Leveraging Semantic Embeddings for Topic Analysis

In [86]:

```python
import re
import warnings

import hdbscan
from langchain_dartmouth.llms import ChatDartmouthCloud
import numpy as np
import pandas as pd
import plotly.express as px
import plotly.io as pio
from sentence_transformers import SentenceTransformer
from umap import UMAP

pio.renderers.default = "iframe"

# Ignore all warnings
warnings.filterwarnings("ignore")
```

In [87]:

```python
df = pd.read_csv("./data/survey_responses.csv")
```

In [88]:

```python
sentence_model = SentenceTransformer("all-MiniLM-L6-v2")
```

In [89]:

```python
df["embeddings"] = sentence_model.encode(df.Response).tolist()
```

In [90]:

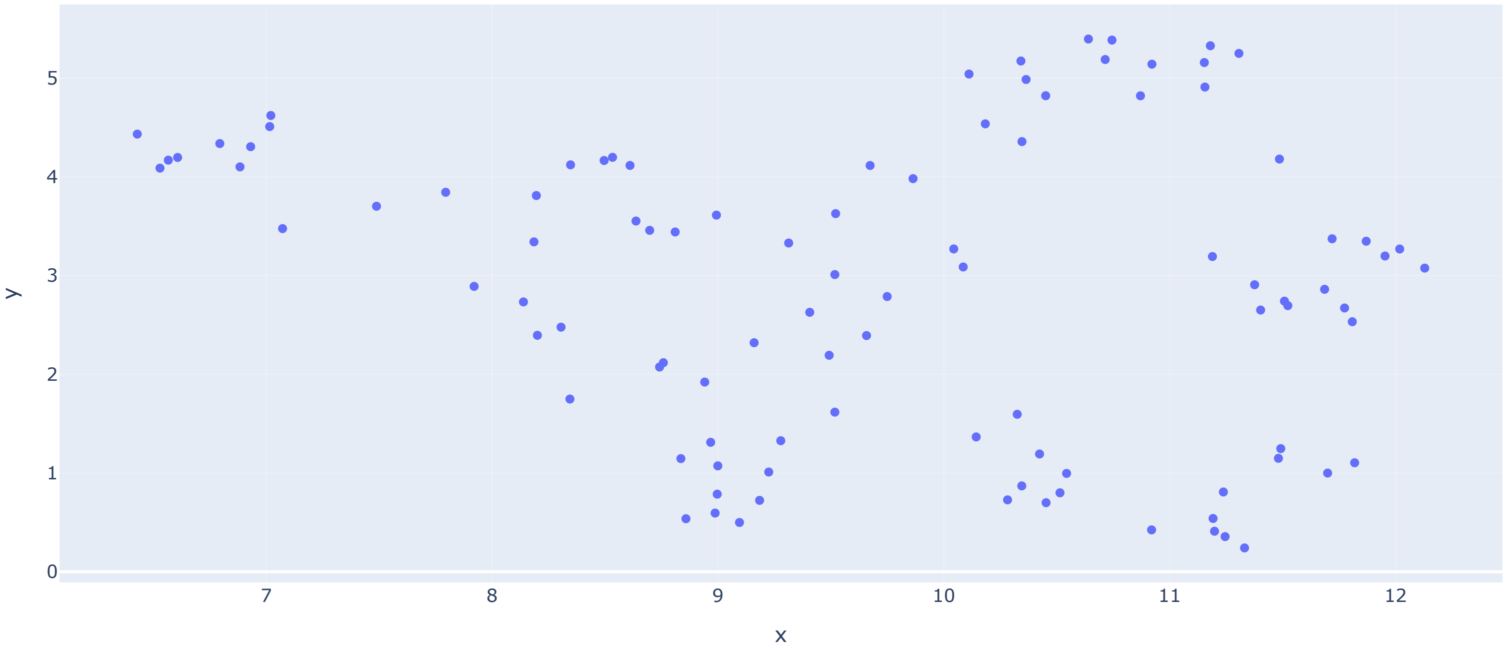
```python
umap_model = UMAP(random_state=5)
```

In [91]:

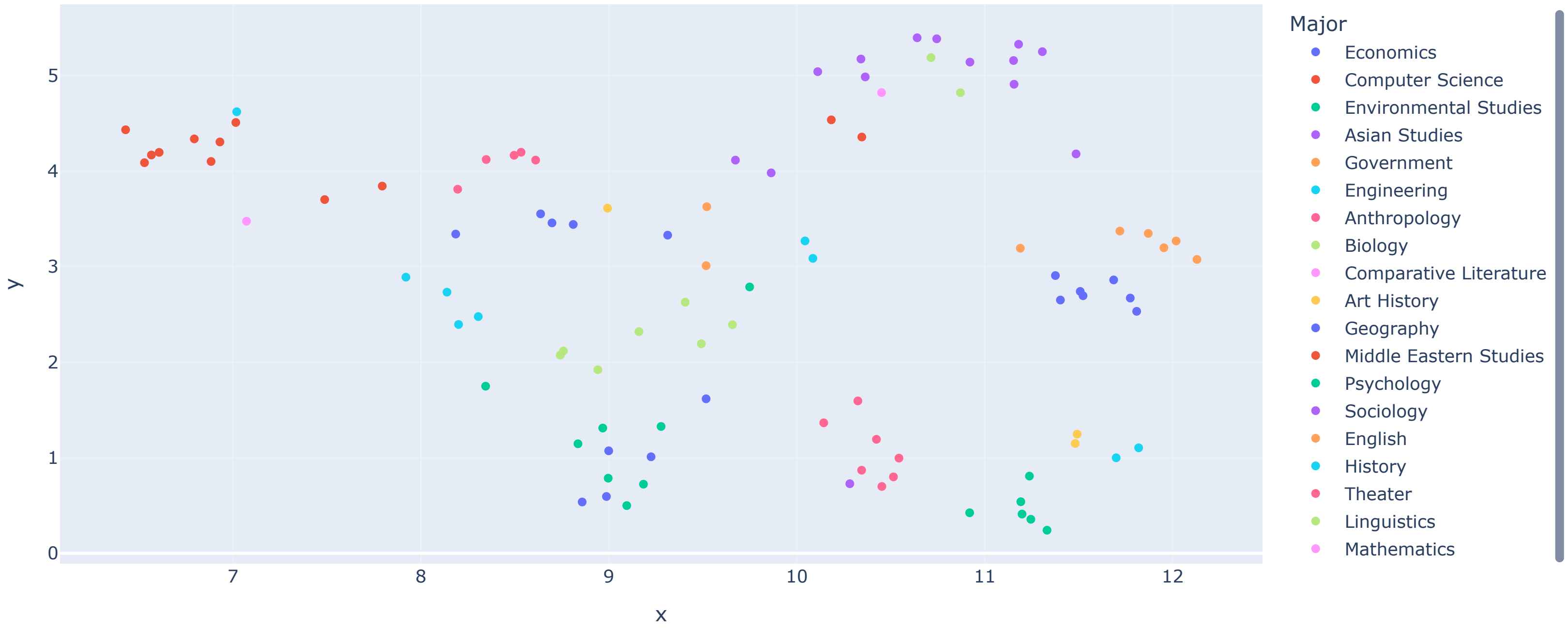
```python
df[["x", "y"]] = umap_model.fit_transform(np.array(df["embeddings"].values.tolist()))
```

In [92]:

```python
fig = px.scatter(df, x="x", y="y", hover_data=["Response"])
fig.show()
```

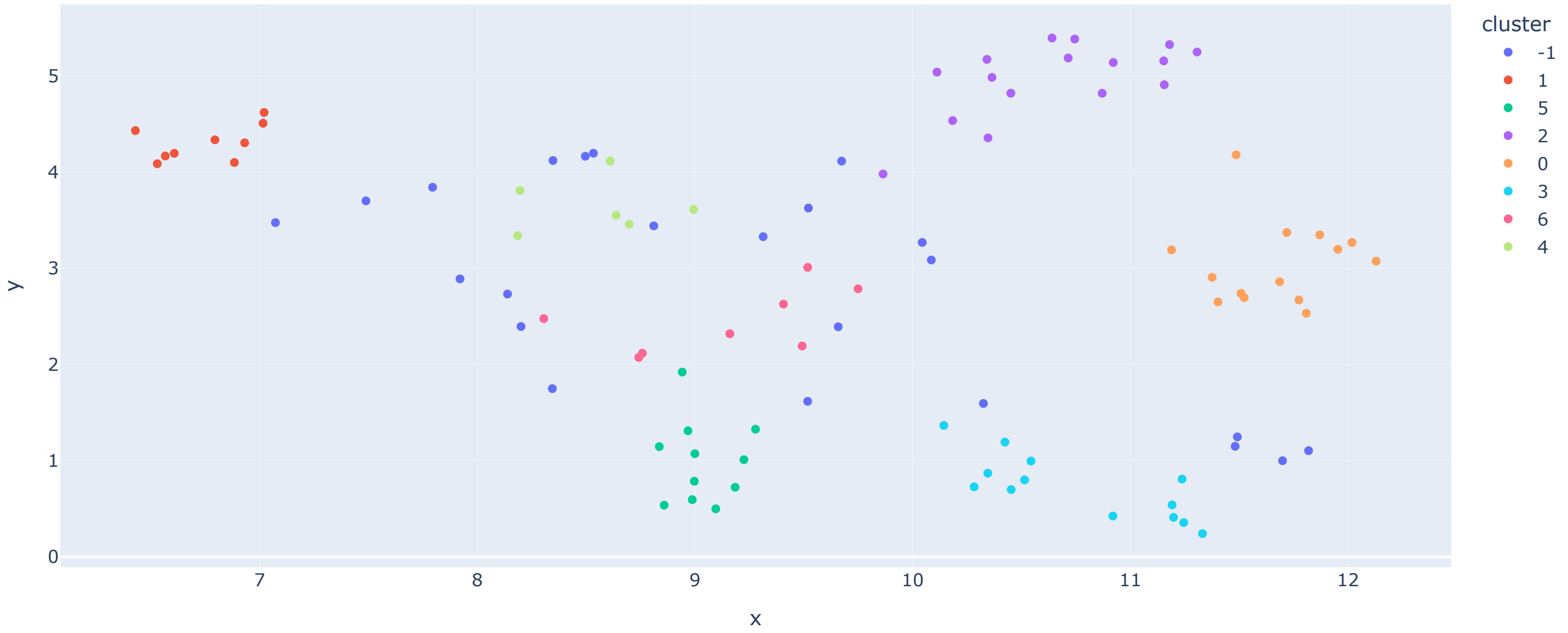


```
In [93]: fig = px.scatter(df, x="x", y="y", hover_data=["Response"], color="Major")
         fig.show()
```



```
In [94]: df["cluster"] = hdbscan.HDBSCAN().fit_predict(df[["x", "y"]]).astype("str")
```

```
In [95]: fig = px.scatter(df, x="x", y="y", hover_data=["Response"], color="cluster")
         fig.show()
```



In [96]:

```python
llm = ChatDartmouthCloud(model_name="openai.gpt-4o-mini-2024-07-18")


def find_cluster_label(responses):
    responses = "\n--\n".join(responses)
    prompt = (
        "The following are responses to the question: "
        "'What do you think was the biggest benefit of the Guarini Exchange Program "
        "for your personal or professional development?' "
        "All of these responses share a common theme or topic, similar to a headline. "
        "Take a few moments to analyze the responses, then identify the most salient topic. "
        "Finally, respond with the topic between the tags <topic_label></topic_label>. "
        "Here are the responses:\n\n"
        f"{responses}"
    )
    response = llm.invoke(prompt)
    label = re.findall(
        pattern=r"<topic_label>(.*)</topic_label>", string=response.content
    )[0]
    return label
```

```python
df["topic"] = None
for cluster in df.cluster.unique():
    if cluster == "-1":
        continue
    subset = df[df.cluster == cluster]
    df.loc[df.cluster == cluster, "topic"] = find_cluster_label(subset.Response)
df
```

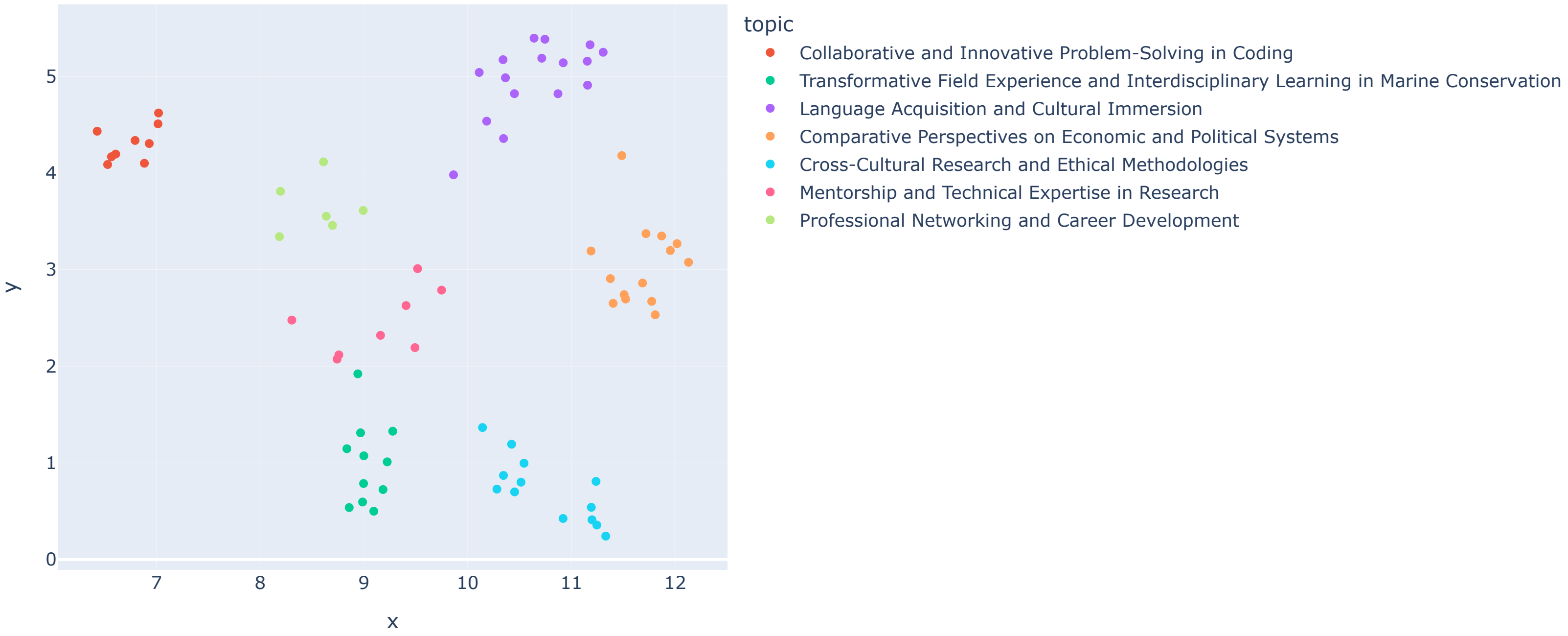
Out[96]:

	Respondent_ID	Major	Response	embeddings	x	y	cluster	topic
0	R001	Economics	The biggest benefit of Guarini Exchange was de...	[-0.02808222733438015, 0.008524204604327679, 0...	9.312801	3.329983	-1	None
1	R002	Computer Science	Learning to code in a different cultural conte...	[-0.023660454899072647, 0.011696777306497097, ...	6.883540	4.101231	1	Collaborative and Innovative Problem-Solving i...
2	R003	Environmental Studies	Studying at Williams-Mystic completely changed...	[-0.013019781559705734, 0.04203595221042633, 0...	8.967394	1.310911	5	Transformative Field Experience and Interdisci...
3	R004	Asian Studies	My time at Waseda Uni in Tokyo improved my Jap...	[-0.004582987632602453, -0.045444514602422714,...	10.744184	5.385358	2	Language Acquisition and Cultural Immersion
4	R005	Government	The Guarini program gave me confidence I never...	[0.0012159398756921291, -0.03586733713746071, ...	9.520711	3.627601	-1	None
...	...	...	...	...	...	...	...	...
95	R096	Psychology	Cross-cultural perspectives on developmental p...	[0.06583299487829208, 0.04757784679532051, -0....	11.198199	0.410375	3	Cross-Cultural Research and Ethical Methodologies
96	R097	Computer Science	AIT Budapest's creative approach to problem-so...	[-0.06781381368637085, 0.07236985117197037, 0....	6.428737	4.433134	1	Collaborative and Innovative Problem-Solving i...
97	R098	Asian Studies	My time at Keio improved my Japanese dramatica...	[-0.0008568129851482809, 0.08134118467569351, ...	11.306201	5.250633	2	Language Acquisition and Cultural Immersion
98	R099	Environmental Studies	The biggest benefit was seeing environmental c...	[0.01006466243416071, 0.08017655462026596, 0.0...	9.277708	1.327294	5	Transformative Field Experience and Interdisci...
99	R100	Government	UCL's comparative approach to political system...	[-0.05885228514671326, -0.0027636357117444277,...	11.953079	3.198130	0	Comparative Perspectives on Economic and Polit...

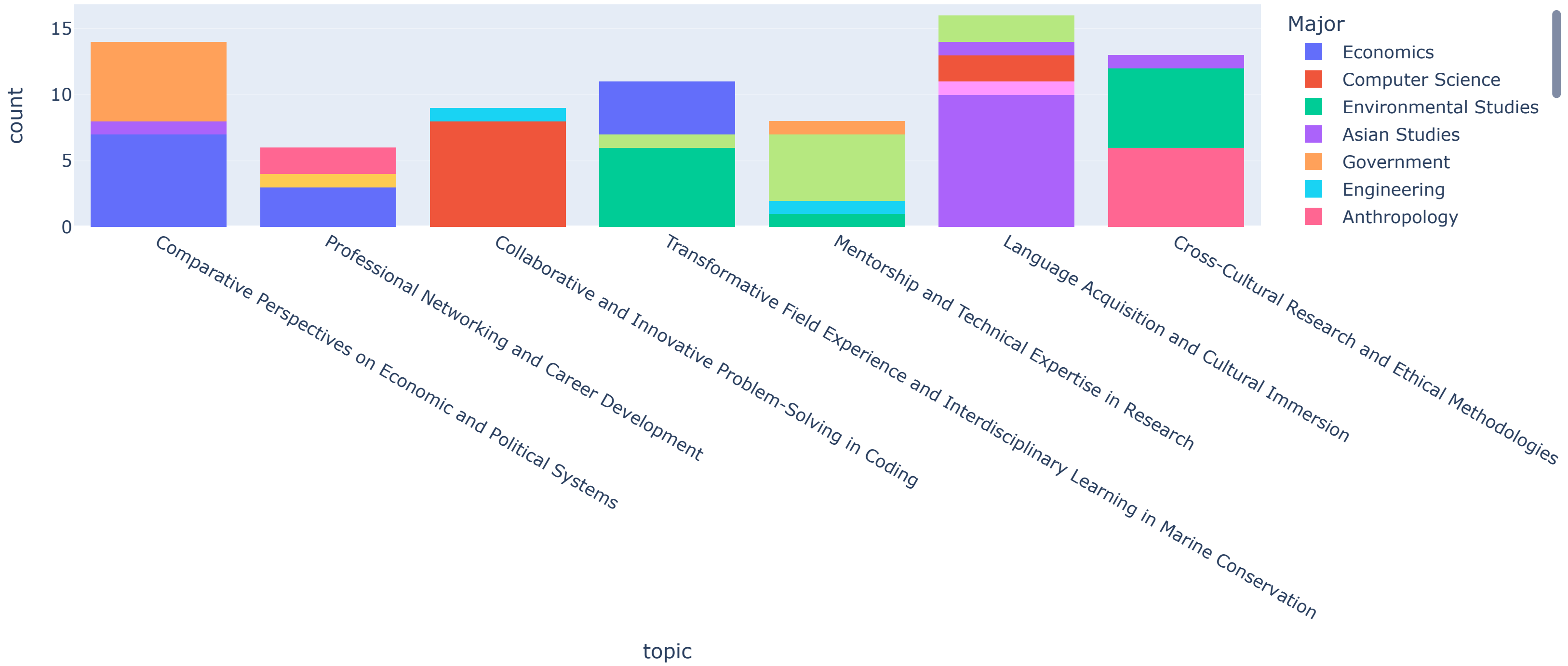
100 rows × 8 columns

In [97]:

```python
fig = px.scatter(df, x="x", y="y", hover_data=["Response"], color="topic")
fig.show()
```



```
In [98]: px.histogram(df, x="topic", color="Major")
```


count
15
10
5
0
topic
Comparative Perspectives on Economic and Political Systems
Professional Networking and Career Development
Collaborative and Innovative Problem-Solving in Coding
Transformative Field Experience and Interdisciplinary Learning in Marine Conservation
Mentorship and Technical Expertise in Research
Language Acquisition and Cultural Immersion
Cross-Cultural Research and Ethical Methodologies
Major
Economics
Computer Science
Environmental Studies
Asian Studies
Government
Engineering
Anthropology