

```
!pip install bertopic
```



Collecting bertopic

Downloading bertopic-0.16.0-py2.py3-none-any.whl (154 kB)

154.1/154.1 kB 2.9 MB/s eta 0:00:00

Requirement already satisfied: numpy>=1.20.0 in /usr/local/lib/python3.10/dist-packages

Collecting hdbscan>=0.8.29 (from bertopic)

Downloading hdbscan-0.8.33.tar.gz (5.2 MB)

5.2/5.2 MB 56.0 MB/s eta 0:00:00

Installing build dependencies ... done

Getting requirements to build wheel ... done

Preparing metadata (pyproject.toml) ... done

Collecting umap-learn>=0.5.0 (from bertopic)

Downloading umap-learn-0.5.5.tar.gz (90 kB)

90.9/90.9 kB 13.4 MB/s eta 0:00:00

Preparing metadata (setup.py) ... done

Requirement already satisfied: pandas>=1.1.5 in /usr/local/lib/python3.10/dist-packages

Requirement already satisfied: scikit-learn>=0.22.2.post1 in /usr/local/lib/python3.10/dist-packages

Requirement already satisfied: tqdm>=4.41.1 in /usr/local/lib/python3.10/dist-packages

Collecting sentence-transformers>=0.4.1 (from bertopic)

Downloading sentence-transformers-2.2.2.tar.gz (85 kB)

86.0/86.0 kB 12.5 MB/s eta 0:00:00

Preparing metadata (setup.py) ... done

Requirement already satisfied: plotly>=4.7.0 in /usr/local/lib/python3.10/dist-packages

Collecting cython<3,>=0.27 (from hdbscan>=0.8.29->bertopic)

Using cached Cython-0.29.36-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (1.3 MB)

Requirement already satisfied: scipy>=1.0 in /usr/local/lib/python3.10/dist-packages

Requirement already satisfied: joblib>=1.0 in /usr/local/lib/python3.10/dist-packages

Requirement already satisfied: python-dateutil>=2.8.1 in /usr/local/lib/python3.10/dist-packages

Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages

Requirement already satisfied: tenacity>=6.2.0 in /usr/local/lib/python3.10/dist-packages

Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-packages

Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.10/dist-packages

Requirement already satisfied: transformers<5.0.0,>=4.6.0 in /usr/local/lib/python3.10/dist-packages

Requirement already satisfied: torch>=1.6.0 in /usr/local/lib/python3.10/dist-packages

Requirement already satisfied: torchvision in /usr/local/lib/python3.10/dist-packages

Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-packages (from sentence-transformers>=0.4.1->bertopic)

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 84.5 MB/s eta 0:00:00

Requirement already satisfied: huggingface-hub>=0.4.0 in /usr/local/lib/python3.10/dist-packages

Requirement already satisfied: numba>=0.51.2 in /usr/local/lib/python3.10/dist-packages

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 8.0 MB/s eta 0:00:00

Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-packages

Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.10/dist-packages

Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages

Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-packages

Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.10/dist-packages

Requirement already satisfied: llvmlite<0.42,>=0.41.0dev0 in /usr/local/lib/python3.10/dist-packages

Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages

Requirement already satisfied: sympy in /usr/local/lib/python3.10/dist-packages (from sentence-transformers>=0.4.1->bertopic)

Requirement already satisfied: networkx in /usr/local/lib/python3.10/dist-packages

Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-packages (from sentence-transformers>=0.4.1->bertopic)

Requirement already satisfied: triton==2.1.0 in /usr/local/lib/python3.10/dist-packages

Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.10/dist-packages

Requirement already satisfied: tokenizers<0.19,>=0.14 in /usr/local/lib/python3.10/dist-packages

Requirement already satisfied: safetensors>0.3.1 in /usr/local/lib/python3.10/dist-packages


!pip install bertopic[visualization]

```
➡ Requirement already satisfied: bertopic[visualization] in /usr/local/lib/python3.10/dist-packages
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
Requirement already satisfied: hdbscan>=0.8.29 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: umap-learn>=0.5.0 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: pandas>=1.1.5 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: scikit-learn>=0.22.2.post1 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: tqdm>=4.41.1 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: sentence-transformers>=0.4.1 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: plotly>=4.7.0 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: cython<3,>=0.27 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: scipy>=1.0 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: joblib>=1.0 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: python-dateutil>=2.8.1 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: tenacity>=6.2.0 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-packages (from bertopic[visualization])
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: transformers<5.0.0,>=4.6.0 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: torch>=1.6.0 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: torchvision in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-packages (from bertopic[visualization])
Requirement already satisfied: sentencepiece in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: huggingface-hub>=0.4.0 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: numba>=0.51.2 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: pynndescent>=0.5 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-packages (from bertopic[visualization])
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (from bertopic[visualization])
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: llvmlite<0.42,>=0.41.0dev0 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from bertopic[visualization])
Requirement already satisfied: sympy in /usr/local/lib/python3.10/dist-packages (from bertopic[visualization])
Requirement already satisfied: networkx in /usr/local/lib/python3.10/dist-packages (from bertopic[visualization])
Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-packages (from bertopic[visualization])
Requirement already satisfied: triton==2.1.0 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: tokenizers<0.19,>=0.14 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: safetensors>=0.3.1 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: click in /usr/local/lib/python3.10/dist-packages (from bertopic[visualization])
Requirement already satisfied: pillow!=8.3.*,>=5.3.0 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: mpmath>=0.19 in /usr/local/lib/python3.10/dist-packages
```


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

```
df=pd.read_csv("all_tickets.csv")
```

```
df.head()
```



	title	body	ticket_type	category	sub_category1	sub_category2	business
0	NaN	hi since recruiter lead permission approve req... icon dear .	1	4	2	21	



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

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



	title	body	ticket_type	category	sub_category1	sub_category2	business_service
--	-------	------	-------------	----------	---------------	---------------	------------------




```
df.shape
```



```
(48549, 9)
```

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

```
df.shape
```



```
(48549, 9)
```

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

```
# 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
```

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

```
df_clean.head()
```



	body
0	hi since recruiter lead permission approve req...
1	icon dear please setup icon per icon engineers...
2	work experience user hi work experience studen...
3	requesting meeting hi please help follow equip...
4	re expire days hi ask help update passwords co...

```
# create model
bert_model = BERTopic(verbose=True)
#convert to list
docs = df_clean.body.to_list()

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

2023-12-09 11:02:21,226 - BERTopic - Embedding - Transforming documents to embeddings	
.gitattributes: 100%	1.18k/1.18k [00:00<00:00, 40.5kB/s]
1_Pooling/config.json:	190/190 [00:00<00:00,
100%	5.96kB/s]
README.md: 100%	10.6k/10.6k [00:00<00:00, 366kB/s]
config.json: 100%	612/612 [00:00<00:00, 19.6kB/s]
config_sentence_transformers.json:	116/116 [00:00<00:00,
100%	5.06kB/s]
data_config.json:	39.3k/39.3k [00:00<00:00,
100%	945kB/s]
pytorch_model.bin:	90.9M/90.9M [00:00<00:00,
100%	133MB/s]
sentence_bert_config.json:	53.0/53.0 [00:00<00:00,
100%	2.98kB/s]
special_tokens_map.json:	112/112 [00:00<00:00,
100%	6.42kB/s]
tokenizer.json: 100%	466k/466k [00:00<00:00, 4.94MB/s]
tokenizer_config.json:	350/350 [00:00<00:00,
100%	14.1kB/s]
train_script.py: 100%	13.2k/13.2k [00:00<00:00, 436kB/s]
vocab.txt: 100%	232k/232k [00:00<00:00, 2.75MB/s]

```

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
```

➡ 0.49108100127495175

```
# 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
```

➡ -5.797941219415932

```
# 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
```

➡ -1.9587533293244106

```
# 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
```

➡ 0.036257978391834346

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



	Topic	Count
1	-1	15232
90	0	1207
113	1	713
93	2	649
25	3	524
77	4	522
43	5	449
319	6	421
10	7	421
172	8	406
62	9	363

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

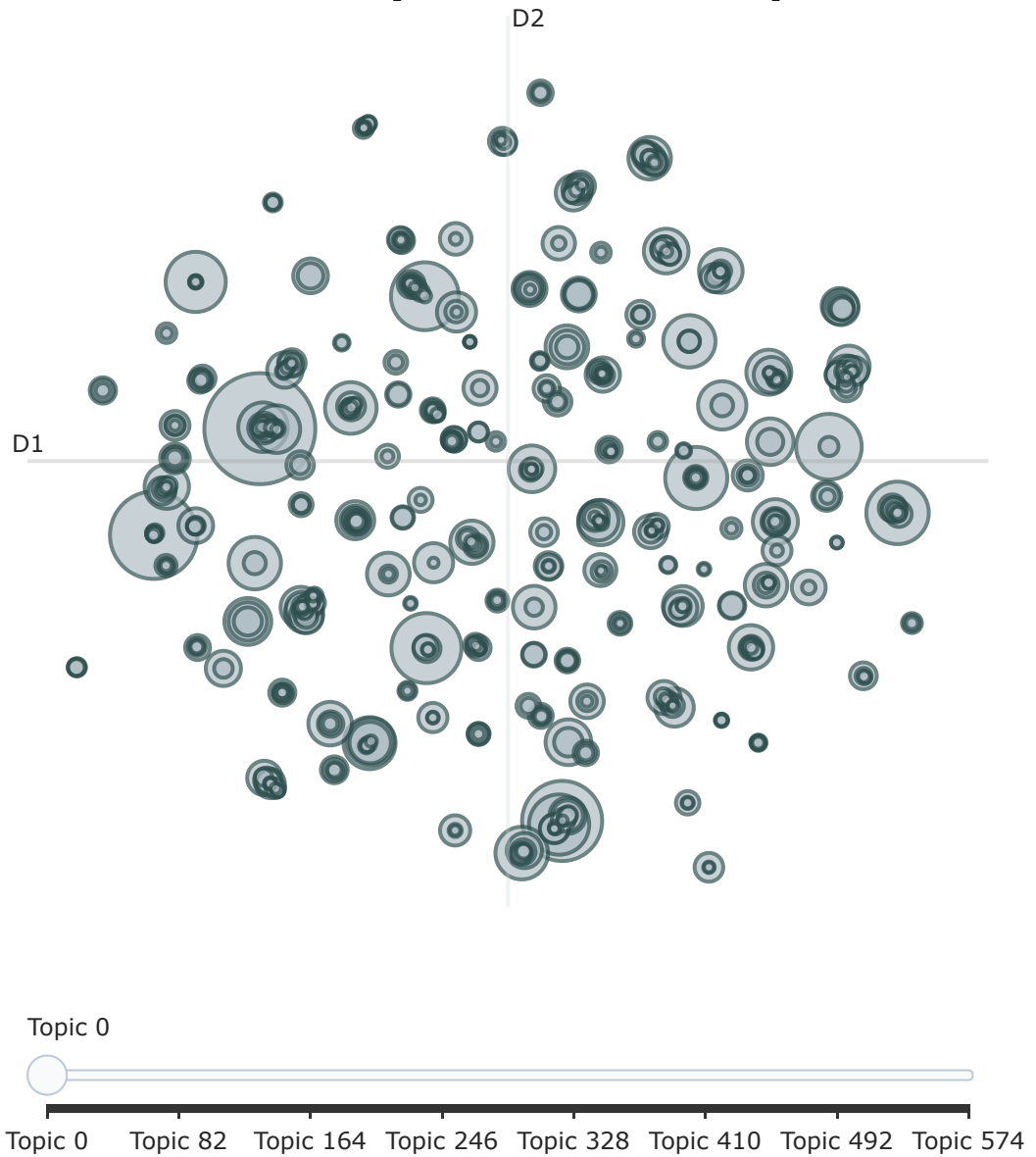


```
[('log', 0.019970476365956),  
 ('logging', 0.01875419266193046),  
 ('wipe', 0.007650576173677811),  
 ('engineer', 0.006278081280945899),  
 ('assign', 0.005904768619089332),  
 ('room', 0.004873396369674367),  
 ('providing', 0.004827055285812126),  
 ('logs', 0.004405521099792103),  
 ('tester', 0.004396604178412055),  
 ('problem', 0.00431233919711506)]
```

```
bert_model.visualize_topics()
```



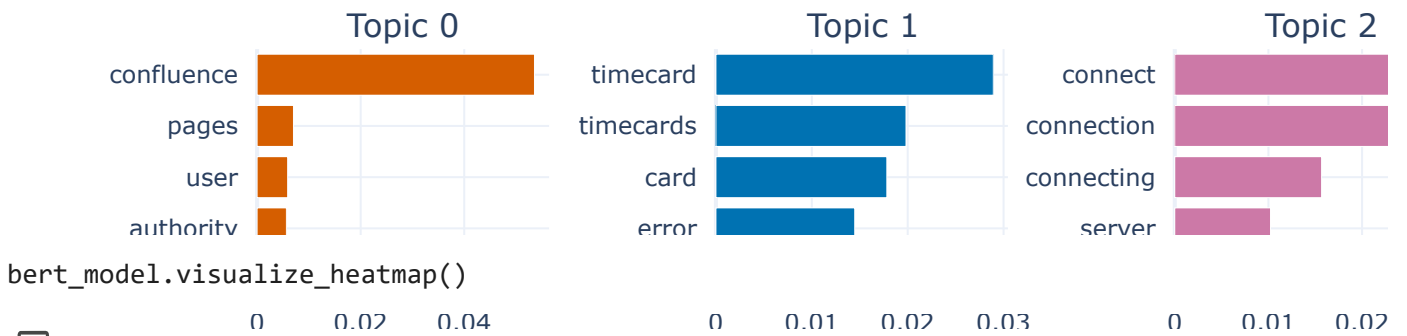
Intertopic Distance Map



```
bert_model.visualize_barchart()
```




Topic Word Scores



Similarity Matrix

