Information Extraction

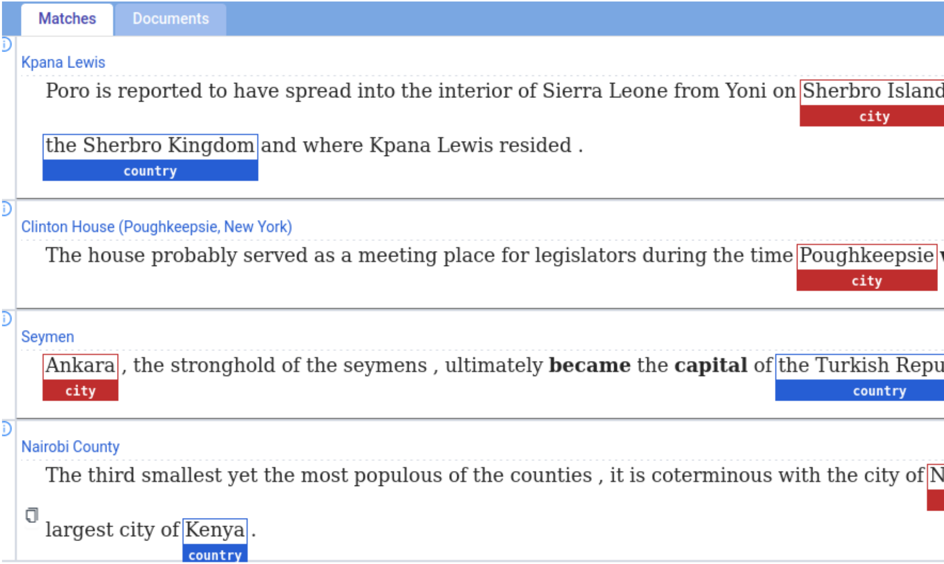
Geo-text data mining

Gosse Bouma

Information Science
Groningen University

May 2023

Information Extraction



Kpana Lewis

Poro is reported to have spread into the interior of Sierra Leone from Yoni on Sherbro Island city

the Sherbro Kingdom country and where Kpana Lewis resided .

Clinton House (Poughkeepsie, New York)

The house probably served as a meeting place for legislators during the time Poughkeepsie city

Seymen

Ankara city , the stronghold of the seymens , ultimately **became** the **capital** of the Turkish Repu country

Nairobi County

The third smallest yet the most populous of the counties , it is coterminous with the city of N

largest city of Kenya country .

Information Extraction

Task and Motivation

- Find relations between entities (and concepts) mentioned in a text
- For large-scale text-mining (scientific literature, news papers, social media)
  - Medical: Relations between diseases and medicines, medicines and symptoms, etc.
  - Political: relationships between politicians (political parties) and political issues
  - Geographical: relationships between events (fires, explosions) and locations
  - ...

Types of IE

- Find and classify relations between concepts and named entities in a text

Supervised IE

- The set of relations of interest is predefined (*capital-of, inhabitants, subdivision-of, located-in, borders*)
- Matching (Syntactic) Patterns can be *learned* from annotated training data

Open IE

- the set of relations is not predefined
- **Distant supervision**: From Wikipedia infoboxes to matching patterns in text on corresponding page

Syntactic patterns for IE

Spike IE system

IE often involves

- finding a verb or noun expressing the relation
- finding entities in some syntactic relation to the verb or noun (subject, object, prepositional phrase)
- Syntactic dependency analysis as in spaCy helps finding such patterns
- Writing patterns is complicated
- Spike supports learning patterns from an annotated example

Assignments

Getting started

- Go to https://spike.apps.allenai.org/datasets/wikipedia/searchwelcome
- Go to **structure search**, and browse the **tutorial**

Assignment 1

- Formulate a query to find **locations that border each other**
- Inspect the results
- Revise the query to make it more accurate (add filters for entities, add notation to match with complex names (*Luang Namtha Province*)
- Revise the query to make it find more instances (i.e. *borders, bordering) (located near/next to/east of/...*)

Assignments

Assignment 2

- Formulate a query to find **location (city) of museums**
- Inspect the results
- Optionally revise the query to make it more accurate and to improve coverage

Assignment 3

- Formulate a query to find **the highest point of a country or region**
- The highest point in a country is usually a mountain. Can you modify the query to find both sentences with the word *mountain* or the word *point*?