .}


INSERT DATA {
  rm:15_Jan_2017 rdf:type rm:Date .
  rm:15_Jan_2017 rm:date "15 Jan 2017" .
  rm:15_Jan_2017 rm:day "Sun" .}


INSERT DATA {
  rm:17_29_23 rdf:type rm:Time .
  rm:17_29_23 rm:time "17:29:23" .
  rm:17_29_23 rm:timezone "PST" .}


INSERT DATA {
  rm:Elon_Musk_Tesla rdf:type rm:Author .
  rm:Elon_Musk_Tesla rm:is rm:Elon_Musk .
  rm:Elon_Musk_Tesla rm:works_for rm:Tesla .

  rm:15_Jan_2017_17_29_23_PST rdf:type rm:Time_and_Date .
  rm:15_Jan_2017_17_29_23_PST rm:at rm:17_29_23 .
  rm:15_Jan_2017_17_29_23_PST rm:on rm:15_Jan_2017 .

  rm:00010 rdf:type rm:News .
  rm:00010 rm:posted_on rm:15_Jan_2017_17_29_23_PST .
  rm:00010 rm:posted_by rm:Elon_Musk_Tesla .
  rm:00010 rm:has rm:Top_Secret .
  rm:00010 rm:part_of rm:Rec_Autos .}
```

5.2 QUERY

V tem delu se nahaja 5 poizvedb, ki so se nama zdele smiselne glede na izbrano zbirko podatkov.

5.2.1 Splošni podatki o novicah

```
SELECT (?subjectd AS ?NEWS) (STR(?news_n) AS ?GROUP)
    (STR(?news_subject) AS ?SUBJECT)
    (STR(?number_of_lines) AS ?LINES)
```

```
WHERE {
  ?subjectn rdf:type rm:Newsgroup .
  ?subjectn rm:name ?news_n .
  ?subjectd rdf:type rm:Data .
  ?subjectd rm:subject ?news_subject .
  ?subjectd rm:number_of_lines ?number_of_lines
}
```

5.2.2 Avtorji, ki so napisali novico dolžine vsaj 15 vrstic

```
SELECT DISTINCT(CONCAT(?fname," ",?lname) AS ?Name)
WHERE {
  ?subjectd rdf:type rm:Data .
  ?subjectd rm:number_of_lines ?number_of_lines .
  ?subjectp rdf:type rm:Person .
  ?subjectp rm:first_name ?fname .
  ?subjectp rm:last_name ?lname .
  FILTER (
    ?number_of_lines > 15
  )
}
```

5.2.3 Število novic, objavljenih na posamezen dan v tednu

```
SELECT ?Day (COUNT(?news) AS ?totalNews)
WHERE {
    ?news rdf:type rm:News .
    ?news rm:has ?data .
    ?news rm:posted_on ?timedate .
    ?timedate rm:on ?date .
    ?date rm:day ?Day .
}
GROUP BY ?Day
```

5.2.4 Imena avtorjev in skupin novic zaposlenih pri organizaciji Microsoft Corporation

```
SELECT (CONCAT(?fname," ",?lname) AS ?Name) ?Newsgroup ?Employer
WHERE {
    ?news rdf:type rm:News .
    ?news rm:part_of ?group .
    ?group rm:name ?Newsgroup .
    ?news rm:posted_by ?author .
    ?author rm:works_for ?employer .
    ?author rm:is ?authorPerson .
    ?authorPerson rm:first_name ?fname .
```

```
    ?authorPerson rm:last_name ?lname .
    ?employer rm:name ?Employer
    FILTER (
        ?employer = rm:MicrosoftCorporation
    )
}
```

5.2.5 Skupina in tema novic, objavljenih ob nedeljah

```
SELECT ?Newsgroup ?Subject ?Day
WHERE {
    ?news rdf:type rm:News .
    ?news rm:has ?data .
    ?news rm:part_of ?newsgroup .
    ?newsgroup rm:name ?Newsgroup .
    ?data rm:subject ?Subject .
    ?news rm:posted_on ?timedate .
    ?timedate rm:on ?date .
    ?date rm:day ?Day .
    FILTER(
      ?Day = "Sun"
    )
}
```

6 Zaključek

V tej nalogi smo se seznanili z ontologijami, iskanjem podatkov s pomočjo regularnih izrazov, vnašanjem podatkov v ontologijo in pisanjem SPARQL poizvedb. Implementirano je avtomatsko branje datotek, ekstrakcija podatkov in vnašanje v ontologijo. Vključenih je nekaj SPARQL poizvedb, ki se nanašajo na podatke iz izbrane zbirke.

References

[1] "Protégé," http://protege.stanford.edu/, accessed: 2016-12-10.

[2] "Apache jena fuseki," https://jena.apache.org/documentation/fuseki2/index.html, accessed: 2016-12-10.