Informationsintegration
Übung 5

SS 2016

Yvonne Lichtblau

Vorstellung Lösungen Übung 4

Wettbewerb

Platz 1: Gruppe 3 (1.58s)
Platz 2: Gruppe 6 (1.94s)
Platz 3: Gruppe 8 (3.11s)

(user time)

Gruppe	1	2	3	4	5	6	8
Web Scraping (Korrektheit)	3			1			5
Web Scarping (Geschwin-digkeit)	5	3			1		
Hierarchical Queries	3		1				5
Query Containment			5			3	1
Summe	11	3	6	1	1	3	11

Assignment 5
Ontology Matching

Übersicht

Schreibt ein (String-Matching basiertes) Programm, welches ein 1:1 Alignment/Matching zwischen der Ontologie für die Anatomie des Menschen und der Ontologie für die Anatomie der Maus berechnet.

Datenset

- Ontology Alignment Evaluation Initiative
  OAEI-2015 Campaign: Teilwettbewerb Anatomie

  http://oaei.ontologymatching.org/2015/anatomy/index.html

- Download des Datensets:
  http://oaei.ontologymatching.org/2015/anatomy/anatomy-dataset.zip

- Drei Dateien:
  human.owl (Ontologie Anatomie Mensch)
  mouse.owl (Ontologie Anatomie Maus)
  reference.rdf (Referenz Alignment)

- Auf Übungs Homepage:
  Java-Programm zum Einlesen der Ontologien

Dateiformate (1)

reference.rdf (RDF: Resource Description Framework)

```
  </map>
-<map>
  -<Cell>
      <entity1 rdf:resource="http://mouse.owl#MA_0000200"/>
      <entity2 rdf:resource="http://human.owl#NCI_C32726"/>
      <measure rdf:datatype="xsd:float">1.0</measure>
      <relation>=</relation>
    </Cell>
  </map>
-<map>
```

Referenzalignment:

- bestehend aus 1516 Korrepondenzen
- 946 davon triviale Korrespondenzen (gleiche normalisierte Strings)

Dateiformate (2)

mouse.owl (OWL: Ontology Web Language, JAVA: OWL-API)

```
<!-- http://mouse.owl#MA_0000200 -->
−<owl:Class rdf:about="http://mouse.owl#MA_0000200">
    <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">cerebellar hemisphere</rdfs:label>
    <rdfs:subClassOf rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
  −<rdfs:subClassOf>
    −<owl:Restriction>
        <owl:onProperty rdf:resource="http://mouse.owl#UNDEFINED_part_of"/>
        <owl:someValuesFrom rdf:resource="http://mouse.owl#MA_0000199"/>
      </owl:Restriction>
    </rdfs:subClassOf>
  </owl:Class>
```

(2744 Konzepte)

human.owl

```
<!-- http://human.owl#NCI_C32726 -->
−<owl:Class rdf:about="http://human.owl#NCI_C32726">
    <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Hemisphere_of_the_Cerebellum</rdfs:label>
    <rdfs:subClassOf rdf:resource="http://human.owl#NCI_C13031"/>
  −<rdfs:subClassOf>
    −<owl:Restriction>
        <owl:onProperty rdf:resource="http://human.owl#UNDEFINED_part_of"/>
        <owl:someValuesFrom rdf:resource="http://human.owl#NCI_C12445"/>
      </owl:Restriction>
    </rdfs:subClassOf>
    <oboInOwl:hasRelatedSynonym rdf:resource="http://human.owl#genid1825"/>
  </owl:Class>
```

(3304 Konzepte)

String Similarity Metrics for Ontology Alignment

- Eine gute Übersicht:
  http://disi.unitn.it/~p2p/RelatedWork/Matching/2015_12_3_957_964.pdf
- **Ähnlichkeit von Wörtern** (z.B. Levenshtein Distanz)
- **Ähnlichkeit von Wortmengen**
  (z.B. Jaccard-Index, TFIDF, Soft-TFIDF)
- **Ähnlichkeit von Konzepten:**
  Konzepte haben oft mehr als eine Bezeichnung (und damit mehr als eine Zeichenkette), Ähnlichkeitsmaße müssen also angepasst werden, z.B. Auswahl des Maximums aller paarweisen Vergleiche der Bezeichnungen

Tipps

- Strings normalisieren
  (Kleinbuchstaben, Entfernung von Bindestrichen …)

- Schritt für Schritt vorgehen, z.B.:
- erst exakte Korrespondenzen
- Auswahl eines 1:1 Alignments
  (bei einer 1:m Beziehung überlegen wie behandeln,
  z.B. Strukturinformationen miteinbeziehen, TFIDF Konfidenz)
- Approximatives Matching auf verbleibenden Konzepten

- Synonyme mit einbeziehen (soweit vorhanden)

Validierung der Alignments (1)

True Positive (TP) Anzahl der Korrespondenzen im berechneteten Alignment, die auch im Referenz Alignments vorkommen.

False Positive (FP) Anzahl der Korrespondenzen im berechneten Alignment, die nicht im Referenz Alignment vorkommen.

False Negatives (FN) Anzahl der Korrespondenzen im Referenz Alignment, die nicht im berechneten Alignment vorkommen.

Validierung der Alignments (2)

$$Precision = \frac{TP}{TP + FP}$$ (Genauigkeit des Alignments)

$$Recall = \frac{TP}{TP + FN}$$

(Trefferquote)

$$Recall+ = \frac{TP_{non-trivial}}{TP_{non-trivial} + FN_{non-trivial}}$$

(Recall abzüglich der 946 exakten Korrespondenzen)

$$F-Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Kriterien zum Bestehen

- F-Score von mindestens 80%
- Recall+ Wert von mindestens 10%

Ergebnisse der Challenge:

Matcher	Runtime	Size	Precision	F-Measure	Recall	Recall+	Coherent
AML	40	1477	0.956	0.944	0.931	0.82	X
COMMAND	63127*	150	0.293	0.053	0.029	0.042	X
CroMatcher	569	1350	0.914	0.861	0.814	0.508	-
DKP-AOM	370	201	0.995	0.233	0.132	0.0	X
DKP-AOM-lite	476	949	0.991	0.763	0.62	0.042	-
GMap	2362**	1344	0.916	0.861	0.812	0.534	-
JarvisOM	217	458	0.365	0.169	0.11	0.01	-
Lily	266	1382	0.87	0.83	0.793	0.513	-
LogMap	24	1397	0.918	0.88	0.846	0.593	X
LogMap-C	49	1084	0.966	0.805	0.691	0.449	X
LogMapBio	895	1549	0.882	0.891	0.901	0.738	X
LogMapLite	20	1147	0.962	0.828	0.728	0.288	-
RSDLWB	22	935	0.959	0.732	0.592	0.0	-
ServOMBI	792	971	0.963	0.752	0.617	0.099	-
YMAP	50	1414	0.928	0.896	0.865	0.647	X
StringEquiv	-	946	0.997	0.766	0.622	0.000	-

Triviale Korrespondenzen

Aufruf und Ausgabe

- Programm muss auf gruenau2 ausführbar sein
- Programm und Aufruf per Email senden
- Ausgabe: eine Korrespondenz des Alignments pro Zeile, comma-separated

```
MA_0002308, NCI_C52930
MA_0002373, NCI_C52991
MA_0001494, NCI_C33157
MA_0000355, NCI_C32421
MA_0002208, NCI_C48947
MA_0001296, NCI_C12346
MA_0000358, NCI_C12392
```

Wettbewerb

Die Gruppe, die den höchsten F-Score erreicht, gewinnt den Wettbewerb!

Abgabe

- Bis Montag, 18.07.2016, 23:59 Uhr

- Per Email an: yvonne.lichtblau@informatik.hu-berlin.de
  (gerne auch Fragen per Email)

- **Abgabe:**
- Programm (ausführbar auf gruenau2)
- Dokumentation (PDF) mit ermittelten Werten für
  Precision, Recall, Recall+ und F-Score

- **Kriterien zum Bestehen der Übung:**
  * F-Score von mindestens 80%
  * Recall+ Wert von mindestens 10%
  * Laufzeit < 20 Minuten

Abgabe

- Bis Montag, 18.07.2016, 23:59 Uhr

- Per Email an: yvonne.lichtblau@informatik.hu-berlin.de
  (gerne auch Fragen per Email)

- **Abgabe:**
- Programm (ausführbar auf gruenau2)
- Dokumentation (PDF) mit ermittelten Werten für
  Precision, Recall, Recall+ und F-Score

- **Kriterien zum Bestehen der Übung:**
  * F-Score von mindestens 80%
  * Recall+ Wert von mindestens 10%
  * Laufzeit < 20 Minuten