In [25]:

```
!pip install datamapplot
```

```
huggingface/tokenizers: The current process just got forked, after parallelism has al
ready been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true | fals
e)
```

```
Requirement already satisfied: datamapplot in /opt/conda/lib/python3.10/site-packages
(0.3.0)
Requirement already satisfied: numpy>=1.21 in /opt/conda/lib/python3.10/site-packages
(from datamapplot) (1.26.4)
Requirement already satisfied: matplotlib>=3.8 in /opt/conda/lib/python3.10/site-pack
ages (from datamapplot) (3.8.4)
Requirement already satisfied: scikit-learn>=1.1 in /opt/conda/lib/python3.10/site-pa
ckages (from datamapplot) (1.2.2)
Requirement already satisfied: pandas>=1.0 in /opt/conda/lib/python3.10/site-packages
(from datamapplot) (2.2.2)
Requirement already satisfied: datashader>=0.16 in /opt/conda/lib/python3.10/site-pac
kages (from datamapplot) (0.16.3)
Requirement already satisfied: colorspacious>=1.1 in /opt/conda/lib/python3.10/site-p
ackages (from datamapplot) (1.1.2)
Requirement already satisfied: scikit-image>=0.22 in /opt/conda/lib/python3.10/site-p
ackages (from datamapplot) (0.22.0)
Requirement already satisfied: numba>=0.56 in /opt/conda/lib/python3.10/site-packages
(from datamapplot) (0.58.1)
Requirement already satisfied: pylabeladjust in /opt/conda/lib/python3.10/site-packag
es (from datamapplot) (0.1.13)
Requirement already satisfied: requests in /opt/conda/lib/python3.10/site-packages (f
rom datamapplot) (2.32.3)
Requirement already satisfied: jinja2 in /opt/conda/lib/python3.10/site-packages (fro
m datamapplot) (3.1.2)
Requirement already satisfied: colorcet in /opt/conda/lib/python3.10/site-packages (f
rom datashader>=0.16->datamapplot) (3.1.0)
Requirement already satisfied: dask in /opt/conda/lib/python3.10/site-packages (from
datashader>=0.16->datamapplot) (2024.7.0)
Requirement already satisfied: multipledispatch in /opt/conda/lib/python3.10/site-pac
kages (from datashader>=0.16->datamapplot) (1.0.0)
Requirement already satisfied: param in /opt/conda/lib/python3.10/site-packages (from
datashader>=0.16->datamapplot) (2.1.1)
Requirement already satisfied: pillow in /opt/conda/lib/python3.10/site-packages (fro
m datashader>=0.16->datamapplot) (9.5.0)
Requirement already satisfied: pyct in /opt/conda/lib/python3.10/site-packages (from
datashader>=0.16->datamapplot) (0.5.0)
Requirement already satisfied: scipy in /opt/conda/lib/python3.10/site-packages (from
datashader>=0.16->datamapplot) (1.11.4)
Requirement already satisfied: toolz in /opt/conda/lib/python3.10/site-packages (from
datashader>=0.16->datamapplot) (0.12.1)
Requirement already satisfied: packaging in /opt/conda/lib/python3.10/site-packages
(from datashader>=0.16->datamapplot) (24.1)
Requirement already satisfied: xarray in /opt/conda/lib/python3.10/site-packages (fro
m datashader>=0.16->datamapplot) (2024.6.0)
Requirement already satisfied: contourpy>=1.0.1 in /opt/conda/lib/python3.10/site-pac
kages (from matplotlib>=3.8->datamapplot) (1.2.0)
Requirement already satisfied: cycler>=0.10 in /opt/conda/lib/python3.10/site-package
s (from matplotlib>=3.8->datamapplot) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /opt/conda/lib/python3.10/site-pa
ckages (from matplotlib>=3.8->datamapplot) (4.47.0)
Requirement already satisfied: kiwisolver>=1.3.1 in /opt/conda/lib/python3.10/site-pa
ckages (from matplotlib>=3.8->datamapplot) (1.4.5)
Requirement already satisfied: pyparsing>=2.3.1 in /opt/conda/lib/python3.10/site-pac
kages (from matplotlib>=3.8->datamapplot) (3.1.1)
Requirement already satisfied: python-dateutil>=2.7 in /opt/conda/lib/python3.10/site
-packages (from matplotlib>=3.8->datamapplot) (2.9.0.post0)
Requirement already satisfied: llvmlite<0.42,>=0.41.0dev0 in /opt/conda/lib/python3.1
```

```
0/site-packages (from numba>=0.56->datamapplot) (0.41.1)
Requirement already satisfied: pytz>=2020.1 in /opt/conda/lib/python3.10/site-package
s (from pandas>=1.0->datamapplot) (2023.3.post1)
Requirement already satisfied: tzdata>=2022.7 in /opt/conda/lib/python3.10/site-packa
ges (from pandas>=1.0->datamapplot) (2023.4)
Requirement already satisfied: networkx>=2.8 in /opt/conda/lib/python3.10/site-packag
es (from scikit-image>=0.22->datamapplot) (3.2.1)
Requirement already satisfied: imageio>=2.27 in /opt/conda/lib/python3.10/site-packag
es (from scikit-image>=0.22->datamapplot) (2.33.1)
Requirement already satisfied: tifffile>=2022.8.12 in /opt/conda/lib/python3.10/site-
packages (from scikit-image>=0.22->datamapplot) (2023.12.9)
Requirement already satisfied: lazy_loader>=0.3 in /opt/conda/lib/python3.10/site-pac
kages (from scikit-image>=0.22->datamapplot) (0.3)
Requirement already satisfied: joblib>=1.1.1 in /opt/conda/lib/python3.10/site-packag
es (from scikit-learn>=1.1->datamapplot) (1.4.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in /opt/conda/lib/python3.10/site
-packages (from scikit-learn>=1.1->datamapplot) (3.2.0)
Requirement already satisfied: MarkupSafe>=2.0 in /opt/conda/lib/python3.10/site-pack
ages (from jinja2->datamapplot) (2.1.3)
Requirement already satisfied: Pyqtree<2.0.0,>=1.0.0 in /opt/conda/lib/python3.10/sit
e-packages (from pylabeladjust->datamapplot) (1.0.0)
Requirement already satisfied: tqdm<5.0.0,>=4.66.2 in /opt/conda/lib/python3.10/site-
packages (from pylabeladjust->datamapplot) (4.66.4)
Requirement already satisfied: charset-normalizer<4,>=2 in /opt/conda/lib/python3.10/
site-packages (from requests->datamapplot) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /opt/conda/lib/python3.10/site-package
s (from requests->datamapplot) (3.6)
Requirement already satisfied: urllib3<3,>=1.21.1 in /opt/conda/lib/python3.10/site-p
ackages (from requests->datamapplot) (1.26.18)
Requirement already satisfied: certifi>=2017.4.17 in /opt/conda/lib/python3.10/site-p
ackages (from requests->datamapplot) (2024.7.4)
Requirement already satisfied: six>=1.5 in /opt/conda/lib/python3.10/site-packages (f
rom python-dateutil>=2.7->matplotlib>=3.8->datamapplot) (1.16.0)
Requirement already satisfied: click>=8.1 in /opt/conda/lib/python3.10/site-packages
(from dask->datashader>=0.16->datamapplot) (8.1.7)
Requirement already satisfied: cloudpickle>=1.5.0 in /opt/conda/lib/python3.10/site-p
ackages (from dask->datashader>=0.16->datamapplot) (2.2.1)
Requirement already satisfied: fsspec>=2021.09.0 in /opt/conda/lib/python3.10/site-pa
ckages (from dask->datashader>=0.16->datamapplot) (2024.5.0)
Requirement already satisfied: partd>=1.4.0 in /opt/conda/lib/python3.10/site-package
s (from dask->datashader>=0.16->datamapplot) (1.4.2)
Requirement already satisfied: pyyaml>=5.3.1 in /opt/conda/lib/python3.10/site-packag
es (from dask->datashader>=0.16->datamapplot) (6.0.1)
Requirement already satisfied: importlib-metadata>=4.13.0 in /opt/conda/lib/python3.1
0/site-packages (from dask->datashader>=0.16->datamapplot) (6.11.0)
Requirement already satisfied: zipp>=0.5 in /opt/conda/lib/python3.10/site-packages
(from importlib-metadata>=4.13.0->dask->datashader>=0.16->datamapplot) (3.17.0)
Requirement already satisfied: locket in /opt/conda/lib/python3.10/site-packages (fro
m partd>=1.4.0->dask->datashader>=0.16->datamapplot) (1.0.0)
```

In [27]:

```
!pip install bertopic==0.16.0 # for auto reduction/merging of topics
```

```
huggingface/tokenizers: The current process just got forked, after parallelism has al
ready been used. Disabling parallelism to avoid deadlocks...
To disable this warning, you can either:
	- Avoid using `tokenizers` before the fork if possible
	- Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true | fals
e)
```

```
Collecting bertopic==0.16.0
  Downloading bertopic-0.16.0-py2.py3-none-any.whl.metadata (21 kB)
Requirement already satisfied: numpy>=1.20.0 in /opt/conda/lib/python3.10/site-packag
es (from bertopic==0.16.0) (1.26.4)
Requirement already satisfied: hdbscan>=0.8.29 in /opt/conda/lib/python3.10/site-pack
ages (from bertopic==0.16.0) (0.8.38.post1)
Requirement already satisfied: umap-learn>=0.5.0 in /opt/conda/lib/python3.10/site-pa
ckages (from bertopic==0.16.0) (0.5.6)
Requirement already satisfied: pandas>=1.1.5 in /opt/conda/lib/python3.10/site-packag
es (from bertopic==0.16.0) (2.2.2)
Requirement already satisfied: scikit-learn>=0.22.2.post1 in /opt/conda/lib/python3.1
0/site-packages (from bertopic==0.16.0) (1.2.2)
Requirement already satisfied: tqdm>=4.41.1 in /opt/conda/lib/python3.10/site-package
s (from bertopic==0.16.0) (4.66.4)
Requirement already satisfied: sentence-transformers>=0.4.1 in /opt/conda/lib/python
3.10/site-packages (from bertopic==0.16.0) (3.0.1)
Requirement already satisfied: plotly>=4.7.0 in /opt/conda/lib/python3.10/site-packag
es (from bertopic==0.16.0) (5.18.0)
Requirement already satisfied: scipy>=1.0 in /opt/conda/lib/python3.10/site-packages
(from hdbscan>=0.8.29->bertopic==0.16.0) (1.11.4)
Requirement already satisfied: joblib>=1.0 in /opt/conda/lib/python3.10/site-packages
(from hdbscan>=0.8.29->bertopic==0.16.0) (1.4.2)
Requirement already satisfied: python-dateutil>=2.8.2 in /opt/conda/lib/python3.10/si
te-packages (from pandas>=1.1.5->bertopic==0.16.0) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in /opt/conda/lib/python3.10/site-package
s (from pandas>=1.1.5->bertopic==0.16.0) (2023.3.post1)
Requirement already satisfied: tzdata>=2022.7 in /opt/conda/lib/python3.10/site-packa
ges (from pandas>=1.1.5->bertopic==0.16.0) (2023.4)
Requirement already satisfied: tenacity>=6.2.0 in /opt/conda/lib/python3.10/site-pack
ages (from plotly>=4.7.0->bertopic==0.16.0) (8.2.3)
Requirement already satisfied: packaging in /opt/conda/lib/python3.10/site-packages
(from plotly>=4.7.0->bertopic==0.16.0) (24.1)
Requirement already satisfied: threadpoolctl>=2.0.0 in /opt/conda/lib/python3.10/site
-packages (from scikit-learn>=0.22.2.post1->bertopic==0.16.0) (3.2.0)
Requirement already satisfied: transformers<5.0.0,>=4.34.0 in /opt/conda/lib/python3.
10/site-packages (from sentence-transformers>=0.4.1->bertopic==0.16.0) (4.42.3)
Requirement already satisfied: torch>=1.11.0 in /opt/conda/lib/python3.10/site-packag
es (from sentence-transformers>=0.4.1->bertopic==0.16.0) (2.1.2+cpu)
Requirement already satisfied: huggingface-hub>=0.15.1 in /opt/conda/lib/python3.10/s
ite-packages (from sentence-transformers>=0.4.1->bertopic==0.16.0) (0.23.4)
Requirement already satisfied: Pillow in /opt/conda/lib/python3.10/site-packages (fro
m sentence-transformers>=0.4.1->bertopic==0.16.0) (9.5.0)
Requirement already satisfied: numba>=0.51.2 in /opt/conda/lib/python3.10/site-packag
es (from umap-learn>=0.5.0->bertopic==0.16.0) (0.58.1)
Requirement already satisfied: pynndescent>=0.5 in /opt/conda/lib/python3.10/site-pac
kages (from umap-learn>=0.5.0->bertopic==0.16.0) (0.5.13)
Requirement already satisfied: filelock in /opt/conda/lib/python3.10/site-packages (f
rom huggingface-hub>=0.15.1->sentence-transformers>=0.4.1->bertopic==0.16.0) (3.13.1)
Requirement already satisfied: fsspec>=2023.5.0 in /opt/conda/lib/python3.10/site-pac
kages (from huggingface-hub>=0.15.1->sentence-transformers>=0.4.1->bertopic==0.16.0)
(2024.5.0)
Requirement already satisfied: pyyaml>=5.1 in /opt/conda/lib/python3.10/site-packages
(from huggingface-hub>=0.15.1->sentence-transformers>=0.4.1->bertopic==0.16.0) (6.0.
1)
Requirement already satisfied: requests in /opt/conda/lib/python3.10/site-packages (f
rom huggingface-hub>=0.15.1->sentence-transformers>=0.4.1->bertopic==0.16.0) (2.32.3)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /opt/conda/lib/python3.1
```

```
0/site-packages (from huggingface-hub>=0.15.1->sentence-transformers>=0.4.1->bertopic
==0.16.0) (4.9.0)
Requirement already satisfied: llvmlite<0.42,>=0.41.0dev0 in /opt/conda/lib/python3.1
0/site-packages (from numba>=0.51.2->umap-learn>=0.5.0->bertopic==0.16.0) (0.41.1)
Requirement already satisfied: six>=1.5 in /opt/conda/lib/python3.10/site-packages (f
rom python-dateutil>=2.8.2->pandas>=1.1.5->bertopic==0.16.0) (1.16.0)
Requirement already satisfied: sympy in /opt/conda/lib/python3.10/site-packages (from
torch>=1.11.0->sentence-transformers>=0.4.1->bertopic==0.16.0) (1.13.0)
Requirement already satisfied: networkx in /opt/conda/lib/python3.10/site-packages (f
rom torch>=1.11.0->sentence-transformers>=0.4.1->bertopic==0.16.0) (3.2.1)
Requirement already satisfied: jinja2 in /opt/conda/lib/python3.10/site-packages (fro
m torch>=1.11.0->sentence-transformers>=0.4.1->bertopic==0.16.0) (3.1.2)
Requirement already satisfied: regex!=2019.12.17 in /opt/conda/lib/python3.10/site-pa
ckages (from transformers<5.0.0,>=4.34.0->sentence-transformers>=0.4.1->bertopic==0.1
6.0) (2023.12.25)
Requirement already satisfied: safetensors>=0.4.1 in /opt/conda/lib/python3.10/site-p
ackages (from transformers<5.0.0,>=4.34.0->sentence-transformers>=0.4.1->bertopic==0.
16.0) (0.4.3)
Requirement already satisfied: tokenizers<0.20,>=0.19 in /opt/conda/lib/python3.10/si
te-packages (from transformers<5.0.0,>=4.34.0->sentence-transformers>=0.4.1->bertopic
==0.16.0) (0.19.1)
Requirement already satisfied: MarkupSafe>=2.0 in /opt/conda/lib/python3.10/site-pack
ages (from jinja2->torch>=1.11.0->sentence-transformers>=0.4.1->bertopic==0.16.0) (2.
1.3)
Requirement already satisfied: charset-normalizer<4,>=2 in /opt/conda/lib/python3.10/
site-packages (from requests->huggingface-hub>=0.15.1->sentence-transformers>=0.4.1->
bertopic==0.16.0) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /opt/conda/lib/python3.10/site-package
s (from requests->huggingface-hub>=0.15.1->sentence-transformers>=0.4.1->bertopic==0.
16.0) (3.6)
Requirement already satisfied: urllib3<3,>=1.21.1 in /opt/conda/lib/python3.10/site-p
ackages (from requests->huggingface-hub>=0.15.1->sentence-transformers>=0.4.1->bertop
ic==0.16.0) (1.26.18)
Requirement already satisfied: certifi>=2017.4.17 in /opt/conda/lib/python3.10/site-p
ackages (from requests->huggingface-hub>=0.15.1->sentence-transformers>=0.4.1->bertop
ic==0.16.0) (2024.7.4)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /opt/conda/lib/python3.10/site-p
ackages (from sympy->torch>=1.11.0->sentence-transformers>=0.4.1->bertopic==0.16.0)
(1.3.0)
Downloading bertopic-0.16.0-py2.py3-none-any.whl (154 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1
54.1/154.1 kB 5.5 MB/s eta 0:00:00
Installing collected packages: bertopic
  Attempting uninstall: bertopic
    Found existing installation: bertopic 0.16.3
    Uninstalling bertopic-0.16.3:
      Successfully uninstalled bertopic-0.16.3
Successfully installed bertopic-0.16.0
```

In [30]:

```python
from bertopic import BERTopic
import pandas as pd
import re
```

Data Description

Scraped from 2002-2023 CAIE A-level Economics past papers

- 110 papers, each with 30 multiple choice questions
- Year 2002, 2017-2023 scraped using `LlamaParse` – higher quality
- Other years scraped using `pymupdf`

In [31]:

```
excel_file_path = '/kaggle/input/bertopic-data/econ_mc_questions.xlsx'

# Read the Excel file into a DataFrame
df = pd.read_excel(excel_file_path)
```

In [32]:

```
df.head()
```

Out[32]:

	year_version	question_no	question
0	202211_11	1	An economy is changing from a planned economy ...
1	202211_11	2	A local council provides a tap for drinking wa...
2	202211_11	3	Food prices in a country increased by 20% in t...
3	202211_11	4	A doctor has very long working hours and a hig...
4	202211_11	5	When will the demand curve for motorcycles shi...

In [33]:

```
df['question'] = df['question'].astype(str)
```

In [38]:

```
def replace_hyphens(text):
    return re.sub(r'(?<!non)-', ' ', text)
```

In [39]:

```
df['question'] = df['question'].apply(replace_hyphens)
```

In [40]:

```
docs = df['question'].tolist()
```

In [43]:

```
len(docs)
```

Out[43]:

```
3030
```

In [17]:
```
from nltk.corpus import stopwords

# Get the list of stopwords for English
stop_words = stopwords.words('english')

stop_words.remove("against")

stop_words = stop_words + ["year", "car", "cars", "ticket", "tickets", "firm", "firms",
                           "change", "changes", "product", "products", "unit", "units",
                          "million", "billion", "yes", "no"]
# Display all stopwords
print(stop_words)
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you'v
e", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his',
'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself',
'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'thi
s', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'bee
n', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an',
'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'at', 'by', 'fo
r', 'with', 'about', 'between', 'into', 'through', 'during', 'before', 'after', 'abov
e', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under',
'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how',
'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no',
'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'w
ill', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 'r
e', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'does
n', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "is
n't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "sh
an't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't",
'wouldn', "wouldn't", 'year', 'car', 'cars', 'ticket', 'tickets', 'firm', 'firms', 'c
hange', 'changes', 'product', 'products', 'unit', 'units']
```

Default

In [51]:
```
topic_model = BERTopic()
```

In [52]:
```
%%time
topics, probs = topic_model.fit_transform(docs)
```

```
CPU times: user 2min 42s, sys: 5.08 s, total: 2min 47s
Wall time: 1min 28s
```

In [17]:
```
topic_model.get_topic_info().shape
```

Out[17]:
```
(68, 5)
```

In [16]:

```
topic_model.get_topic_info().head(20)
```

Out[16]:

	Topic	Count	Name	Representation	Representative_Docs
0	-1	581	-1_the_to_price_in	[the, to, price, in, of, quantity, for, on, is...	[In a free market there is a surplus of a good...
1	0	197	0_production_advantage_possibility_country	[production, advantage, possibility, country, ...	[19 The table shows the output per unit of inp...
2	1	127	1_aggregate_ad_level_real	[aggregate, ad, level, real, output, ad1, ad2,...	[19 The diagram shows a shift in the aggregate...
3	2	100	2_external_social_benefits_costs	[external, social, benefits, costs, private, b...	[14 The government is considering building flo...
4	3	76	3_currency_exchange_its_foreign	[currency, exchange, its, foreign, rate, float...	[Under a system of floating exchange rates, wh...
5	4	74	4_its_terms_prices_export	[its, terms, prices, export, trade, exports, c...	[In which situation will a countrys terms of t...
6	5	69	5_good_public_merit_private	[good, public, merit, private, provided, becau...	[What is a merit good? \n\n A a good w...
7	6	65	6_inflation_year_index_cpi	[inflation, year, index, cpi, fell, rate, was,...	[25 The diagram shows the annual rate of infla...
8	7	57	7_good_rise_decrease_increase	[good, rise, decrease, increase, uncertain, fa...	[11 Good X is a substitute for good Y and a co...
9	8	57	8_maximum_will_price_quantity	[maximum, will, price, quantity, minimum, woul...	[Which statement about maximum and minimum pri...
10	9	56	9_inflation_push_pull_rising	[inflation, push, pull, rising, falling, incre...	[25 What is the most likely cause of cost push...
11	10	54	10_surplus_producer_consumer_area	[surplus, producer, consumer, area, diagram, q...	[Producer surplus is the difference between\n\...
12	11	51	11_planned_economy_mixed_transition	[planned, economy, mixed, transition, market, ...	[2 Which change in economic system is likely t...
13	12	50	12_policy_deficit_payments_currency	[policy, deficit, payments, currency, tariffs,...	[Which measure to correct a balance of payment...

	Topic	Count	Name	Representation	Representative_Docs
14	13	49	13_cars_car_transport_petrol	[cars, car, transport, petrol, public, demand,...	[The demand curve in the diagram shows the rel...
15	14	47	14_current_deficit_account_balance	[current, deficit, account, balance, payments,...	[30 A country at the beginning of a given peri...
16	15	47	15_firm_product_change_price	[firm, product, change, price, percentage, ela...	[The price elasticity of demand for a firm’s p...
17	16	45	16_units_elasticity_10_supplied	[units, elasticity, 10, supplied, price, 15, p...	[A manufacturer progressively reduces the pric...
18	17	45	17_protection_trade_domestic_protectionism	[protection, trade, domestic, protectionism, e...	[20 Which argument is an importer most likely ...
19	18	44	18_cross_good_elasticity_respect	[cross, good, elasticity, respect, demand, pro...	[6 The price of good X rises by 20%. As a resu...

Using saved embedding accelerate training

In [10]:

```python
from sentence_transformers import SentenceTransformer
from sklearn.feature_extraction.text import CountVectorizer
from bertopic.representation import MaximalMarginalRelevance

sentence_model = SentenceTransformer("all-MiniLM-L6-v2")
embeddings = sentence_model.encode(docs, show_progress_bar=True)
```

```
modules.json:   0%|          | 0.00/349 [00:00<?, ?B/s]

config_sentence_transformers.json:   0%|          | 0.00/116 [00:00<?, ?B/s]

README.md:   0%|          | 0.00/10.7k [00:00<?, ?B/s]

sentence_bert_config.json:   0%|          | 0.00/53.0 [00:00<?, ?B/s]

config.json:   0%|          | 0.00/612 [00:00<?, ?B/s]

model.safetensors:   0%|          | 0.00/90.9M [00:00<?, ?B/s]

tokenizer_config.json:   0%|          | 0.00/350 [00:00<?, ?B/s]

vocab.txt:   0%|          | 0.00/232k [00:00<?, ?B/s]

tokenizer.json:   0%|          | 0.00/466k [00:00<?, ?B/s]

special_tokens_map.json:   0%|          | 0.00/112 [00:00<?, ?B/s]

1_Pooling/config.json:   0%|          | 0.00/190 [00:00<?, ?B/s]

Batches:   0%|          | 0/95 [00:00<?, ?it/s]
```

Customize the training process

In [70]:

```python
vectorizer_model = CountVectorizer(stop_words=stop_words, ngram_range=(1, 4), min_df=5)
representation_model = MaximalMarginalRelevance(diversity=0.5)

topic_model_better = BERTopic(
    vectorizer_model=vectorizer_model,
    representation_model=representation_model,
    top_n_words=15,
    min_topic_size=15,
    calculate_probabilities=True
)
```

Model Training

In [71]:

```python
topics_better, prob_better = topic_model_better.fit_transform(docs, embeddings)
```

In [13]:
```
topic_model_better.get_topic_info().shape
```

Out[13]:
```
(49, 5)
```

Inspect topics, their keywords, representative documents

In [106]:
```
topic_model_better.get_topic_info().head(10)
```

Out[106]:

	Topic	Count	Name	Representation	Representative_Docs
0	-1	656	-1_of_price_market_supply	[of, price, market, supply, government, quanti...	[The government fixes a maximum price for whea...
1	0	127	0_aggregate_price level_level_aggregate demand	[aggregate, price level, level, aggregate dema...	[The diagram shows aggregate demand (AD) and a...
2	1	125	1_price elasticity_price elasticity of_elastic...	[price elasticity, price elasticity of, elasti...	[A car manufacturer estimates that the price e...
3	2	103	2_inflation_index_of inflation_rose	[inflation, index, of inflation, rose, rate of...	[25 The table shows a country's rate of inflat...
4	3	99	3_public_good_private_merit	[public, good, private, merit, public goods, e...	[16 Why does the production of public goods ha...
5	4	99	4_external_social_benefits_costs	[external, social, benefits, costs, private, b...	[14 The government is considering building flo...
6	5	98	5_production possibility_possibility_productio...	[production possibility, possibility, producti...	[2\n\nThe production possibility curve for an ...
7	6	97	6_account_balance_current_current account	[account, balance, current, current account, b...	[22 The table shows all of the items on the cu...
8	7	87	7_tax_of tax_income_earners	[tax, of tax, income, earners, 000, specific, ...	[11 The diagram shows the demand curve and sup...
9	8	83	8_shift_curve_supply curve_price of	[shift, curve, supply curve, price of, right, ...	[What could cause a shift in the supply curve ...

Inspect keywords of a topic

In [181]:

```
topic_model_better.get_topic(topic=5)
```

Out[181]:

```
[('account', 0.08284993593106096),
 ('balance', 0.06715781480209881),
 ('current', 0.056899238670662516),
 ('current account', 0.05560458082288543),
 ('balance of', 0.046090604270794325),
 ('services', 0.04318114600371102),
 ('financial', 0.042087919607530236),
 ('balance of payments', 0.03962264355237039),
 ('of payments', 0.03962264355237039),
 ('item', 0.03870281429878804),
 ('payments', 0.03637206951479107),
 ('trade', 0.03176583372821247),
 ('net', 0.029637910751512408),
 ('account balance', 0.027145051553075405),
 ('credit', 0.026143064078806632)]
```

List of Documents, topics they belong to, keywords, related documents

In [176]:

```
topic_model_better.get_document_info(docs)
```

Out[176]:

	Document	Topic	Name	Representation	Representative_Docs
0	An economy is changing from a planned economy ...	15	15_planned_economy_market economy_market	[planned, economy, market economy, market, eco...	[Consumers in country X buy some goods and ser...
1	A local council provides a tap for drinking wa...	2	2_public_good_private_merit	[public, good, private, merit, public goods, p...	[Why does the production of public goods have ...
2	Food prices in a country increased by 20% in t...	14	14_good_of good_rise_increase	[good, of good, rise, increase, price of, fall...	[11 Goods X and Y are complements.\n\nWhat wil...
3	A doctor has very long working hours and a hig...	43	43_opportunity cost_opportunity_cost_train	[opportunity cost, opportunity, cost, train, c...	[1\n\nAn individual has an appointment with hi...
4	When will the demand curve for motorcycles shi...	26	26_public transport_transport_petrol_demand curve	[public transport, transport, petrol, demand c...	[5 Inthe diagram, D,D, shows an individual’s i...
...	...	...	...	...	...
3025	The diagram shows the long run aggregate suppl...	0	0_aggregate_price level_level_aggregate demand	[aggregate, price level, level, aggregate dema...	[The diagram shows aggregate demand (AD) and a...
3026	The government undertakes a policy of financin...	-1	-1_price_of_market_supply	[price, of, market, supply, quantity, demand, ...	[The equilibrium price of a product is $10. Th...
3027	Which combination of initial equilibrium and s...	0	0_aggregate_price level_level_aggregate demand	[aggregate, price level, level, aggregate dema...	[The diagram shows aggregate demand (AD) and a...
3028	A government aims to reduce unemployment throu...	-1	-1_price_of_market_supply	[price, of, market, supply, quantity, demand, ...	[The equilibrium price of a product is $10. Th...
3029	A government wants to use an expansionary mone...	41	41_expansionary_policy_fiscal policy_fiscal	[expansionary, policy, fiscal policy, fiscal, ...	[What would be increased by an expansionary fi...

3030 rows × 8 columns

Evaluating a document using trained models

In [69]:

```
example = ["""A taxi firm raises fares at its busiest times by as much as five times the
      and customers are notified of the changes by mobile (cell) phone.
      What will result from this policy?
      A     It will be less likely that there is a market equilibrium.
      B     Potential customers will have less perfect information.
      C     The market surplus will become a shortage.
      D     The supply of taxi rides will become more price elastic."""]
embeddings_example = sentence_model.encode(example)
topic_example, prob_example = topic_model_better.transform(example, embeddings_example)
topic_model_better.get_topic_info(topic_example[0])
```

```
Batches:   0%|          | 0/1 [00:00<?, ?it/s]
```

Out[69]:

	Topic	Count	Name	Representation	Representative_Docs
0	21	41	21_public transport_transport_petrol_demand curve	[public transport, transport, petrol, demand c...	[5 Inthe diagram, D,D, shows an individual’s i...

In [73]:

```
import gc
gc.collect()
```

Out[73]:

```
1775
```

Keyword distribution for topics

```
In  [63]: topic_model_better.visualize_barchart(top_n_topics=8, n_words = 8)
```

Hierarchical Clustering of Topics

In [72]:

```
hierarchical_topics = topic_model_better.hierarchical_topics(docs)
topic_model_better.visualize_hierarchy(hierarchical_topics=hierarchical_topics, custom_l
```

```
100%|██████████| 48/48 [00:00<00:00, 250.57it/s]
```

Layers and Grouping of topics

In [73]:

```
tree = topic_model_better.get_topic_tree(hierarchical_topics)
print(tree)
```

```
.
├──price_demand_quantity_supply_good
│   ├──elasticity_elasticity of_tax_price elasticity_elasticity of demand
│   │   ├──cross_good_cross elasticity_of demand_elasticity of demand
│   │   │   ├──■──bus_travel_rail_increase of_cross ── Topic: 48
│   │   │   └──good_cross_cross elasticity_elasticity of demand_of demand
│   │   │       ├──■──cross_cross elasticity_good_of demand_price of good
── Topic: 19
│   │   │       └──■──income_income elasticity of_income elasticity_income
elasticity of demand_good ── Topic: 34
│   │   └──tax_elasticity_price elasticity_price elasticity of_elasticity of
│   │       ├──■──tax_of tax_income_earners_000 ── Topic: 6
│   │       └──elasticity_price elasticity_price elasticity of_elasticity of_pr
ice
│   │           ├──■──point_demand_elastic_shows demand_shows demand curve
── Topic: 46
│   │           └──elasticity_price elasticity_price elasticity of_elasticity
of supply_elasticity of
│   │               ├──■──price elasticity_price elasticity of_elasticity_
elasticity of_price ── Topic: 2
│   │               └──■──elasticity_elasticity of supply_of supply_price
elasticity of supply_along ── Topic: 35
│   └──price_price of_quantity_supply_demand
│       ├──quantity_diagram_equilibrium_price_diagram shows
│       │   ├──■──equilibrium_new equilibrium_point_new_coffee ── Topic: 1
6
│       │   └──quantity_diagram_price_surplus_diagram shows
│       │       ├──surplus_diagram_quantity_producer_area
│       │       │   ├──■──surplus_producer_producer surplus_consumer_cons
umer surplus ── Topic: 14
│       │       │   └──■──subsidy_tonnes_agricultural_supply_government
── Topic: 18
│       │       └──■──maximum price_maximum_quantity_price_minimum price ─
─ Topic: 20
│       └──price of_shift_demand_price_curve
│           ├──good_market_of good_price_quantity
│           │   ├──market_10_market supply_kg_market demand
│           │   │   ├──■──individual_must_consumer_market demand_of indus
try ── Topic: 45
│           │   │   └──■──10_market_kg_20_market supply ── Topic: 32
│           │   └──good_of good_decrease_rise_increase
│           │       ├──■──disequilibrium_market_quantity_decrease increase
decrease_increase decrease ── Topic: 42
│           │       └──■──good_of good_rise_uncertain_decrease ── Topic:
15
│           └──shift_curve_price of_demand curve_supply curve
│               ├──shift_curve_price of_supply curve_demand curve
│               │   ├──■──public transport_transport_demand curve_petrol_p
ublic ── Topic: 22
│               │   └──■──shift_curve_supply curve_price of_right ── Top
ic: 7
│               └──■──houses_housing_house_building_2011 ── Topic: 33
└──of_country_goods_rate_production
    ├──production_of_cost_goods_private
    │   ├──inflation_real_aggregate_level_rate
    │   │   ├──policy_expansionary_budget_fiscal_monetary
    │   │   │   ├──■──expansionary_policy_fiscal_fiscal policy_budget ──
```

```
Topic: 38
      │     │     │     └──■───increasing_budget_deflation_policy_fiscal ──── Topi
c: 44
      │     │     └──inflation_real_aggregate_level_price level
      │     │          ├──inflation_aggregate_real_level_price level
      │     │          │     ├──■───aggregate_price level_level_aggregate demand_re
al ──── Topic: 1
      │     │          │     └──inflation_of inflation_index_rate_rate of inflation
      │     │          │          ├──■───inflation_rising_falling_increase_likely
──── Topic: 10
      │     │          │          └──■───inflation_index_of inflation_rose_rate of
inflation ──── Topic: 3
      │     │          └──unemployment_normative_statement_positive_23
      │     │               ├──■───unemployment_23_employment_unemployed_france ─
─ Topic: 23
      │     │               └──■───normative_statement_positive_economic_unemployme
nt ──── Topic: 25
      │     └──production_private_goods_cost_advantage
      │          ├──production_advantage_possibility_production possibility_country
      │          │     ├──payment_transfer_money_of money_paid
      │          │     │     ├──■───payment_transfer_paid_government_made ──── Topi
c: 27
      │          │     │     └──■───of money_money_medium of_medium_medium of excha
nge ──── Topic: 29
      │          │     └──production_advantage_possibility_production possibility_cou
ntry
      │          │          ├──production_advantage_possibility_production possibilit
y_country
      │          │          │     ├──■───opportunity cost_opportunity_cost_train_co
st of ──── Topic: 40
      │          │          │     └──■───production_advantage_possibility_productio
n possibility_country ──── Topic: 0
      │          │          └──labour_enterprise_capital_land_of production
      │          │               ├──■───of labour_labour_workers_productivity_worke
r ──── Topic: 43
      │          │               └──■───enterprise_land_capital_of production_labou
r ──── Topic: 31
      │          └──private_benefits_public_external_social
      │               ├──private_benefits_external_social_costs
      │               │     ├──■───external_social_benefits_costs_private ──── Topi
c: 5
      │               │     └──public_private_good_merit_sector
      │               │          ├──■───public_good_private_merit_provided ──── Top
ic: 4
      │               │          └──■───private_sector_industry_state_public ──── T
opic: 24
      │               └──resources_economy_mechanism_goods_market
      │                    ├──economy_market economy_resources_economic_scarcity
      │                    │     ├──■───economy_market economy_market_economic_econ
omies ──── Topic: 13
      │                    │     └──■───scarcity_resources_problem_economic_allocat
ion ──── Topic: 39
      │                    └──■───mechanism_price mechanism_rationing_resources_fun
ction ──── Topic: 26
      └──us_trade_account_exchange_balance
           ├──account_current account_current_balance_balance of payments
           │     ├──account_balance_current_current account_item
           |     |     ├─■──balance_services_account_account balance_current acco
unt ── Topic: 30
           |     |     └─■──account_item_financial_current_credit ── Topic: 28
           |     └─currency_current account_current_of payments_balance of payments
           |          ├─current account_deficit_current_of payments_balance of payme
nts
           |          |     ├─■──policy_deficit_of payments_balance of payments_p
ayments ── Topic: 12
           |          |     └─■──current account_current_account_deficit_of payme
nts ── Topic: 21
           |          └─■──currency_exchange_exchange rate_floating_floating exch
ange ── Topic: 17
           └─trade_us_tariff_of trade_against
                ├─us_of trade_against_terms of_terms
                |     ├─us_against_uk_exchange_exchange rate
                |     |     ├─■──us_uk_against_exchange_of us ── Topic: 8
                |     |     └─■──of country_trade_country_value_country xs ── T
opic: 47
                |     └─terms of trade_terms of_of trade_prices_terms
                |          ├─■──index_index of_terms of trade_terms of_terms ──
Topic: 37
                |          └─■──prices_terms of trade_of trade_terms of_terms ─
─ Topic: 11
                └─tariff_domestic_world_trade_quota
                     ├─domestic_tariff_world_quota_imports
                     |     ├─■──domestic_tariff_world_of tariff_supply ── Topi
c: 36
                     |     └─■──quota_domestic_imported_protection_tariff ── To
pic: 9
                     └─■──union_free trade_free_trade_countries ── Topic: 41
```

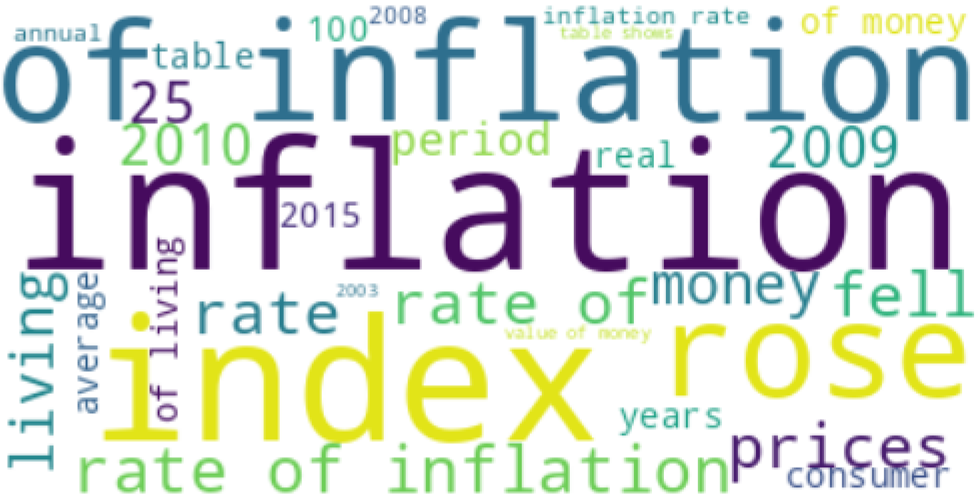
Wordcloud to visualize keywords in a topic

In [75]:

```
from wordcloud import WordCloud
import matplotlib.pyplot as plt

def create_wordcloud(model, topic):
    text = {word: value for word, value in model.get_topic(topic)}
    wc = WordCloud(background_color="white", max_words=1000)
    wc.generate_from_frequencies(text)
    plt.imshow(wc, interpolation="bilinear")
    plt.axis("off")
    plt.show()

# Show wordcloud
create_wordcloud(topic_model_better, topic=3)
```



Visualize Clustering of the orginal embeddings

In [80]:

```
%%time
# Run the visualization with the original embeddings
topic_model_better.visualize_documents(docs, embeddings=embeddings, hide_annotations=Tr
```

```
CPU times: user 20.3 s, sys: 476 ms, total: 20.8 s
Wall time: 13 s
```

Topic Probability Distribution for a document

In [83]:

```
topic_distr, _ = topic_model_better.approximate_distribution(docs)

print(docs[1])
```

```
A local council provides a tap for drinking water in a town.     Would this make drin
king water a free good?     A     No, because it is possible to exclude some people f
rom using the tap.     B     No, because it requires the use of scarce resources.
C     Yes, because it is available to all passers by.     D     Yes, because it is im
possible to charge for it.
```

In [84]:

```
topic_model_better.visualize_distribution(topic_distr[1], width=1000, custom_labels=Tru
```

Identify relative importance of words in determining topics for a document

In [101]:

```
topic_distr, topic_token_distr = topic_model_better.approximate_distribution(docs, calcu

df = topic_model_better.visualize_approximate_distribution(docs[1], topic_token_distr[1
df
```

Out[101]:

	local	council	provides	tap	for	drinking
0_production_advantage_possibility_production possibility	0.000	0.000	0.000	0.000	0.000	0.000
4_public_good_private_merit	0.107	0.107	0.107	0.107	0.000	0.000
7_shift_curve_supply curve_price of	0.000	0.000	0.000	0.000	0.000	0.000
13_economy_market economy_market_economic	0.000	0.000	0.000	0.000	0.000	0.000
15_good_of good_rise_uncertain	0.000	0.000	0.000	0.000	0.000	0.000
19_cross_cross elasticity_good_of demand	0.000	0.000	0.000	0.000	0.000	0.000
20_maximum price_maximum_quantity_price	0.000	0.000	0.000	0.000	0.000	0.000
23_unemployment_23_employment_unemployed	0.000	0.000	0.000	0.000	0.000	0.000
26_mechanism_price mechanism_rationing_resources	0.000	0.000	0.000	0.000	0.000	0.000
34_income_income elasticity of_income elasticity_income elasticity of demand	0.000	0.000	0.000	0.000	0.000	0.000
39_scarcity_resources_problem_economic	0.000	0.000	0.000	0.000	0.000	0.000
41_union_free trade_free_trade	0.000	0.000	0.000	0.000	0.000	0.000
42_disequilibrium_market_quantity_decrease increase decrease	0.000	0.000	0.000	0.000	0.000	0.000
43_of labour_labour_workers_productivity	0.000	0.000	0.000	0.000	0.000	0.000