

Information Extraction

Geo-text data mining

Gosse Bouma

Information Science
Groningen University

May 2023

Information Extraction

Matches	Documents	Download Results
<p>Kpana Lewis</p>	<p>Poro is reported to have spread into the interior of Sierra Leone from Yoni on Sherbro Island, which was the capital of the Sherbro Kingdom and where Kpana Lewis resided .</p>	
<p>Clinton House (Poughkeepsie, New York)</p>	<p>The house probably served as a meeting place for legislators during the time Poughkeepsie was capital of New York in 1777 .</p>	
<p>Seymen</p>	<p>Ankara, the stronghold of the seymens , ultimately became the capital of the Turkish Republic in 1923 .</p>	
<p>Nairobi County</p>	<p>The third smallest yet the most populous of the counties , it is coterminous with the city of Nairobi, which is also the capital and largest city of Kenya .</p>	

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/search`
- 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*?

ChatGPT

Become a Prompt Engineer



Geocoding with chatGPT

G.

Geocoding text is the process of detecting place names and other geographical locations in text and assigning specific geographic coordinates to them or a specific id in a database such as geonames. Can you identify the geographical entities in the following text:



"Groningen was established more than 950 years ago and gained city rights in 1245.[7] Due to its relatively isolated location from the then successive Dutch centres of power (Utrecht, The Hague, Brussels), Groningen was historically reliant on itself and nearby regions. As a Hanseatic city, it was part of the North German trade network, but later it mainly became a regional market centre. At the height of its power in the 15th century, Groningen could be considered an independent city-state and it remained autonomous until the French era.[8]

Geocoding with chatGPT



Sure! Here are the geographical entities that I could identify in the given text:

- Groningen: a city in the northern part of the Netherlands
- Utrecht: a city in the central part of the Netherlands
- The Hague: a city on the western coast of the Netherlands and the seat of the Dutch government
- Brussels: the capital city of Belgium and the European Union
- North German: referring to the region in the northern part of Germany
- 15th century: a time period in history, roughly from 1401 to 1500
- French era: a historical period in the Netherlands during which it was under French rule, roughly from 1795 to 1813.

Geocoding with chatGPT

G.

What are the geographical coordinates of Groningen?



The geographical coordinates of Groningen are approximately 53.2194° N, 6.5665° E.

G.

Which geographic coordinate system are you using?



As an AI language model, I do not have a specific default geographic coordinate system that I use. However, the most commonly used coordinate system for representing the location of places on Earth is the WGS84 (World Geodetic System 1984) coordinate system, which is based on latitude and longitude measurements. The coordinates I provided for Groningen in my previous response are based on the WGS84 coordinate system.

Hands-on

Geocoding

- Test chatGPT's capacity for identifying geographical names in text with some prompt and input of your own
- Find a text in which an ambiguous geo name occurs. Think of a prompt that would test whether chatGPT identifies the correct reference location
- Can we use chatGPT for (geographical) information extraction?

spacy-llm and LLM

spacy-llm: Integrating LLMs into structured NLP pipelines



build failing pypi v0.1.2 code style black

This package integrates Large Language Models (LLMs) into [spacy](#), featuring a modular system for **fast prototyping** and **prompting**, and turning unstructured responses into **robust outputs** for various NLP tasks, **no training data** required.

- Serializable `llm` **component** to integrate prompts into your pipeline
- **Modular functions** to define the **task** (prompting and parsing) and **backend** (model to use)
- Support for **hosted APIs** and self-hosted **open-source models**
- Integration with `MiniChain` and `LangChain`
- Access to **OpenAI API**, including GPT-4 and various GPT-3 models
- Built-in support for **open-source Dolly** models hosted on Hugging Face
- Usage examples for **Named Entity Recognition** and **Text Classification**
- Easy implementation of **your own functions** via [spacy's registry](#) for custom prompting, parsing and model integrations

Zero-shot Classification with spaCy-llm and GPT-3

config.cfg:

```
[nlp]
lang = "en"
pipeline = ["llm"]

[components]

[components.llm]
factory = "llm"

[components.llm.task]
@llm_tasks = "spacy.TextCat.v1"
labels = COPIES_SOLD, COUNTRY_OF_ORIGIN

[components.llm.backend]
@llm_backends = "spacy.REST.v1"
api = "OpenAI"
config = {"model": "gpt-3.5-turbo", "temperature": 0.3}
```

Zero-shot Classification with GPT-3

```
from spacy import util

import os
import openai

openai.api_key = os.getenv("OPENAI_API_KEY")

config = util.load_config("config.cfg")
nlp = util.load_model_from_config(config, auto_fill=True)

q = "From what country is Call of Duty?"

analysis = nlp(q),
print(analysis.cats)
```

Zero-shot Classification with GPT-3

What is the country of origin of the game World of Tanks?

```
{'COPIES_SOLD': 0.0, 'COUNTRY_OF_ORIGIN': 1.0}
```

In what country did The Witcher 3: Wild Hunt originate?

```
{'COPIES_SOLD': 0.0, 'COUNTRY_OF_ORIGIN': 1.0}
```

From what country is Call of Duty?

```
{'COPIES_SOLD': 0.0, 'COUNTRY_OF_ORIGIN': 1.0}
```

What country does Final Fantasy come from?

```
{'COPIES_SOLD': 0.0, 'COUNTRY_OF_ORIGIN': 1.0}
```

Which country is Fall Guys from?

```
{'COPIES_SOLD': 0.0, 'COUNTRY_OF_ORIGIN': 1.0}
```

What is Clash of Clans' country of origin?

```
{'COPIES_SOLD': 0.0, 'COUNTRY_OF_ORIGIN': 1.0}
```

How many copies has Minecraft sold?

```
{'COPIES_SOLD': 1.0, 'COUNTRY_OF_ORIGIN': 0.0}
```

How many times has the PlayStation 5 been sold?

```
{'COPIES_SOLD': 1.0, 'COUNTRY_OF_ORIGIN': 0.0}
```

How many units of Terraria were sold?

```
{'COPIES_SOLD': 1.0, 'COUNTRY_OF_ORIGIN': 0.0}
```

How many units did the Nintendo Wii sell?