Technologie semantyczne i sieci społecznościowe

Projekt zaliczeniowy: Akwizycja danych semantycznych

Anna Cudzich
Sławomir Wałkowski
Bartosz Kosarzycki

Oryginalna treść specyfikacji:

Celem projektu jest przetłumaczenie danych semi-strukturalnych dostępnych w **Wikipedii** do postaci semantycznej. Zadanie polega na pobraniu kategorii American Inventors z wikipedii i przepisaniu zawartych tam danych do modelu **RDF** przy użyciu zaprojektowanej przez siebie ontologii. Ontologia powinna umożliwiać reprezentowanie podstawowych faktów (**X urodził się w Y**, X wynalazł Z, X studiował w M, X współpracował z N, X mieszkał w P, X działał w latach Q-R, itp.) Wynikiem projektu powinno być odnalezienie łańcuchów zdarzeń i własności wiążących te osoby (*vide*: **James Burke, "Connections"**)

Co zamierzamy zrobić?

• parsować tekst Wikipedii w poszukiwaniu fraz odpowiadającym danym trójkom RDFowym (np. X urodził się w Y)

• brać pod uwagę możliwość wystąpienia różnych określeń językowych na jedno pojęcie np.

- X urodził się w
- miastem rodzinnym X było
- X od urodzenia mieszkał w etc.

- <Poznań>
- <Poznan>
- <Posen>

• wytworzyć plik RDF z trójkami według specyfikacji:

RDF = Resource Description Framework
http://www.w3.org/RDF/

- **Zdania w postaci** trójek **(ang. *triples*):**
  <subject, predicate, object>

• umożliwić odczyt danych (stworzyć odpowiednią ontologię). Czyli wydawanie zapytań typu:
Podaj wszystkie informacje o X

• predefiniowane, spójne, pojedyncze określenia na dane połączenie – generowanie wyniku zapytania

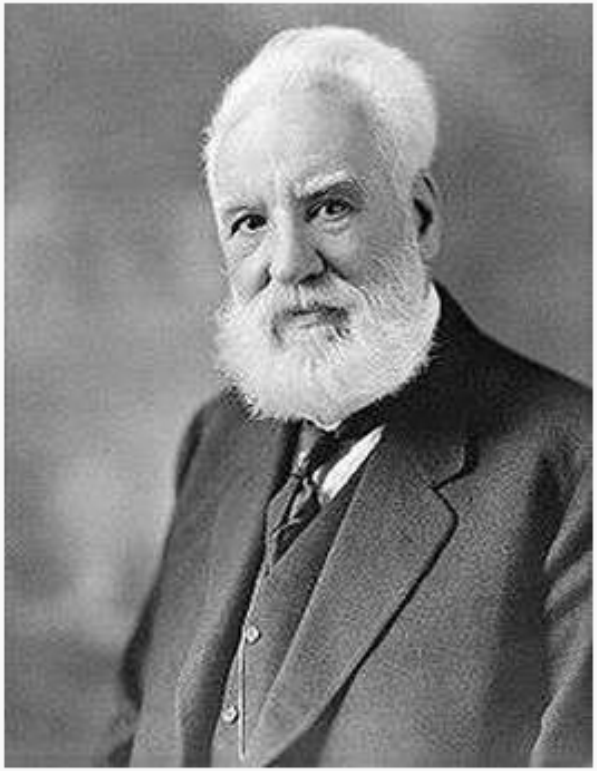
Pobieranie stron WWW:

- wykorzystanie programów „wget” bądź FlashGet

Parsowanie:

- Częściowy podział na tokeny (~słowa)
- Analiza współwystąpień charakterystycznych słów i wzorców
- Być może wykorzystanie wyrażeń regularnych
- wykorzystanie parsera HTML „HTMLParser”.

Alexander Graham Bell



Portrait of Alexander Graham Bell
c.1914–1919

Born	March 3, 1847 Edinburgh, Scotland, UK
Died	August 2, 1922 (aged 75) Beinn Bhreagh, Nova Scotia, Canada
Cause of death	Complications from Diabetes

Wykorzystanie danych częściowo ustrukturalizowanych:

- Tabelki w Wikipedii

Zapis wyników parsowania do pliku RDF (któraś z opcji):

- Wykorzystanie bibliotek do generowania RDF bezpośrednio z kodu, np. Jena (Java API)
- Ewentualnie: tworzenie pliku RDF zwykłymi bibliotekami XML

Przykładowy wynik parsowania (RDF file)

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:si="http://www.w3schools.com/rdf/">
    <rdf:Description rdf:about="http://www.w3schools.com">
      <si:subject>X</si:subject>
      <si:wasBornIn>Poznań</si:wasBornIn>
    </rdf:Description>
</rdf:RDF>
```

Wynik przykładowego zapytania?

- Describe X

W formie tekstowej:

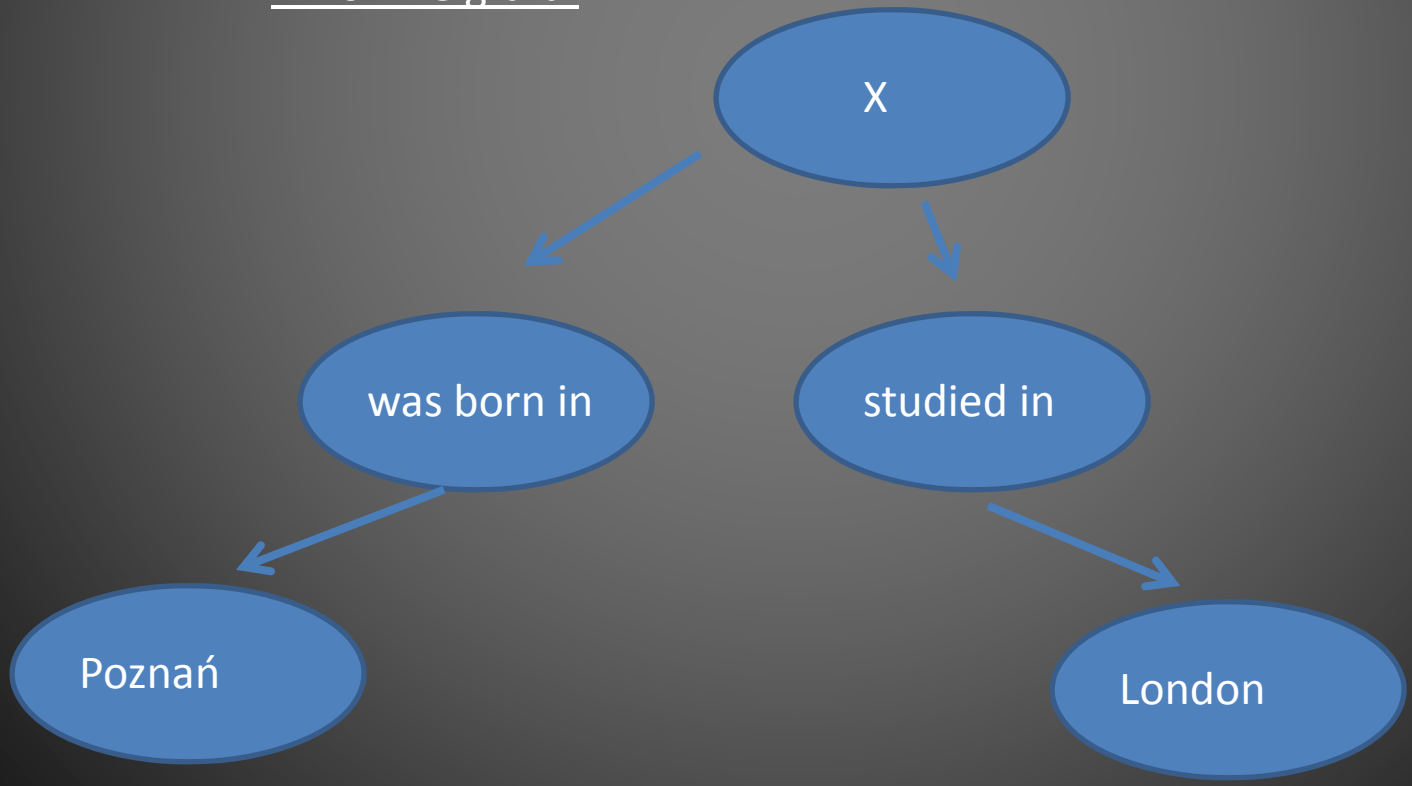
- X „was born in year” Y(year)
- X „was born in” Z(place)
- X „studied” M (major)

etc

W formie pliku RDF:

```
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:si="http://www.w3schools.com/rdf/">
  <rdf:Description rdf:about="http://www.w3schools.com">
    <si:subject>X</si:subject>
    <si:wasBornIn>Poznań</si:wasBornIn>
    <si:wasBornInYear>1987</si:wasBornInYear>
    <si:studiedIn>Poznań</si:studiedIn>
  </rdf:Description>
</rdf:RDF>
```

W formie grafu:



Zapytania do pliku RDF (któraś z opcji):

- SPARQL
- Silnik Oracle

Dziękujemy za uwagę!