INFO5731 Assignment: 4

This exercise will provide a valuable learning experience in working with text data and extracting features using various topic modeling algorithms. Key concepts such as Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA) and BERTopic.

Expectations:

- Students are expected to complete the exercise during lecture period to meet the active participation criteria of the course.
- Use the provided *.ipynb* document to write your code & respond to the questions. Avoid generating a new file.
- Write complete answers and run all the cells before submission.
- Make sure the submission is "clean"; *i.e.*, no unnecessary code cells.
- Once finished, allow shared rights from top right corner (*see Canvas for details*).

Total points: 100

NOTE: The output should be presented well to get **full points**

Late submissions will have a penalty of 10% of the marks for each day of late submission, and no requests will be answered. Manage your time accordingly.

Question 1 (20 Points)

Dataset: 20 Newsgroups dataset

Dataset Link: https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html

Consider Random 2000 rows only

Generate K=10 topics by using LDA and LSA, then calculate coherence score and determine the optimized K value by the coherence score. Further, summarize and visualize each topics in you own words.

```
!pip install gensim
!pip uninstall -y numpy
!pip install numpy==1.24.4 --force-reinstall --no-cache-dir
```

```
Collecting gensim
  Downloading gensim-4.3.3-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl
Collecting numpy<2.0,>=1.18.5 (from gensim)
  Downloading numpy-1.26.4-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 61.0/61.0 kB 435.9 kB/s eta 0:00:00
Collecting scipy<1.14.0,>=1.7.0 (from gensim)
  Downloading scipy-1.13.1-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 60.6/60.6 kB 1.0 MB/s eta 0:00:00
Requirement already satisfied: smart-open>=1.8.1 in /usr/local/lib/python3.11/dist-pa
Requirement already satisfied: wrapt in /usr/local/lib/python3.11/dist-packages (from
Downloading gensim-4.3.3-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 26.7/26.7 MB 21.5 MB/s eta 0:00:00
Downloading numpy-1.26.4-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 18.3/18.3 MB 19.8 MB/s eta 0:00:00
Downloading scipy-1.13.1-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 38.6/38.6 MB 6.1 MB/s eta 0:00:00
Installing collected packages: numpy, scipy, gensim
  Attempting uninstall: numpy
    Found existing installation: numpy 2.0.2
    Uninstalling numpy-2.0.2:
      Successfully uninstalled numpy-2.0.2
  Attempting uninstall: scipy
    Found existing installation: scipy 1.14.1
    Uninstalling scipy-1.14.1:
      Successfully uninstalled scipy-1.14.1
Successfully installed gensim-4.3.3 numpy-1.26.4 scipy-1.13.1
Found existing installation: numpy 1.26.4
Uninstalling numpy-1.26.4:
  Successfully uninstalled numpy-1.26.4
Collecting numpy==1.24.4
  Downloading numpy-1.24.4-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl
Downloading numpy-1.24.4-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 17.3/17.3 MB 76.9 MB/s eta 0:00:00
Installing collected packages: numpy
ERROR: pip's dependency resolver does not currently take into account all the package
jax 0.5.2 requires numpy>=1.25, but you have numpy 1.24.4 which is incompatible.
pymc 5.21.2 requires numpy>=1.25.0, but you have numpy 1.24.4 which is incompatible.
treescope 0.1.9 requires numpy>=1.25.2, but you have numpy 1.24.4 which is incompatib
tensorflow 2.18.0 requires numpy<2.1.0,>=1.26.0, but you have numpy 1.24.4 which is i
blosc2 3.2.1 requires numpy>=1.26, but you have numpy 1.24.4 which is incompatible.
jaxlib 0.5.1 requires numpy>=1.25, but you have numpy 1.24.4 which is incompatible.
Successfully installed numpy-1.24.4
```

```
from sklearn.datasets import fetch_20newsgroups
import random
import pandas as pd
data = fetch_20newsgroups(subset='all', remove=('headers', 'footers', 'quotes'))
random.seed(42)
SampleIndecies = random.sample(range(len(data.data)), 2000)
SmpldData = [data.data[i] for i in SampleIndecies]
df = pd.DataFrame(SmpldData, columns=["text"])
```

```
import nltk
from nltk.corpus import stopwords
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
```

```
from nltk.stem import WordNetLemmatizer
import re
nltk.download('stopwords')
nltk.download('wordnet')
StpWrds = set(stopwords.words('english'))
Lemmatizer = WordNetLemmatizer()
def DtaPreProcess(text):
    text = re.sub(r'\W+', ' ', text.lower())
    Tokens = text.split()
    Tokens = [lemmatizer.lemmatize(word) for word in tokens if word not in stop_words and
    return " ".join(Tokens)
df['CleanedText'] = df['text'].apply(DtaPreProcess)
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
[nltk_data] Downloading package wordnet to /root/nltk_data...
```

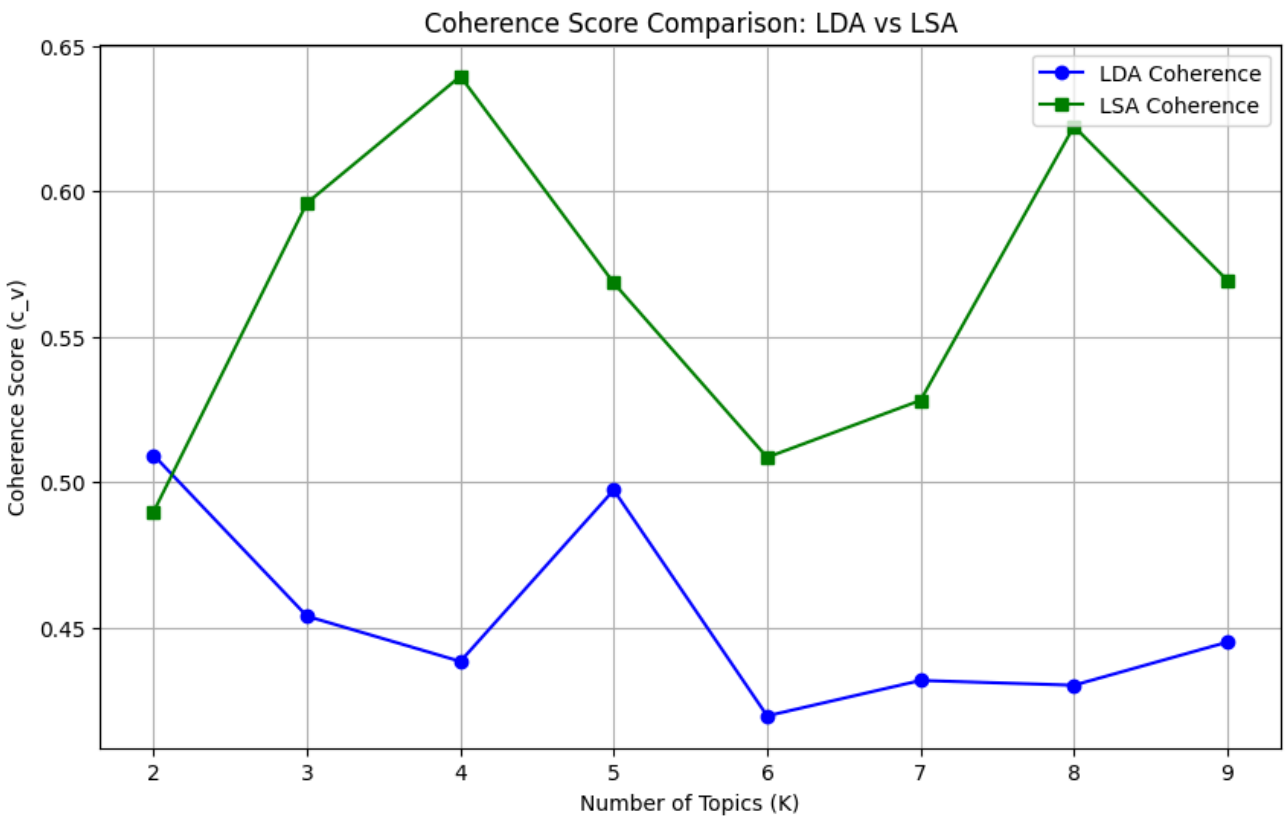
```
from sklearn.decomposition import LatentDirichletAllocation, TruncatedSVD
from gensim.models.coherencemodel import CoherenceModel
from gensim.corpora.dictionary import Dictionary
import gensim
import numpy as np
DocsTknized = [doc.split() for doc in df['CleanedText']]
Dict = Dictionary(DocsTknized)
TextCorpus = [Dict.doc2bow(i) for i in tokenized_docs]
Vector = CountVectorizer(max_df=0.95, min_df=2)
TFre = Vector.fit_transform(df['CleanedText'])
Vect_TfIdf = TfidfVectorizer(max_df=0.95, min_df=2)
Matrix_TFIDF = Vect_TfIdf.fit_transform(df['cleaned'])
TM_LDA = LatentDirichletAllocation(n_components=10, random_state=42)
TpicsLDA = TM_LDA.fit_transform(TFre)
TM_LSA = TruncatedSVD(n_components=10, random_state=42)
TpicsLSA = TM_LSA.fit_transform(tfidf)
```

```
def Coherence_Calc(model_type, texts, Dict, corpus, start=2, limit=10, step=1):
    Cohr_Scrs = []
    for j in range(start, limit, step):
        if model_type == 'lda':
            model = gensim.models.LdaModel(corpus=corpus, id2word=Dict, num_topics=j, ran
        elif model_type == 'lsa':
            model = gensim.models.LsiModel(corpus=corpus, id2word=Dict, num_topics=j)
        Model_Coherence = CoherenceModel(model=model, texts=texts, dictionary=Dict, coher
        Cohr_Scrs.append((j, Model_Coherence.get_coherence()))
    return Cohr_Scrs
LDA_Cohr = Coherence_Calc('lda', DocsTknized, Dict, TextCorpus)
LSA_Cohr = Coherence_Calc('lsa', DocsTknized, Dict, TextCorpus)
```

```
WARNING:gensim.models.ldamodel:too few updates, training might not converge; consider
WARNING:gensim.models.ldamodel:too few updates, training might not converge; consider
WARNING:gensim.models.ldamodel:too few updates, training might not converge; consider
WARNING:gensim.models.ldamodel:too few updates, training might not converge; consider
```

```
WARNING:gensim.models.ldamodel:too few updates, training might not converge; consider
WARNING:gensim.models.ldamodel:too few updates, training might not converge; consider
WARNING:gensim.models.ldamodel:too few updates, training might not converge; consider
WARNING:gensim.models.ldamodel:too few updates, training might not converge; consider
```

```
import matplotlib.pyplot as plt
LDA_j, LDA_Scrs = zip(*LDA_Cohr)
LSA_j, LSA_Scrs = zip(*LSA_Cohr)
plt.figure(figsize=(10, 6))
plt.plot(LDA_j, LDA_Scrs, marker='o', label='LDA Coherence', color='blue')
plt.plot(LSA_j, LSA_Scrs, marker='s', label='LSA Coherence', color='green')
plt.xlabel("Number of Topics (K)")
plt.ylabel("Coherence Score (c_v)")
plt.title("Coherence Score Comparison: LDA vs LSA")
plt.legend()
plt.grid(True)
plt.show()
```




```
from gensim.models import LsiModel
BestMLSA = LsiModel(corpus=TextCorpus, id2word=Dict, num_topics=4)
topics = BestMLSA.print_topics(num_topics=5, num_words=10)
```

```
for i, j in enumerate(topics):
    print(f"Topic {i+1}: {j}")
```

```
Topic 1: (0, '-0.226*"president" + -0.219*"stephanopoulos" + -0.195*"program" + -0.18
Topic 2: (1, '-0.312*"stephanopoulos" + 0.274*"entry" + -0.261*"president" + 0.257*"f
Topic 3: (2, '0.646*"entry" + -0.189*"data" + -0.175*"available" + -0.164*"image" + 0
Topic 4: (3, '0.438*"stephanopoulos" + -0.264*"administration" + -0.242*"russian" + -
```

BERTopic

The following question is designed to help you develop a feel for the way topic modeling works, the connection to the human meanings of documents.

Dataset from **assignment-3** (text dataset) .

> Dont use any custom datasets.

> Dataset must have 1000+ rows, no duplicates and null values

Question 2 (20 Points)

Q2) **Generate K=10 topics by using BERTopic and then find optimal K value by the coherence score. Interpret each topic and visualize with suitable style.**

```
!pip install 'numpy>=1.24'
#!pip install --upgrade jax bertopic
```

```
Collecting numpy>=1.24
  Downloading numpy-2.2.4-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 62.0/62.0 kB 2.3 MB/s eta 0:00:00
Downloading numpy-2.2.4-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (1
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 16.4/16.4 MB 21.4 MB/s eta 0:00:00
Installing collected packages: numpy
  Attempting uninstall: numpy
    Found existing installation: numpy 1.23.5
    Uninstalling numpy-1.23.5:
      Successfully uninstalled numpy-1.23.5
ERROR: pip's dependency resolver does not currently take into account all the package
gensim 4.3.3 requires numpy<2.0,>=1.18.5, but you have numpy 2.2.4 which is incompati
tensorflow 2.18.0 requires numpy<2.1.0,>=1.26.0, but you have numpy 2.2.4 which is in
numba 0.60.0 requires numpy<2.1,>=1.22, but you have numpy 2.2.4 which is incompatibl
Successfully installed numpy-2.2.4
```

```
!pip install --upgrade numpy --quiet
!pip uninstall -y bertopic
```

```
!pip install bertopic[all] --quiet
```

```
Found existing installation: bertopic 0.17.0
Uninstalling bertopic-0.17.0:
  Successfully uninstalled bertopic-0.17.0
WARNING: bertopic 0.17.0 does not provide the extra 'all'
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 60.9/60.9 kB 2.2 MB/s eta 0:00:00
    ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 19.5/19.5 MB 32.2 MB/s eta 0:00:00
ERROR: pip's dependency resolver does not currently take into account all the package
gensim 4.3.3 requires numpy<2.0,>=1.18.5, but you have numpy 2.0.2 which is incompati
```

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from bertopic import BERTopic
from gensim.models.coherencemodel import CoherenceModel
from gensim.corpora import Dictionary
```

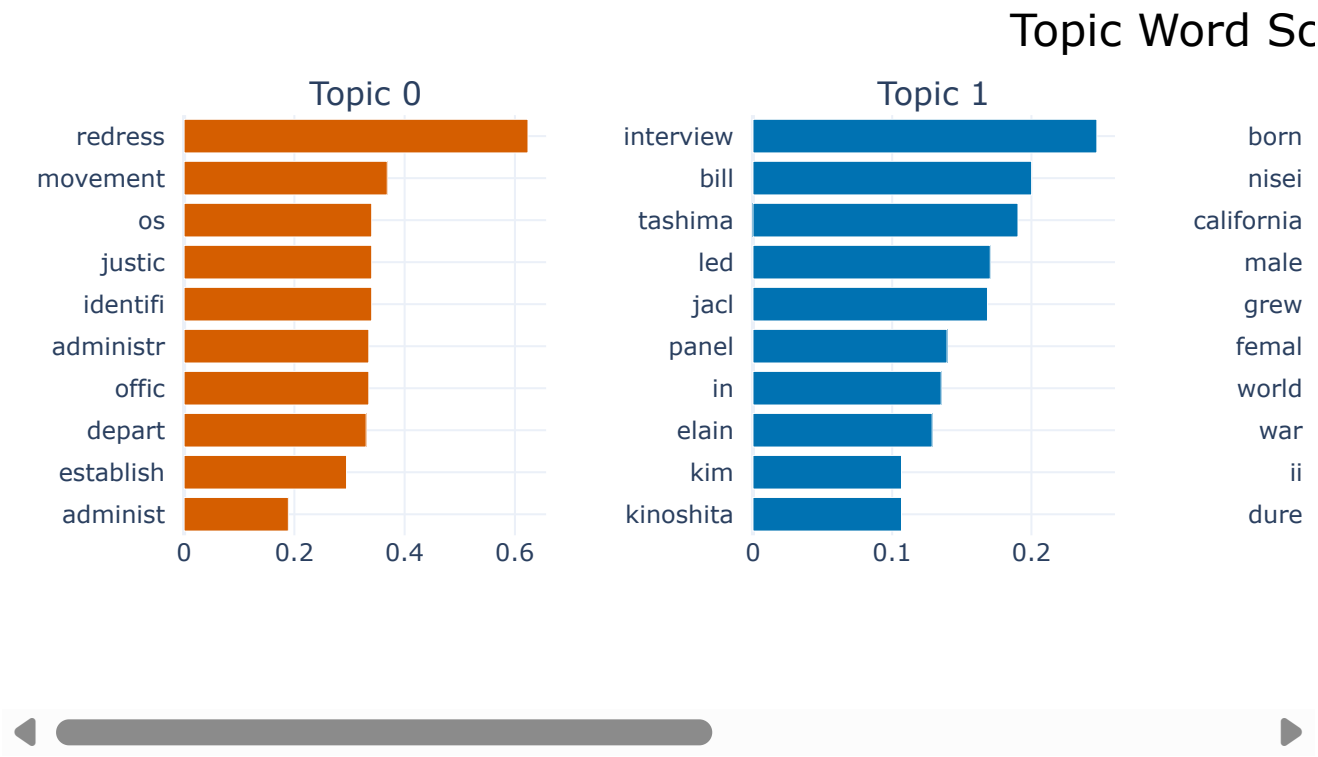
```
k = 10
df = pd.read_csv('/content/Narrators_Information_Cleaned.csv', usecols=['CleanedDetails']
NarrDetails = df.CleanedDetails.to_list()
df.head()
```

	CleanedDetails
0	nisei femal born may selleck washington spent ...
1	nisei male born june seattl washington grew ar...
2	nisei femal born octob seattl washington famil...
3	nisei femal born juli boyl height california a...
4	sansei male born march torranc california grew...

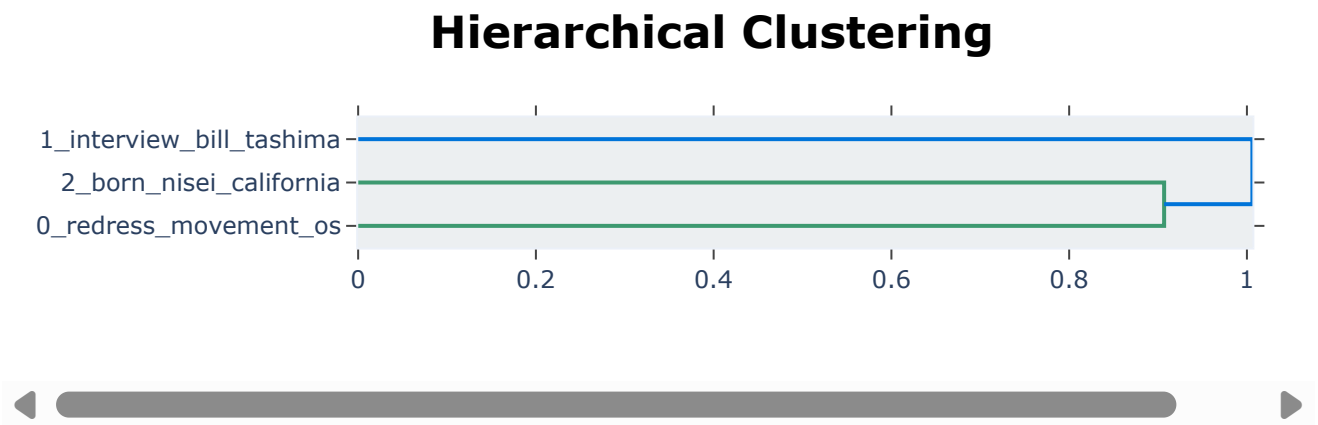
```
TM_Bert = BERTopic(nr_topics=k)
Tpcs, Prblts = Berttopic_model.fit_transform(NarrDetails)
Berttopic_model.get_topic_info()
```

	Topic	Count	Name	Representation	Representative_Docs
0	0	12	0_redress_movement_os_justic	[redress, movement, os, justic, identifi, admi...	[born honolulu hawaii dure redress movement de...
				[interview. bill.	[. . . . .

```
Berttopic_model.visualize_barchart(top_n_topics=10, n_words = 40, width = 300, height = 3
```




```
Berttopic_model.visualize_hierarchy(top_n_topics=10, width = 700, height = 700)
```




```
pip install gensim
```

```
Requirement already satisfied: gensim in /usr/local/lib/python3.11/dist-packages (4.3
Requirement already satisfied: numpy<2.0,>=1.18.5 in /usr/local/lib/python3.11/dist-p
Requirement already satisfied: scipy<1.14.0,>=1.7.0 in /usr/local/lib/python3.11/dist
Requirement already satisfied: smart-open>=1.8.1 in /usr/local/lib/python3.11/dist-pa
Requirement already satisfied: wrapt in /usr/local/lib/python3.11/dist-packages (from
```

```
pip install --upgrade h5py
```

```
Requirement already satisfied: h5py in /usr/local/lib/python3.11/dist-packages (3.13.
Requirement already satisfied: numpy>=1.19.3 in /usr/local/lib/python3.11/dist-packag
```

```
!pip install --upgrade jax jaxlib
```

```
Requirement already satisfied: jax in /usr/local/lib/python3.11/dist-packages (0.5.3)
Requirement already satisfied: jaxlib in /usr/local/lib/python3.11/dist-packages (0.5
Requirement already satisfied: ml_dtypes>=0.4.0 in /usr/local/lib/python3.11/dist-pac
Requirement already satisfied: numpy>=1.25 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: opt_einsum in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: scipy>=1.11.1 in /usr/local/lib/python3.11/dist-packag
```

```
from gensim.models import CoherenceModel
from gensim.corpora import Dictionary
def Coherence_Calc(docs, Min_Tpcs=2, Max_Tpcs=10):
    Cohr_Scrs = []
    for j in range(Min_Tpcs, Max_Tpcs + 1):
        TM = BERTopic(nr_topics=j)
        Tpcs, Scrs = TM.fit_transform(docs)
        Tpc_Keywrds = [
            [i for i, j in TM.get_topic(topic)]
            for tpc in TM.get_topics().keys()
            if tpc != -1
        ]
        DocsTknized = [doc.split() for doc in docs]
        Dict = Dictionary(DocsTknized)
        Chr_Model = CoherenceModel(
            topics=topic_keywords,
            dictionary=Dict,
            texts=DocsTknized,
            coherence='c_v'
        )
        score = Chr_Model.get_coherence()
        Chr_Scrs.append((num_topics, score))

        print(f"Topics={num_topics}, Coherence Score={score:.4f}")
    return Chr_Scrs
Chr_Scrs = Coherence_Calc(NarrDetails, min_topics=2, max_topics=20)
```

```
Topics=2, Coherence Score=0.7109
Topics=3, Coherence Score=0.6907
Topics=4, Coherence Score=0.8084
Topics=5, Coherence Score=0.6740
Topics=6, Coherence Score=0.7934
Topics=7, Coherence Score=0.6324
Topics=8, Coherence Score=0.6525
Topics=9, Coherence Score=0.6298
Topics=10, Coherence Score=0.6108
Topics=11, Coherence Score=0.6078
Topics=12, Coherence Score=0.5842
Topics=13, Coherence Score=0.7934
Topics=14, Coherence Score=0.6032
Topics=15, Coherence Score=0.7934
Topics=16, Coherence Score=0.8069
```

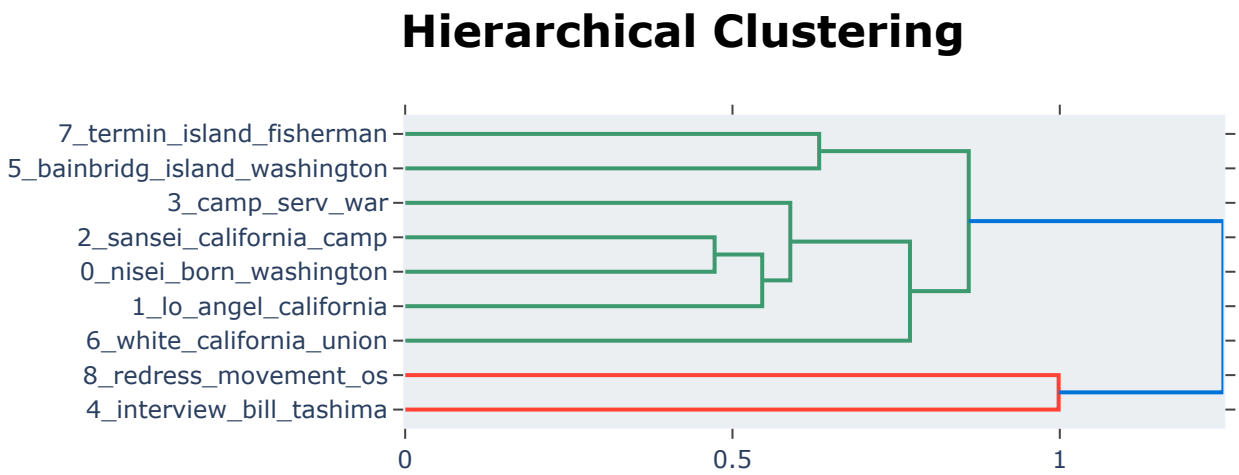
```
Topics=17, Coherence Score=0.8084
Topics=18, Coherence Score=0.6393
Topics=19, Coherence Score=0.8084
Topics=20, Coherence Score=0.8069
```

```
Best_Count = 10
Final_TM = BERTopic(nr_topics=Best_Count)
Final_Tpcs, Final_Prbls = Final_TM.fit_transform(NarrDetails)
```

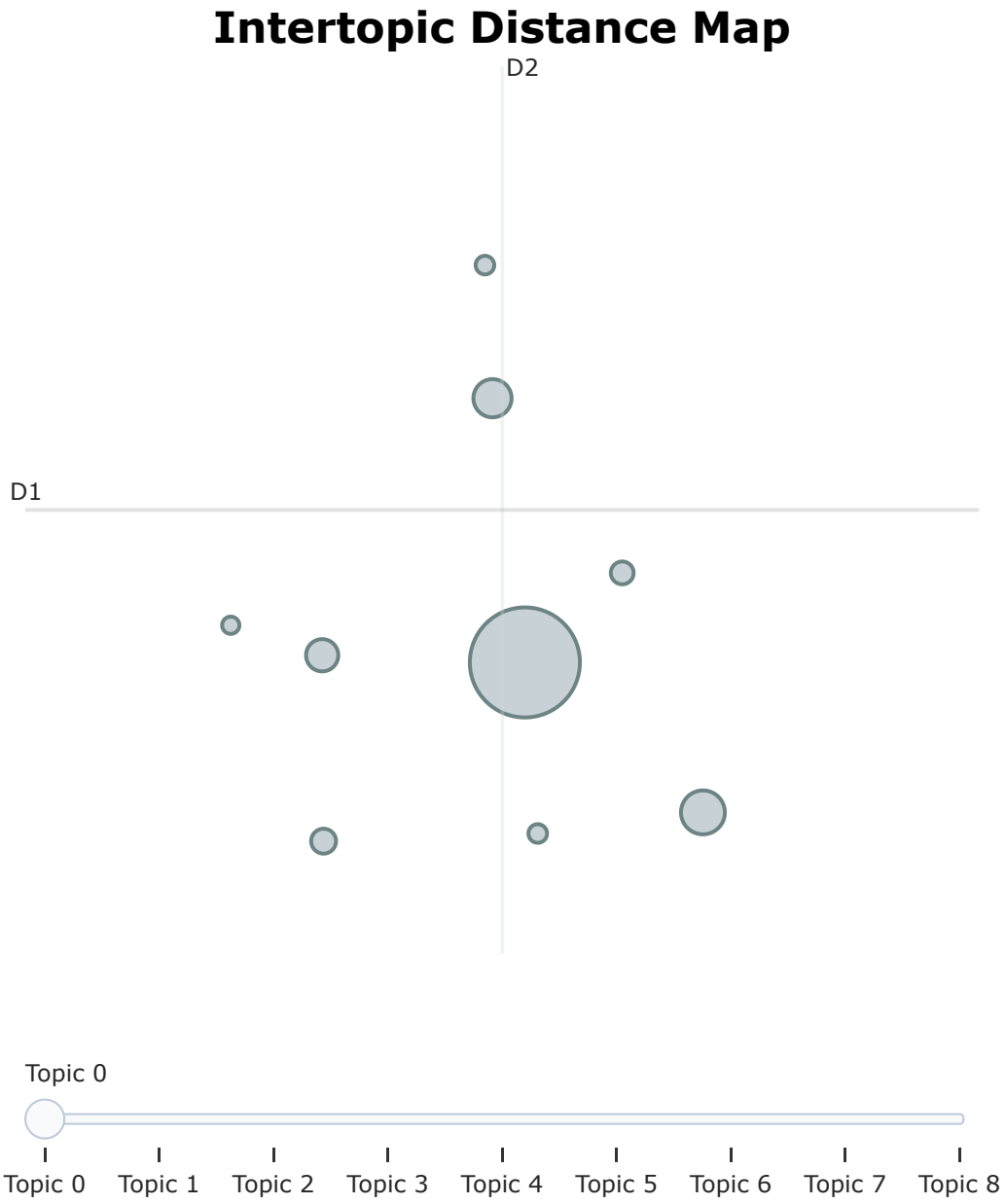
```
def Eval_Cohr(documents, min_topics=2, max_topics=20):
    scores = []
    for k in range(min_topics, max_topics + 1):
        TM = BERTopic(nr_topics=num_topics)
        Tpcs, Probs = TM.fit_transform(documents)
        Tpc_Trms = [list(dict(TM.get_topic(k)).keys()) for i in range(num_topics)]
        Dict = Dictionary([i for i in Tpc_Trms])
        Text_Crpus = [Dict.doc2bow(terms) for terms in Tpc_Trms]
        Chr_Model = CoherenceModel(
            topics=Tpc_Trms,
            texts=[doc.split() for doc in documents],
            dictionary=Dict,
            coherence='c_v'
        )
        scores.append((num_topics, Chr_Model.get_coherence()))
    return scores
```

```
TM = BERTopic(nr_topics=Best_Count)
Tpcs, Prbs = model.fit_transform(NarrDetails)
print("\nTopic Interpretation (Top Words):")
for topic_num in range(best_topic_count):
    print(f"Topic {topic_num}:")
    print(model.get_topic(topic_num))
    print("\n")
TM.visualize_topics()
TM.visualize_barchart(top_n_topics=12, n_words=10, width=350, height=350)
TM.visualize_hierarchy(top_n_topics=12, width=700, height=700)
```

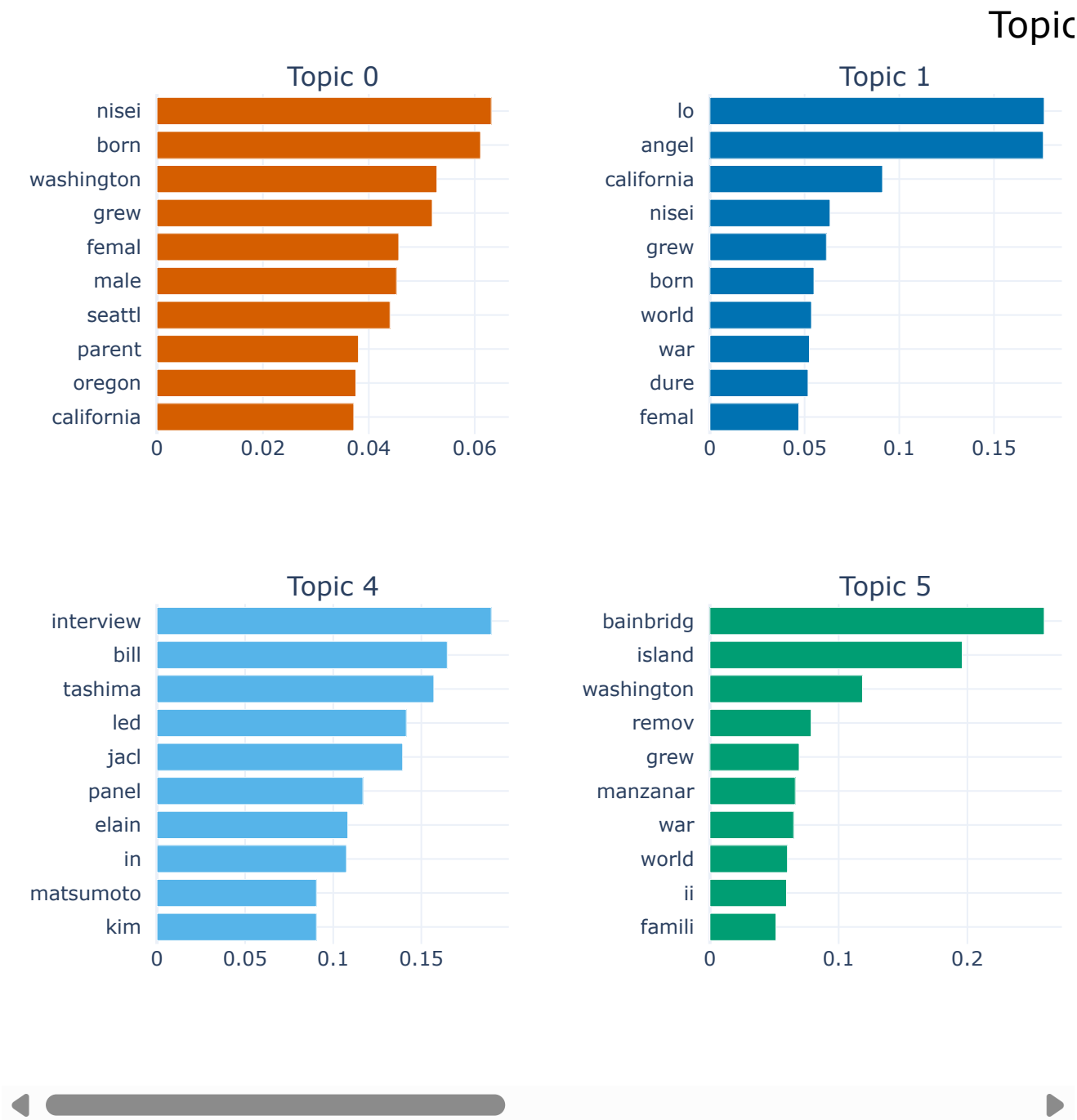
```
Topic Interpretation (Top Words):
Topic 0:
[('nisei', 0.06312162631700156), ('born', 0.061102051671154414), ('washington', 0.


Topic 1:
[('lo', 0.17653052001175715), ('angel', 0.17606553670192612), ('california', 0.091


Topic 2:
[('sansei', 0.1640773161398449), ('california', 0.0890586695619215), ('camp', 0.08


Topic 3:
[('camp', 0.08424442995518266), ('serv', 0.08377513908521746), ('war', 0.080085874


Topic 4:
[('interview', 0.18966418991933756), ('bill', 0.16468641652305405), ('tashima', 0.


Topic 5:
[('bainbridg', 0.2596602516818111), ('island', 0.19599294052433727), ('washington'


Topic 6:
[('white', 0.21150897891386705), ('california', 0.09050442718467996), ('union', 0.


Topic 7:
[('termin', 0.3384169414036539), ('island', 0.24928994896386528), ('fisherman', 0.


Topic 8:
[('redress', 0.4859369965810396), ('movement', 0.2959822669247629), ('os', 0.27427


Topic 9:
False
```



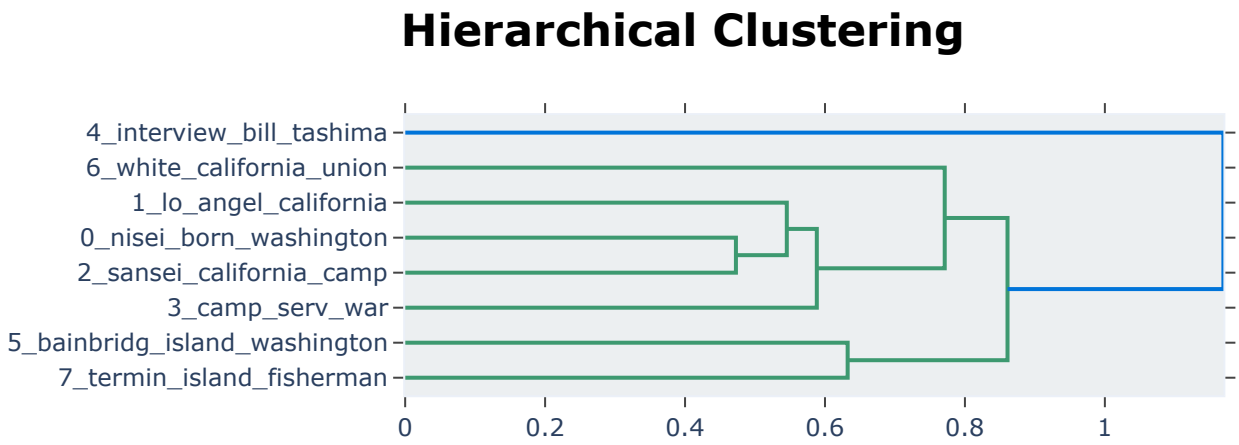
```
TM.visualize_topics()
```




```
TM.visualize_barchart(top_n_topics=8, n_words = 10, width = 350, height = 350)
```




```
TM.visualize_hierarchy(top_n_topics=8, width = 700, height = 700)
```



Question 3 (25 points)

Dataset Link: 20 Newsgroup Dataset (Random 2000 values)

Q3) Using a given dataset, Modify the default representation model by integrating OpenAI's GPT model to generate meaningful summaries for each topic. Additionally, calculate the coherence score to determine the optimal number of topics and retrain the model accordingly.

Usefull Link:
https://maartengr.github.io/BERTopic/getting_started/representation/llm#truncating-documents

```
import pandas as pd
import random
from sklearn.datasets import fetch_20newsgroups
data = fetch_20newsgroups(subset='all', remove=('headers', 'footers', 'quotes'))
Smpled_Data = random.sample(data.data, 2000)
df = pd.DataFrame(Smpled_Data, columns=['text'])
print(df.head())
```

```
                                                text
0  \nWasn't there an 85,000 New York at Cleveland...
1  \n\nThis is vague, so I am posting it in case ...
2  \nIsn't that just a variation of the "Achilles...
3  Sumatriptan(Imitrex) just became available in ...
4  \nI did say *any* invader, didn't I?  What do ...
```

```
import re
import nltk
from nltk.corpus import stopwords
```

```
from nltk.stem import WordNetLemmatizer
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')
nltk.download('punkt_tab')
Stp_Words = set(stopwords.words('english'))
Lmtizer = WordNetLemmatizer()
def Dta_Pre_Prcs(text):
    text = text.lower()
    text = re.sub(r'[^a-z\s]', '', text)
    tkns = nltk.word_tokenize(text)
    tkns = [Lmtizer.lemmatize(word) for word in tkns if word not in Stp_Words and len(wor
    return " ".join(tokens)
df['cleaned'] = df['text'].apply(Dta_Pre_Prcs)
print(df[['text', 'cleaned']].head())
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data]   Package punkt_tab is already up-to-date!
                                                  text  \
0  \nWasn't there an 85,000 New York at Cleveland...
1  \n\nThis is vague, so I am posting it in case ...
2  \nIsn't that just a variation of the "Achilles...
3  Sumatriptan(Imitrex) just became available in ...
4  \nI did say *any* invader, didn't I?  What do ...

                                              cleaned
0                     wasnt york cleveland game late
1  vague posting case anyone else know recall rea...
2  isnt variation achilles turtle paradox state a...
3  sumatriptanimitrex became available subcutaneo...
4  invader didnt want perhaps neural design count...
```

```
from gensim import corpora
texts = [doc.split() for doc in df['cleaned']]
Dict = corpora.Dictionary(texts)
Crpus = [Dict.doc2bow(text) for text in texts]
print(f"Sample dictionary tokens: {Dict.token2id}")
print(f"Sample corpus: {Crpus[0][:20]}")
```

```
Sample dictionary tokens: {'cleveland': 0, 'game': 1, 'late': 2, 'wasnt': 3, 'york':
Sample corpus: [(0, 1), (1, 1), (2, 1), (3, 1), (4, 1)]
```

```
pip install numpy==1.24.4
```

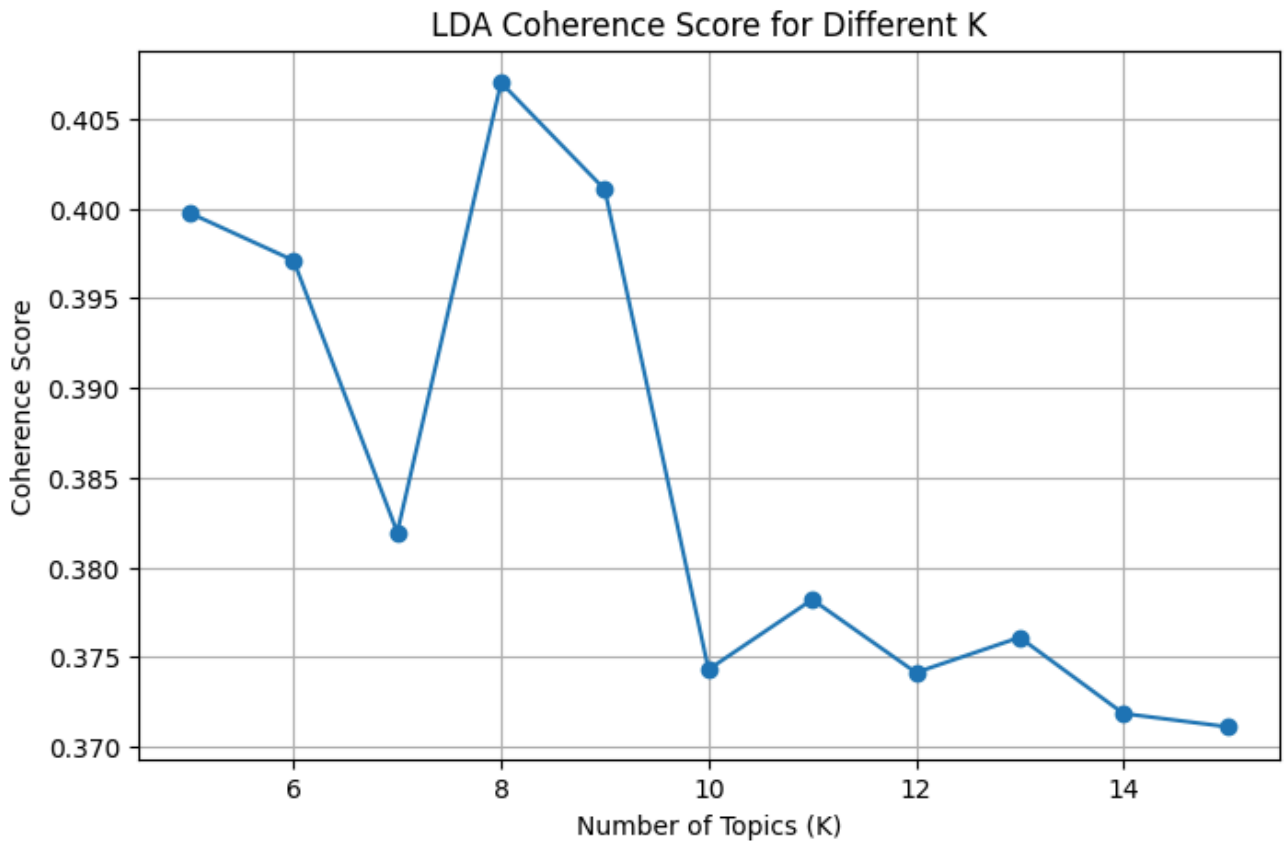
```
Collecting numpy==1.24.4
  Downloading numpy-1.24.4-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl
Downloading numpy-1.24.4-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 17.3/17.3 MB 38.4 MB/s eta 0:00:00
Installing collected packages: numpy
  Attempting uninstall: numpy
    Found existing installation: numpy 1.26.4
    Uninstalling numpy-1.26.4:
      Successfully uninstalled numpy-1.26.4
ERROR: pip's dependency resolver does not currently take into account all the package
jaxlib 0.5.3 requires numpy>=1.25, but you have numpy 1.24.4 which is incompatible.
jax 0.5.3 requires numpy>=1.25, but you have numpy 1.24.4 which is incompatible.
pymc 5.21.2 requires numpy>=1.25.0, but you have numpy 1.24.4 which is incompatible.
treescope 0.1.9 requires numpy>=1.25.2, but you have numpy 1.24.4 which is incompatib
tensorflow 2.18.0 requires numpy<2.1.0,>=1.26.0, but you have numpy 1.24.4 which is i
blosc2 3.2.1 requires numpy>=1.26, but you have numpy 1.24.4 which is incompatible.
Successfully installed numpy-1.24.4
```

```
from gensim.models import LdaModel, CoherenceModel
import matplotlib.pyplot as plt
Chr_Scrs = []
for j in range(5, 16):
    LDA_TM = LdaModel(corpus=Crpus, id2word=Dict, num_topics=j, random_state=42)
    Chr_Mdl = CoherenceModel(model=LDA_TM, texts=texts, dictionary=Dict, coherence='c_v')
    Chrnc = Chr_Mdl.get_coherence()
    Chr_Scrs.append((j, coherence))
    print(f"K={j}, Coherence Score={Chrnc:.4f}")
```

```
WARNING:gensim.models.ldamodel:too few updates, training might not converge; consider
WARNING:gensim.models.ldamodel:too few updates, training might not converge; consider
K=5, Coherence Score=0.3997
WARNING:gensim.models.ldamodel:too few updates, training might not converge; consider
K=6, Coherence Score=0.3971
WARNING:gensim.models.ldamodel:too few updates, training might not converge; consider
K=7, Coherence Score=0.3819
WARNING:gensim.models.ldamodel:too few updates, training might not converge; consider
K=8, Coherence Score=0.4070
WARNING:gensim.models.ldamodel:too few updates, training might not converge; consider
K=9, Coherence Score=0.4011
WARNING:gensim.models.ldamodel:too few updates, training might not converge; consider
K=10, Coherence Score=0.3743
WARNING:gensim.models.ldamodel:too few updates, training might not converge; consider
K=11, Coherence Score=0.3782
WARNING:gensim.models.ldamodel:too few updates, training might not converge; consider
K=12, Coherence Score=0.3741
WARNING:gensim.models.ldamodel:too few updates, training might not converge; consider
K=13, Coherence Score=0.3761
WARNING:gensim.models.ldamodel:too few updates, training might not converge; consider
K=14, Coherence Score=0.3718
K=15, Coherence Score=0.3711
```

```
k_vals, scores = zip(*coherence_scores)
plt.figure(figsize=(8, 5))
```

```
plt.plot(k_vals, scores, marker='o')
plt.xlabel("Number of Topics (K)")
plt.ylabel("Coherence Score")
plt.title("LDA Coherence Score for Different K")
plt.grid(True)
plt.show()
K_Best = max(Chr_Scrs, key=lambda x: x[1])[0]
print(f"\nBest K based on coherence: {K_Best}")
```




```
Best K based on coherence: 8
```

```
LDA_TM = LdaModel(corpus=Crpus, id2word=Dict, num_topics=K_Best, random_state=42)
Tpcs = LDA_TM.show_topics(num_topics=K_Best, num_words=10, formatted=False)
for idx, topic in topics:
    KeyWrds = [word for word, prob in topic]
    print(f"Topic {idx+1}: {', '.join(KeyWrds)}")
```

```
WARNING:gensim.models.ldamodel:too few updates, training might not converge; consider
Topic 1: would, like, image, think, people, also, dont, know, make, well
Topic 2: image, would, time, dont, also, people, like, jpeg, thing, even
Topic 3: would, time, used, also, drive, like, system, even, right, dont
Topic 4: would, dont, know, time, much, also, people, year, data, system
Topic 5: maxaxaxaxaxaxaxaxaxaxaxaxaxaxaxaxax, know, would, dont, image, system, also, pec
Topic 6: would, also, people, good, image, know, file, first, dont, time
Topic 7: image, would, dont, window, file, jpeg, also, system, problem, card
Topic 8: would, like, dont, know, file, maxaxaxaxaxaxaxaxaxaxaxaxaxaxaxaxax, time, people
```

```
import openai
openai.api_key = "sk-proj-j9OVQiRJkRDbPXK7_RynlrKu9Mr8ZMMiongcCyVQ70s1R0umNoboWcBCvgrhcD7
def Tpc_Smmry(keywords):
    prompt = f"Generate a short, meaningful summary for a topic based on these keywords:
    Rsp = openai.ChatCompletion.create(
        model="gpt-3.5-turbo",
        messages=[{"role": "user", "content": prompt}],
        max_tokens=50
    )
    return Rsp.choices[0].message.content.strip()
print("\n=== GPT Summaries ===")
for i, j in topics:
    KeyWrds = [word for word, prob in j]
    Smmry = Tpc_Smmry(KeyWrds)
    print(f"Topic {idx+1}: {Smmry}")
```

```
=== GPT Summaries ===
Topic 1: People often like to think about how they would like to present themselves i
Topic 2: Images, like JPEG files, hold a powerful influence over people and can evoke
Topic 3: The importance of efficiently allocating time and resources in a system, whe
Topic 4: Many people would like to know more about the data system, but they don't ha
Topic 5: The topic explores the use of the system Maxaxaxaxaxaxaxaxaxaxaxaxaxaxaxax and
Topic 6: First time users should know that having a good image file is essential beca
Topic 7: This topic explores the problem of not being able to view an image file in j
Topic 8: People who would like to know how to file an image may not know that there i
```

Question 4 (35 Points)

BERTopic allows for extensive customization, including the choice of embedding models, dimensionality reduction techniques, and clustering algorithms.

Dataset Link: 20 Newsgroup Dataset (Random 2000 values)

4)

4.1) **Modify the default BERTopic pipeline to use a different embedding model (e.g., Sentence-Transformers) and a different clustering algorithm (e.g., DBSCAN instead of HDBSCAN).

4.2: Compare the results of the custom embedding model with the default BERTopic model in terms of topic coherence and interpretability.

4.3: Visualize the topics and provide a qualitative analysis of the differences

**

Usefull Link :https://www.pinecone.io/learn/bertopic/

```
import pandas as pd
import random
from sklearn.datasets import fetch_20newsgroups
data = fetch_20newsgroups(subset='all', remove=('headers', 'footers', 'quotes'))
Smpled_Data = random.sample(data.data, 2000)
df3 = pd.DataFrame(Smpled_Data, columns=['text'])
print(dataframe_3.head())
```

```
                                                text
0  \nAbsolutely.  Unfortunately, most of them hav...
1  AT&T also puts out two new products for window...
2  :>>\n:>> As someone else has pointed out, why ...
3  \n\nWell I agree with you in the sense that th...
4  I am trying to obtain a HI-FI copy of Guns N' ...
```

```
!pip install bertopic
```

```
Requirement already satisfied: bertopic in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: hdbscan>=0.8.29 in /usr/local/lib/python3.11/dist-p
Requirement already satisfied: numpy>=1.20.0 in /usr/local/lib/python3.11/dist-pac
Requirement already satisfied: pandas>=1.1.5 in /usr/local/lib/python3.11/dist-pac
Requirement already satisfied: plotly>=4.7.0 in /usr/local/lib/python3.11/dist-pac
Requirement already satisfied: scikit-learn>=1.0 in /usr/local/lib/python3.11/dist
Requirement already satisfied: sentence-transformers>=0.4.1 in /usr/local/lib/pyth
Requirement already satisfied: tqdm>=4.41.1 in /usr/local/lib/python3.11/dist-pack
Requirement already satisfied: umap-learn>=0.5.0 in /usr/local/lib/python3.11/dist
Requirement already satisfied: scipy>=1.0 in /usr/local/lib/python3.11/dist-packag
Requirement already satisfied: joblib>=1.0 in /usr/local/lib/python3.11/dist-packa
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-pack
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-pa
Requirement already satisfied: tenacity>=6.2.0 in /usr/local/lib/python3.11/dist-p
Requirement already satisfied: packaging in /usr/local/lib/python3.11/dist-package
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.11/d
Requirement already satisfied: transformers<5.0.0,>=4.41.0 in /usr/local/lib/pytho
Requirement already satisfied: torch>=1.11.0 in /usr/local/lib/python3.11/dist-pac
Requirement already satisfied: huggingface-hub>=0.20.0 in /usr/local/lib/python3.1
Requirement already satisfied: Pillow in /usr/local/lib/python3.11/dist-packages (
Requirement already satisfied: numba>=0.51.2 in /usr/local/lib/python3.11/dist-pac
Requirement already satisfied: pynndescent>=0.5 in /usr/local/lib/python3.11/dist-
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.11/dist-
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packa
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python
Requirement already satisfied: llvmlite<0.44,>=0.43.0dev0 in /usr/local/lib/python
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages (
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.4.127 in /usr/local/lib/
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.4.127 in /usr/local/li
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.4.127 in /usr/local/lib/
Requirement already satisfied: nvidia-cudnn-cu12==9.1.0.70 in /usr/local/lib/pytho
Requirement already satisfied: nvidia-cublas-cu12==12.4.5.8 in /usr/local/lib/pyth
Requirement already satisfied: nvidia-cufft-cu12==11.2.1.3 in /usr/local/lib/pytho
Requirement already satisfied: nvidia-curand-cu12==10.3.5.147 in /usr/local/lib/py
Requirement already satisfied: nvidia-cusolver-cu12==11.6.1.9 in /usr/local/lib/py
Requirement already satisfied: nvidia-cusparse-cu12==12.3.1.170 in /usr/local/lib/
```

```
Requirement already satisfied: nvidia-cusparselt-cu12==0.6.2 in /usr/local/lib/pyt
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in /usr/local/lib/python3.
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in /usr/local/lib/python
Requirement already satisfied: nvidia-nvjitlink-cu12==12.4.127 in /usr/local/lib/p
Requirement already satisfied: triton==3.2.0 in /usr/local/lib/python3.11/dist-pac
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/dist-pac
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.11/dis
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.11/dist
Requirement already satisfied: tokenizers<0.22,>=0.21 in /usr/local/lib/python3.11
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.11/dis
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-p
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-pack
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dis
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dis
```

```
!pip install openai==0.27.8
```

```
Collecting openai==0.27.8
  Downloading openai-0.27.8-py3-none-any.whl.metadata (13 kB)
Requirement already satisfied: requests>=2.20 in /usr/local/lib/python3.11/dist-packa
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-packages (from
Requirement already satisfied: aiohttp in /usr/local/lib/python3.11/dist-packages (fr
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-package
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-p
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-p
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in /usr/local/lib/python3.11/d
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.11/dist-pac
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-packag
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.11/dist-pa
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.11/dist-
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.11/dist-pac
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.11/dist-pa
Downloading openai-0.27.8-py3-none-any.whl (73 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 73.6/73.6 kB 2.5 MB/s eta 0:00:00
Installing collected packages: openai
  Attempting uninstall: openai
    Found existing installation: openai 0.28.0
    Uninstalling openai-0.28.0:
      Successfully uninstalled openai-0.28.0
```

```
!pip install 'numpy>=1.24'
```

```
Requirement already satisfied: numpy>=1.24 in /usr/local/lib/python3.11/dist-packages
```

```
!pip install --upgrade numpy --quiet
!pip uninstall -y bertopic
!pip install bertopic[all] --quiet
```

```
ERROR: pip's dependency resolver does not currently take into account all the package
gensim 4.3.3 requires numpy<2.0,>=1.18.5, but you have numpy 2.2.4 which is incompati
tensorflow 2.18.0 requires numpy<2.1.0,>=1.26.0, but you have numpy 2.2.4 which is in
numba 0.60.0 requires numpy<2.1,>=1.22, but you have numpy 2.2.4 which is incompatibl
```

```
Found existing installation: bertopic 0.17.0
Uninstalling bertopic-0.17.0:
  Successfully uninstalled bertopic-0.17.0
WARNING: bertopic 0.17.0 does not provide the extra 'all'
ERROR: pip's dependency resolver does not currently take into account all the package
gensim 4.3.3 requires numpy<2.0,>=1.18.5, but you have numpy 2.0.2 which is incompati
```

```
!pip install --upgrade jax jaxlib
```

```
Requirement already satisfied: jax in /usr/local/lib/python3.11/dist-packages (0.5.3)
Requirement already satisfied: jaxlib in /usr/local/lib/python3.11/dist-packages (0.5
Requirement already satisfied: ml_dtypes>=0.4.0 in /usr/local/lib/python3.11/dist-pac
Requirement already satisfied: numpy>=1.25 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: opt_einsum in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: scipy>=1.11.1 in /usr/local/lib/python3.11/dist-packag
```

```
!pip install bertopic[all]
!pip install --upgrade sentence-transformers
!pip install --upgrade jax jaxlib
```

```
Requirement already satisfied: joblib>=1.2.0 in /usr/local/lib/python3.11/dist-pac
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.11/d
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-p
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-pack
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dis
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dis
Downloading sentence_transformers-4.0.2-py3-none-any.whl (340 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 340.6/340.6 kB 6.2 MB/s eta 0:00:00
Installing collected packages: sentence-transformers
  Attempting uninstall: sentence-transformers
    Found existing installation: sentence-transformers 3.4.1
    Uninstalling sentence-transformers-3.4.1:
      Successfully uninstalled sentence-transformers-3.4.1
Successfully installed sentence-transformers-4.0.2
Requirement already satisfied: jax in /usr/local/lib/python3.11/dist-packages (0.5
Requirement already satisfied: jaxlib in /usr/local/lib/python3.11/dist-packages (
Requirement already satisfied: ml_dtypes>=0.4.0 in /usr/local/lib/python3.11/dist-
Requirement already satisfied: numpy>=1.25 in /usr/local/lib/python3.11/dist-packa
Requirement already satisfied: opt_einsum in /usr/local/lib/python3.11/dist-packag
Requirement already satisfied: scipy>=1.11.1 in /usr/local/lib/python3.11/dist-pac
```

```
!pip install --upgrade jax jaxlib
!pip install --upgrade tensorflow
```

```
Requirement already satisfied: jax in /usr/local/lib/python3.11/dist-packages (0.5
Requirement already satisfied: jaxlib in /usr/local/lib/python3.11/dist-packages (
Requirement already satisfied: ml_dtypes>=0.4.0 in /usr/local/lib/python3.11/dist-
Requirement already satisfied: numpy>=1.25 in /usr/local/lib/python3.11/dist-packa
Requirement already satisfied: opt_einsum in /usr/local/lib/python3.11/dist-packag
Requirement already satisfied: scipy>=1.11.1 in /usr/local/lib/python3.11/dist-pac
Requirement already satisfied: tensorflow in /usr/local/lib/python3.11/dist-packag
Collecting tensorflow
  Downloading tensorflow-2.19.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x8
Requirement already satisfied: absl-py>=1.0.0 in /usr/local/lib/python3.11/dist-pa
Requirement already satisfied: astunparse>=1.6.0 in /usr/local/lib/python3.11/dist
Requirement already satisfied: flatbuffers>=24.3.25 in /usr/local/lib/python3.11/d
Requirement already satisfied: gast!=0.5.0,!=0.5.1,!=0.5.2,>=0.2.1 in /usr/local/l
Requirement already satisfied: google-pasta>=0.1.1 in /usr/local/lib/python3.11/di
Requirement already satisfied: libclang>=13.0.0 in /usr/local/lib/python3.11/dist-
Requirement already satisfied: opt-einsum>=2.3.2 in /usr/local/lib/python3.11/dist
Requirement already satisfied: packaging in /usr/local/lib/python3.11/dist-package
Requirement already satisfied: protobuf!=4.21.0,!=4.21.1,!=4.21.2,!=4.21.3,!=4.21.
Requirement already satisfied: requests<3,>=2.21.0 in /usr/local/lib/python3.11/di
Requirement already satisfied: setuptools in /usr/local/lib/python3.11/dist-packag
Requirement already satisfied: six>=1.12.0 in /usr/local/lib/python3.11/dist-packa
Requirement already satisfied: termcolor>=1.1.0 in /usr/local/lib/python3.11/dist-
Requirement already satisfied: typing-extensions>=3.6.6 in /usr/local/lib/python3.
Requirement already satisfied: wrapt>=1.11.0 in /usr/local/lib/python3.11/dist-pac
Requirement already satisfied: grpcio<2.0,>=1.24.3 in /usr/local/lib/python3.11/di
Collecting tensorboard~=2.19.0 (from tensorflow)
  Downloading tensorboard-2.19.0-py3-none-any.whl.metadata (1.8 kB)
Requirement already satisfied: keras>=3.5.0 in /usr/local/lib/python3.11/dist-pack
Requirement already satisfied: numpy<2.2.0,>=1.26.0 in /usr/local/lib/python3.11/d
Requirement already satisfied: h5py>=3.11.0 in /usr/local/lib/python3.11/dist-pack
Collecting ml-dtypes<1.0.0,>=0.5.1 (from tensorflow)
  Downloading ml_dtypes-0.5.1-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_
Requirement already satisfied: tensorflow-io-gcs-filesystem>=0.23.1 in /usr/local/
Requirement already satisfied: wheel<1.0,>=0.23.0 in /usr/local/lib/python3.11/dis
Requirement already satisfied: rich in /usr/local/lib/python3.11/dist-packages (fr
Requirement already satisfied: namex in /usr/local/lib/python3.11/dist-packages (f
Requirement already satisfied: optree in /usr/local/lib/python3.11/dist-packages (
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-pack
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dis
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dis
Requirement already satisfied: markdown>=2.6.8 in /usr/local/lib/python3.11/dist-p
Requirement already satisfied: tensorboard-data-server<0.8.0,>=0.7.0 in /usr/local
Requirement already satisfied: werkzeug>=1.0.1 in /usr/local/lib/python3.11/dist-p
Requirement already satisfied: MarkupSafe>=2.1.1 in /usr/local/lib/python3.11/dist
Requirement already satisfied: markdown-it-py>=2.2.0 in /usr/local/lib/python3.11/
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in /usr/local/lib/python3.1
Requirement already satisfied: mdurl~=0.1 in /usr/local/lib/python3.11/dist-packag
Downloading tensorflow-2.19.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 644.9/644.9 MB 1.3 MB/s eta 0:00:00
Downloading ml_dtypes-0.5.1-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 4.7/4.7 MB 49.4 MB/s eta 0:00:00
Downloading tensorboard-2.19.0-py3-none-any.whl (5.5 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 5.5/5.5 MB 50.0 MB/s eta 0:00:00
Installing collected packages: ml-dtypes, tensorboard, tensorflow
  Attempting uninstall: ml-dtypes
    Found existing installation: ml-dtypes 0.4.1
    Uninstalling ml-dtypes-0.4.1:
      Successfully uninstalled ml-dtypes-0.4.1
  Attempting uninstall: tensorboard
```

```
    Found existing installation: tensorboard 2.18.0
    Uninstalling tensorboard-2.18.0:
      Successfully uninstalled tensorboard-2.18.0
  Attempting uninstall: tensorflow
    Found existing installation: tensorflow 2.18.0
    Uninstalling tensorflow-2.18.0:
      Successfully uninstalled tensorflow-2.18.0
ERROR: pip's dependency resolver does not currently take into account all the pack
tensorflow-text 2.18.1 requires tensorflow<2.19,>=2.18.0, but you have tensorflow
tf-keras 2.18.0 requires tensorflow<2.19,>=2.18, but you have tensorflow 2.19.0 wh
Successfully installed ml-dtypes-0.5.1 tensorboard-2.19.0 tensorflow-2.19.0
```

```
!pip install openai==0.27.8
```

```
Requirement already satisfied: openai==0.27.8 in /usr/local/lib/python3.11/dist-packa
Requirement already satisfied: requests>=2.20 in /usr/local/lib/python3.11/dist-packa
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-packages (from
Requirement already satisfied: aiohttp in /usr/local/lib/python3.11/dist-packages (fr
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-package
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-p
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-p
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in /usr/local/lib/python3.11/d
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.11/dist-pac
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-packag
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.11/dist-pa
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.11/dist-
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.11/dist-pac
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.11/dist-pa
```

```
!pip install openai --upgrade
```

```
Requirement already satisfied: openai in /usr/local/lib/python3.11/dist-packages (0.2
Collecting openai
  Downloading openai-1.71.0-py3-none-any.whl.metadata (25 kB)
Requirement already satisfied: anyio<5,>=3.5.0 in /usr/local/lib/python3.11/dist-pack
Requirement already satisfied: distro<2,>=1.7.0 in /usr/local/lib/python3.11/dist-pac
Requirement already satisfied: httpx<1,>=0.23.0 in /usr/local/lib/python3.11/dist-pac
Requirement already satisfied: jiter<1,>=0.4.0 in /usr/local/lib/python3.11/dist-pack
Requirement already satisfied: pydantic<3,>=1.9.0 in /usr/local/lib/python3.11/dist-p
Requirement already satisfied: sniffio in /usr/local/lib/python3.11/dist-packages (fr
Requirement already satisfied: tqdm>4 in /usr/local/lib/python3.11/dist-packages (fro
Requirement already satisfied: typing-extensions<5,>=4.11 in /usr/local/lib/python3.1
Requirement already satisfied: idna>=2.8 in /usr/local/lib/python3.11/dist-packages (
Requirement already satisfied: certifi in /usr/local/lib/python3.11/dist-packages (fr
Requirement already satisfied: httpcore==1.* in /usr/local/lib/python3.11/dist-packag
Requirement already satisfied: h11<0.15,>=0.13 in /usr/local/lib/python3.11/dist-pack
Requirement already satisfied: annotated-types>=0.6.0 in /usr/local/lib/python3.11/di
Requirement already satisfied: pydantic-core==2.33.1 in /usr/local/lib/python3.11/dis
Requirement already satisfied: typing-inspection>=0.4.0 in /usr/local/lib/python3.11/
Downloading openai-1.71.0-py3-none-any.whl (598 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 599.0/599.0 kB 10.9 MB/s eta 0:00:00
Installing collected packages: openai
  Attempting uninstall: openai
    Found existing installation: openai 0.27.8
```

```
    Uninstalling openai-0.27.8:
      Successfully uninstalled openai-0.27.8
Successfully installed openai-1.71.0
```

```
!pip install --upgrade openai --quiet
```

```
from bertopic import BERTopic
from sklearn.cluster import DBSCAN
from sentence_transformers import SentenceTransformer
from sklearn.feature_extraction.text import CountVectorizer
Embdng_M = SentenceTransformer("all-MiniLM-L6-v2")
```

```
Embdngs = Embdng_M.encode(df3['text'].tolist(), show_progress_bar=True)
DBscan_M = DBSCAN(eps=0.3, min_samples=3, metric='cosine')
```

```
Batches: 100%                                                    63/63 [03:00<00:00,  1.68it/s]
```

```
TM = BERTopic(
    embedding_model=Embdng_M,
    hdbscan_model=DBscan_M,
    vectorizer_model=CountVectorizer(ngram_range=(1, 2)),
    verbose=True
)
Tpcs, probs = TM.fit_transform(dataframe_3['text'], Embdngs)
```

```
2025-04-08 03:41:30,987 - BERTopic - Dimensionality - Fitting the dimensionality redu
2025-04-08 03:41:59,505 - BERTopic - Dimensionality - Completed ✓
2025-04-08 03:41:59,507 - BERTopic - Cluster - Start clustering the reduced embedding
2025-04-08 03:41:59,622 - BERTopic - Cluster - Completed ✓
2025-04-08 03:41:59,645 - BERTopic - Representation - Fine-tuning topics using repres
2025-04-08 03:42:02,810 - BERTopic - Representation - Completed ✓
```

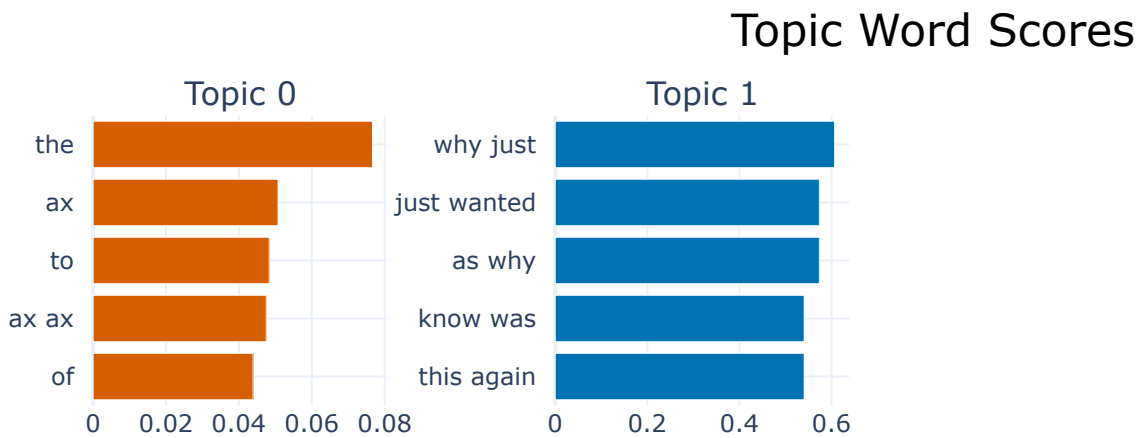
```
print(TM.get_topic_info())
for i in topic_model.get_topics().keys():
    print(f"Topic {i}: {topic_model.get_topic(i)}")
```

```
   Topic  Count                                   Name  \
0      0   1938                      0_the_ax_to_ax ax
1      1     62  1_why just_just wanted_as why_know was

                                        Representation  \
0    [the, ax, to, ax ax, of, and, in, is, that, it]
1  [why just, just wanted, as why, know was, this...

                                   Representative_Docs
0  [\n[ stuff deleted ]\n   |> Are you calling na...
1  [\nSuch as?, \nNot this again.\n, I just wante...
Topic 0: [('the', np.float64(0.07663833162992664)), ('ax', np.float64(0.0507332504185
Topic 1: [('why just', np.float64(0.6067108212902631)), ('just wanted', np.float64(0.
```

```
Tpc_Info = TM.get_topic_info()
print(Tpc_Info)
```

```
   Topic  Count                                   Name  \
0      0   1938                      0_the_ax_to_ax ax
1      1     62  1_why just_just wanted_as why_know was

                                       Representation  \
0    [the, ax, to, ax ax, of, and, in, is, that, it]
1  [why just, just wanted, as why, know was, this...

                                  Representative_Docs
0  [\n[ stuff deleted ]\n   |> Are you calling na...
1  [\nSuch as?, \nNot this again.\n, I just wante...
```

```
TM.visualize_barchart(top_n_topics=5)
```

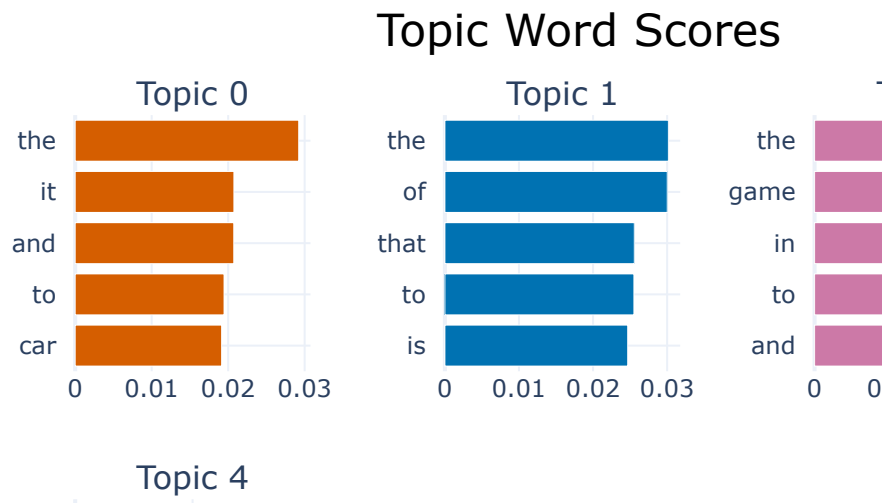



```
Embdngs = Embdng_M.encode(df3['text'].tolist(), show_progress_bar=True)
```

Batches: 100% 63/63 [03:08<00:00, 1.46it/s]

```
TM_Def = BERTopic()
Tpc, Prb = TM_Def.fit_transform(df3['text'])
```

```
TM_Def.visualize_barchart(top_n_topics=5)
```


Topic Word Scores
Topic 0
the
it
and
to
car
0
0.01
0.02
0.03
Topic 1
the
of
that
to
is
0
0.01
0.02
0.03
the
game
in
to
and
0
0
Topic 4