Problem Definition

Extract metadata related to

1. The Economic and Sector work conducted by the Bank
2. From January 1st, 2010 to July 1st 2024
3. Use documents written in English

Libraries

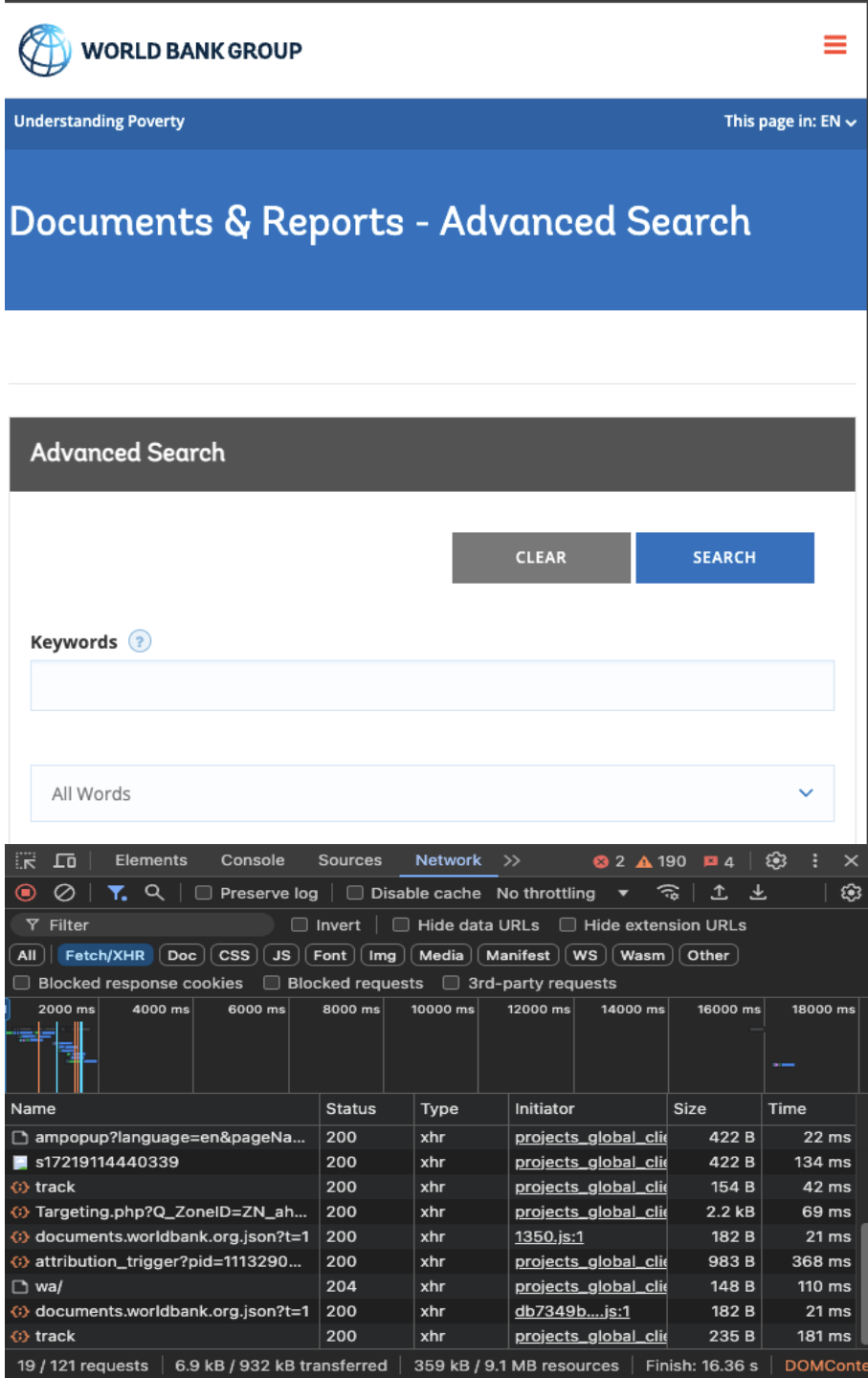
- **import requests** for sending HTTP requests
- **import pandas as pd** for data manipulation and analysis
- **import re** for handling regular expressions and string pattern matching
- **from tqdm import tqdm** to display progress bars for loops and iterations
- **import spacy** for Natural Language Processing (NLP)
- **import plotly.express as px** A high-level interface for interactive data visualizations
- **import matplotlib.pyplot as plt** for creating 2D plots and visualizations

APIs

```
api_endpoint = 'https://search.worldbank.org/api/v2/wds'

params = {
'format': 'json',
'str_docdt': '2010-01-01',
'end_docdt': '2024-07-01',
'majdocty_key': '906674',
'lang_key':'120701',
}
```

2 Date From
2 Date To
1 Economic and Sector Work (ESW)
3 English



Data Preparation

id	title	year
34366193	Suriname - Poverty and Equity Assessment	2024-07-01T00:00:00Z
34369413	Serbia Policy Notes 2024	2024-06-30T00:00:00Z
34351433	Maldives - Country Climate and Development Report	2024-06-30T00:00:00Z
34354089	Microfinance in the Kyrgyz Republic Evaluation Study	2024-06-28T00:00:00Z
34354258	Operational Note - How to use the Analytical Tools	2024-06-28T00:00:00Z

Data Pre-Processing

Better

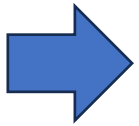
- **Exclude missing values:** found 6 titles = None
- **Remove unnecessary spaces and carriage returns:** double space & '\n'

Worse

- **Make lower cases:** double space & '\n'
- **Filter non-alphanumeric characters: '-', ':', etc.**

Title - Example

Maldives - Country Climate and\n De...



Maldives - Country Climate and Developmen...

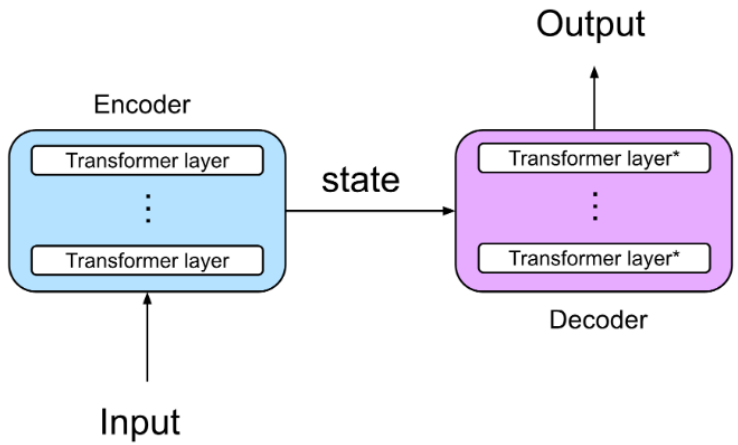
Training

spaCy

Models

- **en_core_web_sm**
  smaller & lighter & faster

- **en_core_web_trf**
  bigger & heavier & slower
  transfermer-based model



Labels

LOC
Location

+

GPE
Geopolitical
Entity

Prediction Rate

SpaCy - en_core_web_sm

52%

1522 out of 2920

SpaCy - en_core_web_trf

77%

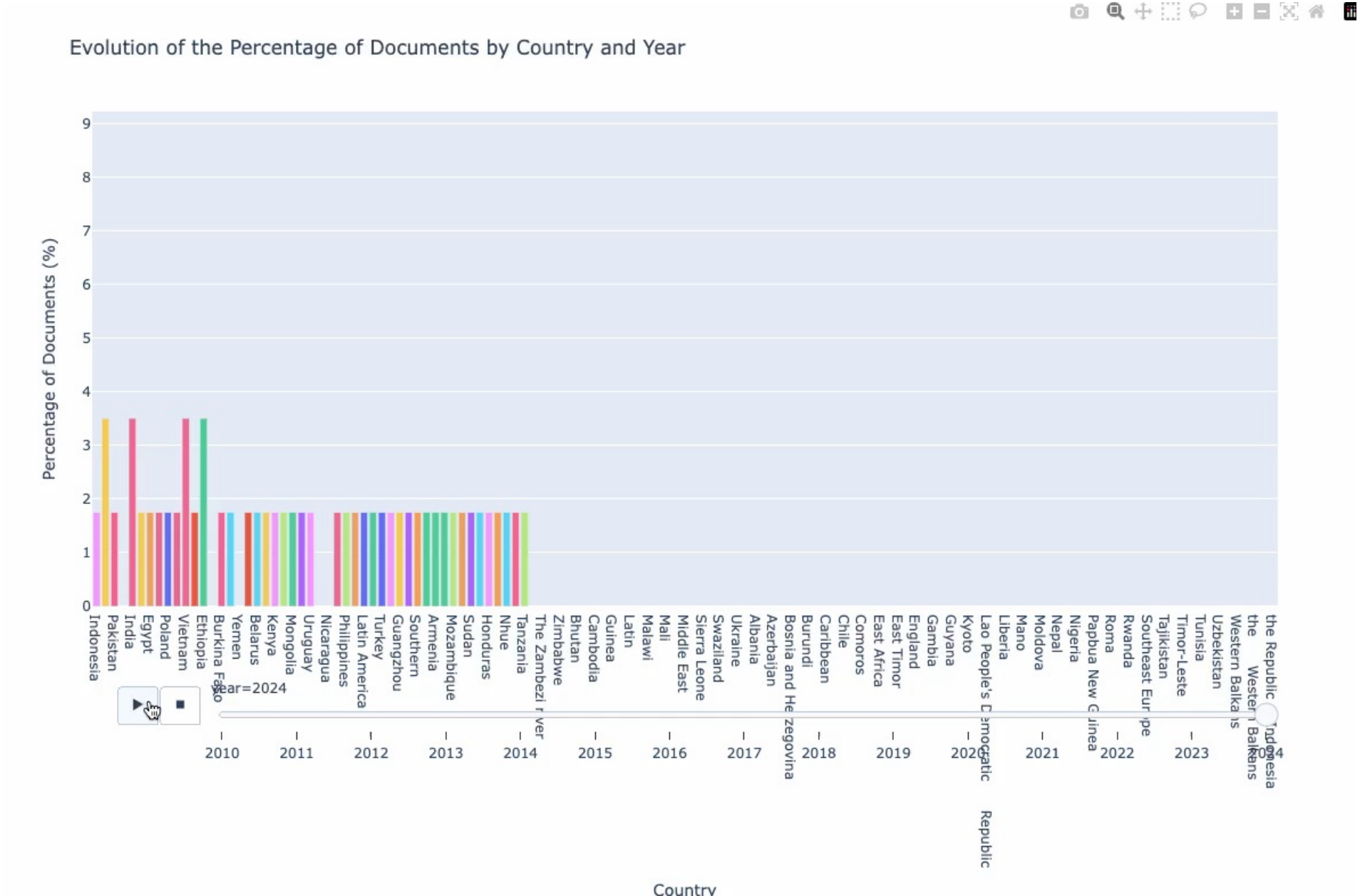
2270 out of 2920

nltk - locationtagger

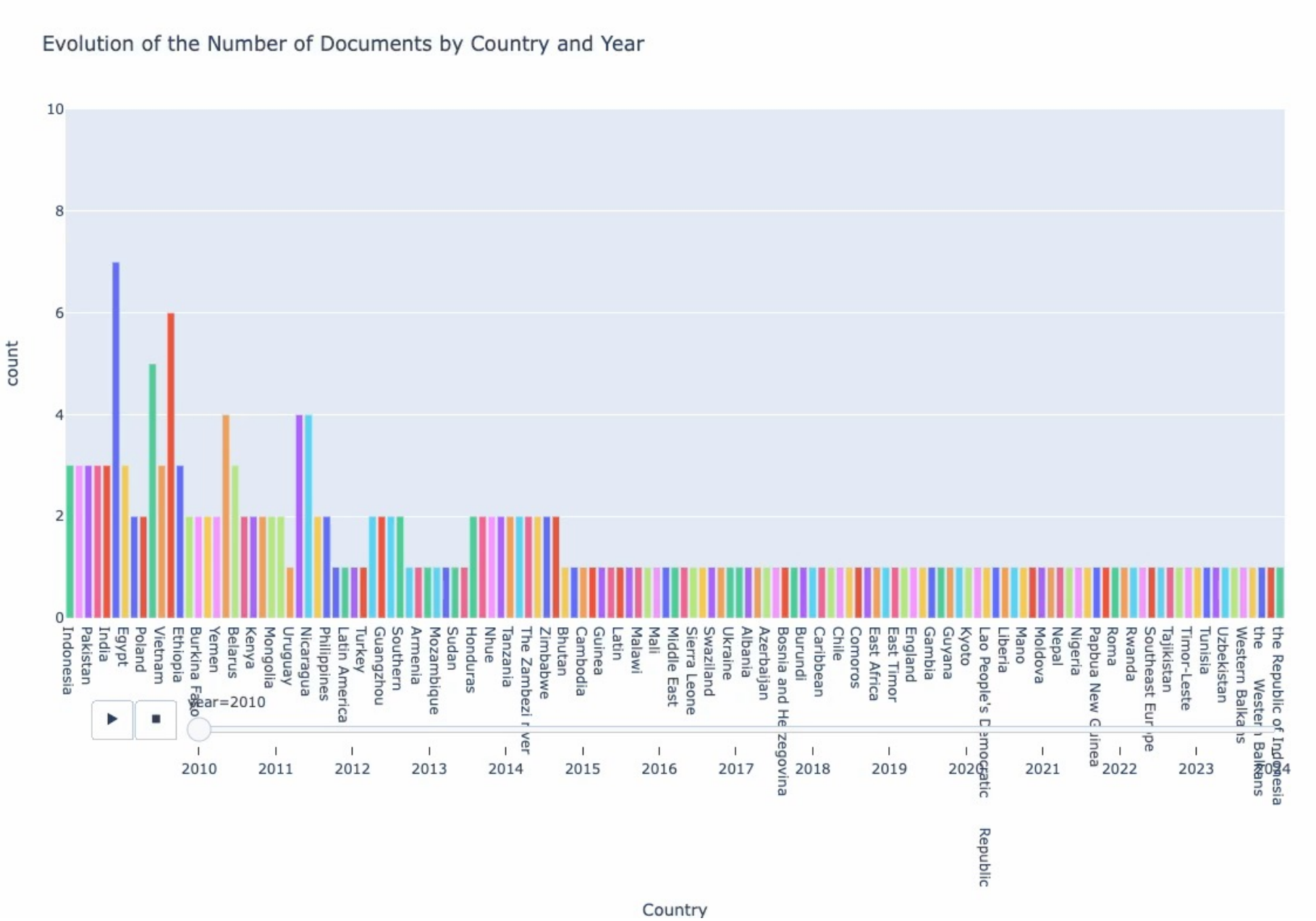
60%

1756 out of 2920

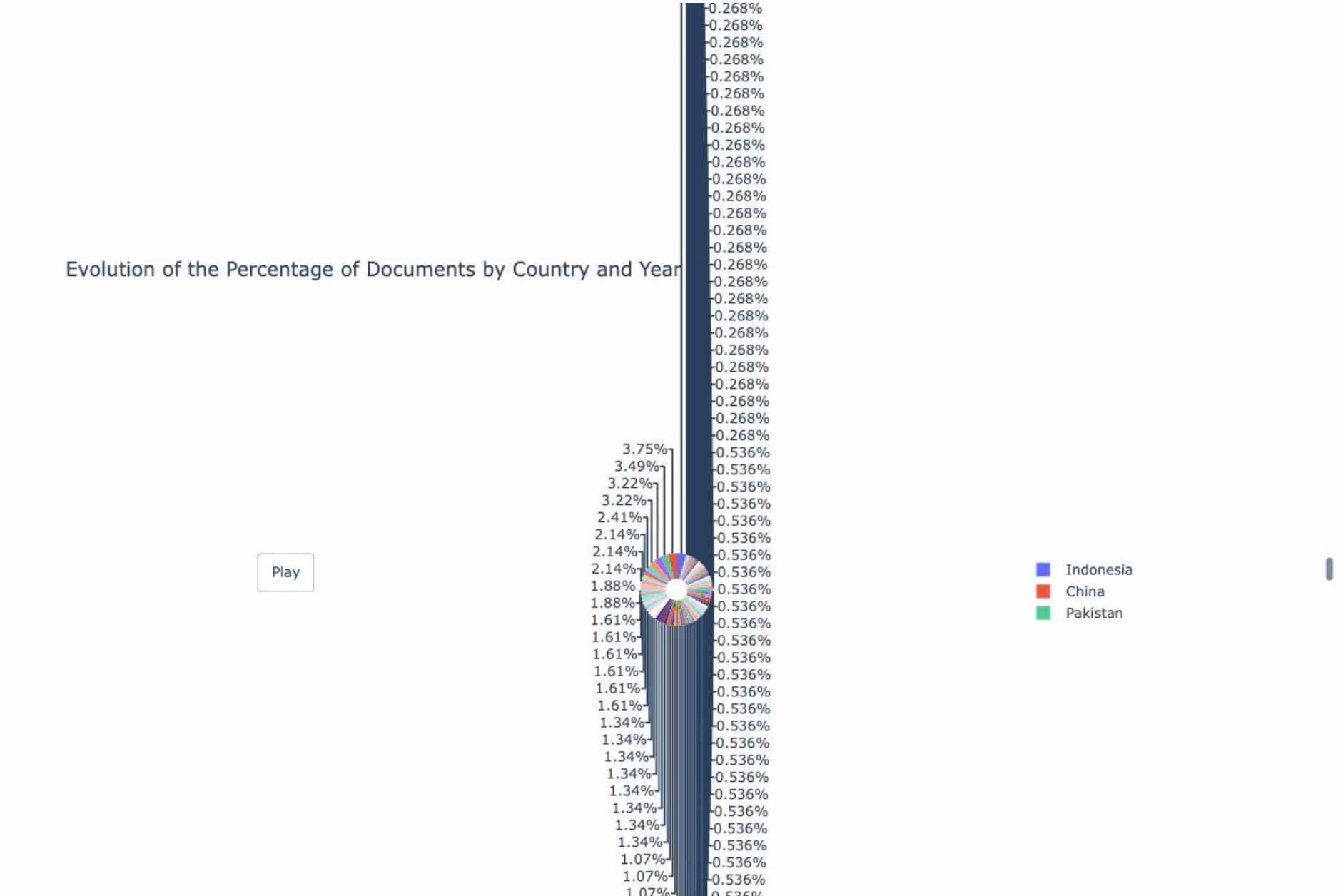
Data Visualization



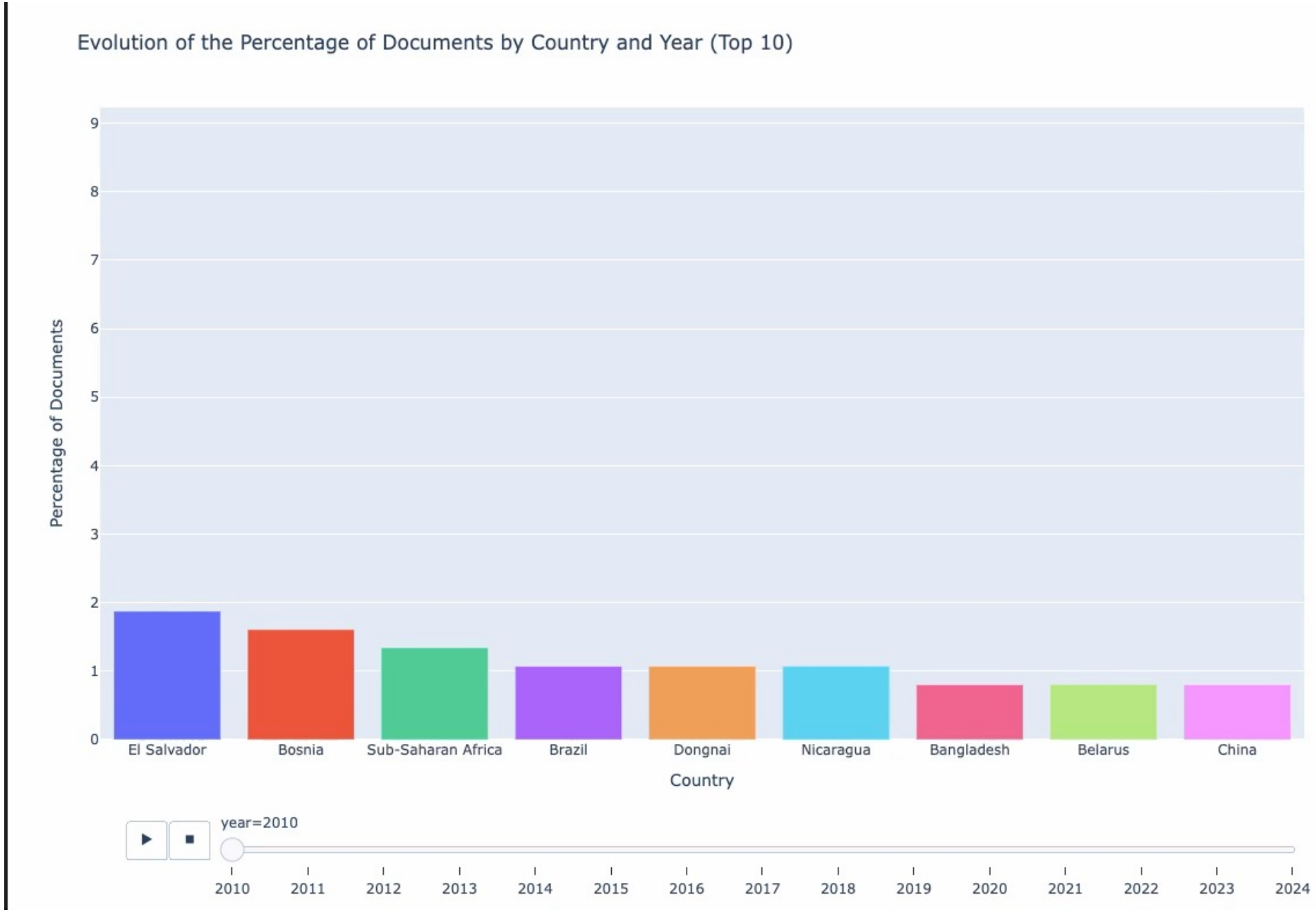
Data Visualization



Data Visualization



Data Visualization

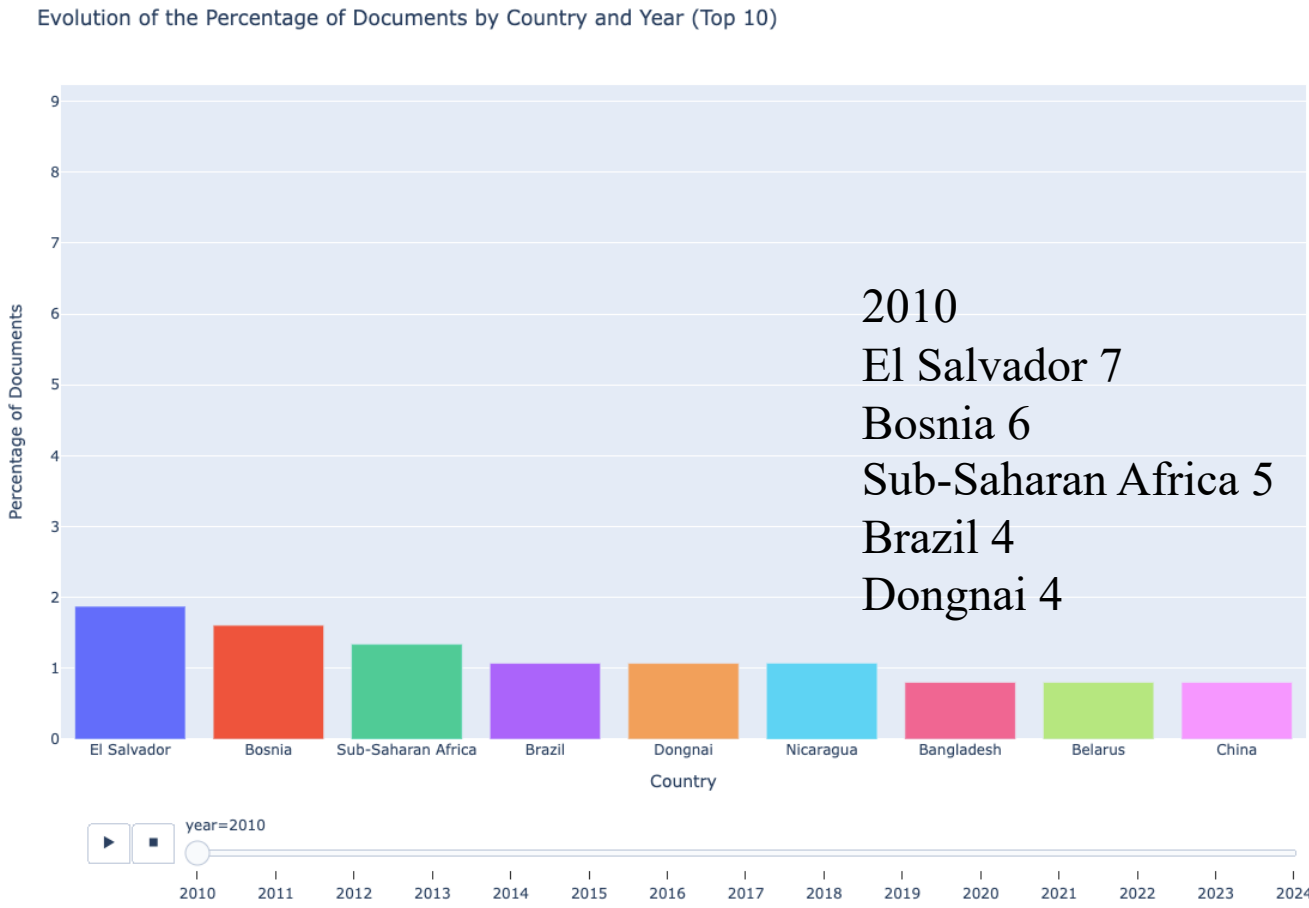


Validation

1. Percentage - Summation

2010 100.0
2011 100.0
2012 100.0
2013 100.0
2014 100.0
2015 100.0
2016 100.0
2017 100.0
2018 100.0
2019 100.0
2020 100.0
2021 100.0
2022 100.0
2023 100.0
2024 100.0

2. Counts - Subset of Data



2010
El Salvador 7
Bosnia 6
Sub-Saharan Africa 5
Brazil 4
Dongnai 4