CWordTM (0.7.4) Toolkit Usage on BBC News

This Jupyter notebook demonstrates how to use the package "CWordTM" on the BBC News:

1. Meta Information Features
2. Utility Features
3. Text Visualization - Word Cloud
4. Text Summarization
5. Pivot Table
6. Topic Modeling - LDA, BERTopic and NMF

CWordTM Toolkit's Documentation: https://cwordtm.readthedocs.io

In [1]:
```
import warnings
warnings.filterwarnings('ignore')
```

1. Meta Information Features

In [2]:
```
import cwordtm
from cwordtm import *
```

In [3]:
```
# Show brief module information
print(meta.get_module_info())
```

```
The member information of the module 'cwordtm'
1. Submodule meta:
   addin (func, *, timing=False, code=0)
   addin_all (modname='cwordtm', *, timing=False, code=0)
   addin_all_functions (submod, *, timing=False, code=0)
   get_function (mod_name, submodules, func_name, *, timing=False, code=0)
   get_module_info (detailed=False, *, timing=False, code=0)
   get_submodule_info (submodname, detailed=False, *, timing=False, code=0)
   import_module (name, package=None, *, timing=False, code=0)
   wraps (wrapped, assigned=('__module__', '__name__', '__qualname__', '__doc__', '__annotations__'), updated=('__di
ct__',), *, timing=False, code=0)
2. Submodule pivot:
   pivot (df, value='text', category='category', *, timing=False, code=0)
   stat (df, chi=False, *, timing=False, code=0)
3. Submodule quot:
   extract_quotation (text, quot_marks, *, timing=False, code=0)
   match_text (target, sent_tokens, lang, threshold, n=5, *, timing=False, code=0)
   match_verse (i, ot_list, otdf, df, book, chap, verse, lang, threshold, *, timing=False, code=0)
   show_quot (target, source='ot', lang='en', threshold=0.5, *, timing=False, code=0)
   tokenize (sentence, *, timing=False, code=0)
4. Submodule ta:
   get_sent_scores (sentences, diction, sent_len, *, timing=False, code=0) -> dict
   get_sentences (docs, lang='en', *, timing=False, code=0)
   get_summary (sentences, sent_weight, threshold, sent_len, *, timing=False, code=0)
   pos_tag (tokens, tagset=None, lang='eng', *, timing=False, code=0)
   preprocess_sent (text, *, timing=False, code=0)
   sent_tokenize (text, language='english', *, timing=False, code=0)
   split_chi_sentences (text, *, timing=False, code=0)
   summary_chi (docs, weight=1.5, sent_len=8, *, timing=False, code=0)
   summary_en (docs, sent_len=8, *, timing=False, code=0)
   word_tokenize (text, language='english', preserve_line=False, *, timing=False, code=0)
5. Submodule tm:
   BTM (doc_file, num_topics, chi=False, embed=True)
   LDA (doc_file, num_topics, chi=False)
   NMF (doc_file, num_topics, chi=False)
   btm_process (doc_file, num_topics=10, source=0, text_col='text', doc_size=0, cat=0, chi=False, group=True, eval=F
alse, web_app=False, *, timing=False, code=0)
   lda_process (doc_file, num_topics=10, source=0, text_col='text', doc_size=0, cat=0, chi=False, group=True, eval=F
alse, web_app=False, *, timing=False, code=0)
   load_bible (textfile, cat=0, group=True, *, timing=False, code=0)
   load_text (textfile, doc_size=0, text_col='text', *, timing=False, code=0)
   ngrams (sequence, n, *, timing=False, code=0, **kwargs)
   nmf_process (doc_file, num_topics=10, source=0, text_col='text', doc_size=0, cat=0, chi=False, group=True, eval=F
alse, web_app=False, *, timing=False, code=0)
   pprint (object, stream=None, indent=1, width=80, depth=None, *, compact=False, sort_dicts=True, underscore_number
s=False, timing=False, code=0)
   process_text (doc, *, timing=False, code=0)
6. Submodule util:
   add_chi_vocab (*, timing=False, code=0)
   bible_cat_info (lang='en', *, timing=False, code=0)
   chi_sent_terms (text, *, timing=False, code=0)
   chi_stops (*, timing=False, code=0)
   clean_sentences (sentences, *, timing=False, code=0)
   clean_text (df, text_col='text', *, timing=False, code=0)
   extract (df, testament=-1, category='', book=0, chapter=0, verse=0, *, timing=False, code=0)
   extract2 (df, filter='', *, timing=False, code=0)
   get_diction (docs, *, timing=False, code=0)
   get_diction_chi (docs, *, timing=False, code=0)
   get_diction_en (docs, *, timing=False, code=0)
   get_list (df, column='book', *, timing=False, code=0)
   get_sent_terms (text, *, timing=False, code=0)
   get_text (df, text_col='text', *, timing=False, code=0)
   get_text_list (df, text_col='text', *, timing=False, code=0)
   group_text (df, column='chapter', *, timing=False, code=0)
   is_chi (*, timing=False, code=0)
   load_csv (file_obj, doc_size=0, info=False, *, timing=False, code=0)
   load_text (file_obj, doc_size=0, info=False, *, timing=False, code=0)
   load_word (ver='web.csv', nr=0, info=False, *, timing=False, code=0)
   preprocess_text (text, *, timing=False, code=0)
   remove_noise (text, noise_list, *, timing=False, code=0)
   reset_rows (*, timing=False, code=0)
   set_lang (lang='en', *, timing=False, code=0)
   set_rows (n=None, *, timing=False, code=0)
   word_tokenize (text, language='english', preserve_line=False, *, timing=False, code=0)
7. Submodule version:
8. Submodule viz:
   chi_wordcloud (docs, figsize=(15, 10), bg='white', image=0, web_app=False, *, timing=False, code=0)
   plot_cloud (wordcloud, figsize, web_app=False, *, timing=False, code=0)
   show_wordcloud (docs, clean=False, figsize=(12, 8), bg='white', image=0, web_app=False, *, timing=False, code=0)
```

In [4]:
```
# Show detailed module information of a submodule
print(meta.get_submodule_info("viz", detailed=True))
```

```
The function(s) of the submodule 'cwordtm.viz':

def chi_wordcloud(docs, figsize=(15, 10), bg='white', image=0, web_app=False):
    """Prepare and show a Chinese wordcloud

    :param docs: The collection of Chinese documents for preparing a wordcloud,
        default to None
    :type docs: pandas.DataFrame
    :param figsize: Size (width, height) of word cloud, default to (15, 10)
    :type figsize: tuple, optional
    :param bg: The background color (name) of the wordcloud, default to 'white'
    :type bg: str, optional
    :param image: The filename of the presribed image as the mask of the wordcloud,
        or 1/2/3/4 for using an internal image (heart / disc / triangle / arrow),
        default to 0 (No image mask)
    :type image: int or str or BytesIO, optional
    :param web_app: The flag indicating the function is initiated from a web
        application, default to False
    :type web_app: bool
    :return: The wordcloud figure
    :rtype: matplotlib.pyplot.figure
    """

    util.set_lang('chi')
    diction = util.get_diction(docs)

    masks = ['heart.jpg', 'disc.jpg', 'triangle.jpg', 'arrow.jpg']

    if image == 0:
        mask = None
    elif image in [1, 2, 3, 4]:  # Internal image file
        img_file = files('cwordtm.images').joinpath(masks[image-1])
        mask = np.array(Image.open(img_file))
    elif isinstance(image, str) and len(image) > 0:
        mask = np.array(Image.open(image))
    elif isinstance(image, BytesIO):
        mask = np.array(Image.open(BytesIO(image.getvalue())))
    else:
        mask = None

    font_file = files('cwordtm.data').joinpath('msyh.ttc')
    # wordcloud = WordCloud(background_color=bg, colormap='Set2',
    wordcloud = WordCloud(background_color=bg, colormap='rainbow',
                          mask=mask, font_path=str(font_file)) \
                    .generate_from_frequencies(frequencies=diction)

    return plot_cloud(wordcloud, figsize=figsize, web_app=web_app)

def plot_cloud(wordcloud, figsize, web_app=False):
    """Plot the prepared 'wordcloud'

    :param wordcloud: The WordCloud object for plotting, default to None
    :type wordcloud: WordCloud object
    :param figsize: Size (width, height) of word cloud, default to None
    :type figsize: tuple
    :param web_app: The flag indicating the function is initiated from a web
        application, default to False
    :type web_app: bool
    :return: The wordcloud figure
    :rtype: matplotlib.pyplot.figure
    """

    fig = plt.figure(figsize=figsize)
    plt.axis("off");
    plt.imshow(wordcloud)
    if web_app: return fig

def show_wordcloud(docs, clean=False, figsize=(12, 8), bg='white', image=0, web_app=False):
    """Prepare and show a wordcloud

    :param docs: The collection of documents for preparing a wordcloud,
        default to None
    :type docs: pandas.DataFrame
    :param clean: The flag whether text preprocessing is needed,
        default to False
    :type clean: bool, optional
    :param figsize: Size (width, height) of word cloud, default to (12, 8)
    :type figsize: tuple, optional
    :param bg: The background color (name) of the wordcloud, default to 'white'
    :type bg: str, optional
    :param image: The filename of the presribed image as the mask of the wordcloud,
        or 1/2/3/4 for using an internal image (heart / disc / triangle / arrow),
        default to 0 (No image mask)
    :type image: int or str or BytesIO, optional
    :param web_app: The flag indicating the function is initiated from a web
        application, default to False
```

```
    :type web_app: bool
    :return: The wordcloud figure
    :rtype: matplotlib.pyplot.figure
    """

    masks = ['heart.jpg', 'disc.jpg', 'triangle.jpg', 'arrow.jpg']

    if image == 0:
        mask = None
    elif image in [1, 2, 3, 4]:  # Internal image file
        img_file = files('cwordtm.images').joinpath(masks[image-1])
        mask = np.array(Image.open(img_file))
    elif isinstance(image, str) and len(image) > 0:
        mask = np.array(Image.open(image))
    elif isinstance(image, BytesIO):
        mask = np.array(Image.open(BytesIO(image.getvalue())))
    else:
        mask = None

    if isinstance(docs, pd.DataFrame):
        docs = ' '.join(list(docs.text.astype(str)))
    elif isinstance(docs, pd.Series):
        docs = ' '.join(list(docs.astype(str)))
    elif isinstance(docs, list) or isinstance(docs, np.ndarray):
        docs = ' '.join(str(doc) for doc in docs)

    if clean:
        docs = util.preprocess_text(docs)

    # wordcloud = WordCloud(background_color=bg, colormap='Set2', mask=mask) \
    wordcloud = WordCloud(background_color=bg, colormap='rainbow', mask=mask) \
                    .generate(docs)

    return plot_cloud(wordcloud, figsize=figsize, web_app=web_app)
```

In [5]:
```
# Show execution time
bbc_news = "BBC/BBC News Train.csv"
df = util.load_csv(bbc_news, timing=True)
```

```
Finished 'load_csv' in 0.0977 secs
```

In [6]:
```
# Execute and show code
df = util.load_csv(bbc_news, code=1)
```

```
def load_csv(file_obj, doc_size=0, info=False):
    """Loads a CSV file with a "text" column.

    :param file_obj: The prescribed file path from which the text is loaded,
        or a BytesIO object from Streamlit's file_uploader, default to None
    :type file_obj: str or io.BytesIO
    :param doc_size: The number of documents to be loaded, 0 represents all documents,
        or the range (tuple) of documents to be processed, default to 0
    :type doc_size: int, tuple, optional
    :param info: The flag whether the dataset information is shown,
        default to False
    :type info: bool, optional
    :return: The collection of text with the prescribed number of rows loaded
    :rtype: pandas.DataFrame
    """

    # print("Loading file '%s' ..." %filepath)
    if isinstance(file_obj, BytesIO):
        fname = file_obj.name
    else:
        fname = str(file_obj)

    if fname.lower().endswith('csv'):
        df = pd.read_csv(file_obj, encoding='utf-8')
    else:  # text file
        if isinstance(file_obj, BytesIO):
            stringio = StringIO(file_obj.getvalue().decode("utf-8"))
            lines = list(stringio.read().split('\n'))
        else:
            tf = open(file_obj, encoding='utf-8')
            lines = [line.strip() for line in tf.readlines()]

        df = pd.DataFrame({'text': lines})

    if isinstance(doc_size, int):
        if doc_size > 0:
            df = df.iloc[:doc_size]
    elif isinstance(doc_size, tuple):
        df = df.iloc[doc_size[0]-1:doc_size[1]]

    noise_list = ['\u3000', '— ', '•']
    for noise in noise_list:
        df['text'] = df['text'].str.replace(noise, '')

    if info:
        print("\nDataset Information:")
        df.info()

    return df
```

In [7]:
```
# Show code without execution
df = util.load_csv(bbc_news, code=2)
```

```
def load_csv(file_obj, doc_size=0, info=False):
    """Loads a CSV file with a "text" column.

    :param file_obj: The prescribed file path from which the text is loaded,
        or a BytesIO object from Streamlit's file_uploader, default to None
    :type file_obj: str or io.BytesIO
    :param doc_size: The number of documents to be loaded, 0 represents all documents,
        or the range (tuple) of documents to be processed, default to 0
    :type doc_size: int, tuple, optional
    :param info: The flag whether the dataset information is shown,
        default to False
    :type info: bool, optional
    :return: The collection of text with the prescribed number of rows loaded
    :rtype: pandas.DataFrame
    """

    # print("Loading file '%s' ..." %filepath)
    if isinstance(file_obj, BytesIO):
        fname = file_obj.name
    else:
        fname = str(file_obj)

    if fname.lower().endswith('csv'):
        df = pd.read_csv(file_obj, encoding='utf-8')
    else:  # text file
        if isinstance(file_obj, BytesIO):
            stringio = StringIO(file_obj.getvalue().decode("utf-8"))
            lines = list(stringio.read().split('\n'))
        else:
            tf = open(file_obj, encoding='utf-8')
            lines = [line.strip() for line in tf.readlines()]

        df = pd.DataFrame({'text': lines})

    if isinstance(doc_size, int):
        if doc_size > 0:
            df = df.iloc[:doc_size]
    elif isinstance(doc_size, tuple):
        df = df.iloc[doc_size[0]-1:doc_size[1]]

    noise_list = ['\u3000', '— ', '•']
    for noise in noise_list:
        df['text'] = df['text'].str.replace(noise, '')

    if info:
        print("\nDataset Information:")
        df.info()

    return df
```

In [8]:
```
# Add timing and code reveal features to some other function
from importlib_resources import files
files = meta.addin(files)
files(code=2)
```

```
@package_to_anchor
def files(anchor: Optional[Anchor] = None) -> Traversable:
    """
    Get a Traversable resource for an anchor.
    """
    return from_package(resolve(anchor))
```

2. Utility Features

Load BBC News

In [9]:
```
df = util.load_csv(bbc_news, info=True)
```

```
Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1490 entries, 0 to 1489
Data columns (total 3 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   articleId  1490 non-null   int64
 1   text       1490 non-null   object
 2   category   1490 non-null   object
dtypes: int64(1), object(2)
memory usage: 35.0+ KB
```

Preprocessing Text

In [10]:
```
import nltk
nltk.download('punkt_tab', quiet=True)
```

Out[10]:
```
True
```

In [11]:
```
text_list = util.get_text_list(df.iloc[:500], text_col='text')
text = util.preprocess_text(text_list)
```

3. Text Visualization - Word Cloud

In [12]:
```
# White background with no image mask
viz.show_wordcloud(text)
```

```
C:\Dev\Anaconda3\envs\aiml\lib\site-packages\wordcloud\wordcloud.py:106: MatplotlibDeprecationWarning: The get_cmap function was deprecated in Matplotlib 3.7 and will be removed two minor releases later. Use ``matplotlib.colormaps[name]`` or ``matplotlib.colormaps.get_cmap(obj)`` instead.
  self.colormap = plt.cm.get_cmap(colormap)
```

In [13]:
```
# Black background with the prescribed image as the mask
viz.show_wordcloud(text, bg='black', image=2)  # Disc
```

```
C:\Dev\Anaconda3\envs\aiml\lib\site-packages\wordcloud\wordcloud.py:106: MatplotlibDeprecationWarning: The get_cmap function was deprecated in Matplotlib 3.7 and will be removed two minor releases later. Use ``matplotlib.colormaps[name]`` or ``matplotlib.colormaps.get_cmap(obj)`` instead.
  self.colormap = plt.cm.get_cmap(colormap)
```

4. Text Summarization

In [14]:
```
import nltk
nltk.download('averaged_perceptron_tagger_eng', quiet=True)
```

Out[14]:
```
True
```

In [15]:
```
news = df.iloc[0]['text']  # "df" stores previously loaded news
ta.summary_en(news)
```

Out[15]:
```
['prosecution lawyers have argued that mr ebbers orchestrated a series of accounting tricks at worldcom  ordering em
ployees to hide expenses and inflate revenues to meet wall street earnings estimates.',
 'but ms cooper  who now runs her own consulting business  told a jury in new york on wednesday that external audito
rs arthur andersen had approved worldcom s accounting in early 2001 and 2002. she said andersen had given a  green l
ight  to the procedures and practices used by worldcom.',
 'ms cooper also said that during shareholder meetings mr ebbers often passed over technical questions to the compan
y s finance chief  giving only  brief  answers himself.',
 'cynthia cooper  worldcom s ex-head of internal accounting  alerted directors to irregular accounting practices at
the us telecoms giant in 2002. her warnings led to the collapse of the firm following the discovery of an $11bn (?']
```

5. Pivot Table

In [16]:
```
pivot.pivot(df)
```

Out[16]:

	text
category	
business	336
entertainment	273
politics	274
sport	346
tech	261
Total	1490

In [17]:
```
len(df)
```

Out[17]:
```
1490
```

6. Topic Modeling

In [18]:
```
import warnings
warnings.filterwarnings('ignore')
```

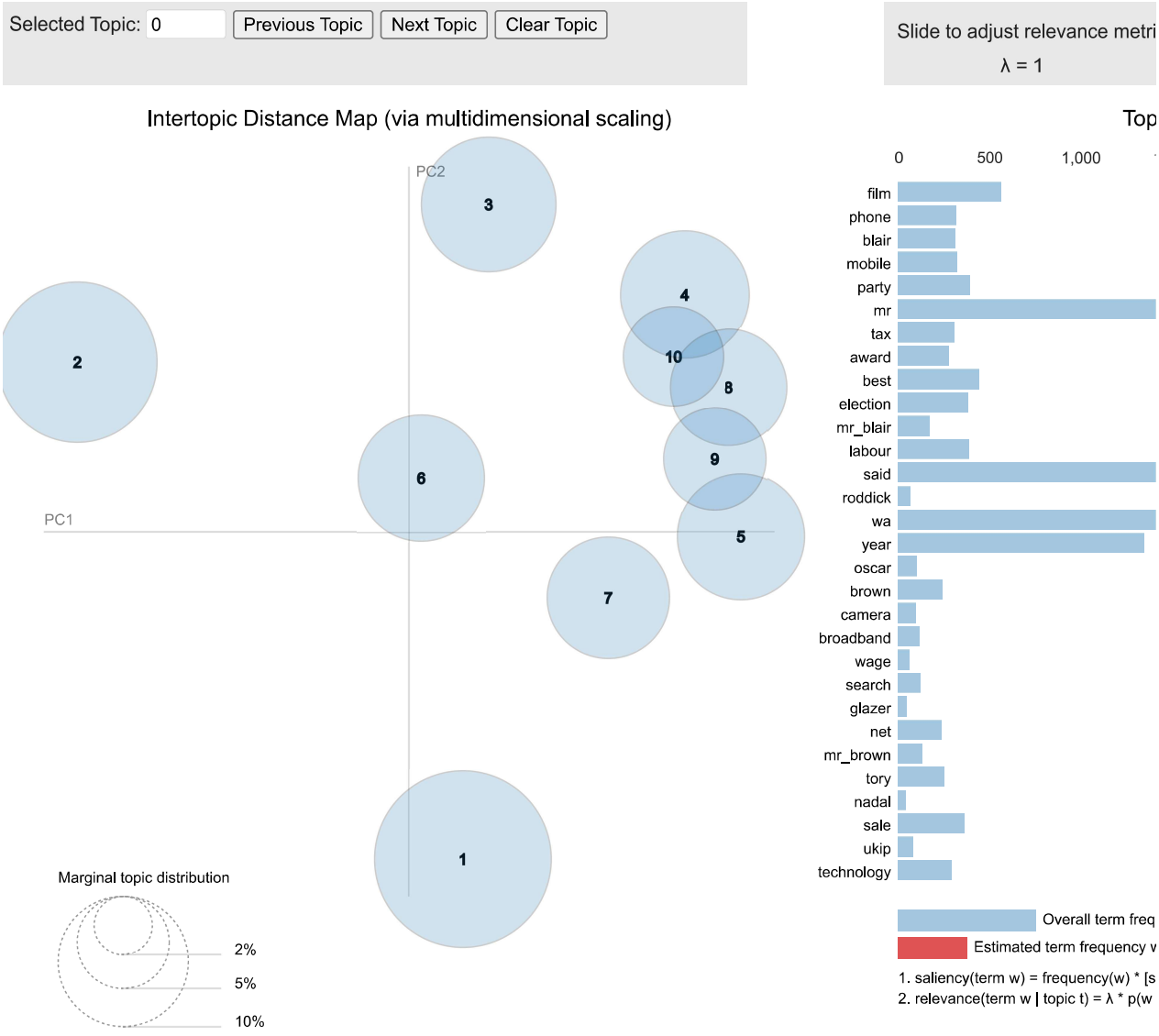
LDA Modeling

In [19]:
```
lda = tm.lda_process(bbc_news, source=1, text_col='text', eval=True, timing=True)
```

```
Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1491 entries, 0 to 1490
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   text    1491 non-null   object
dtypes: object(1)
memory usage: 11.8+ KB
Corpus loaded!
Text preprocessed!
Text trained!
You may execute the following commands to show the visualization:
    import pyLDAvis
    pyLDAvis.display(lda.vis_data)

Visualization prepared!

Topics from LDA Model:
[(0,
  '0.004*"wa" + 0.004*"said" + 0.003*"ha" + 0.003*"year" + 0.001*"film" + '
  '0.001*"mobile" + 0.001*"people" + 0.001*"new" + 0.001*"sale" + '
  '0.001*"game"'),
 (1,
  '0.005*"said" + 0.005*"wa" + 0.004*"ha" + 0.003*"film" + 0.002*"year" + '
  '0.002*"best" + 0.002*"award" + 0.002*"people" + 0.001*"new" + 0.001*"mr"'),
 (2,
  '0.008*"said" + 0.006*"mr" + 0.005*"ha" + 0.005*"wa" + 0.003*"year" + '
  '0.002*"people" + 0.002*"party" + 0.002*"tax" + 0.002*"labour" + '
  '0.002*"say"'),
 (3,
  '0.005*"said" + 0.003*"wa" + 0.003*"ha" + 0.002*"phone" + 0.002*"people" + '
  '0.002*"year" + 0.002*"mobile" + 0.002*"new" + 0.001*"uk" + 0.001*"game"'),
 (4,
  '0.007*"said" + 0.006*"wa" + 0.003*"ha" + 0.002*"year" + 0.002*"new" + '
  '0.001*"people" + 0.001*"mr" + 0.001*"game" + 0.001*"world" + 0.001*"time"'),
 (5,
  '0.004*"said" + 0.004*"ha" + 0.004*"wa" + 0.002*"year" + 0.002*"people" + '
  '0.002*"mr" + 0.001*"time" + 0.001*"say" + 0.001*"film" + 0.001*"game"'),
 (6,
  '0.006*"said" + 0.005*"ha" + 0.004*"wa" + 0.003*"mr" + 0.003*"year" + '
  '0.001*"new" + 0.001*"business" + 0.001*"company" + 0.001*"time" + '
  '0.001*"club"'),
 (7,
  '0.005*"said" + 0.003*"ha" + 0.002*"wa" + 0.002*"year" + 0.002*"new" + '
  '0.001*"mr" + 0.001*"net" + 0.001*"uk" + 0.001*"people" + 0.001*"game"'),
 (8,
  '0.007*"said" + 0.006*"wa" + 0.003*"ha" + 0.003*"mr" + 0.002*"year" + '
  '0.002*"people" + 0.001*"blair" + 0.001*"say" + 0.001*"time" + '
  '0.001*"world"'),
 (9,
  '0.007*"wa" + 0.006*"said" + 0.005*"ha" + 0.003*"new" + 0.002*"mr" + '
  '0.002*"year" + 0.002*"game" + 0.002*"people" + 0.001*"government" + '
  '0.001*"time"')]

Model Evaluation Scores:
  Coherence: 0.6542712327796304
  Perplexity: -11.237065066541756
  Topic diversity: 0.0007468373723517731
  Topic size distribution: 0.006172839506172839

Finished 'lda_process' in 161.5518 secs
```

In [20]:
```
# LDA Model Visualization
import pyLDAvis
pyLDAvis.display(lda.vis_data)
```

Out[20]:



(2)

Save LDA Model

In [21]:
```
lda.save("models/lda_bbc.gensim")
```

```
LDA model has been stored in 'models/lda_bbc.gensim'.
```

Load LDA Model

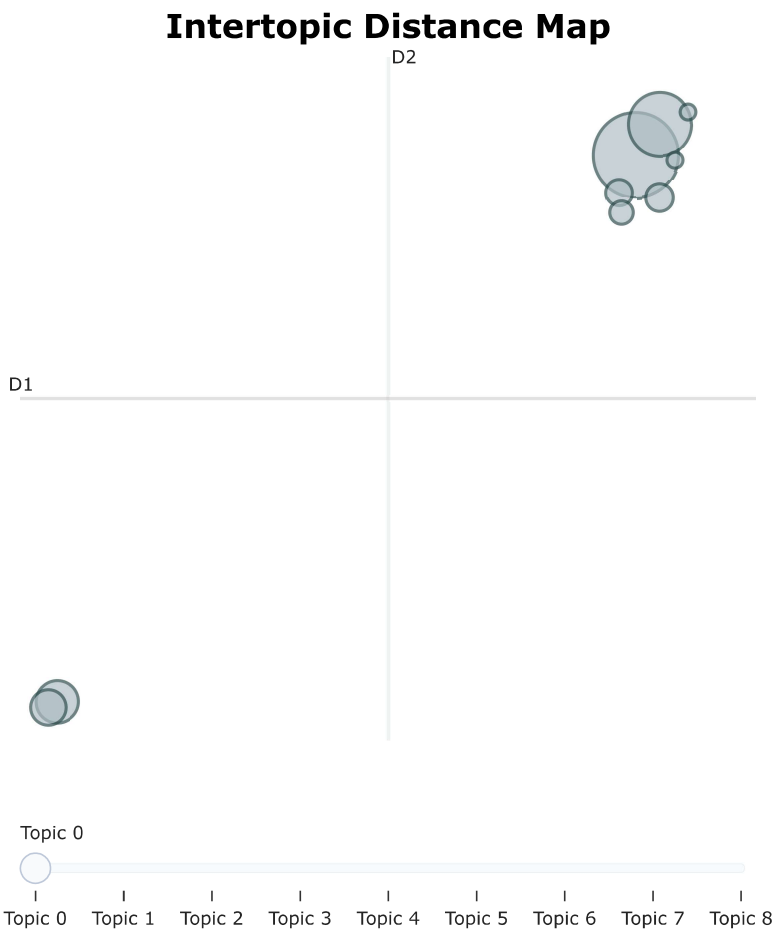
In [22]:
```
lda2 = tm.LDA("", lda.num_topics)
lda2.model = lda2.load("models/lda_bbc.gensim")
lda2.show_topics()
```

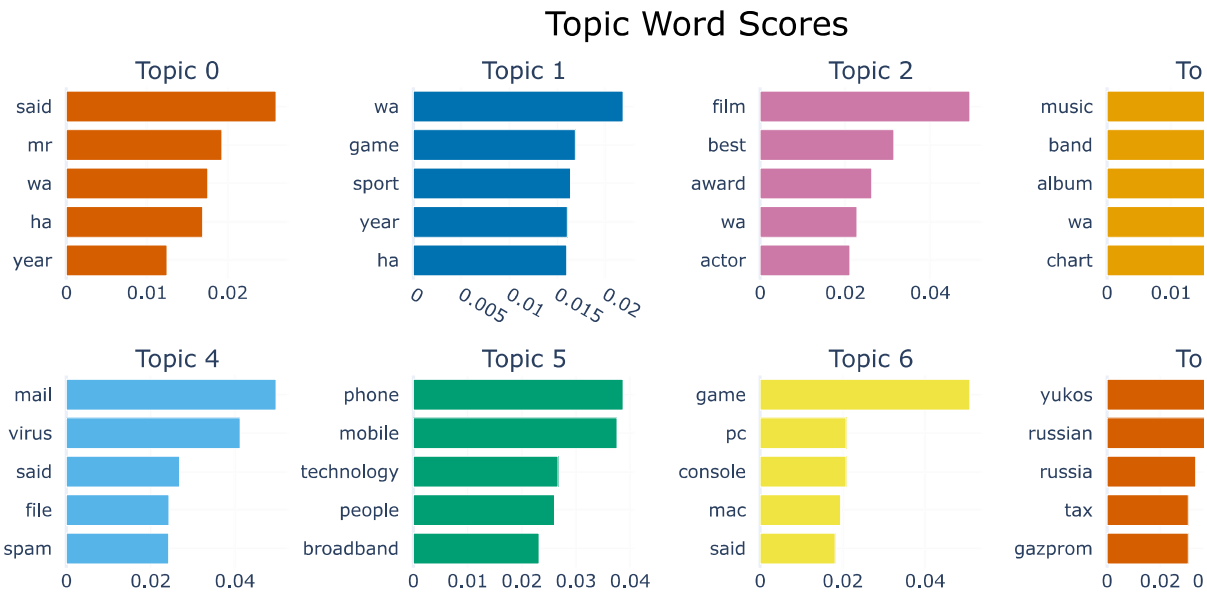
```
Topics from LDA Model:
[(0,
  '0.004*"wa" + 0.004*"said" + 0.003*"ha" + 0.003*"year" + 0.001*"film" + '
  '0.001*"mobile" + 0.001*"people" + 0.001*"new" + 0.001*"sale" + '
  '0.001*"game"'),
 (1,
  '0.005*"said" + 0.005*"wa" + 0.004*"ha" + 0.003*"film" + 0.002*"year" + '
  '0.002*"best" + 0.002*"award" + 0.002*"people" + 0.001*"new" + 0.001*"mr"'),
 (2,
  '0.008*"said" + 0.006*"mr" + 0.005*"ha" + 0.005*"wa" + 0.003*"year" + '
  '0.002*"people" + 0.002*"party" + 0.002*"tax" + 0.002*"labour" + '
  '0.002*"say"'),
 (3,
  '0.005*"said" + 0.003*"wa" + 0.003*"ha" + 0.002*"phone" + 0.002*"people" + '
  '0.002*"year" + 0.002*"mobile" + 0.002*"new" + 0.001*"uk" + 0.001*"game"'),
 (4,
  '0.007*"said" + 0.006*"wa" + 0.003*"ha" + 0.002*"year" + 0.002*"new" + '
  '0.001*"people" + 0.001*"mr" + 0.001*"game" + 0.001*"world" + 0.001*"time"'),
 (5,
  '0.004*"said" + 0.004*"ha" + 0.004*"wa" + 0.002*"year" + 0.002*"people" + '
  '0.002*"mr" + 0.001*"time" + 0.001*"say" + 0.001*"film" + 0.001*"game"'),
 (6,
  '0.006*"said" + 0.005*"ha" + 0.004*"wa" + 0.003*"mr" + 0.003*"year" + '
  '0.001*"new" + 0.001*"business" + 0.001*"company" + 0.001*"time" + '
  '0.001*"club"'),
 (7,
  '0.005*"said" + 0.003*"ha" + 0.002*"wa" + 0.002*"year" + 0.002*"new" + '
  '0.001*"mr" + 0.001*"net" + 0.001*"uk" + 0.001*"people" + 0.001*"game"'),
 (8,
  '0.007*"said" + 0.006*"wa" + 0.003*"ha" + 0.003*"mr" + 0.002*"year" + '
  '0.002*"people" + 0.001*"blair" + 0.001*"say" + 0.001*"time" + '
  '0.001*"world"'),
 (9,
  '0.007*"wa" + 0.006*"said" + 0.005*"ha" + 0.003*"new" + 0.002*"mr" + '
  '0.002*"year" + 0.002*"game" + 0.002*"people" + 0.001*"government" + '
  '0.001*"time"')]
```

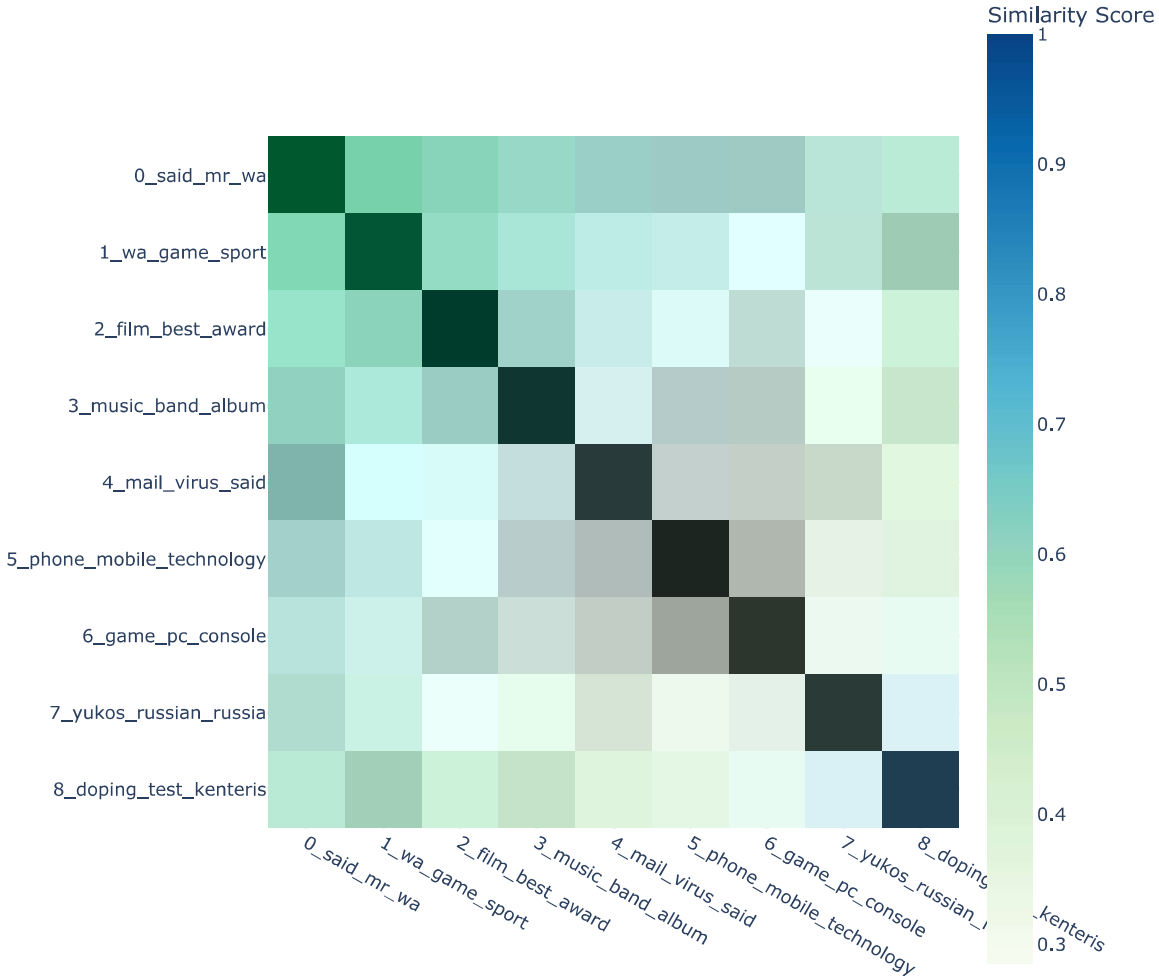
BERTopic Modeling

In [23]:
```
btm = tm.btm_process(bbc_news, source=1, text_col='text', eval=True, timing=True)
```

```
Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1491 entries, 0 to 1490
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   text    1491 non-null   object
dtypes: object(1)
memory usage: 11.8+ KB
Corpus loaded!
Text preprocessed!
Text trained!

Topics from BERTopic Model:
Topic 0: said | mr | wa | ha | year | government | election | bn | labour | business
Topic 1: wa | game | sport | year | ha | england | said | time | win | half
Topic 2: film | best | award | wa | actor | oscar | star | year | entertainment | ha
Topic 3: music | band | album | wa | chart | song | single | year | said | singer
Topic 4: mail | virus | said | file | spam | anti | software | security | site | user
Topic 5: phone | mobile | technology | people | broadband | said | camera | service | digital | gadget
Topic 6: game | pc | console | mac | said | nintendo | machine | computer | gaming | mini
Topic 7: yukos | russian | russia | tax | gazprom | oil | company | ha | bn | court
Topic 8: doping | test | kenteris | iaaf | conte | greek | drug | thanou | sprinter | sport

Model Evaluation Scores:
  Coherence: 0.6442308990836441

BERTopic Model Visualization:
```


Intertopic Distance Map
D2
D1
Topic 0
Topic 0
Topic 1
Topic 2
Topic 3
Topic 4
Topic 5
Topic 6
Topic 7
Topic 8


Topic Word Scores
Topic 0
said
mr
wa
ha
year
0
0.01
0.02
Topic 1
wa
game
sport
year
ha
0
0.005
0.01
0.015
0.02
Topic 2
film
best
award
wa
actor
0
0.02
0.04
To
music
band
album
wa
chart
0
0.01
Topic 4
mail
virus
said
file
spam
0
0.02
0.04
Topic 5
phone
mobile
technology
people
broadband
0
0.01
0.02
0.03
0.04
Topic 6
game
pc
console
mac
said
0
0.02
0.04
To
yukos
russian
russia
tax
gazprom
0
0.02
0




```
Finished 'btm_process' in 356.9247 secs
<Figure size 1000x800 with 0 Axes>
```

Save BERTopic Model

In [24]:
```
btm.save("models/bertopic_bbc.pickle")
```

```
2025-02-01 21:07:42,245 - BERTopic - WARNING: When you use `pickle` to save/load a BERTopic model,please make sure t
hat the environments in which you saveand load the model are **exactly** the same. The version of BERTopic,its depen
dencies, and python need to remain the same.
BERTopic model has been stored in 'models/bertopic_bbc.pickle'.
```

Load BERTopic Model

In [25]:
```
btm2 = tm.BTM("", btm.num_topics)
btm2.model = btm2.load("models/bertopic_bbc.pickle")
btm2.show_topics()
```

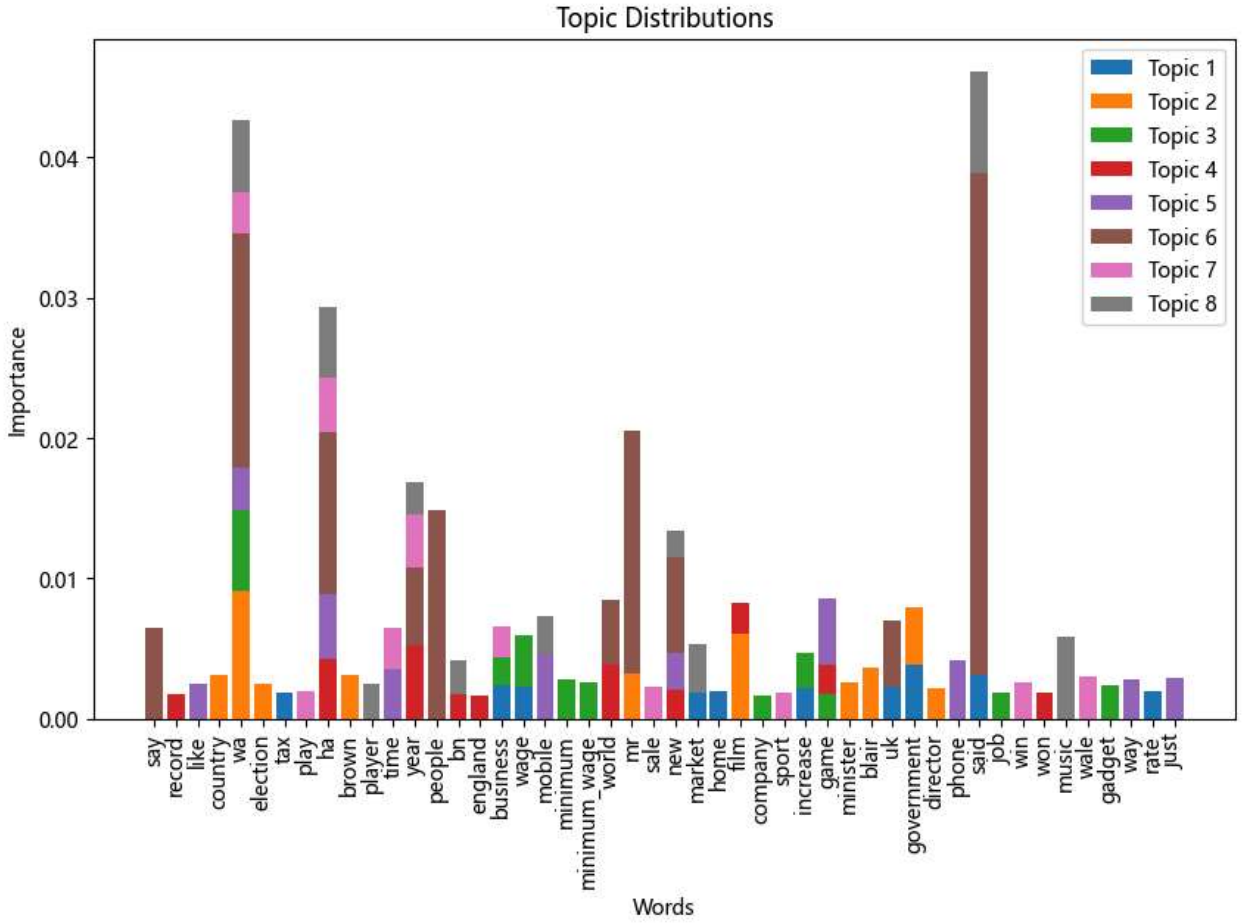
```
Topics from BERTopic Model:
Topic 0: said | mr | wa | ha | year | government | election | bn | labour | business
Topic 1: wa | game | sport | year | ha | england | said | time | win | half
Topic 2: film | best | award | wa | actor | oscar | star | year | entertainment | ha
Topic 3: music | band | album | wa | chart | song | single | year | said | singer
Topic 4: mail | virus | said | file | spam | anti | software | security | site | user
Topic 5: phone | mobile | technology | people | broadband | said | camera | service | digital | gadget
Topic 6: game | pc | console | mac | said | nintendo | machine | computer | gaming | mini
Topic 7: yukos | russian | russia | tax | gazprom | oil | company | ha | bn | court
Topic 8: doping | test | kenteris | iaaf | conte | greek | drug | thanou | sprinter | sport
```

NMF Modeling

In [26]:
```
nmf = tm.nmf_process(bbc_news, num_topics=8, source=1, text_col='text', eval=True, timing=True, code=1)
```

```
Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1491 entries, 0 to 1490
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   text    1491 non-null   object
dtypes: object(1)
memory usage: 11.8+ KB
Corpus loaded!
Text preprocessed!
Text trained!

Topics-Words from NMF Model:
Topic 1:
government (0.003865)
said (0.003107)
business (0.002423)
uk (0.002309)
wage (0.002272)
increase (0.002213)
home (0.001990)
rate (0.001914)
tax (0.001876)
market (0.001846)

Topic 2:
wa (0.009117)
film (0.006064)
government (0.004103)
blair (0.003639)
mr (0.003203)
brown (0.003116)
country (0.003104)
minister (0.002579)
election (0.002519)
director (0.002199)

Topic 3:
wa (0.005719)
wage (0.003683)
minimum (0.002841)
minimum_wage (0.002636)
increase (0.002527)
gadget (0.002402)
business (0.001981)
job (0.001821)
game (0.001800)
company (0.001688)

Topic 4:
year (0.005225)
ha (0.004286)
world (0.003879)
film (0.002146)
new (0.002034)
game (0.002009)
won (0.001836)
bn (0.001797)
record (0.001763)
england (0.001687)

Topic 5:
game (0.004776)
ha (0.004571)
mobile (0.004550)
phone (0.004207)
time (0.003503)
wa (0.003047)
just (0.002898)
way (0.002805)
new (0.002623)
like (0.002527)

Topic 6:
said (0.035792)
mr (0.017338)
wa (0.016696)
people (0.014887)
ha (0.011548)
new (0.006891)
say (0.006505)
year (0.005508)
uk (0.004654)
world (0.004638)
```

```
Topic 7:
ha (0.003863)
year (0.003770)
wale (0.003024)
wa (0.002976)
time (0.002970)
win (0.002640)
sale (0.002329)
business (0.002211)
play (0.001927)
sport (0.001856)

Topic 8:
said (0.007264)
music (0.005812)
wa (0.005166)
ha (0.005112)
market (0.003508)
mobile (0.002733)
player (0.002507)
year (0.002380)
bn (0.002363)
new (0.001842)
```



```
Model Evaluation Scores:
  Coherence: 0.5686655564831191
  Topic diversity: 0.0007091336767379371
  Topic size distribution: 0.001074258090506244

Finished 'nmf_process' in 77.9429 secs

def nmf_process(doc_file, num_topics=10, source=0, text_col='text', doc_size=0, cat=0, chi=False, group=True, eval=F
alse, web_app=False):
    """Pipelines the NMF modeling.

    :param doc_file: The filename of the prescribed text file to be loaded,
        or a BytesIO object from Streamlit's file_uploader, default to None
    :type doc_file: str or io.BytesIO
    :param num_topics: The number of topics to be modeled, default to 10
    :type num_topics: int, optional
    :param source: The source of the prescribed document file ('doc_file'),
        where 0 refers to internal store of the package and 1 to external file,
        default to 0
    :type source: int, optional
    :param text_col: The name of the text column to be extracted, default to 'text'
    :type text_col: str, optional
    :param doc_size: The number of documents to be processed, 0 represents all documents,
        or the range (tuple) of documents to be processed, default to 0
    :type doc_size: int, tuple, optional
    :param cat: The category indicating a subset of the Scripture to be loaded, where
        0 stands for the whole Bible, 1 for OT, 2 for NT, or one of the ten categories
        ['tor', 'oth', 'ket', 'map', 'mip', 'gos', 'nth', 'pau', 'epi', 'apo'] (See
        the package's internal file 'data/book_cat.csv'), default to 0
    :type cat: int or str, optional
    :param chi: The flag indicating whether the text is processed as Chinese (True)
        or English (False), default to False
    :type chi: bool, optional
    :param group: The flag indicating whether the loaded text is grouped by chapter,
        default to True
    :type group: bool, optional
    :param eval: The flag indicating whether the model evaluation results will be shown,
        default to False
    :type eval: bool, optional
    :param web_app: The flag indicating the function is initiated from a web application,
        default to False
    :type web_app: bool
    :return: The pipelined NMF
    :rtype: cwordtm.tm.NMF object
    """

    nmf = NMF(doc_file, num_topics, chi)
    if source == 0:
        nmf.docs = load_bible(nmf.doc_file, cat=cat, group=group)
    else:
        nmf.docs = load_text(nmf.doc_file, doc_size, text_col)

    print("Corpus loaded!")

    if chi:
        nmf.preprocess_chi()
    else:
        nmf.preprocess()
    print("Text preprocessed!")

    nmf.fit()
    print("Text trained!")
    nmf.show_topics_words()
    nmf.viz(web_app)

    if eval:
        print("\nModel Evaluation Scores:")
        nmf.evaluate()

    return nmf

>> cwordtm.tm.NMF
class NMF:
    """The NMF object for Non-negative Matrix Factorization (NMF) modeling.

    :cvar num_topics: The number of topics to be modeled, default to 10
    :vartype num_topics: int
    :ivar doc_file: The filename of the text file to be processed
    :vartype doc_file: str
    :ivar chi: The flag indicating whether the processed text is in Chinese or not,
        True stands for Traditional Chinese or False for English
    :vartype chi: bool
    :ivar num_topics: The number of topics set for the topic model
    :vartype num_topics: int
    :ivar docs: The collection of the original documents to be processed
    :vartype docs: pandas.DataFrame or list
```

```
    :ivar pro_docs: The collection of documents, in form of list of lists of words
        after text preprocessing
    :vartype pro_docs: list
    :ivar dictionary: The dictionary of word ids with their tokenized words
        from preprocessed documents ('pro_docs')
    :vartype dictionary: gensim.corpora.Dictionary
    :ivar corpus: The list of documents, where each document is a list of tuples
        (word id, word frequency in the particular document)
    :vartype corpus: list
    :ivar model: The NMF model object
    :vartype model: gensim.models.Nmf
    :ivar figures: The list of model visualization figures
    :vartype figures: list(matplotlib.pyplot.figure)
    """

    def __init__(self, doc_file, num_topics, chi=False):
        """Constructor method.
        """

        self.doc_file = doc_file
        self.num_topics = num_topics
        self.chi = chi
        self.docs = None
        self.pro_docs = None
        self.dictionary = None
        self.corpus = None
        self.model = None
        self.figures = []


    def preprocess(self):
        """Process the original English documents (cwordtm.tm.NMF.docs)
        by invoking cwordtm.tm.process_text, and build a dictionary
        and a corpus from the preprocessed documents for the NMF model.
        """

        self.pro_docs = [process_text(doc) for doc in self.docs]

        for i, doc in enumerate(self.pro_docs):
            self.pro_docs[i] += ["_".join(w) for w in ngrams(doc, 2)]
            # self.pro_docs[i] += ["_".join(w) for w in ngrams(doc, 3)]

        # Create a dictionary and corpus for the NMF model
        self.dictionary = corpora.Dictionary(self.pro_docs)
        self.corpus = [self.dictionary.doc2bow(doc) for doc in self.pro_docs]


    def preprocess_chi(self):
        """Process the original Chinese documents (cwordtm.tm.NMF.docs)
        by tokenizing text, removing stopwords, and building a dictionary
        and a corpus from the preprocessed documents for the NMF model.
        """

        # Build stop words
        stop_file = files('cwordtm.data').joinpath("tc_stopwords_2.txt")
        stopwords = [k[:-1] for k in open(stop_file, encoding='utf-8')\
                     .readlines() if k != '']

        # Tokenize the Chinese text using Jieba
        dict_file = files('cwordtm.data').joinpath("user_dict_4.txt")
        jieba.load_userdict(str(dict_file))
        docs = [jieba.cut(doc) for doc in self.docs]

        # Replace special characters
        docs = [[word.replace('\u3000', ' ') for word in doc] \
                                    for doc in docs]

        # Remove stop words
        self.pro_docs = [' '.join([word for word in doc if word not in stopwords]) \
                                        for doc in docs]

        self.pro_docs = [doc.split() for doc in self.pro_docs]

        # Create a dictionary and corpus
        self.dictionary = corpora.Dictionary(self.pro_docs)
        self.corpus = [self.dictionary.doc2bow(doc) for doc in self.pro_docs]


    def fit(self):
        """Build the NMF model with the created corpus and dictionary.
        """

        self.model = models.Nmf(self.corpus,
                                num_topics=self.num_topics)
```

```
    def show_topics_words(self):
        """Shows the topics with their keywords from the built NMF model.
        """

        print("\nTopics-Words from NMF Model:")
        for topic_id in range(self.num_topics):
            topic_words = self.model.show_topic(topic_id, topn=10)
            print(f"Topic {topic_id+1}:")
            for word_id, prob in topic_words:
                word = self.dictionary[int(word_id)]
                print("%s (%.6f)" %(word, prob))
            print()


    def viz(self, web_app=False):
        """Plot the topic distributions as a stacked bar chart for the built NMF model.

        :param web_app: The flag indicating the function is initiated from a web
            application, default to False
        :type web_app: bool
        """

        # Build a list of word ids from the built topics
        word_ids = []
        for topic_id in range(self.num_topics):
            topic_words = self.model.show_topic(topic_id, topn=10)
            for word_id, _ in topic_words:
                word_ids.append(int(word_id))

        word_ids = list(set(word_ids))

        # Create a topic distribution table
        topic_dist = np.zeros((self.num_topics, len(word_ids)))
        for topic_id in range(self.num_topics):
            topic_words = self.model.show_topic(topic_id, topn=10)
            for word_id, prob in topic_words:
                topic_dist[topic_id, word_ids.index(int(word_id))] = prob

        # Build a list of distinct words from the word id list
        word_list = []
        for i in range(len(word_ids)):
            word_list.append(self.dictionary[word_ids[i]])

        # Plot the topic distributions
        matplotlib.rcParams['font.family'] = ['Microsoft YaHei']
        fig = plt.figure(figsize=(10, 6))

        bottom = np.zeros(len(word_list))
        for i, topic in enumerate(topic_dist):
            plt.bar(word_list, topic, width=0.8, bottom=bottom, label=f"Topic {i+1}")
            bottom += topic

        plt.xticks(range(len(word_list)), word_list, rotation=90)
        plt.title("Topic Distributions")
        plt.xlabel("Words")
        plt.ylabel("Importance")
        plt.legend(loc="best")
        plt.show()
        if web_app: self.figures.append(fig)


    def evaluate(self):
        """Computes and outputs the coherence score, topic diversity,
        and topic size distribution.
        """

        # Compute coherence score
        coherence_model = CoherenceModel(model=self.model,
                                         texts=self.pro_docs,
                                         dictionary=self.dictionary,
                                         coherence='c_v')
        print(f"  Coherence: {coherence_model.get_coherence()}")

        # Compute topic diversity
        topic_sizes = [len(self.model[self.corpus[i]]) for i in range(len(self.corpus))]
        total_docs = sum(topic_sizes)
        topic_diversity = sum([(size/total_docs)**2 for size in topic_sizes])
        print(f"  Topic diversity: {topic_diversity}")

        # Compute topic size distribution
        # topic_sizes = [len(self.model[self.corpus[i]]) for i in range(len(self.corpus))]
        topic_size_distribution = max(topic_sizes) / sum(topic_sizes)
        print(f"  Topic size distribution: {topic_size_distribution}\n")


    def save(self, file):
```

```
        """Saves the built NMF model to the specified file.

        :param file: The name of the file to store the built model, default to None
        :type file: str
        """

        if file is None or len(file.strip())==0:
            print("No valid filename has been specifid!")
            return

        base_name = file.split('.')[0]
        model_file = base_name + '_model.gensim'
        dict_file = base_name + '_dictionary.gensim'
        self.model.save(model_file)
        self.dictionary.save(dict_file)
        # corpora.MmCorpus.serialize(base_name+'_corpus.mm', self.corpus)
        print(f"NMF model has been saved: {model_file!r} and {dict_file!r}")


    def load(self, file):
        """Loads the stored NMF model from the specified file.

        :param file: The name of the file to be loaded, default to None
        :type file: str
        :return: The loaded NMF model and the loaded dictionary of the NMF's corpus
        :rtype: gensim.models.Nmf, gensim.corpora.Dictionary
        """

        if file is None or len(file.strip())==0:
            print("No valid filename has been specifid!")
            return

        base_name = file.split('.')[0]
        model_file = base_name + '_model.gensim'
        dict_file = base_name + '_dictionary.gensim'
        try:
            loaded_model = models.Nmf.load(model_file)
            loaded_dict = corpora.Dictionary.load(dict_file)
        except:
            print("Moldel file or dictionary file cannot be loaded!")
            return

        return loaded_model, loaded_dict
```

```
>> cwordtm.tm.load_bible
def load_bible(textfile, cat=0, group=True):
    """Loads and returns the Bible Scripture from the prescribed internal
    file ('textfile').

    :param textfile: The package's internal Bible text from which the text is loaded,
        either World English Bible ('web.csv') or Chinese Union Version (Traditional)
        ('cuv.csv'), default to None
    :type textfile: str
    :param cat: The category indicating a subset of the Scripture to be loaded, where
        0 stands for the whole Bible, 1 for OT, 2 for NT, or one of the ten categories
        ['tor', 'oth', 'ket', 'map', 'mip', 'gos', 'nth', 'pau', 'epi', 'apo'] (See
        the package's internal file 'data/book_cat.csv'), default to 0
    :type cat: int or str, optional
    :param group: The flag indicating whether the loaded text is grouped by chapter,
        default to True
    :type group: bool, optional
    :return: The collection of Scripture loaded
    :rtype: pandas.DataFrame
    """

    # textfile = "web.csv"
    scfile = files('cwordtm.data').joinpath(textfile)
    print("Loading Bible '%s' ..." %scfile)
    df = pd.read_csv(scfile)

    cat_list = ['tor', 'oth', 'ket', 'map', 'mip',\
                'gos', 'nth', 'pau', 'epi', 'apo']
    cat = str(cat)
    if cat == '1' or cat == 'ot':
        df = util.extract(df, testament=0)
    elif cat == '2' or cat == 'nt':
        df = util.extract(df, testament=1)
    elif cat in cat_list:
        df = util.extract(df, category=cat)

    if group:
        # Group verses into chapters
        df = df.groupby(['book_no', 'chapter'])\
                        .agg({'text': lambda x: ' '.join(x)})\
                .reset_index()
```

```
    df.text = df.text.str.replace('  ', '')
    return list(df.text)

>> cwordtm.tm.load_text
def load_text(textfile, doc_size=0, text_col='text'):
    """Loads and returns the list of documents from the prescribed file ('textfile').

    :param textfile: The prescribed text file from which the text is loaded,
        default to None
    :type textfile: str
    :param nr: The number of rows of text to be loaded; 0 represents all rows,
        default to 0
    :type nr: int, optional
    :param doc_size: The number of documents to be processed, 0 represents all documents,
        or the range (tuple) of documents to be processed, default to 0
    :type doc_size: int, tuple, optional
    :param text_col: The name of the text column to be extracted, default to 'text'
    :type text_col: str, optional
    :return: The list of documents loaded
    :rtype: list
    """

    docs = util.load_text(textfile, doc_size, text_col)
    return list(docs[text_col])
```