Técnicas de Extracción de Información

José Ignacio Fernández Villamor

Grupo de Sistemas Inteligentes
Departamento de Ingeniería de Sistemas Telemáticos
Universidad Politécnica de Madrid
jifv@dit.upm.es

Marzo de 2012

Extracción de información

- Extraer datos de recursos no estructurados
  - Web hecha para humanos, sin APIs a veces
  - Solución: screen scraping.

```
<div class="apertura-a3 noticia2010 clearfix">
  <div class="doc">
    <div class="lead">
      <div class="overhead">
      <h3 class="headline">
          <a title="España acepta el ajuste de otros 5.000 millones
          que exige la UE" href="/20120313/economia/abci-espana-
          guindos-acepta-deficit-201203130942.html">España acepta el
          ajuste de otros 5.000 millones que exige la UE</a>
        </h3>
        <div class="subhead">El Eurogrupo ha fijado el objetivo de
        déficit en el 5,3%, a medio camino entre el 4,4% exigido
        inicialmente y el 5,8% propuesto por España</div>
      <div class="numcoment">
      <ul class="link-appl">
      </div>
```

Ejemplos


Rastreator.com
Tu Comparador de Seguros


acierto.com
el comparador líder en ahorro


eDreams
viajamos contigo


skyscanner


trivago


KAYAK

Screen Scraping

- Extracción de información de HTML plano
- Librerías:
  - **Java**: jSoup, WebHarvest
  - **JavaScript**: jQuery, jsdom
  - **Ruby**: Mechanize, Scrappy
  - **Python**: Mechanize, BeautifulSoup

JavaScrape:
http://javascrape.herokuapp.com

Proceso de scraping

- Abrir página HTML:
  - `get("http://www.marca.com", function(html){...});`
- "Parsear" HTML:
  - `var doc = $(html);`
- Buscar información:
  - `var datos = doc.find("h2")`
  - `var dato = datos.first().text();`
- Devolver la información:
  - `println(dato);`

Tareas

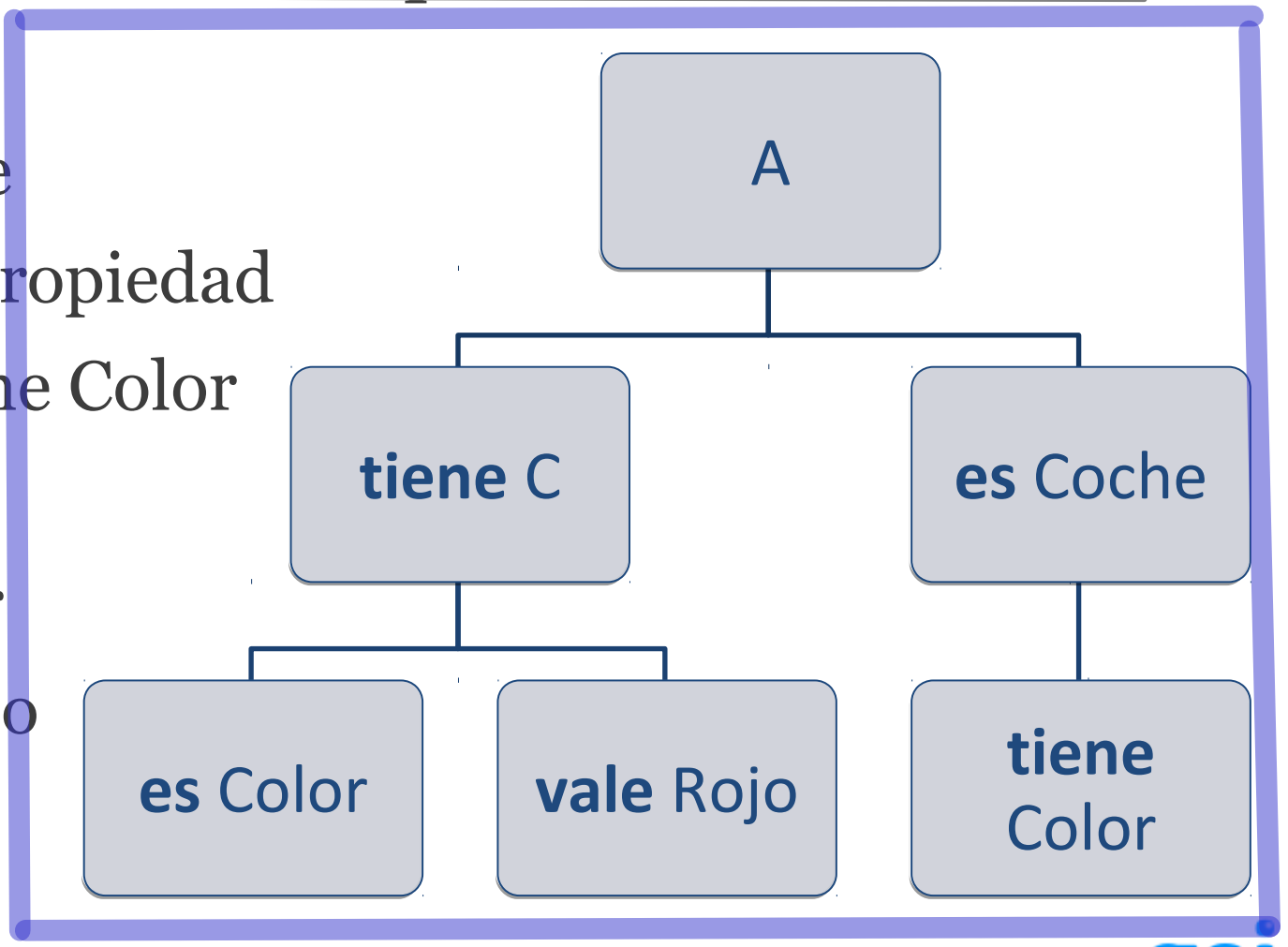
- http://javascrape.heroku.com
- Sacar los titulares de un periódico. Fácil
- Sacar el contenido de cada una de las noticias de la portada de un periódico. Medio
- Buscar la noticia más comentada de http://barrapunto.com. Difícil

Web Semántica

- Dotar de significado (semántica) a la información:
  - Cada sistema proporciona datos con significado
  - La información puede ser procesada automáticamente
  - Permite razonar sobre la información
  - Información extensible, interoperable
- Screen Scraping ya no es necesario

Tripletas

- Cualquier información se puede definir usando tripletas:
  - A es coche
  - Color es propiedad
  - Coche tiene Color
  - A tiene C
  - C es Color
  - C vale Rojo



Resource Description Framework

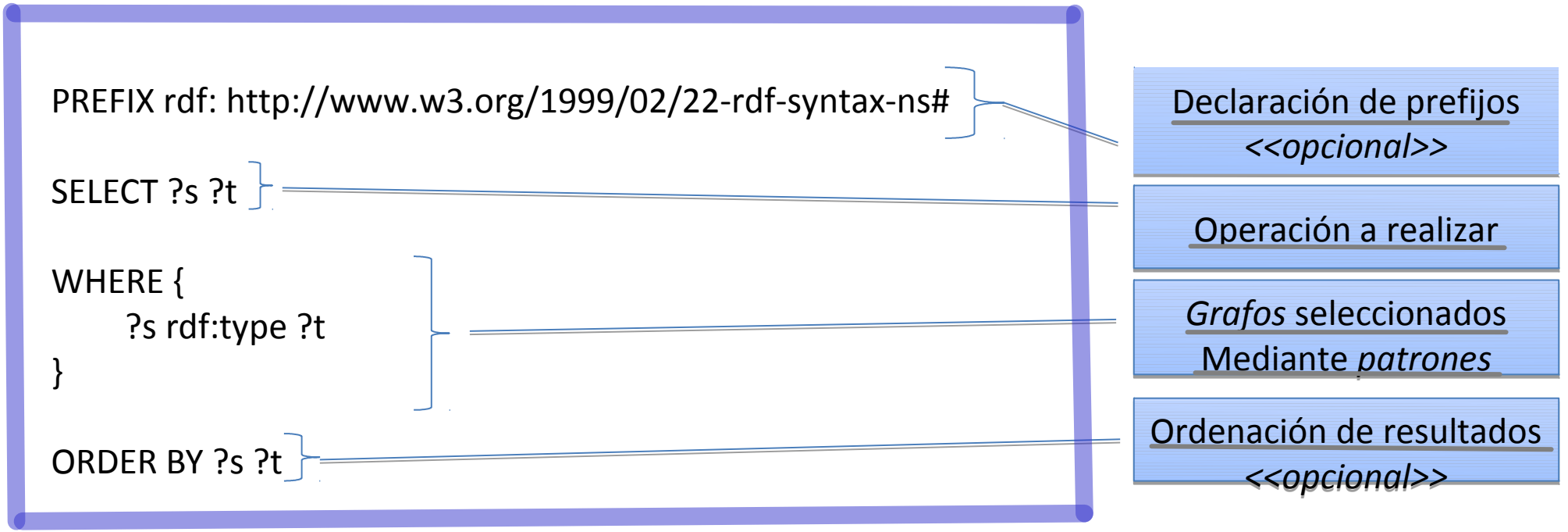
- Lenguaje de la Web Semántica
- Utiliza tripletas
- Cada elemento puede ser:
  - URI. Ej: http://elmundo.es
  - Literal (cadena de texto). Ej: “Hola a todos”
  - Nodo en blanco. Ej: _:bnode123
- Para abreviar URIs se usan prefijos:
  - http://dbpedia.org/property/derivatives
  - dbpprop:derivatives

Tripletas RDF

D dbpedia.org/page/The_Lord_of_the_Rings

	▪ dbpedia:The_Return_of_the_King ▪ 'Volumes:''
dbpprop:country	▪ United Kingdom
dbpprop:genre	▪ dbpedia:High_fantasy ▪ Adventure novel
dbpprop:imageCaption	▪ Tolkien's own cover designs for the three volumes
dbpprop:language	▪ English
dbpprop:mediaType	▪ Print
dbpprop:name	▪ The Lord of the Rings
dbpprop:pages	▪ 1216 (xsd:integer)
dbpprop:precededBy	▪ dbpedia:The_Hobbit
dbpprop:pubDate	▪ 1954 (xsd:integer)
dbpprop:publisher	▪ dbpedia:Allen_&_Unwin
dbpprop:wikiPageUsesTemplate	▪ dbpedia:Template:Infobox_book_series
dcterms:subject	▪ category:Fantasy_books_by_series ▪ category:1950s_fantasy_novels ▪ category:Sequel_novels ▪ category:The_Lord_of_the_Rings ▪ category:English_novels ▪ category:1954_novels ▪ category:Monomyths ▪ category:High_fantasy_novels ▪ category:Middle-earth_books

SPARQL (I)


PREFIX rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
SELECT ?s ?t
WHERE {
?s rdf:type ?t
}
ORDER BY ?s ?t
Declaración de prefijos
<<opcional>>
Operación a realizar
Grafos seleccionados
Mediante patrones
Ordenación de resultados
<<opcional>>

SPARQL (II)

```
?uri foaf:nick ?Nick.
?uri foaf:name ?Name.
```

Concatenación estándar

.

```
?uri foaf:name ?Name;
      foaf:nick ?Nick;
      foaf:img ?Foto;
```

Compartiendo Primer término

;

```
?c foaf:knows [
      foaf:nick ?bNick
].
```

Usando anónimamente un término

[ ]

```
{ ?a foaf:knows ?b }
UNION
{ ?b foaf:knows ?a }
```

Unión de grafos

{ } UNION {}

```
OPTIONAL {
   ?uri foaf:img ?Foto
}
```

Ejecución opcional

OPTIONAL { }

```
FILTER ( ?a != ?b )
```

Filtrado de soluciones

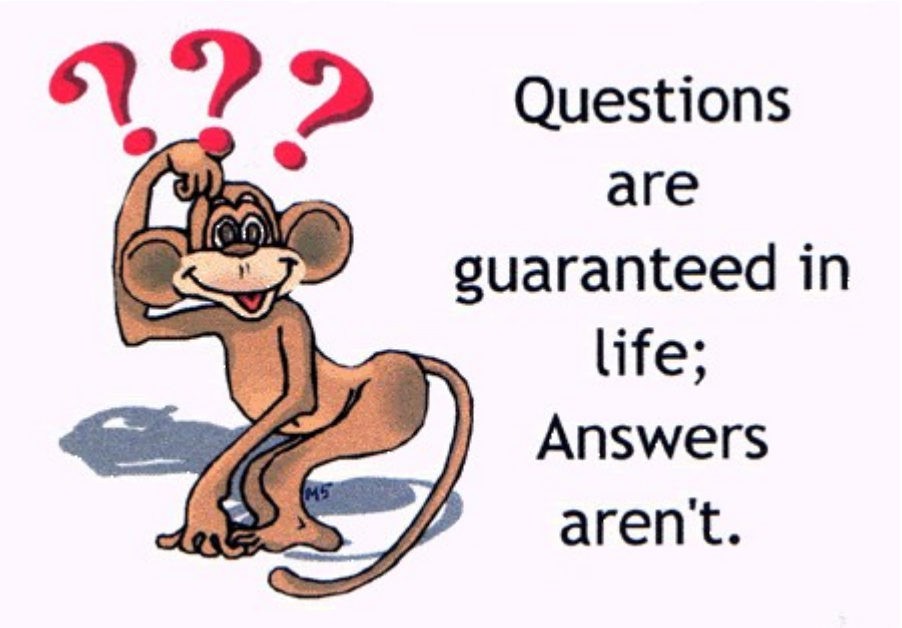
FILTER ( )

Los comentarios empiezan por

#

Tareas

- http://dbpedia.org/sparql
- Investigar tripletas en: http://dbpedia.org/page/XYZ
- ¿Qué libros ha escrito Shakespeare?
  - Usar dbpprop:author, dbpedia:William_Shakespeare
- ¿Cuáles son las capitales de Europa?
  - Investigar http://dbpedia.org/page/Madrid, por ejemplo
- ¿Qué tienen en común Cervantes y Shakespeare?
  - Usar dbpedia:Miguel_de_Cervantes, dbpedia:William_Shakespeare
- ¿Qué estilos de rock existen? Ojo a subestilos
  - dbpedia:Rock_music, dbpprop:derivatives



José Ignacio Fernández Villamor

jifv@dit.upm.es
Twitter: @cespino

(¡thx Jota por las transpas de SPARQL!)