Addressing Geographical Biases in Language Models: A Practical Tutorial

Rémy Decoupes and Maguelonne Teisseire

JRU TETIS - Montpellier - France
Territories, Environment, Remote Sensing, Spatial Information

A Joint Research Unit

based in the Remote Sensing Centre
Montpellier - France

A Joint Research Unit

Aims of this tutorial

- **Improve your awareness on biases related to geographical knowledge in LMs**
  - How to detect them?
  - How to evaluate them?
  - How to leverage them in various NLP tasks?

and enhance your critical thinking

Aims of this tutorial

- **Improve your awareness on biases related to geographical knowledge in LMs**
  - How to detect them?
  - How to evaluate them?
  - How to leverage them in various NLP tasks?

  > Not only on LM and LLMs

  > Not only on Spatial Information

and enhance your critical thinking

- **Two steps**
  - Overview of Concepts
  - Practical session with notebooks

- **Interactive process**
  - Running all the code together
  - Ask questions whenever you want

Outline

- **Part 1 - Introduction - Concept Definitions**
  - Large language Model
  - 5 Key Biases identified in NLP
  - Geographical Knowledge from Text
- **Part 2 - Experiments with LMs and LLMs**
  - The chosen LMs
  - Spatial representation in LLMs
- **Part 3 - How to Assess Disparities?**
  - Presentation of 4 Indicators
- **Part 4 - Practical Session**
  - Preliminaries
  - Steps to follow
  - Going further with new LLMs

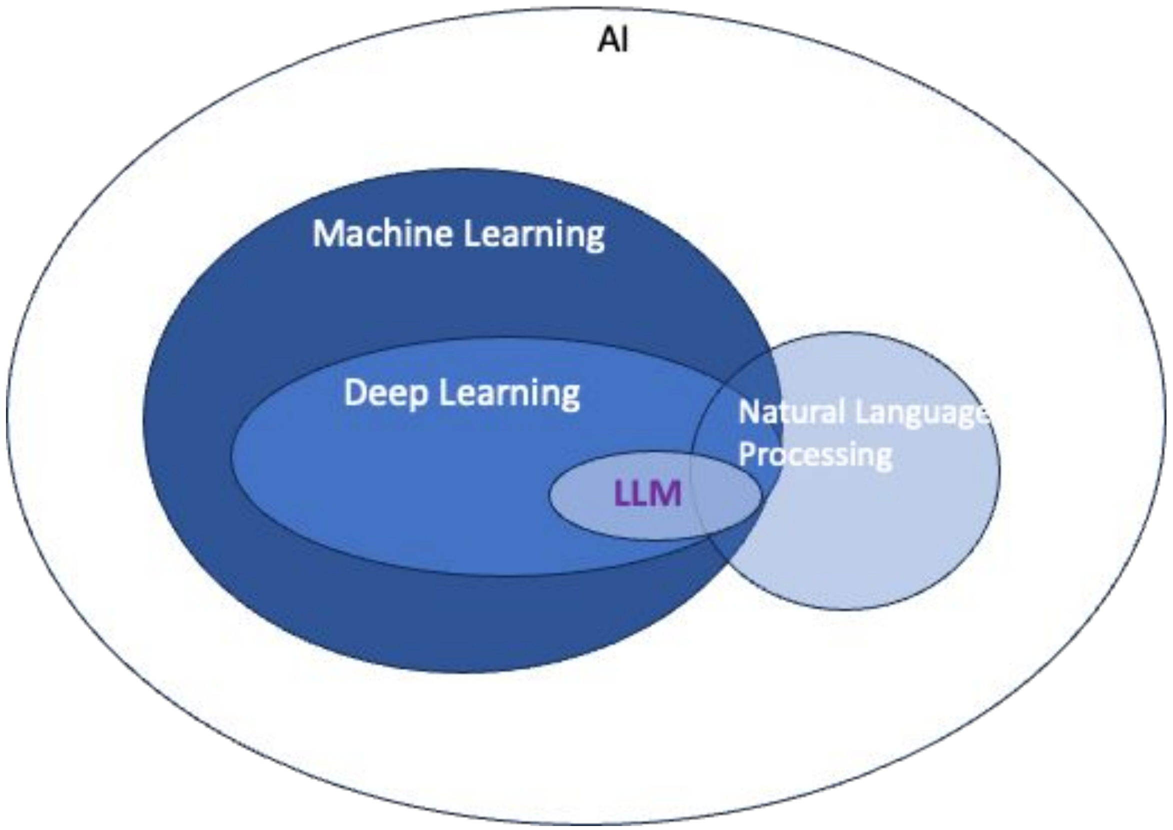
Outline

- **Part 1 - Introduction - Concept Definitions (MT)**
  - Large language Model
  - 5 Key Biases identified in NLP
  - Geographical Knowledge from Text
- **Part 2 - Experiments with LMs and LLMs (RD)**
  - The chosen LMs
  - Spatial representation in LLMs
- **Part 3 - How to Assess Disparities? (RD)**
  - Presentation of 4 Indicators
- **Part 4 - Practical Session (MT + RD)**
  - Preliminaries
  - Steps to follow
  - Going further with new LLMs

Outline

- Part 1 - Introduction - Concept Definitions
  - Large language Model
  - 5 Key Biases identified in NLP
  - Geographical Knowledge from Text
- **Part 2 - Experiments with LMs and LLMs**
  - The chosen LMs
  - Spatial representation in LLMs
- **Part 3 - How to Assess Disparities?**
  - Presentation of 4 Indicators
- **Part 4 - Practical Session**
  - Preliminaries
  - Steps to follow
  - Going further with new LLMs

Large language Model Overview



Large language Model Overview

- **Definition:**
  - LLMs are advanced artificial intelligence models **pre-trained** on extensive text datasets and **fine-tuned** for specific tasks
- **Training Process:**
  - They analyze vast amounts of text to understand statistical patterns between words, phrases, and sentences
- **Text Generation:**
  - Leveraging learned relationships, LLMs generate text resembling that of their training data

LLMs have the capability to produce human-like text output, revolutionizing various applications from natural language processing to content generation

Large language Model - Concept Overview

- **LLMs are commonly constructed using a transformer architecture**
  - Transformers are **a specific type of neural network** tailored for **natural language processing tasks** such as machine translation, text generation, and sentiment analysis

- **IA generative**

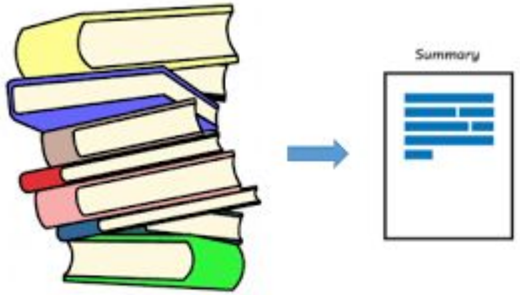
- **Attention mechanism**

- **Embedding**

(pre-training, fine tuning …)

Large language Model in Pictures

well-known NLP tasks solved by LMs:



Text Summarization

Text Classification

Question Answering

Large language Model in Pictures

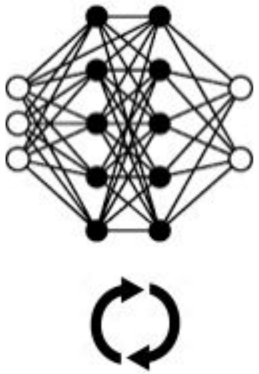
IA generative: new content generation

Large language Model in Pictures

IA generative: too talkative?

Large language Model in Pictures

Pre-training



Predict the next token

Part I - Introduction

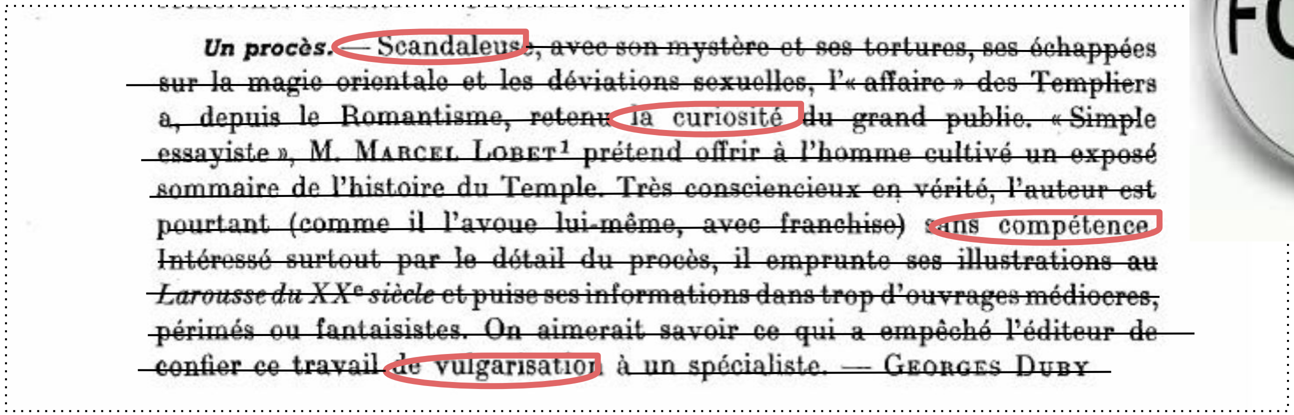
Large language Model in Pictures

Attention mechanism

Un procès. — Scandaleuse, avec son mystère et ses tortures, ses échappées sur la magie orientale et les déviations sexuelles, l'« affaire » des Templiers a, depuis le Romantisme, retenu la curiosité du grand public. « Simple essayiste », M. Marcel Lobet[1] prétend offrir à l'homme cultivé un exposé sommaire de l'histoire du Temple. Très consciencieux en vérité, l'auteur est pourtant (comme il l'avoue lui-même, avec franchise) sans compétence. Intéressé surtout par le détail du procès, il emprunte ses illustrations au *Larousse du XXe siècle* et puise ses informations dans trop d'ouvrages médiocres, périmés ou fantaisistes. On aimerait savoir ce qui a empêché l'éditeur de confier ce travail de vulgarisation à un spécialiste. — Georges Duby

Large language Model in Pictures

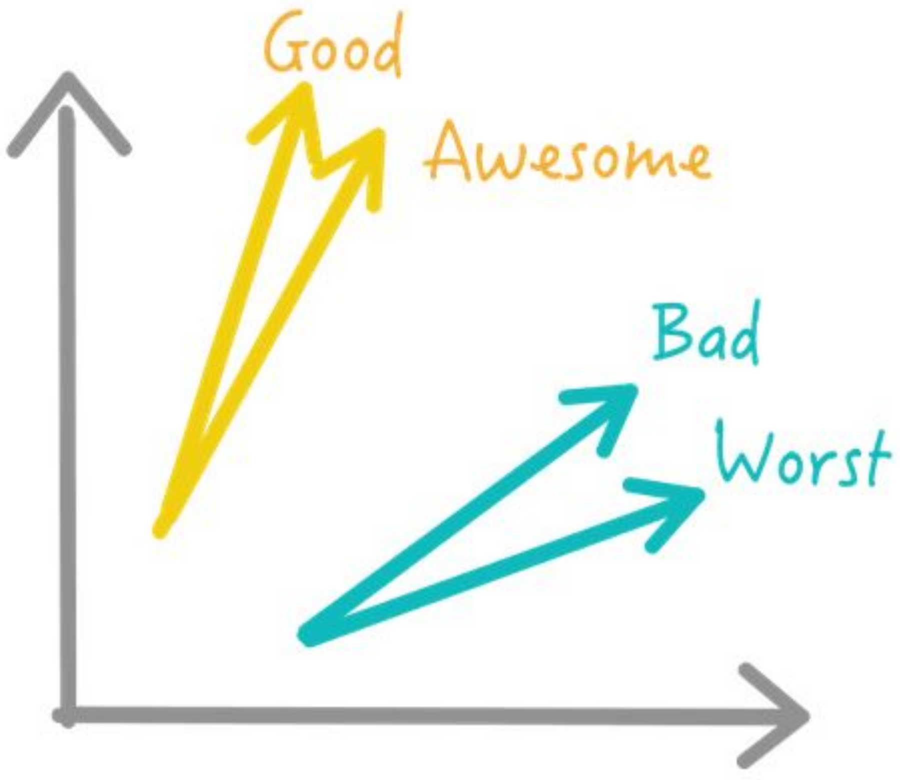
Attention mechanism



Un procès. — Scandaleuse, avec son mystère et ses tortures, ses échappées sur la magie orientale et les déviations sexuelles, l'« affaire » des Templiers a, depuis le Romantisme, retenu la curiosité du grand public. « Simple essayiste », M. Marcel Lobet[1] prétend offrir à l'homme cultivé un exposé sommaire de l'histoire du Temple. Très consciencieux en vérité, l'auteur est pourtant (comme il l'avoue lui-même, avec franchise) sans compétence. Intéressé surtout par le détail du procès, il emprunte ses illustrations au *Larousse du XXe siècle* et puise ses informations dans trop d'ouvrages médiocres, périmés ou fantaisistes. On aimerait savoir ce qui a empêché l'éditeur de confier ce travail de vulgarisation à un spécialiste. — Georges Duby

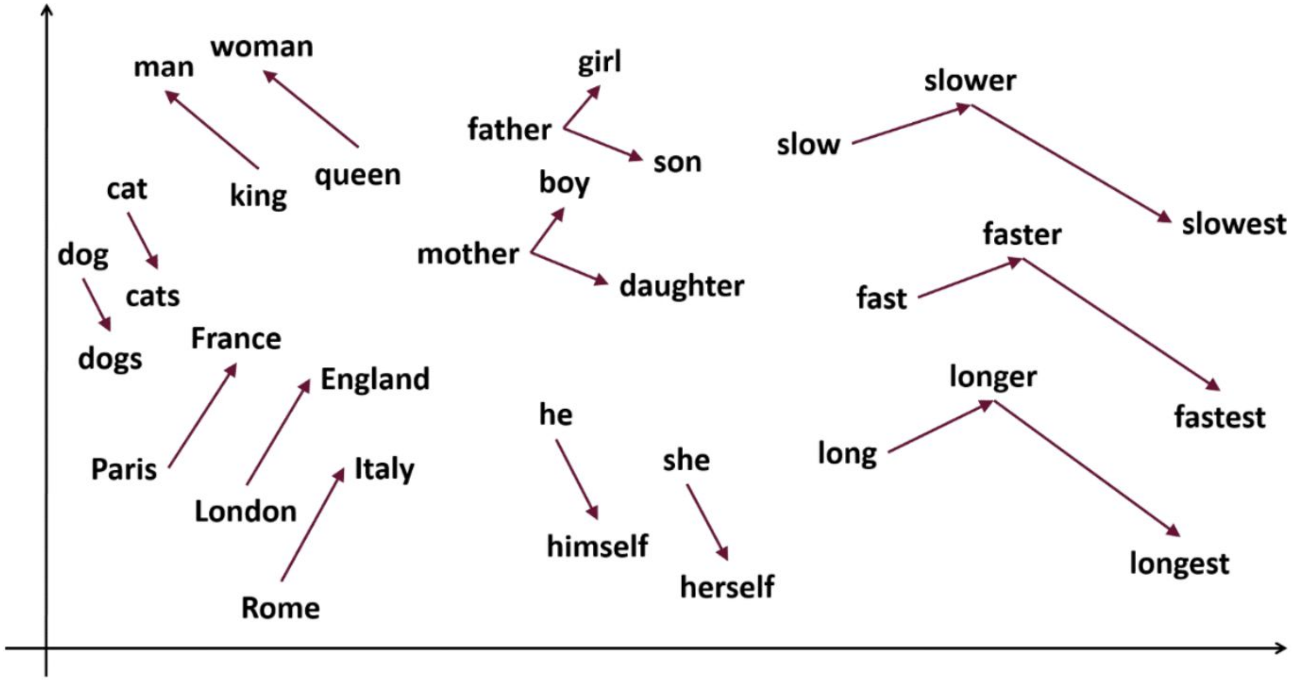
Large language Model in Pictures

Embedding



Large language Model in Pictures

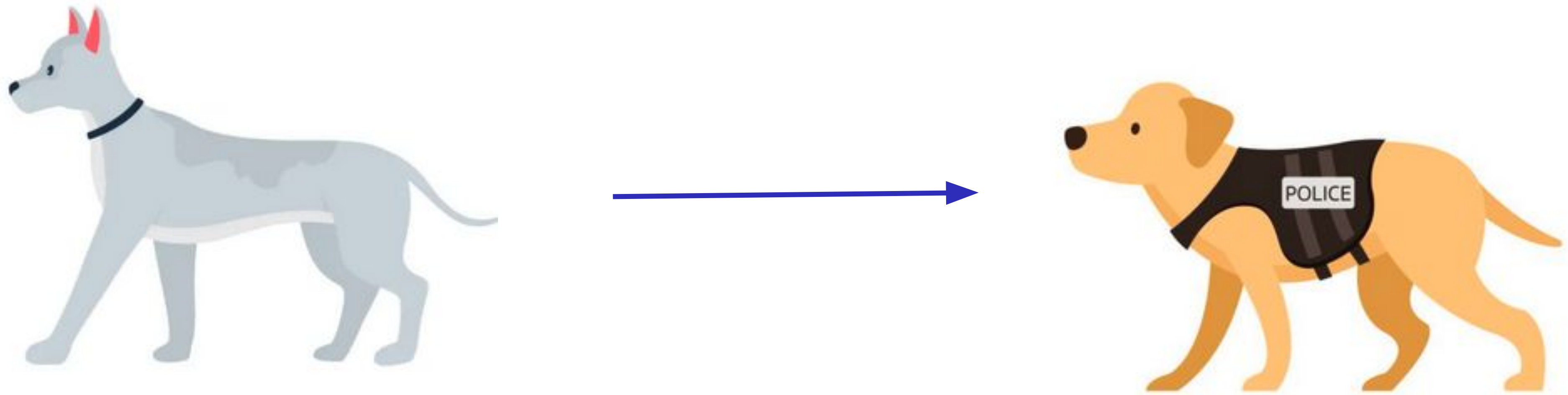
Embedding



2-dimensional representation of the results of a “word embeddings” algorithm

Large language Model in Pictures

Fine tuning



Part I - Introduction

5 Key Biases identified in NLP

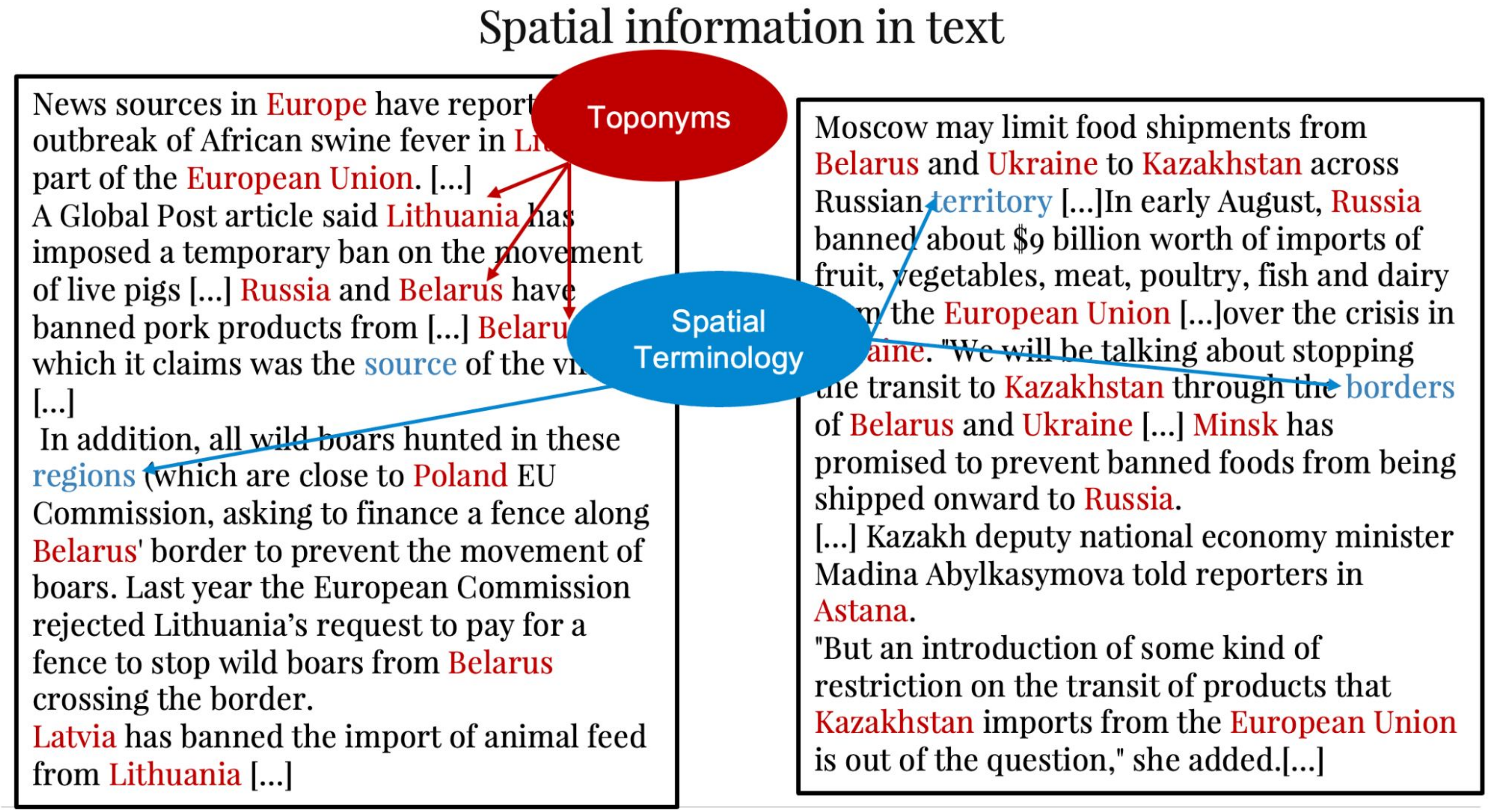
(1) the data
(2) the annotation process
(3) the input representations
(4) the models
(5) the research conceptualization

Know them to better manage them

[Five sources of bias in natural language processing - Hovy - 2021]

Part I - Introduction

Geographical Knowledge



Which Biases for Spatial Information?

(1) the data
(2) the annotation process
(3) the input representations
(4) the models
(5) the research conceptualization

Know them to better manage them

Plan of the tutorial

- **Part 1 - Introduction - Concept Definitions**
  - Large language Model
  - 5 Key Biaises identified in NLP
  - Geographical Knowledge from Text
- **Part 2 - Experiments with LMs and LLMs**
  - The chosen LMs
  - Spatial representation in LMs
- **Part 3 - How to Assess Disparities?**
  - Presentation of 4 Indicators
- **Part 4 - Practical Session**
  - Preliminaries
  - Steps to follow
  - Going further with new LLMs

Part 2 - Experiments with LLMs

The chosen LMs

3 kinds of Language Models:

1. Small Language Model (SLM) or Encoder-based LM
2. LLMs in local inference
3. LLMs through a remote API

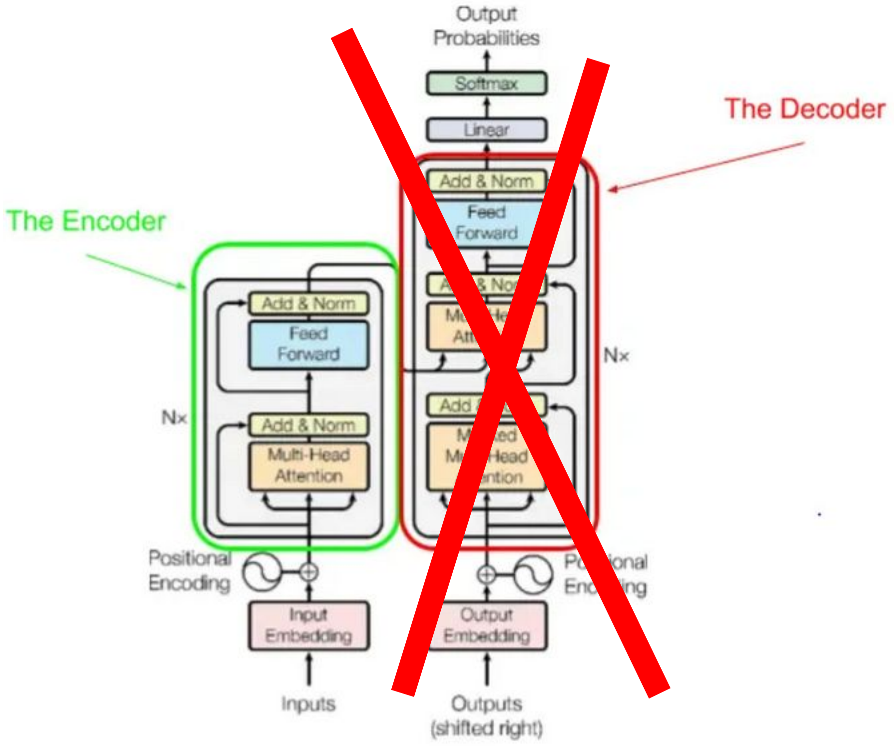
Part 2 - Experiments with LLMs

The chosen LMs

3 kinds of Language Models:

1. **Small Language Model (SLM) or Encoder-based LM**
   a. BERT and BERT multilingual
   b. RoBERTa and XML RoBERTa
2. **LLMs in local inference**

3. **LLMs through a remote API**

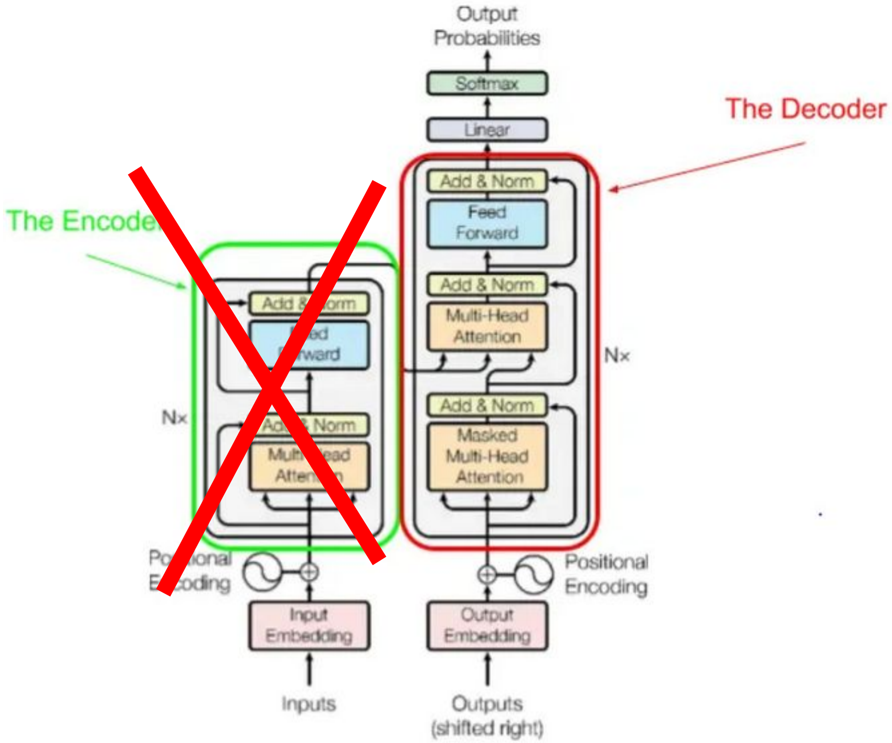


Attention is all you need - 2017

The chosen LMs

3 kinds of Language Models:

1. **Small Language Model (SLM) or Encoder-based LM**
   a. BERT and BERT multilingual
   b. RoBERTa and XML RoBERTa
2. **LLMs in local inference**
   a. Mistral-7B (instruct)
   b. Llama-2-7B (chat)
3. **LLMs through a remote API**



Attention is all you need - 2017

The chosen LMs

3 kinds of Language Models:

1. **Small Language Model (SLM) or Encoder-based LM**
   a. BERT and BERT multilingual
   b. RoBERTa and XML RoBERTa
2. **LLMs in local inference**
   a. Mistral-7B (instruct)
   b. Llama-2-7B (chat)
3. **LLMs through a remote API**
   a. GPT-3.5

Part 2 - Experiments with LLMs

The chosen LMs

3 kinds of Language Models:

1. **Small Language Model (SLM) or Encoder-based LM**
    a. BERT and BERT multilingual
    b. RoBERTa and XML RoBERTa
2. **LLMs in local inference**
    a. Mistral-7B (instruct)
    b. Llama-2-7B (chat)
3. **LLMs through a remote API**
    a. GPT-3.5

=> But we can go further and use new trending LLMs like Llama-3 or Phi-3 or others !

Spatial representation

The spatial dimensions of locations are encoded in the embeddings

Spatial representation

The spatial dimensions of locations are encoded in the embeddings

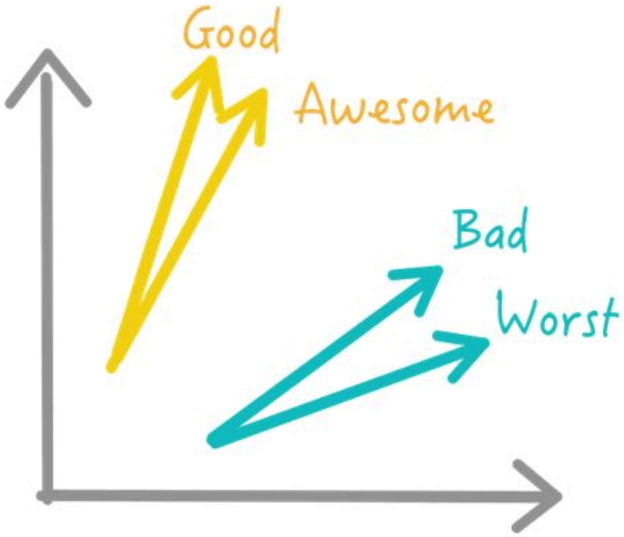
Maguelonne said:

Part 2 - Experiments with LLMs

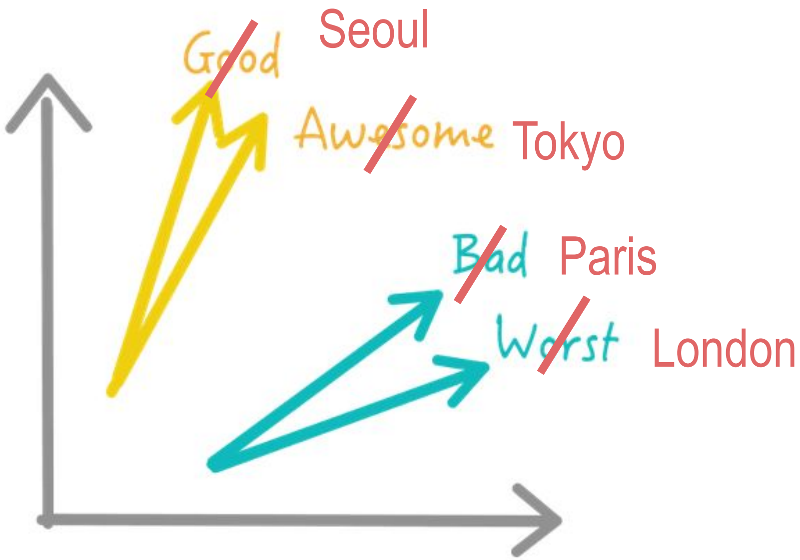
Spatial representation

The spatial dimensions of locations are encoded in the embeddings

Maguelonne said:



=> Could we see correlation between semantic and geographical distances ?



Outline

- **Part 1 - Introduction - Concept Definitions**
  - Large language Model
  - 5 Key Biaises identified in NLP
  - Geographical Knowledge from Text
- **Part 2 - Experiments with LMs and LLMs**
  - The chosen LMs
  - Spatial representation in LMs
- **Part 3 - How to Assess Disparities?**
  - Presentation of 4 Indicators
- **Part 4 - Practical Session**
  - Preliminaries
  - Steps to follow
  - Going further with new LLMs

Part 3 - How to Assess Disparities?

Presentation of the 4 Indicators

To assess Geographical Knowledge disparities worldwide:

1. **Evaluate geographical knowledge disparities worldwide**
2. **Assess indirectly the amount of geographical information used in their training**
3. **Is there a correlation between semantic distance and geographic distance ?**
4. **Do some countries exhibit semantic isolation?**

Part 3 - How to Assess Disparities?

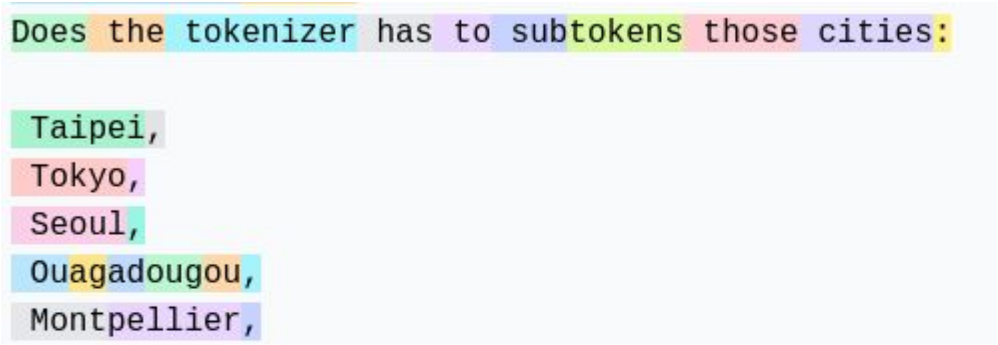
Presentation of the 4 Indicators

To assess Geographical Knowledge disparities worldwide:

1. **Evaluate geographical knowledge disparities worldwide**
   a. Could LLMs find the country when given its capital ?
2. **Assess indirectly the amount of geographical information used in their training**
3. **Is there a correlation between semantic distance and geographic distance ?**
4. **Do some countries exhibit semantic isolation?**

Part 3 - How to Assess Disparities?

Presentation of the 4 Indicators



https://tiktokenizer.vercel.app/

To assess Geographical Knowledge disparities worldwide:

1. **Evaluate geographical knowledge disparities worldwide**
   a. Could LLMs find the country when given its capital ?
2. **Assess indirectly the amount of geographical information used in their training**
   a. How many capitals are in their tokenizer vocabulary ?
3. **Is there a correlation between semantic distance and geographic distance ?**
4. **Do some countries exhibit semantic isolation?**

Presentation of the 4 Indicators

To assess Geographical Knowledge disparities worldwide:

1. **Evaluate geographical knowledge disparities worldwide**
   a. Could LLMs find the country when given its capital ?
2. **Assess indirectly the amount of geographical information used i**
   a. How many capitals are in their tokenizer vocabulary ?
3. **Is there a correlation between semantic distance and geographic distance ?**
   a. Make a correlation plot between pair of cities
4. **Do some countries exhibit semantic isolation?**

Part 3 - How to Assess Disparities?

Presentation of the 4 Indicators

To assess Geographical Knowledge disparities worldwide:

1. **Evaluate geographical knowledge disparities worldwide**
   a. Could LLMs find the country when given its capital ?
2. **Assess indirectly the amount of geographical information used in their training**
   a. How many capitals are in their tokenizer vocabulary ?
3. **Is there a correlation between semantic distance and geographic distance ?**
   a. Make a correlation plot between pair of cities
4. **Do some countries exhibit semantic isolation?**
   a. Compare the average semantic distance between one capital to the others worldwide

Plan of the tutorial

- **Part 1 - Introduction - Concept Definitions**
  - Large language Model
  - 5 Key Biaises identified in NLP
  - Geographical Knowledge from Text
- **Part 2 - Experiments with LMs and LLMs**
  - The chosen LMs
  - Spatial representation in LLMs
- **Part 3 - How to Assess Disparities?**
  - Presentation of 4 Indicators
- **Part 4 - Practical Session**
  - Preliminaries
  - Steps to follow
  - Going further with new LLMs

Part 4 - Practical Session

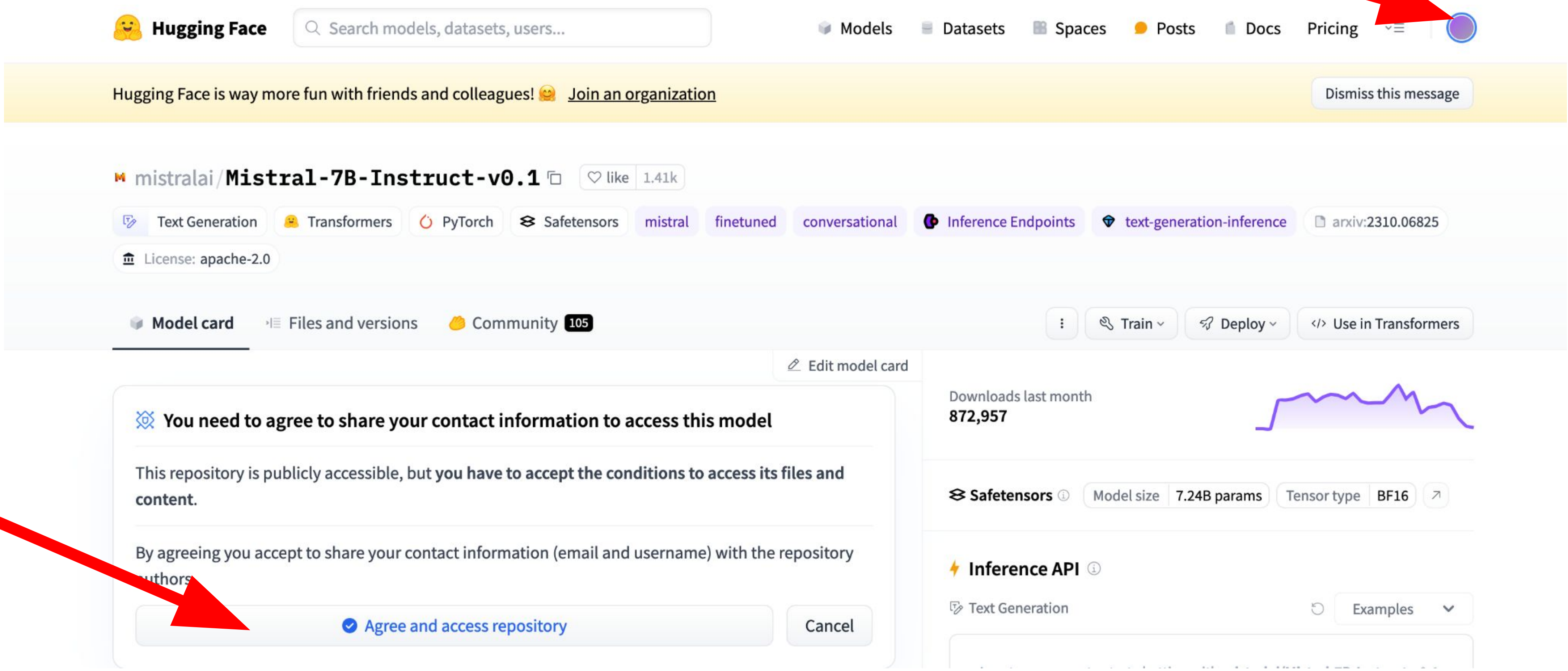
Preliminaries

- **create your account to obtain an API key for:**
  - Hugging face
    https://huggingface.co/
  - Open AI
    https://platform.openai.com/

Copy and keep the secret key value from OpenAI as you will not be able to access twice

Part 4 - Practical Session

To access repository for Mistral



Part 4 - Practical Session

Steps to follow

- Access to the github

  https://github.com/tetis-nlp/geographical-biases-in-llms

- For each step, wait for the guidelines, the explanation and the questions (in blue)

- Open experiments are suggested at the end of the provide notebook

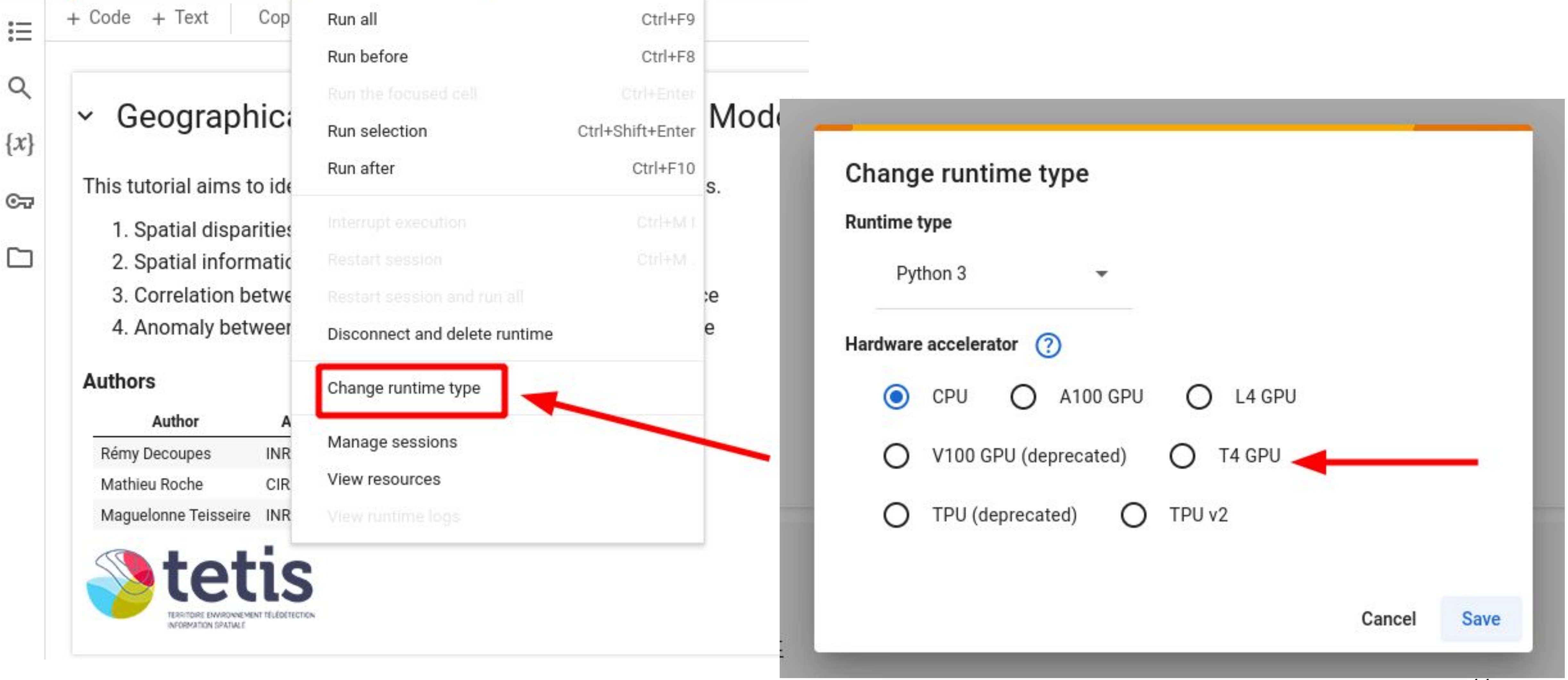
Never never use back in your google colab window

Part 4 - Practical Session

Step 1 - Up to 10 mn

1. Spatial disparities in geographical knowledge. Open in Colab
2. Spatial information coverage in training datasets. Open in Colab
3. Correlation between geographic distance and semantic distance. Open in Colab
4. Anomaly between geographical distance and semantic distance. Open in Colab

Part 4 - Practical Session



tetis

Part 4 - Practical Session

Step 1 - Continue

1. **Install Models and fill the parameters with your own API keys**

```
[ ] # Installation
    !pip install -U bitsandbytes
    !pip install transformers==4.37.2
    !pip install -U git+https://github.com/huggingface/peft.git
    !pip install -U git+https://github.com/huggingface/accelerate.git
    !pip install openai==0.28
```

Be patient

```
[ ] import getpass

    HF_API_TOKEN = getpass.getpass(prompt="Your huggingFace API Key")
    OPENAI_API_KEY = getpass.getpass(prompt="Your OpenAI API Key")
```

Part 4 - Practical Session

Step 1 - Continue

- **2 different types of language models are used:**
  Small Language Model (SLM) and Large Language Model (LLM)

  How many models per category are provided?

- **Install the geo datasets and associated libraries**

Geo datasets

Retrieve all country information needed through a python library `countryinfo`:

- Country Name
- Capital
- Region / subregion
- Coordinates

```
[ ]  !pip install countryinfo
```

...

Part 4 - Practical Session

Step 1 - Continue

`CountryInfo` **List of countries with their associated capitals**

access to the rows

How many countries are in the dataframe?

Now ready to start with the differents models and the ground truth

Step 1 - Continue

The questions are **Q1)** From which country Taipei is the capital?
Q2) Same question for all capitals of the data frame

- **Let's start with Small Language Model (SLM)**

1.1 SLMs

Roberta-base: **Q1 with Taipei**

Step 1 - Continue

The questions are **Q1)** From which country Taipei is the capital?
Q2) Same question for all capitals of the data frame

- **Let's start with Small Language Model (SLM)**

1.1 SLMs

Roberta-base: **Q2**

The ground truth

Predicted by the model

access to the rows for Q2

```
df_countries["predicted_country_from_capital"] = df_countries["Capital"].apply(self_masking).str.lower()
df_countries
```

	Country	Capital	Region	Subregion	Coordinates	predicted_country_from_capital
0	turkey	Ankara	Asia	Western Asia	POLYGON ((36.91313 41.33536, 38.34766 40.94859...	armenia
1	kuwait	Kuwait City	Asia	Western Asia	POLYGON ((47.97452 29.97582, 48.18319 29.53448...	kuwait
2	bermuda	Hamilton	Americas	Northern America	POLYGON ((-64.77997 32.30720, -64.78733 32.303...	canada
3	uzbekistan	Tashkent	Asia	Central Asia	POLYGON ((66.51861 37.36278, 66.54615 37.97469...	azerbaijan
4	guinea	Conakry	Africa	Western Africa	POLYGON ((-8.43930 7.68604, -8.72212 7.71167, ...	armenia

Have a quick look to correct and incorrect results

Part 4 - Practical Session

Step 1 - Continue

The questions are **Q1)** From which country Taipei is the capital?
Q2) Same question for all capitals in the data frame

Let's start with Small Language Model (SLM)

1.1 SLMs

- **Roberta-base**

See the % of good prediction per continent

```
df_countries[f"correct"] = df_countries['Country'] == df_countries[f"predicted_country_from_capital"]
```

```
df_countries.plot("correct", cmap="RdYlGn")
```

Plot the results

What is your feeling about correct and incorrect results?

Part 4 - Practical Session

Step 1 - Continue

The questions are **Q1)** From which country Taipei is the capital?
Q2) Same question for all capitals in the data frame

Let's move to Local LLMs

1.2 Local LLMs

Be patient

- **Mistral**
  - **Quantization phase**
  - **Q1** `city = "Tapei"`

Part 4 - Practical Session

Step 1 - Continue

The questions are **Q1)** From which country Taipei is the capital?
Q2) Same question for all capitals in the data frame

Let’s move to Local LLMs

1.2 Local LLMs

- **Mistral**
  - **Q2**
```
def self_masking(city):
```
    -
```
    messages = [
```

Be patient

What is your analysis about correct and incorrect results compare to SLM?

access to the rows

	Country	Capital	Region	Subregion	Coordinates	predicted_country_from_capital
0	suriname	Paramaribo	Americas	South America	POLYGON ((-57.14744 5.97315, -55.94932 5.77288...	the country corresponding to the capital param...
1	bahrain	Manama	Asia	Western Asia	None	bahrain
2	cuba	Havana	Americas	Caribbean	POLYGON ((-82.26815 23.18861, -81.40446 23.117...	cuba
3	senegal	Dakar	Africa	Western Africa	POLYGON ((-16.71373 13.59496, -17.12611 14.373...	senegal
				a	None	<s> [inst] name the country corresponding to i...
				..	...	...

and plot the results

Step 1 - Continue

The questions are Q1) From which country Taipei is the capital?
Q2) Same question for all capitals in the data frame

Last model with remote LLMs

1.3 Remote LLMs

- OpenAi
  - Q1
    ```
    Import openai
    openai.api_key = OPENAI_API_KEY
    city = "Taipei"
    ```
  - Q2
    ```
    def self_masking(city):
    ```
      - `messages = [`
  -

What are the specificities of remote LLMs?

Part 4 - Practical Session

Step 2 - Up to 15 mn

1. Spatial disparities in geographical knowledge. Open in Colab
2. Spatial information coverage in training datasets. Open in Colab
3. Correlation between geographic distance and semantic distance. Open in Colab
4. Anomaly between geographical distance and semantic distance. Open in Colab

Part 4 - Practical Session

Step 2 - Continue

- **New session: Install libraries (as part of Step 1)**

```
from transformers import AutoTokenizer …
```

and

```
import pandas as pd
```

Part 4 - Practical Session

Step 2 - Continue

- **The questions are** **Q1)** Is Taipei in the vocabulary of the model?
  Q2) Same question for all capitals of the data frame

Let's start with Small Language Model (SLM)

1.1 SLMs

- **Roberta-base**

```
model_name = "roberta-base"

tokenizer = AutoTokenizer.from_pretrained('roberta-base')

print(f"Size of {model_name} vocabulary: {len(tokenizer.get_vocab())}")
tokenizer.get_vocab()
```

Have a quick look to the Roberta vocabulary

How many words?
What do you observe on the cutted words?

Step 2 - Continue

The questions are

Q1) Is Taipei in the vocabulary of the model?
Q2) Same question for all capitals of the data frame

- **Roberta-base**

2.1.1 Example

Q1)

```
[7]  city = "Taipei"
     print(f"Is {city} (without uppercase) in vocab ?: {str.lower(city) in tokenizer.get_vocab() or str.lower('Ġ' + city) in tokenizer.get_vocab()}")
     print(f"Is {city} (with uppercase) in vocab ?: {city in tokenizer.get_vocab() or str('Ġ' + city) in tokenizer.get_vocab()}")
```

Is the result surprising?

Try with another capital (London)

Step 2 - Continue

The questions are

- **Roberta-base**

Q1) Is Taipei in the vocabulary of the model?
Q2) Same question for all capitals of the data frame

2.1.2 Worldwide

Q2)
```
df_countries["in_vocab"] = df_countries["Capital"].apply(in_vocab)
df_countries
```

Take time to explore the results

access to the rows

```
df_countries.plot("in_vocab", cmap="RdYlGn")
```

What is your analysis per continent?

Plot the results

Step 2 - Continue

Let's move to Local LLMs

2.2 Local LLMs

- **Mistral**

```
model_name = "mistralai/Mistral-7B-Instruct-v0.1"

tokenizer = AutoTokenizer.from_pretrained(model_name, token=HF_API_TOKEN)

print(f"Size of {model_name} vocabulary: {len(tokenizer.get_vocab())}")
tokenizer.get_vocab()
```

Have a quick look to the Mistral vocabulary

How many words?
What do you observe on the cutted words?
Compare to Roberta?

Step 2 - Continue

The questions are **Q1)** Is Taipei in the vocabulary of the model?
Q2) Same question for all capitals of the data frame

- **Mistral**

2.2.1 Example

Q1)

```
city = "Taipei"
print(f"Is {city} (without uppercase) in vocab ?: {str.lower(city) in tokenizer.get_vocab() or str.lower('Ġ' + city) in tokenizer.get_vocab()}")
print(f"Is {city} (with uppercase) in vocab ?: {city in tokenizer.get_vocab() or str('Ġ' + city) in tokenizer.get_vocab()}")
```

Is the result surprising?
Try with another capital (London)

tetis

Part 4 - Practical Session

Step 2 - Continue

The questions are

Q1) Is Taipei in the vocabulary of the model?
Q2) Same question for all capitals of the data frame

- **Mistral**

2.2.2 Worldwide

Q2)

```
df_countries["in_vocab"] = df_countries["Capital"].apply(in_vocab)
accuracy_by_continent = df_countries.groupby('Region')[f"in_vocab"].mean() * 100
accuracy_by_continent
```

access to the rows

Take time to explore the results

```
df_countries.plot("in_vocab", cmap="RdYlGn")
```

What is your analysis per continent? Compare to SLM?

Plot the results

Part 4 - Practical Session

Step 2 - Continue

The questions are

Q1) Is Taipei in the vocabulary of the model?
Q2) Same question for all capitals of the data frame

Last model with remote LLMs

2.3 Remote LLMs

```
!pip install tiktoken
!pip install openai
```

2.3.1 Example

Q1)

```
city = "Taipei"
tokenizer.encode(city)
```

What's happened?

Try with London, what is the difference? (in terms of number of token)

Step 2 - Continue

The questions are

Q1) Is Taipei in the vocabulary of the model?
Q2) Same question for all capitals of the data frame

Last model with remote LLMs

2.3 Remote LLMs

2.3.2 Worldwide

Q2)

```
df_countries["in_vocab"] = df_countries["Capital"].apply(in_vocab)
accuracy_by_continent = df_countries.groupby('Region')[f"in_vocab"].mean()  * 100
accuracy_by_continent
```

Take time to explore the results per continent

```
df_countries.plot("in_vocab", cmap="RdYlGn")
```

What is your analysis per continent? Compare to SLM and local LLM?

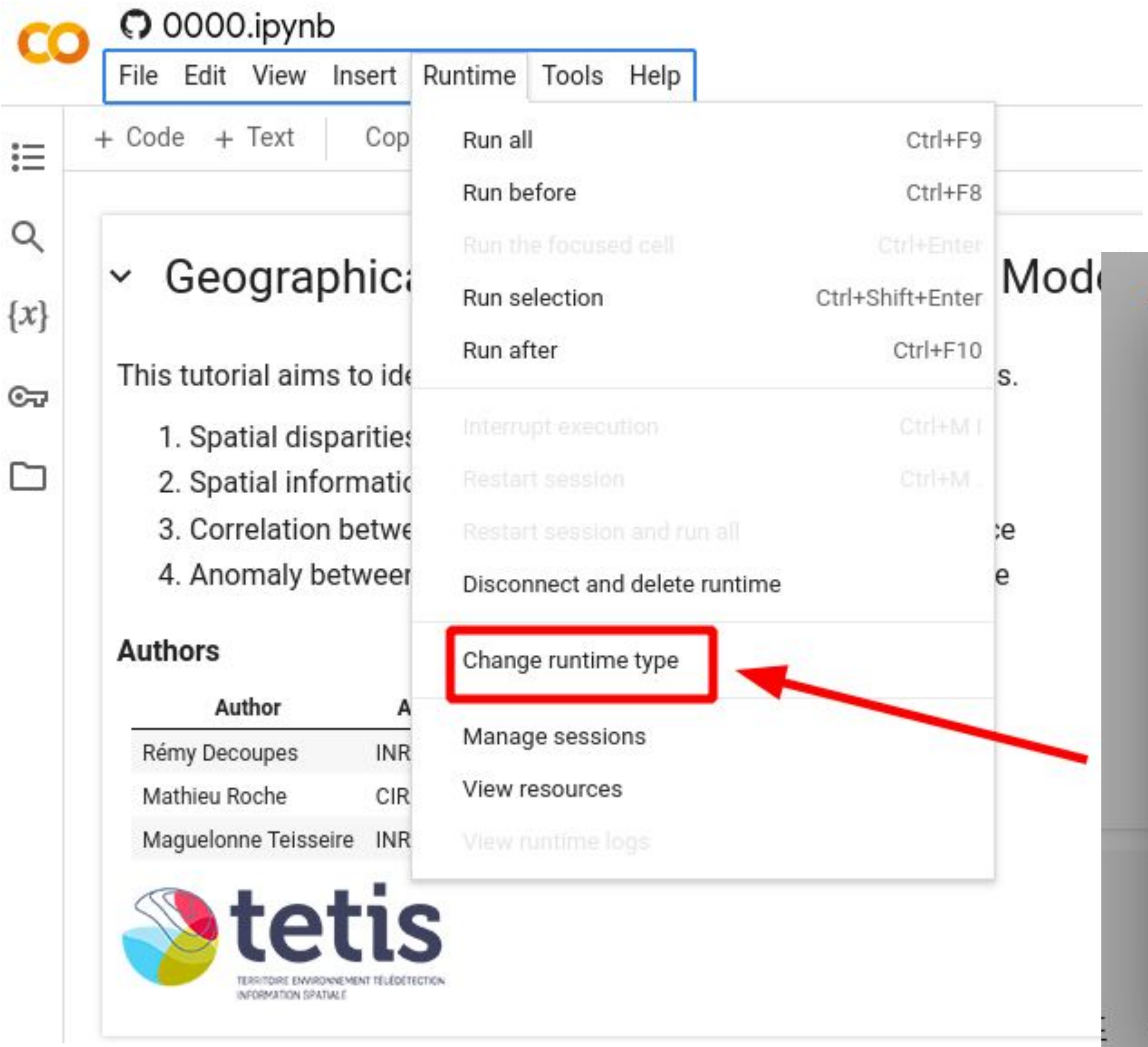
Plot the results

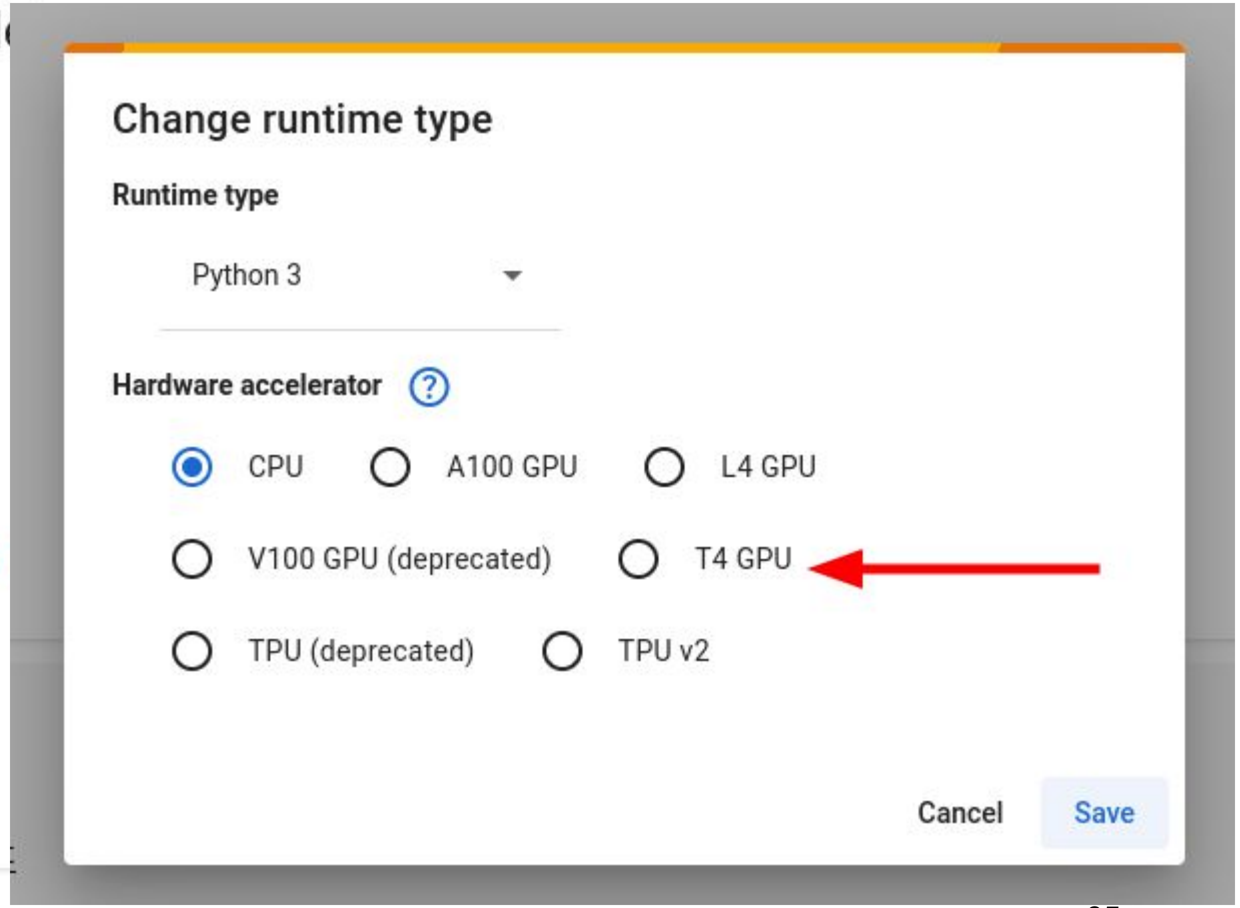
Part 4 - Practical Session

Step 3 - Up to 15 mn

1. Spatial disparities in geographical knowledge. Open in Colab
2. Spatial information coverage in training datasets. Open in Colab
3. Correlation between geographic distance and semantic distance. Open in Colab
4. Anomaly between geographical distance and semantic distance. Open in Colab

Part 4 - Practical Session





Step 3 - Continue

- **New session: Install libraries (as part of Step 1)**

```
!pip install transformers==4.37.2
```

```
from transformers import AutoTokenizer, AutoModelForCausalLM, AutoModel …
```

and Captials coordinates

```
from geopy.geocoders import Nominatim
from shapely.geometry import Point
```

Be patient

Step 3 - Continue

```
df_countries
```

	Country	Capital	Region	Subregion	Coordinates	capital_coordinates
0	east timor	Dili	Asia	South-Eastern Asia	POLYGON ((124.96868 -8.89279, 125.08625 -8.656...	POINT (125.57841 -8.55368)
1	tunisia	Tunis	Africa	Northern Africa	POLYGON ((9.48214 30.30756, 9.05560 32.10269, ...	POINT (10.18578 36.80021)
2	netherlands	Amsterdam	Europe	Western Europe	POLYGON ((6.07418 53.51040, 6.90514 53.48216, ...	POINT (4.89245 52.37308)

```
df_countries.plot(ax=ax, color="red")
```

See the plot of capitals

Step 3 - Continue

The questions are

Q1) Similarity distance of word embedding representation between Taipei and 5 other cities

Q2) Same question for all pairs of capital of the data frame

- **Roberta-base**

3.1.1 Example

```
def word_embedding(input_text):
 try:
```

See the embedding of Taipei

Q1)

```
from sklearn.metrics.pairwise import cosine_similarity
from geopy.distance import geodesic
```

Take time to compare the two measures (cosinus and geodistance), are they correlated?

Step 3 - Continue

The questions are

Q1) Similarity distance of word embedding representation between Taipei and 5 other cities
Q2) Same question for all pairs of capital of the data frame

- **Roberta-base**

3.1.2 Worldwide

Q2)

```
def compute_geo_distance(df):
coordinates = df["capital coordinates"].tolist()
num_city = len(coordinates)
```

Explore the plot, do you see any correlation?

Let's see per continent

Explore the plots, do you see better correlation?

Part 4 - Practical Session

Step 3 - Continue

Q1) Similarity distance of word embedding representation between Taipei and 5 other cities
Q2) Same question for all pairs of capital of the data frame

Same experiments with Local LLMs
3.2 Local LLMs

- **Mistral**

```
tokenizer = AutoTokenizer.from_pretrained("mistralai/Mistral-7B-Instruct-v0.1", token=HF_API_TOKEN)
model = AutoModel.from_pretrained("mistralai/Mistral-7B-Instruct-v0.1", token=HF_API_TOKEN)
```

Step 3 - Continue

Q1) Similarity distance of word embedding representation between Taipei and 5 other cities
Q2) Same question for all pairs of capital of the data frame

3.2 Local LLMs

3.2.1 Example

```
def word_embedding(input_text):
try:
    input_ids = tokenizer.encode(input_text, return_tensors="pt")
    with torch.no_grad():
```

See the embedding of Taipei

Q1)

```
def word_embedding(input_text):
 try:      input_ids = tokenizer.encode(input_text, return_tensors="pt") with torch.no_grad():...
```

Take time to see the embeddings and compare the cosinus similarity with SLM, what do you constat?

Have a look to the dimension of the embedding

Part 4 - Practical Session

Step 3 - Continue

Q1) Similarity distance of word embedding representation between Taipei and 5 other cities
Q2) Same question for all pairs of capital of the data frame

3.2 Local LLMs

3.2.2 Worldwide

Q2)
```
df_countries["capital_embedding_tensor"] = df_countries["Capital"].apply(word_embedding)
```

Let's see per continent

Explore the plots, do you see better correlations?

Part 4 - Practical Session

Step 3 - Continue

The questions are **Q1)** Similarity distance of word embedding representation between Taipei and 5 other cities

Q2) Same question for all pairs of capital of the data frame

Last model with remote LLMs

3.3 Remote LLMs

```
import openai from langchain.embeddings import OpenAIEmbeddings
```

Part 4 - Practical Session

Step 3 - Continue

The questions are **Q1)** Similarity distance of word embedding representation between Taipei and 5 other cities
Q2) Same question for all pairs of capital of the data frame

Last model with remote LLMs

3.3.1 Example

Q1)
```
def word_embedding(input_text):
 return np.array(model.embed_documents([input_text])[0])
```

Take time to see the embeddings

and compare the cosinus similarity with SLM and local LLM, what do you constat?

Have a look to the dimension of the embedding

Step 3 - Continue

The questions are **Q1)** Similarity distance of word embedding representation between Taipei and 5 other cities
Q2) Same question for all pairs of capital of the data frame

Last model with remote LLMs

3.3.2 Worldwide

Q2)

```
for region in df_countries["Region"].unique():
  print(region)
     df = df_countries[df_countries["Region"] == region]
```

Let's see per continent

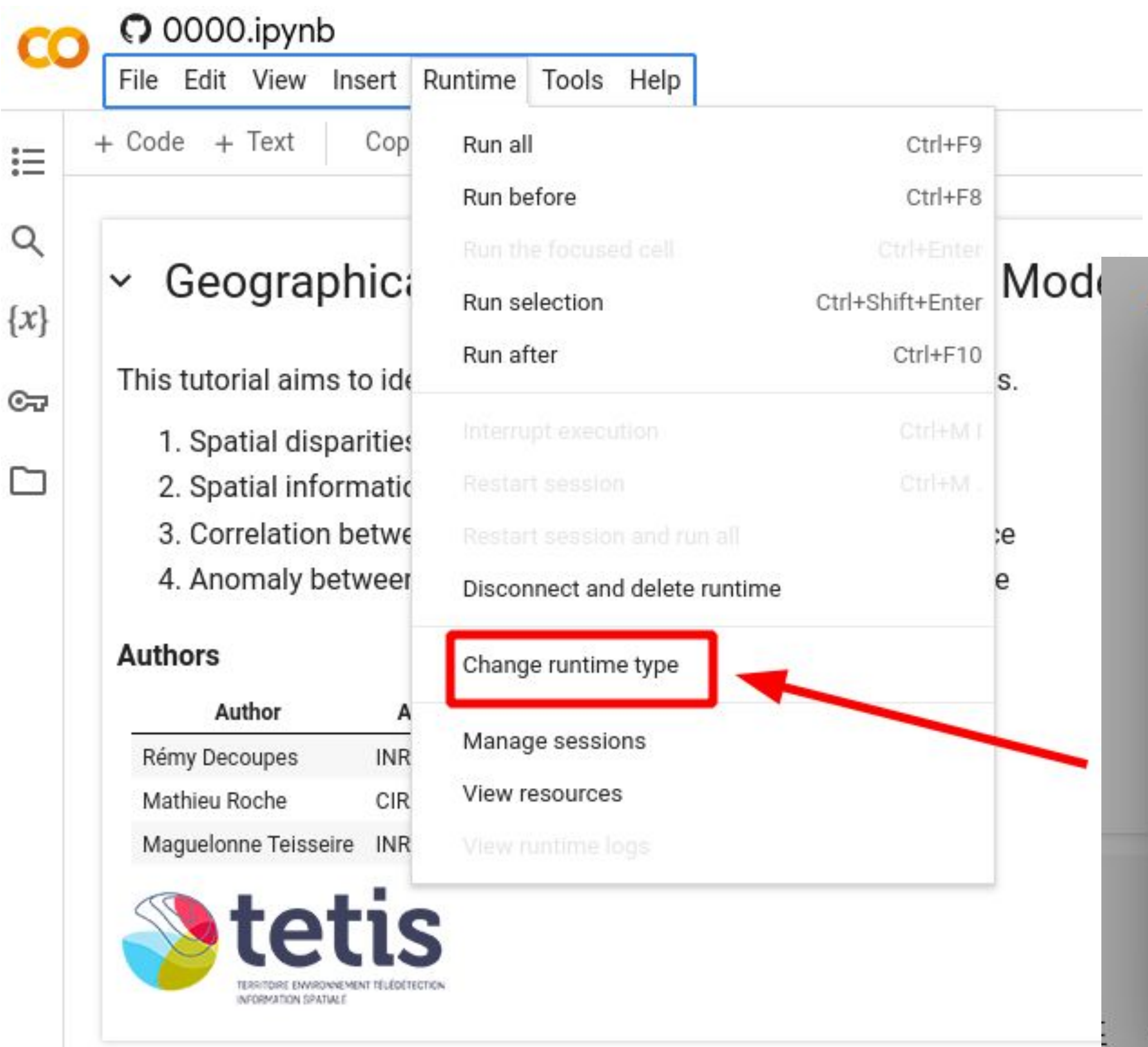
Explore the plots, do you see better correlations?

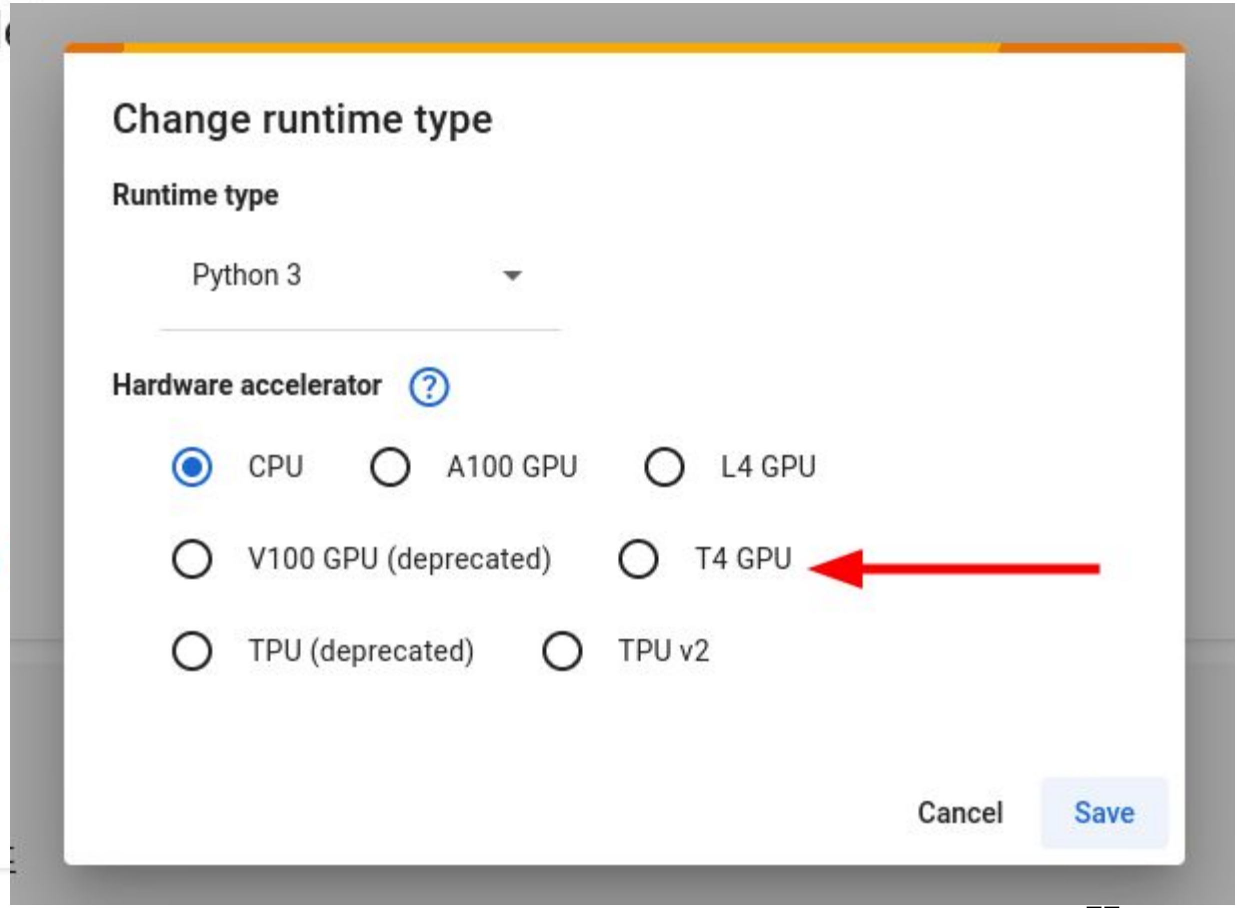
Part 4 - Practical Session

Step 4 - Up to 10 mn

1. Spatial disparities in geographical knowledge. Open in Colab
2. Spatial information coverage in training datasets. Open in Colab
3. Correlation between geographic distance and semantic distance. Open in Colab
4. Anomaly between geographical distance and semantic distance. Open in Colab

Part 4 - Practical Session





Part 4 - Practical Session

Step 4 - Continue

- **New session: Install libraries (as part of Step 1)** `!pip install countryinfo …`

Step 4 - Continue

The question is Q1) What the average semantic distance between one capital to the others worldwide

4.1 SLMs and Local LLMs

access to the rows

	average_semantic_distance	Country	Region	Subregion	Coordinates	capital_embedding_tensor	capital_embedding
	0.080545	macau	Asia	Eastern Asia	None	[tensor(-0.0539), tensor(0.0910), tensor(-0.01...	[-0.05388526, 0.09104285, -0.01820982, -0.1035...
	0.080545	heard island and mcdonald islands			None	[tensor(-0.0539), tensor(0.0910), tensor(-0.01...	[-0.05388526, 0.09104285, -0.01820982, -0.1035...
	0.080545	macau	Asia	Eastern Asia	None	[tensor(-0.0539), tensor(0.0910), tensor(-0.01...	[-0.05388526, 0.09104285, -0.01820982, -0.1035...
	0.080545	heard island and mcdonald islands			None	[tensor(-0.0539), tensor(0.0910), tensor(-0.01...	[-0.05388526, 0.09104285, -0.01820982, -0.1035...
Abu Dhabi	0.047251	united arab emirates	Asia	Western Asia	POLYGON ((51.57952 24.24550, 51.75744 24.29407...	[tensor(-0.1716), tensor(0.1237), tensor(0.033...	[-0.17156275, 0.12369239, 0.03390613, 0.091912...

Explore the cities and the associated average distance to the others

```
import matplotlib.pyplot as plt
```

See the plot

Step 4 - Continue

- **The question is** Q1) What the average semantic distance between one capital to the others worldwide

4.2 Remote LLMs

access to the rows

	average_semantic_distance	Country	Region	Subregion	Coordinates	capital_embedding_tensor	capital_embedding
	0.291749	macau	Asia	Eastern Asia	None	[tensor(-0.0539), tensor(0.0910), tensor(-0.01...	[0.0015624701186355266, -0.01700555921035143, ...
	0.291749	heard island and mcdonald islands			None	[tensor(-0.0539), tensor(0.0910), tensor(-0.01...	[0.0015624701186355266, -0.01700555921035143, ...
	0.291749	macau	Asia	Eastern Asia	None	[tensor(-0.0539), tensor(0.0910), tensor(-0.01...	[0.0015624701186355266, -0.01700555921035143, ...
	0.291749	heard island and mcdonald islands			None	[tensor(-0.0539), tensor(0.0910), tensor(-0.01...	[0.0015624701186355266, -0.01700555921035143, ...
Abu Dhabi	0.197620	united arab emirates	Asia	Western Asia	POLYGON ((51.57952 24.24550, 51.75744 24.29407	[tensor(-0.1716), tensor(0.1237), tensor(0.033...	[-0.006773835961625818, -0.0002240014294191519...

Explore the cities and the associated average distance to the others - compare with SLM and Local LLM

```
import matplotlib.pyplot as plt
```

See the plot compare to SLM and local LLM

tetis

For references - See https://arxiv.org/abs/2404.17401

arXiv > cs > arXiv:2404.17401

Search...

Help | Advanc

Computer Science > Computation and Language

[Submitted on 26 Apr 2024]

Evaluation of Geographical Distortions in Language Models: A Crucial Step Towards Equitable Representations

Rémy Decoupes, Roberto Interdonato, Mathieu Roche, Maguelonne Teisseire, Sarah Valentin

Language models now constitute essential tools for improving efficiency for many professional tasks such as writing, coding, or learning. For this reason, it is imperative to identify inherent biases. In the field of Natural Language Processing, five sources of bias are well-identified: data, annotation, representation, models, and research design. This study focuses on biases related to geographical knowledge. We explore the connection between geography and language models by highlighting their tendency to misrepresent spatial information, thus leading to distortions in the representation of geographical distances. This study introduces four indicators to assess these distortions, by comparing geographical and semantic distances. Experiments are conducted from these four indicators with ten widely used language models. Results underscore the critical necessity of inspecting and rectifying spatial biases in language models to ensure accurate and equitable representations.

Subjects: **Computation and Language (cs.CL)**

Cite as: arXiv:2404.17401 **[cs.CL]**

(or arXiv:2404.17401v1 **[cs.CL]** for this version)

https://doi.org/10.48550/arXiv.2404.17401

Outline

- **Part 1 - Introduction - Concept Definitions**
  - Large language Model
  - 5 Key Biases identified in NLP
  - Geographical Knowledge from Text
- **Part 2 - Experiments with LMs and LLMs**
  - The chosen LMs
  - Spatial representation in LLMs
- **Part 3 - How to Assess Disparities?**
  - Presentation of 4 Indicators
- **Part 4 - Practical Session**
  - Preliminaries
  - Steps to follow

Going further with new LLMs

tetis

Going further with new LLMs

1. **Predict Country from capital**
   a. Optimize the prompts for LLMs to make easy the parsing
   b. Propose other basic questions around the geography
2. **Vocabulary**
   a. Compare the proportions for subtokens between LLMs and SLMs
3. **Correlation Semantic - Geo**
   a. Clusterisation of countries based on their embedding
4. **Geo disparities**
   a. Data visualization: which cities are in the center semantic space

Addressing Geographical Biases in Language Models: A Practical Tutorial

This study was partially funded by EU grant 874850 MOOD and is catalogued as MOOD099.

Rémy Decoupes and Maguelonne Teisseire

JRU TETIS - Montpellier - France
Territories, Environment, Remote Sensing, Spatial Information