```
!pip install bertopic
```

```
Requirement already satisfied: joblib>=1.0 in /usr/local/lib/python3.10/dist-packages (from hdbscan>=0.8.29->bertopic) (1.3.2)
Requirement already satisfied: python-dateutil>=2.8.1 in /usr/local/lib/python3.10/dist-packages (from pandas>=1.1.5->bertopic) (
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas>=1.1.5->bertopic) (2023.3.pos
Requirement already satisfied: tenacity>=6.2.0 in /usr/local/lib/python3.10/dist-packages (from plotly>=4.7.0->bertopic) (8.2.3)
Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-packages (from plotly>=4.7.0->bertopic) (23.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.10/dist-packages (from scikit-learn>=0.22.2.post1->
Requirement already satisfied: transformers<5.0.0,>=4.6.0 in /usr/local/lib/python3.10/dist-packages (from sentence-transformers>
Requirement already satisfied: torch>=1.6.0 in /usr/local/lib/python3.10/dist-packages (from sentence-transformers>=0.4.1->bertop
Requirement already satisfied: torchvision in /usr/local/lib/python3.10/dist-packages (from sentence-transformers>=0.4.1->bertopi
Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-packages (from sentence-transformers>=0.4.1->bertopic) (3.8
Collecting sentencepiece (from sentence-transformers>=0.4.1->bertopic)
  Downloading sentencepiece-0.1.99-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (1.3 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.3/1.3 MB 47.9 MB/s eta 0:00:00
Requirement already satisfied: huggingface-hub>=0.4.0 in /usr/local/lib/python3.10/dist-packages (from sentence-transformers>=0.4
Requirement already satisfied: numba>=0.51.2 in /usr/local/lib/python3.10/dist-packages (from umap-learn>=0.5.0->bertopic) (0.58.
Collecting pynndescent>=0.5 (from umap-learn>=0.5.0->bertopic)
  Downloading pynndescent-0.5.11-py3-none-any.whl (55 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 55.8/55.8 kB 6.6 MB/s eta 0:00:00
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.4.0->sentence-transfo
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.4.0->sentence
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.4.0->sentence-transfo
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.4.0->sentence-tran
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.4.0
Requirement already satisfied: llvmlite<0.42,>=0.41.0dev0 in /usr/local/lib/python3.10/dist-packages (from numba>=0.51.2->umap-le
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.8.1->pandas>=1.1.5->b
Requirement already satisfied: sympy in /usr/local/lib/python3.10/dist-packages (from torch>=1.6.0->sentence-transformers>=0.4.1-
Requirement already satisfied: networkx in /usr/local/lib/python3.10/dist-packages (from torch>=1.6.0->sentence-transformers>=0.4
Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-packages (from torch>=1.6.0->sentence-transformers>=0.4.1
Requirement already satisfied: triton==2.1.0 in /usr/local/lib/python3.10/dist-packages (from torch>=1.6.0->sentence-transformers
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.10/dist-packages (from transformers<5.0.0,>=4.6.0->sen
Requirement already satisfied: tokenizers<0.19,>=0.14 in /usr/local/lib/python3.10/dist-packages (from transformers<5.0.0,>=4.6.0
Requirement already satisfied: safetensors>=0.3.1 in /usr/local/lib/python3.10/dist-packages (from transformers<5.0.0,>=4.6.0->se
Requirement already satisfied: click in /usr/local/lib/python3.10/dist-packages (from nltk->sentence-transformers>=0.4.1->bertopi
Requirement already satisfied: pillow!=8.3.*,>=5.3.0 in /usr/local/lib/python3.10/dist-packages (from torchvision->sentence-trans
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.10/dist-packages (from jinja2->torch>=1.6.0->sentence-tr
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests->huggingface-hu
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests->huggingface-hub>=0.4.0->se
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests->huggingface-hub>=0.4
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests->huggingface-hub>=0.4
Requirement already satisfied: mpmath>=0.19 in /usr/local/lib/python3.10/dist-packages (from sympy->torch>=1.6.0->sentence-transf
Building wheels for collected packages: hdbscan, sentence-transformers, umap-learn
  Building wheel for hdbscan (pyproject.toml) ... done
  Created wheel for hdbscan: filename=hdbscan-0.8.33-cp310-cp310-linux_x86_64.whl size=3039180 sha256=ace212252b7c218aab6c1e11a04
  Stored in directory: /root/.cache/pip/wheels/75/0b/3b/dc4f60b7cc455efaefb62883a7483e76f09d06ca81cf87d610
  Building wheel for sentence-transformers (setup.py) ... done
  Created wheel for sentence-transformers: filename=sentence_transformers-2.2.2-py3-none-any.whl size=125923 sha256=57dea818f5e47
  Stored in directory: /root/.cache/pip/wheels/62/f2/10/1e606fd5f02395388f74e7462910fe851042f97238cbbd902f
  Building wheel for umap-learn (setup.py) ... done
  Created wheel for umap-learn: filename=umap_learn-0.5.5-py3-none-any.whl size=86832 sha256=ee5f04cb415dfb2609f094733117bce5519f
  Stored in directory: /root/.cache/pip/wheels/3a/70/07/428d2b58660a1a3b431db59b806a10da736612ebbc66c1bcc5
Successfully built hdbscan sentence-transformers umap-learn
Installing collected packages: sentencepiece, cython, pynndescent, hdbscan, umap-learn, sentence-transformers, bertopic
  Attempting uninstall: cython
    Found existing installation: Cython 3.0.6
    Uninstalling Cython-3.0.6:
      Successfully uninstalled Cython-3.0.6
Successfully installed bertopic-0.16.0 cython-0.29.36 hdbscan-0.8.33 pynndescent-0.5.11 sentence-transformers-2.2.2 sentencepiece
```

```
!pip install bertopic[visualization]
```

```
Requirement already satisfied: bertopic[visualization] in /usr/local/lib/python3.10/dist-packages (0.16.0)
WARNING: bertopic 0.16.0 does not provide the extra 'visualization'
Requirement already satisfied: numpy>=1.20.0 in /usr/local/lib/python3.10/dist-packages (from bertopic[visualization]) (1.23.5)
Requirement already satisfied: hdbscan>=0.8.29 in /usr/local/lib/python3.10/dist-packages (from bertopic[visualization]) (0.8.33)
Requirement already satisfied: umap-learn>=0.5.0 in /usr/local/lib/python3.10/dist-packages (from bertopic[visualization]) (0.5.5)
Requirement already satisfied: pandas>=1.1.5 in /usr/local/lib/python3.10/dist-packages (from bertopic[visualization]) (1.5.3)
Requirement already satisfied: scikit-learn>=0.22.2.post1 in /usr/local/lib/python3.10/dist-packages (from bertopic[visualization])
Requirement already satisfied: tqdm>=4.41.1 in /usr/local/lib/python3.10/dist-packages (from bertopic[visualization]) (4.66.1)
Requirement already satisfied: sentence-transformers>=0.4.1 in /usr/local/lib/python3.10/dist-packages (from bertopic[visualization
Requirement already satisfied: plotly>=4.7.0 in /usr/local/lib/python3.10/dist-packages (from bertopic[visualization]) (5.15.0)
Requirement already satisfied: cython<3,>=0.27 in /usr/local/lib/python3.10/dist-packages (from hdbscan>=0.8.29->bertopic[visualizat
Requirement already satisfied: scipy>=1.0 in /usr/local/lib/python3.10/dist-packages (from hdbscan>=0.8.29->bertopic[visualization]
Requirement already satisfied: joblib>=1.0 in /usr/local/lib/python3.10/dist-packages (from hdbscan>=0.8.29->bertopic[visualization
Requirement already satisfied: python-dateutil>=2.8.1 in /usr/local/lib/python3.10/dist-packages (from pandas>=1.1.5->bertopic[visua
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas>=1.1.5->bertopic[visualization]
Requirement already satisfied: tenacity>=6.2.0 in /usr/local/lib/python3.10/dist-packages (from plotly>=4.7.0->bertopic[visualizatic
Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-packages (from plotly>=4.7.0->bertopic[visualization]) (2
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.10/dist-packages (from scikit-learn>=0.22.2.post1->ber
Requirement already satisfied: transformers<5.0.0,>=4.6.0 in /usr/local/lib/python3.10/dist-packages (from sentence-transformers>=0
Requirement already satisfied: torch>=1.6.0 in /usr/local/lib/python3.10/dist-packages (from sentence-transformers>=0.4.1->bertopic
Requirement already satisfied: torchvision in /usr/local/lib/python3.10/dist-packages (from sentence-transformers>=0.4.1->bertopic[v
Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-packages (from sentence-transformers>=0.4.1->bertopic[visuali
Requirement already satisfied: sentencepiece in /usr/local/lib/python3.10/dist-packages (from sentence-transformers>=0.4.1->bertopic
Requirement already satisfied: huggingface-hub>=0.4.0 in /usr/local/lib/python3.10/dist-packages (from sentence-transformers>=0.4.1
Requirement already satisfied: numba>=0.51.2 in /usr/local/lib/python3.10/dist-packages (from umap-learn>=0.5.0->bertopic[visualizat
Requirement already satisfied: pynndescent>=0.5 in /usr/local/lib/python3.10/dist-packages (from umap-learn>=0.5.0->bertopic[visuali
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.4.0->sentence-transforme
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.4.0->sentence-tr
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.4.0->sentence-transforme
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.4.0->sentence-transfo
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.4.0->s
Requirement already satisfied: llvmlite<0.42,>=0.41.0dev0 in /usr/local/lib/python3.10/dist-packages (from numba>=0.51.2->umap-lear
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.8.1->pandas>=1.1.5->bert
Requirement already satisfied: sympy in /usr/local/lib/python3.10/dist-packages (from torch>=1.6.0->sentence-transformers>=0.4.1->be
Requirement already satisfied: networkx in /usr/local/lib/python3.10/dist-packages (from torch>=1.6.0->sentence-transformers>=0.4.1
Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-packages (from torch>=1.6.0->sentence-transformers>=0.4.1->b
Requirement already satisfied: triton==2.1.0 in /usr/local/lib/python3.10/dist-packages (from torch>=1.6.0->sentence-transformers>=
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.10/dist-packages (from transformers<5.0.0,>=4.6.0->senten
Requirement already satisfied: tokenizers<0.19,>=0.14 in /usr/local/lib/python3.10/dist-packages (from transformers<5.0.0,>=4.6.0->s
Requirement already satisfied: safetensors>=0.3.1 in /usr/local/lib/python3.10/dist-packages (from transformers<5.0.0,>=4.6.0->sente
Requirement already satisfied: click in /usr/local/lib/python3.10/dist-packages (from nltk->sentence-transformers>=0.4.1->bertopic[v
Requirement already satisfied: pillow!=8.3.*,>=5.3.0 in /usr/local/lib/python3.10/dist-packages (from torchvision->sentence-transfor
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.10/dist-packages (from jinja2->torch>=1.6.0->sentence-trans
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests->huggingface-hub>=
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests->huggingface-hub>=0.4.0->sente
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests->huggingface-hub>=0.4.0
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests->huggingface-hub>=0.4.0
Requirement already satisfied: mpmath>=0.19 in /usr/local/lib/python3.10/dist-packages (from sympy->torch>=1.6.0->sentence-transform
```

```
import pandas as pd
import numpy as np
import json
from bertopic import BERTopic
```

```
# Opening JSON file
f = open('complaints-2021-05-14_08_16_.json')
datafile = json.load(f)
df = pd.json_normalize(datafile)
```

```
df.head()
```

	_index	_type	_id	_score	_source.tags	_source.zip_code	_source.complaint_id	_source.issue	_source.date_received
0	complaint-public-v2	complaint	3211475	0.0	None	90301	3211475	Attempts to collect debt not owed	2019-04-13T12:00:00-05:00
1	complaint-public-v2	complaint	3229299	0.0	Servicemember	319XX	3229299	Written notification about debt	2019-05-01T12:00:00-05:00
2	complaint-public-v2	complaint	3199379	0.0	None	77069	3199379	Other features, terms, or problems	2019-04-02T12:00:00-05:00
3	complaint-public-v2	complaint	2673060	0.0	None	48066	2673060	Trouble during payment process	2017-09-13T12:00:00-05:00
4	complaint-public-v2	complaint	3203545	0.0	None	10473	3203545	Fees or interest	2019-04-05T12:00:00-05:00

5 rows × 22 columns

```
#Assign nan in place of blanks in the body column
df[df.loc[:, '_source.complaint_what_happened'] == ''] = np.nan
```

```
# Check if blank values still exist
df[df.loc[:, '_source.complaint_what_happened'] == '']
```

	_index	_type	_id	_score	_source.tags	_source.zip_code	_source.complaint_id	_source.issue	_source.date_received	_source.st

0 rows × 22 columns

```
df.shape
```

```
(78313, 22)
```

```
#Remove all rows where body column is nan
df = df[~df['_source.complaint_what_happened'].isnull()]
```

```
df.shape
```

```
(21072, 22)
```

```python
# Convert body column to string for performing text operations
df['_source.complaint_what_happened'] = df['_source.complaint_what_happened'].astype(str)
```

```python
# Write your function here to clean the text and remove all the unnecessary elements.
def clean_text(sent):
    sent = sent.lower() # Text to lowercase
    pattern = '[^\w\s]' # Removing punctuation
    sent = re.sub(pattern, '', sent)
    pattern = '\w*\d\w*' # Removing words with numbers in between
    sent = re.sub(pattern, '', sent)
    return sent
```

```python
import re
df_clean = pd.DataFrame(df['_source.complaint_what_happened'].apply(clean_text))
# df_clean.columns = ['complaint_what_happened']
```

```python
df_clean.head()
```

	_source.complaint_what_happened
1	good morning my name is xxxx xxxx and i apprec...
2	i upgraded my xxxx xxxx card in and was told ...
10	chase card was reported on however fraudulent...
11	on while trying to book a xxxx xxxx ticket ...
14	my grand son give me check for i deposit it i...

```python
# create model
bert_model = BERTopic(verbose=True)
#convert to list
docs = df_clean['_source.complaint_what_happened'].to_list()
```

```python
#bert_model.fit_transform(docs)
topics, probabilities = bert_model.fit_transform(docs)
```

```
2023-12-12 18:16:04,432 - BERTopic - Embedding - Transforming documents to embeddings.
.gitattributes: 100%                                  1.18k/1.18k [00:00<00:00, 48.8kB/s]
1_Pooling/config.json: 100%                           190/190 [00:00<00:00, 1.73kB/s]
README.md: 100%                                       10.6k/10.6k [00:00<00:00, 169kB/s]
config.json: 100%                                     612/612 [00:00<00:00, 6.95kB/s]
config_sentence_transformers.json: 100%               116/116 [00:00<00:00, 918B/s]
data_config.json: 100%                                39.3k/39.3k [00:00<00:00, 418kB/s]
pytorch_model.bin: 100%                               90.9M/90.9M [00:01<00:00, 92.0MB/s]
sentence_bert_config.json: 100%                       53.0/53.0 [00:00<00:00, 453B/s]
special_tokens_map.json: 100%                         112/112 [00:00<00:00, 1.20kB/s]
tokenizer.json: 100%                                  466k/466k [00:00<00:00, 1.89MB/s]
tokenizer_config.json: 100%                           350/350 [00:00<00:00, 4.66kB/s]
train_script.py: 100%                                 13.2k/13.2k [00:00<00:00, 160kB/s]
vocab.txt: 100%                                       232k/232k [00:00<00:00, 1.41MB/s]
modules.json: 100%                                    349/349 [00:00<00:00, 3.33kB/s]
Batches: 100%                                         659/659 [53:44<00:00, 3.20it/s]
2023-12-12 19:10:02,449 - BERTopic - Embedding - Completed ✓
2023-12-12 19:10:02,454 - BERTopic - Dimensionality - Fitting the dimensionality reduction algorithm
2023-12-12 19:11:01,722 - BERTopic - Dimensionality - Completed ✓
2023-12-12 19:11:01,726 - BERTopic - Cluster - Start clustering the reduced embeddings
2023-12-12 19:11:06,211 - BERTopic - Cluster - Completed ✓
2023-12-12 19:11:06,232 - BERTopic - Representation - Extracting topics from clusters using representation models.
2023-12-12 19:11:17,159 - BERTopic - Representation - Completed ✓
```

```python
import gensim.corpora as corpora
from gensim.models.coherencemodel import CoherenceModel

# Preprocess documents
```

```
cleaned_docs = bert_model._preprocess_text(docs)

# Extract vectorizer and tokenizer from BERTopic
vectorizer = bert_model.vectorizer_model
tokenizer = vectorizer.build_tokenizer()

# Extract features for Topic Coherence evaluation
words = vectorizer.get_feature_names_out()
tokens = [tokenizer(doc) for doc in cleaned_docs]
dictionary = corpora.Dictionary(tokens)
corpus = [dictionary.doc2bow(token) for token in tokens]
topic_words = [[words for words, _ in bert_model.get_topic(topic)]
               for topic in range(len(set(topics))-1)]

# Evaluate
coherence_model = CoherenceModel(topics=topic_words,
                                 texts=tokens,
                                 corpus=corpus,
                                 dictionary=dictionary,
                                 coherence='c_v')
coherence = coherence_model.get_coherence()
coherence
```

```
0.5304336847668605
```

```
# Evaluate
u_mass_coherence_model = CoherenceModel(topics=topic_words,
                                 texts=tokens,
                                 corpus=corpus,
                                 dictionary=dictionary,
                                 coherence='u_mass')
u_mass_coherence = u_mass_coherence_model.get_coherence()
u_mass_coherence
```

```
-2.0991738393963684
```

```
# Evaluate
c_uci_coherence_model = CoherenceModel(topics=topic_words,
                                 texts=tokens,
                                 corpus=corpus,
                                 dictionary=dictionary,
                                 coherence='c_uci')
c_uci_coherence = c_uci_coherence_model.get_coherence()
c_uci_coherence
```

```
-0.16214295288344877
```

```
# Evaluate
c_npmi_coherence_model = CoherenceModel(topics=topic_words,
                                 texts=tokens,
                                 corpus=corpus,
                                 dictionary=dictionary,
                                 coherence='c_npmi')
c_npmi_coherence = c_npmi_coherence_model.get_coherence()
c_npmi_coherence
```

```
0.052055372170235786
```

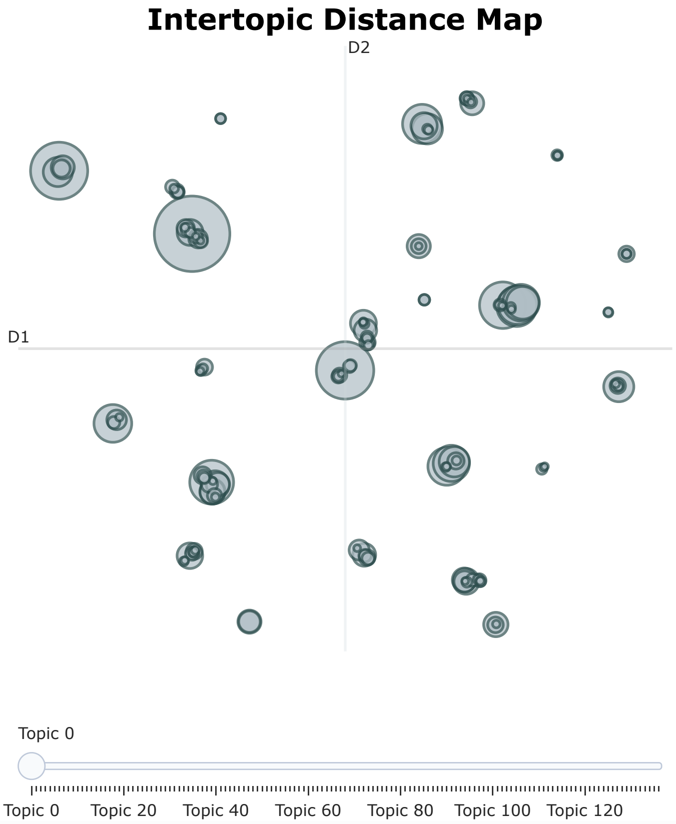
```
bert_model.get_topic_freq().head(11)
```

	Topic	Count
0	-1	9239
24	0	1214
7	1	702
14	2	695
9	3	461
53	4	409
8	5	336
42	6	326
52	7	322
1	8	304
13	9	303

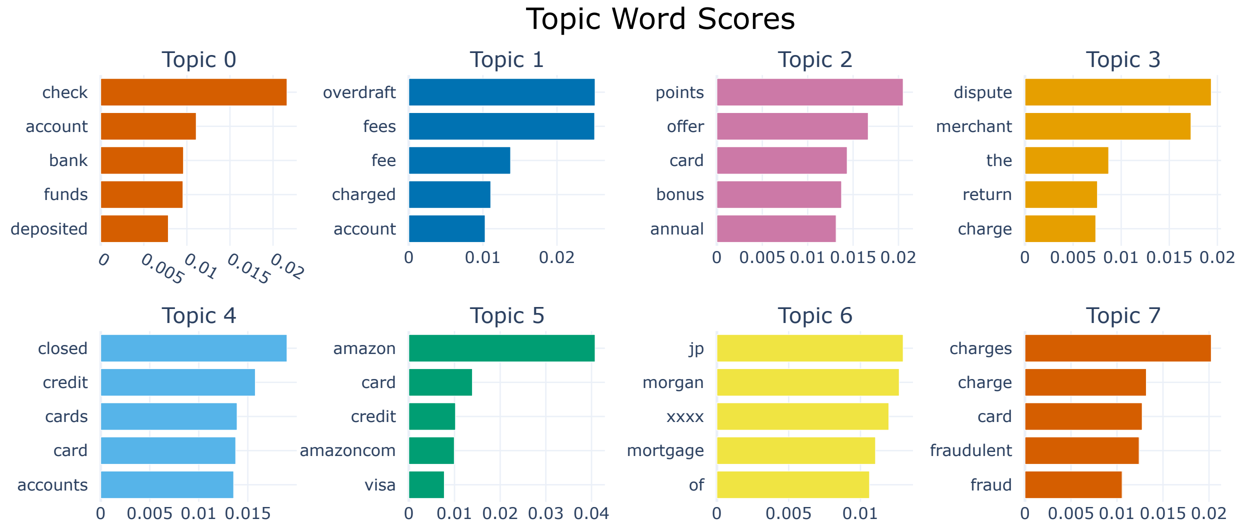
```
bert_model.get_topic(7)
```

```
[('charges', 0.020244036713973217),
 ('charge', 0.013185950293615205),
 ('card', 0.012735738048743625),
 ('fraudulent', 0.012402159623782147),
 ('fraud', 0.010552823126942617),
 ('xxxx', 0.007860165663693287),
 ('that', 0.007591016239874271),
 ('on', 0.007371886854056404),
 ('chase', 0.007042722292726619),
 ('were', 0.006956869461476829)]
```

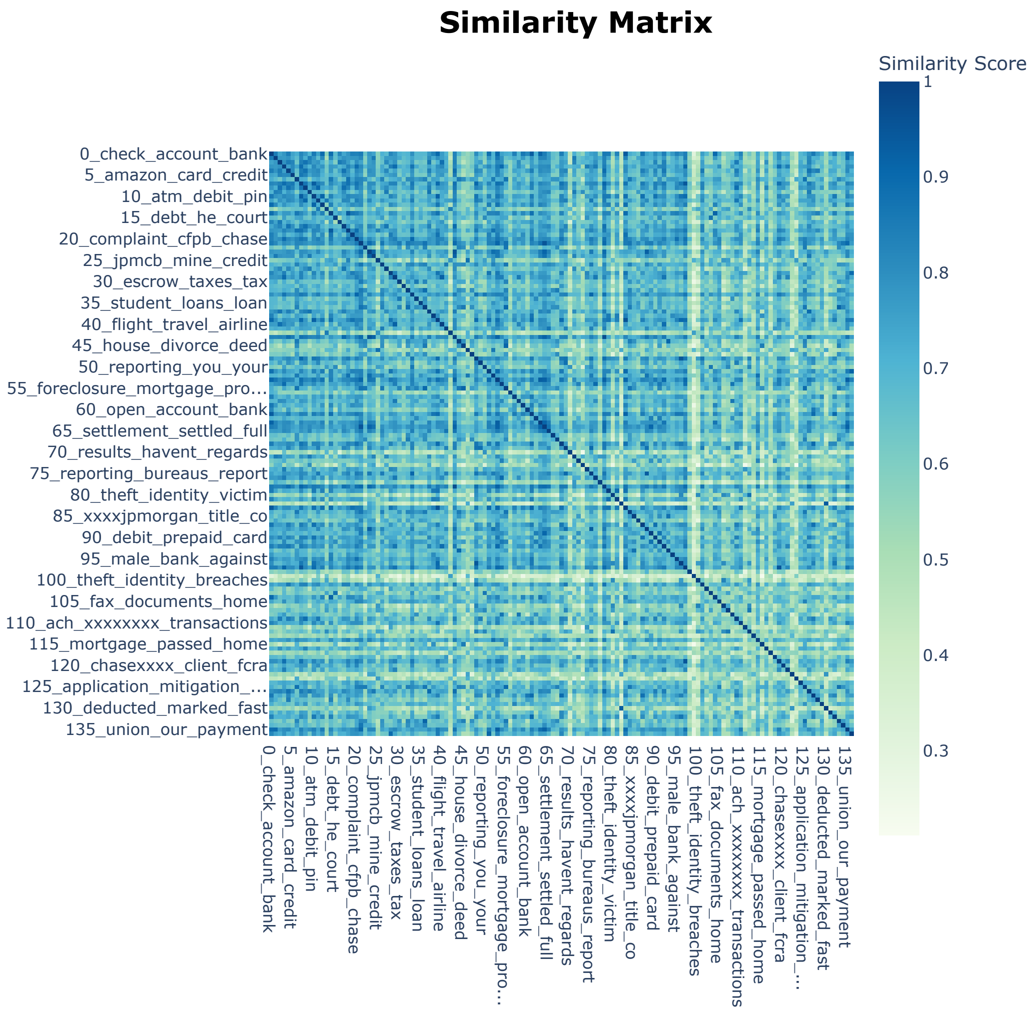
```
bert_model.visualize_topics()
```



```
bert_model.visualize_barchart()
```



```
bert_model.visualize_heatmap()
```



```
bert_model.save("complaints_bertmodel")
```

```
2023-12-12 19:22:43,081 - BERTopic - WARNING: When you use `pickle` to save/load a BERTopic model,please make sure that the environm
```

Start coding or generate with AI.