Modelado de temas con PyCaret

[https://towardsdatascience.com/topic-modeling-on-pycaret-2ce0c65ba3ff (https://towardsdatascience.com/topic-modeling-on-pyca](https://towardsdatascience.com/topic-modeling-on-pycaret-2ce0c65ba3ff)

Instalación de la biblioteca y modelos de lenguaje

In [4]:
```
#pip install pycaret
#python -m spacy download en_core_web_sm
#python -m textblob.download_corpora
```

Importar PyCaret

In [3]:
```
from pycaret.nlp import *
```

1. Importar datos

In [5]:
```
import pandas as pd
```

In [14]:
```
path = "C:/Users/Arceus/Desktop/Themes.csv"

df = pd.read_csv(path)
df.head(5)
```

```
c:\users\arceus\appdata\local\programs\python\python38\lib\site-packages\IPython\core\interactiveshell.py:3441: D
option on import or set low_memory=False.
  exec(code_obj, self.user_global_ns, self.user_ns)
```

Out[14]:

	Unnamed: 0	GLOBALEVENTID	SQLDATE	MonthYear	Year	FractionDate	Actor1Code	Actor1Name	Actor1CountryCode	Actor1Known
0	0	943928286.0	20200901.0	202009.0	2020.0	2020.6603	MEX	MEXICO	MEX	
1	1	943932111.0	20200901.0	202009.0	2020.0	2020.6603	MEX	MEXICO	MEX	
2	2	943946299.0	20200901.0	202009.0	2020.0	2020.6603	MEX	YUCATAN PENINSULA	MEX	
3	3	943946300.0	20200901.0	202009.0	2020.0	2020.6603	MEX	YUCATAN PENINSULA	MEX	
4	4	943952149.0	20200901.0	202009.0	2020.0	2020.6603	MEX	MEXICO	MEX	

5 rows × 89 columns

In [15]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 47680 entries, 0 to 47679
Data columns (total 89 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   Unnamed: 0                   47680 non-null  int64  
 1   GLOBALEVENTID                47680 non-null  float64
 2   SQLDATE                      47680 non-null  float64
 3   MonthYear                    47680 non-null  float64
 4   Year                         47680 non-null  float64
 5   FractionDate                 47680 non-null  float64
 6   Actor1Code                   47680 non-null  object 
 7   Actor1Name                   47680 non-null  object 
 8   Actor1CountryCode            47680 non-null  object 
 9   Actor1KnownGroupCode         0 non-null      float64
 10  Actor1EthnicCode             0 non-null      float64
 11  Actor1Religion1Code          13 non-null     object 
 12  Actor1Religion2Code          5 non-null      object 
 13  Actor1Type1Code              5579 non-null   object 
 14  Actor1Type2Code              174 non-null    object 
 15  Actor1Type3Code              0 non-null      float64
 16  Actor2Code                   33289 non-null  object 
 17  Actor2Name                   33289 non-null  object 
 18  Actor2CountryCode            22264 non-null  object 
 19  Actor2KnownGroupCode         137 non-null    object 
 20  Actor2EthnicCode             183 non-null    object 
 21  Actor2Religion1Code          170 non-null    object 
 22  Actor2Religion2Code          73 non-null     object 
 23  Actor2Type1Code              14474 non-null  object 
 24  Actor2Type2Code              745 non-null    object 
 25  Actor2Type3Code              26 non-null     object 
 26  IsRootEvent                  47680 non-null  float64
 27  EventCode                    47680 non-null  float64
 28  EventBaseCode                47680 non-null  float64
 29  EventRootCode                47680 non-null  float64
 30  QuadClass                    47680 non-null  float64
 31  GoldsteinScale               47680 non-null  float64
 32  NumMentions                  47680 non-null  float64
 33  NumSources                   47680 non-null  float64
 34  NumArticles                  47680 non-null  float64
 35  AvgTone                      47680 non-null  float64
 36  Actor1Geo_Type              47680 non-null  float64
 37  Actor1Geo_FullName          47512 non-null  object 
 38  Actor1Geo_CountryCode       47514 non-null  object 
 39  Actor1Geo_ADM1Code          47514 non-null  object 
 40  Actor1Geo_ADM2Code          27032 non-null  object 
 41  Actor1Geo_Lat               47512 non-null  float64
 42  Actor1Geo_Long              47512 non-null  float64
 43  Actor1Geo_FeatureID         47514 non-null  object 
 44  Actor2Geo_Type              47680 non-null  float64
 45  Actor2Geo_FullName          33228 non-null  object 
 46  Actor2Geo_CountryCode       33229 non-null  object 
 47  Actor2Geo_ADM1Code          33229 non-null  object 
 48  Actor2Geo_ADM2Code          14135 non-null  object 
 49  Actor2Geo_Lat               33228 non-null  float64
 50  Actor2Geo_Long              33228 non-null  float64
 51  Actor2Geo_FeatureID         33229 non-null  object 
 52  ActionGeo_Type              47680 non-null  float64
 53  ActionGeo_FullName          47512 non-null  object 
 54  ActionGeo_CountryCode       47514 non-null  object 
 55  ActionGeo_ADM1Code          47514 non-null  object 
 56  ActionGeo_ADM2Code          23111 non-null  object 
 57  ActionGeo_Lat               47512 non-null  float64
 58  ActionGeo_Long              47512 non-null  float64
 59  ActionGeo_FeatureID         47514 non-null  object 
 60  DATEADDED                   47680 non-null  float64
 61  SOURCEURL                   47680 non-null  object 
 62  GKGRECORDID                 47680 non-null  object 
 63  DATE                        47680 non-null  float64
 64  SourceCollectionIdentifier  47680 non-null  float64
 65  SourceCommonName            47680 non-null  object 
 66  DocumentIdentifier          47680 non-null  object 
 67  Counts                      21380 non-null  object 
 68  V2Counts                    21380 non-null  object 
 69  Themes                      47680 non-null  object 
 70  V2Themes                    47680 non-null  object 
 71  Locations                   47491 non-null  object 
 72  V2Locations                 47487 non-null  object 
 73  Persons                     45127 non-null  object 
 74  V2Persons                   44999 non-null  object 
 75  Organizations               44738 non-null  object 
 76  V2Organizations             43918 non-null  object 
 77  V2Tone                      47680 non-null  object 
 78  Dates                       28992 non-null  object 
 79  GCAM                        47680 non-null  object 
 80  SharingImage                38352 non-null  object 
 81  RelatedImages               9385 non-null   object 
 82  SocialImageEmbeds           2047 non-null   object 
 83  SocialVideoEmbeds           19136 non-null  object 
 84  Quotations                  16521 non-null  object 
 85  AllNames                    47571 non-null  object 
 86  Amounts                     44780 non-null  object 
 87  TranslationInfo             0 non-null      float64
 88  Extras                      47680 non-null  object 
dtypes: float64(31), int64(1), object(57)
memory usage: 32.4+ MB
```

2. Configuración del entorno

In [16]:

```
nlp = setup(data = df, target = 'Themes', custom_stopwords = [ 'rt', 'https', 'http', 'co', 'amp', 'the', ' the',
```

Description	Value
session_id	7904
Documents	47680
Vocab Size	3150
Custom Stopwords	True

3. Creación del modelo

In [17]:

```
lda = create_model('lda', num_topics = 6, multi_core = True)
```

4. Asignación del modelo

In [18]:

```
df_lda = assign_model(lda)
```

```
In [19]: df_lda.head(10)
```

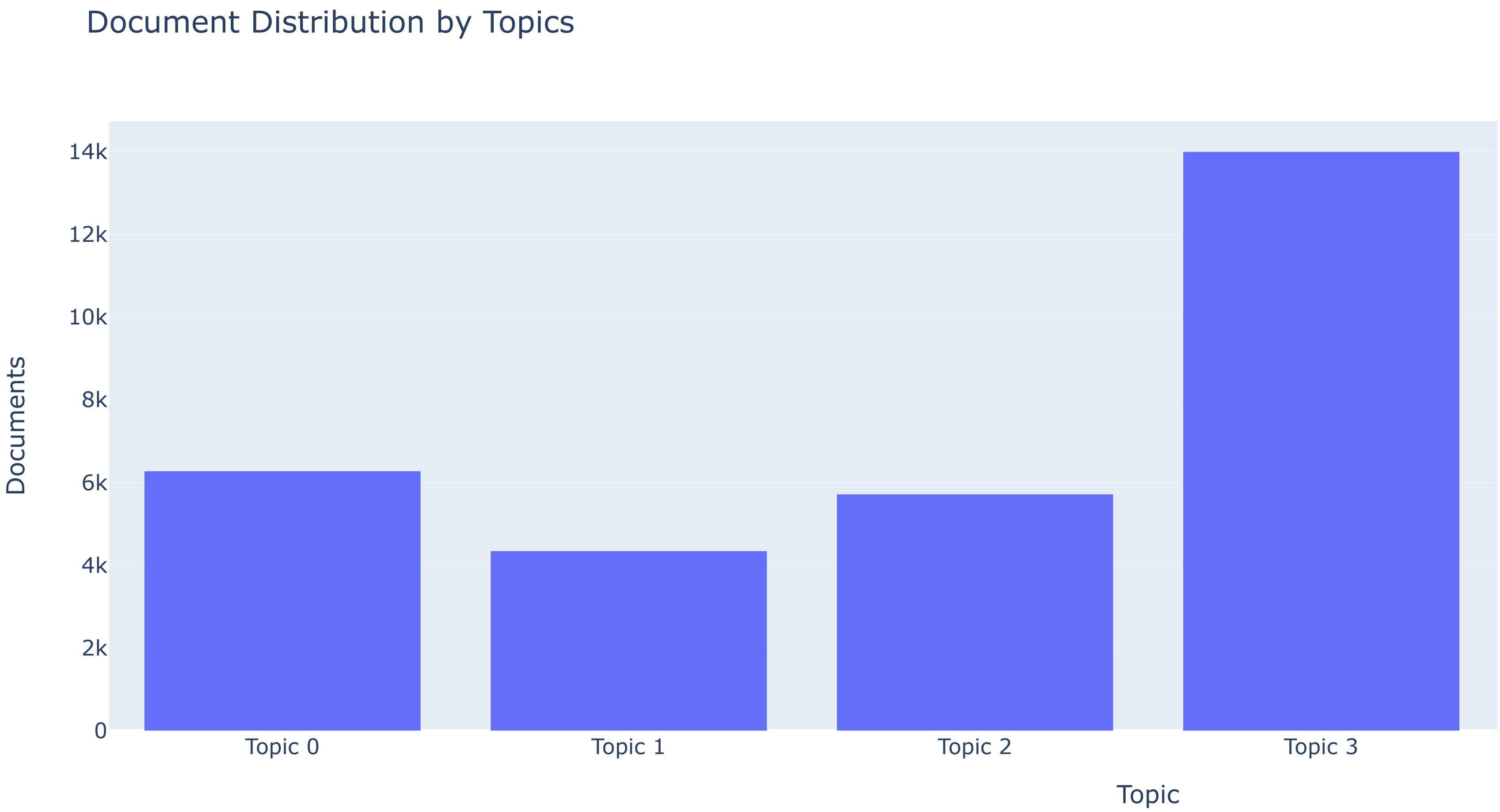
Out[19]:

	Unnamed: 0	GLOBALEVENTID	SQLDATE	MonthYear	Year	FractionDate	Actor1Code	Actor1Name	Actor1CountryCode	Actor1Known
0	0	943928286.0	20200901.0	202009.0	2020.0	2020.6603	MEX	MEXICO	MEX	
1	1	943932111.0	20200901.0	202009.0	2020.0	2020.6603	MEX	MEXICO	MEX	
2	2	943946299.0	20200901.0	202009.0	2020.0	2020.6603	MEX	YUCATAN PENINSULA	MEX	
3	3	943946300.0	20200901.0	202009.0	2020.0	2020.6603	MEX	YUCATAN PENINSULA	MEX	
4	4	943952149.0	20200901.0	202009.0	2020.0	2020.6603	MEX	MEXICO	MEX	
5	5	943952150.0	20200901.0	202009.0	2020.0	2020.6603	MEX	MEXICO CITY	MEX	
6	6	943952153.0	20200901.0	202009.0	2020.0	2020.6603	MEX	MEXICAN	MEX	
7	7	943952155.0	20200901.0	202009.0	2020.0	2020.6603	MEXCVL	MEXICAN	MEX	
8	8	943972675.0	20200901.0	202009.0	2020.0	2020.6603	MEX	MEXICO	MEX	
9	9	943984683.0	20200901.0	202009.0	2020.0	2020.6603	MEX	MEXICO	MEX	

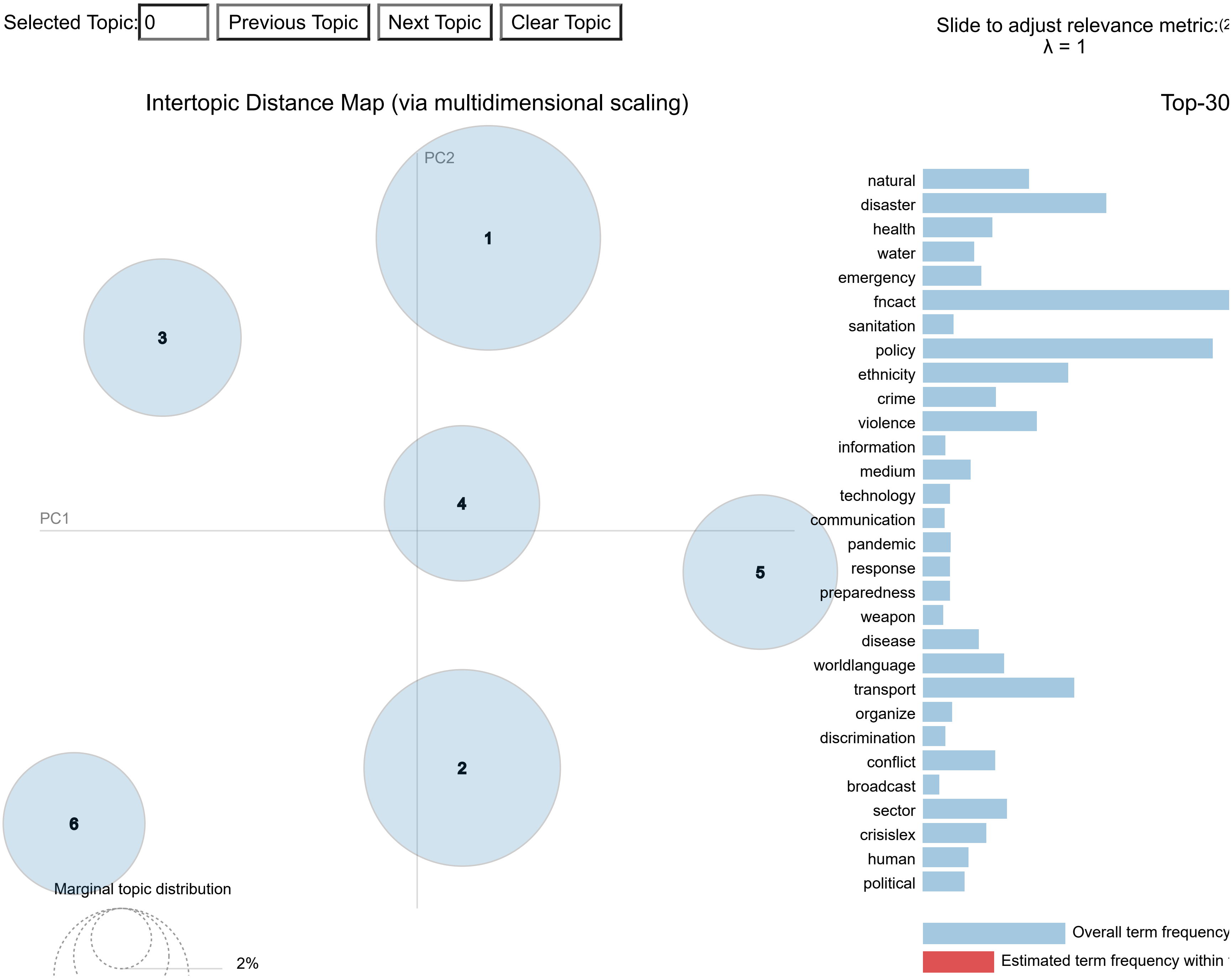
10 rows × 97 columns

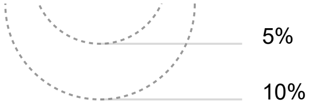
Visualización de datos

In [20]:
```
plot_model(lda, plot='topic_distribution')
```



```
In [21]: plot_model(lda, plot='topic_model')
```

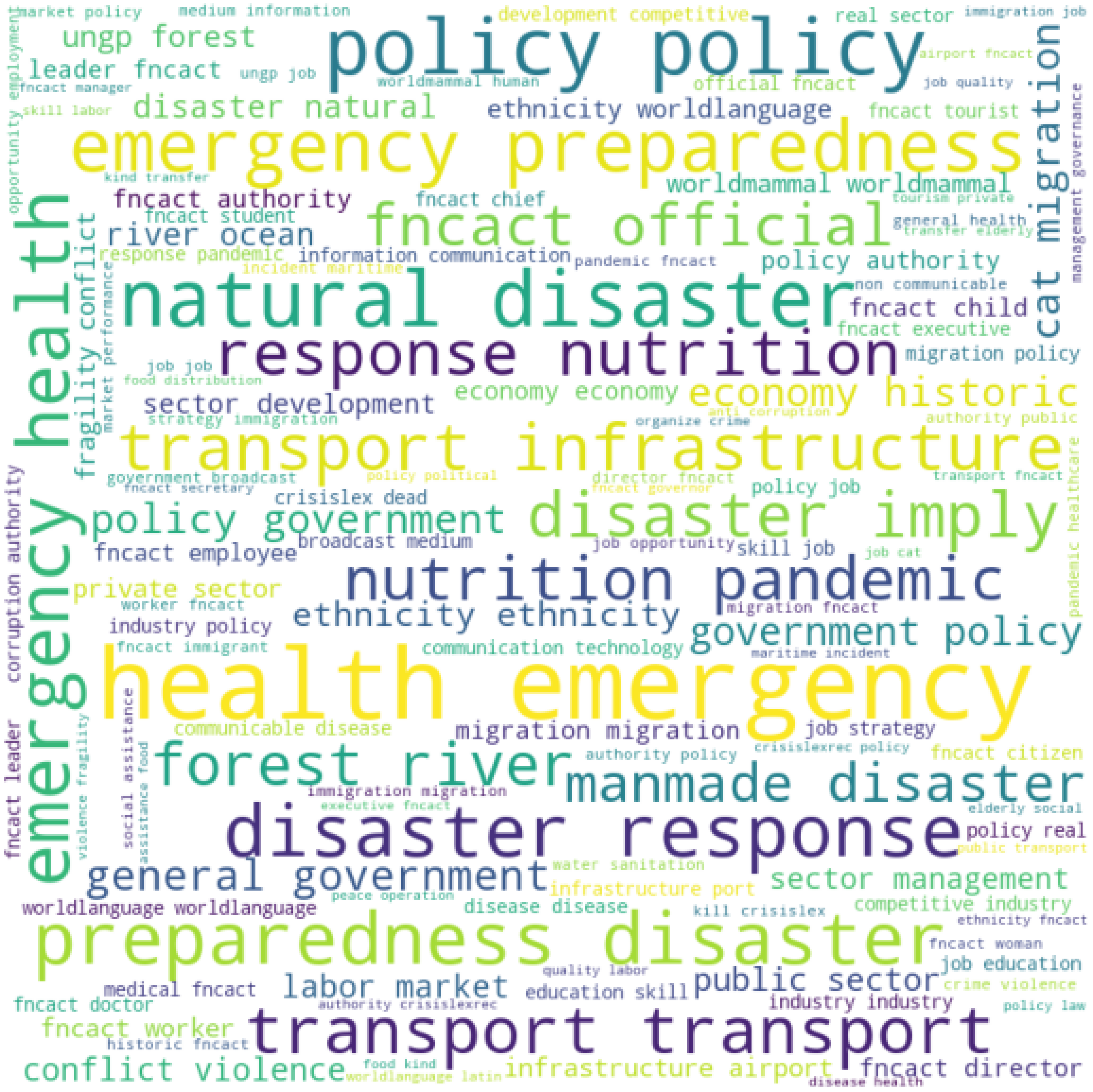




1. saliency(term w) = frequency(w) * [sum_t

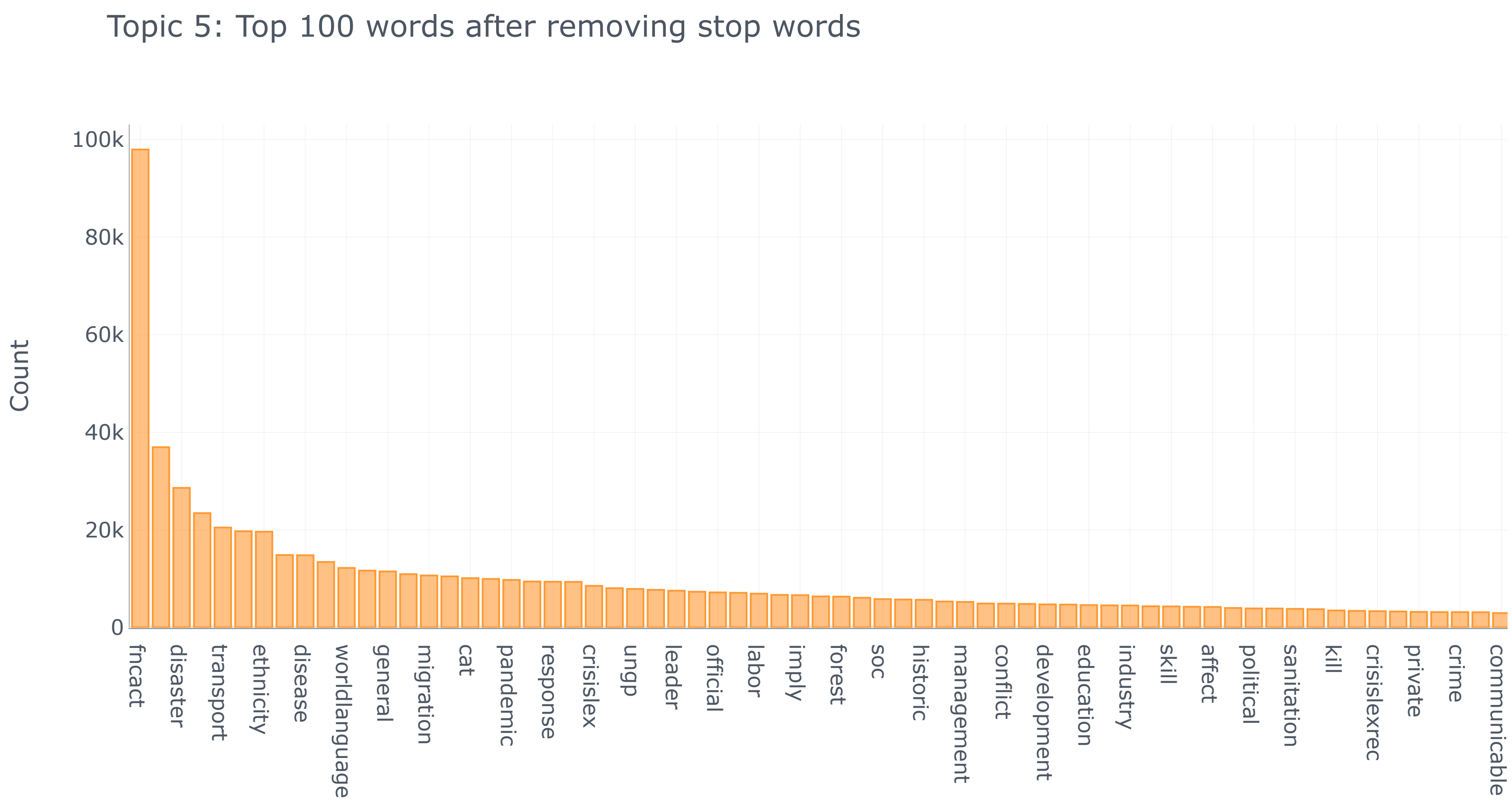
2. relevance(term w | topic t) = λ * p(w | t) + (

```
In [22]: plot_model(lda, plot='wordcloud', topic_num = 'Topic 5')
```

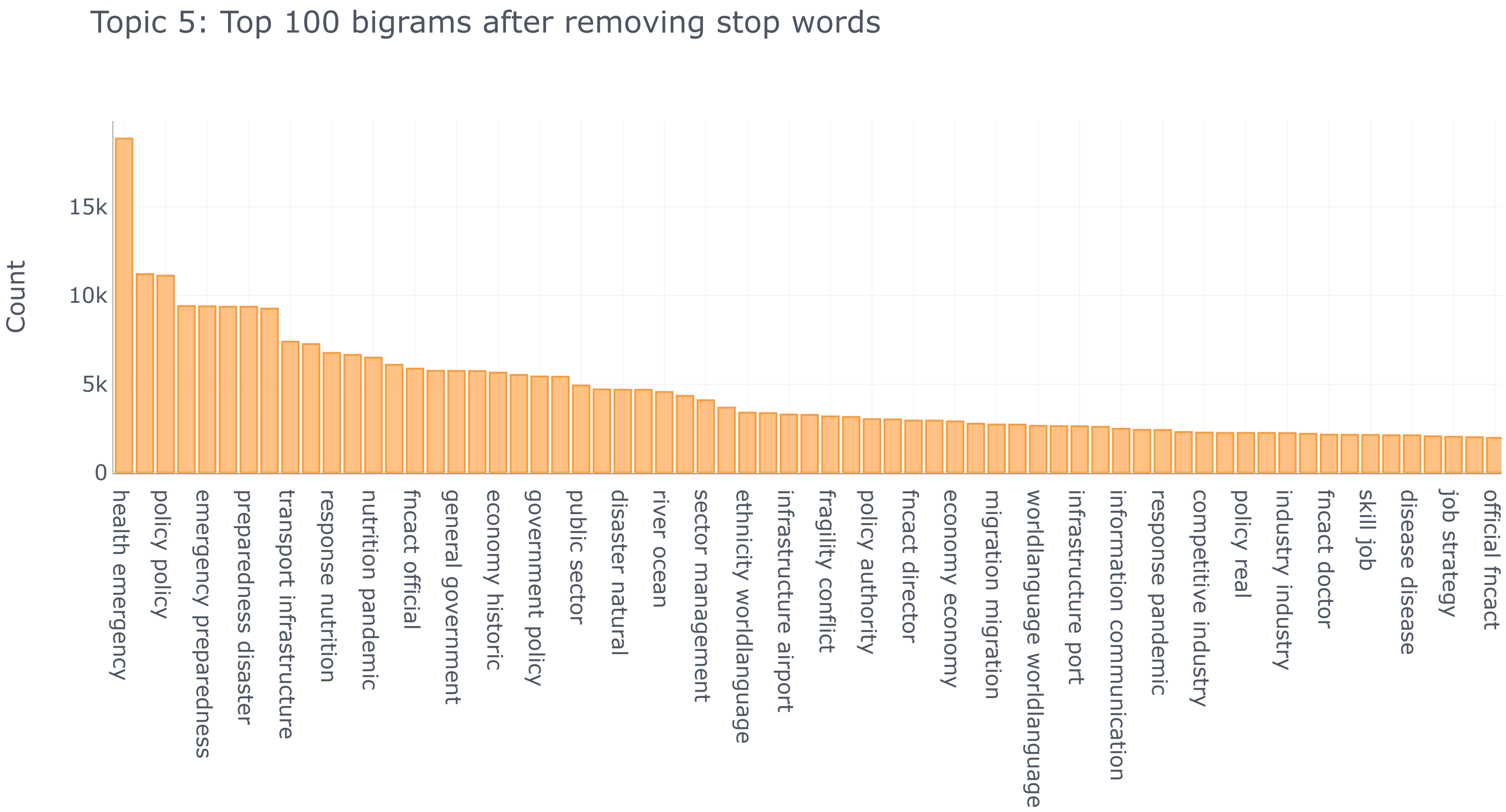


In [23]:

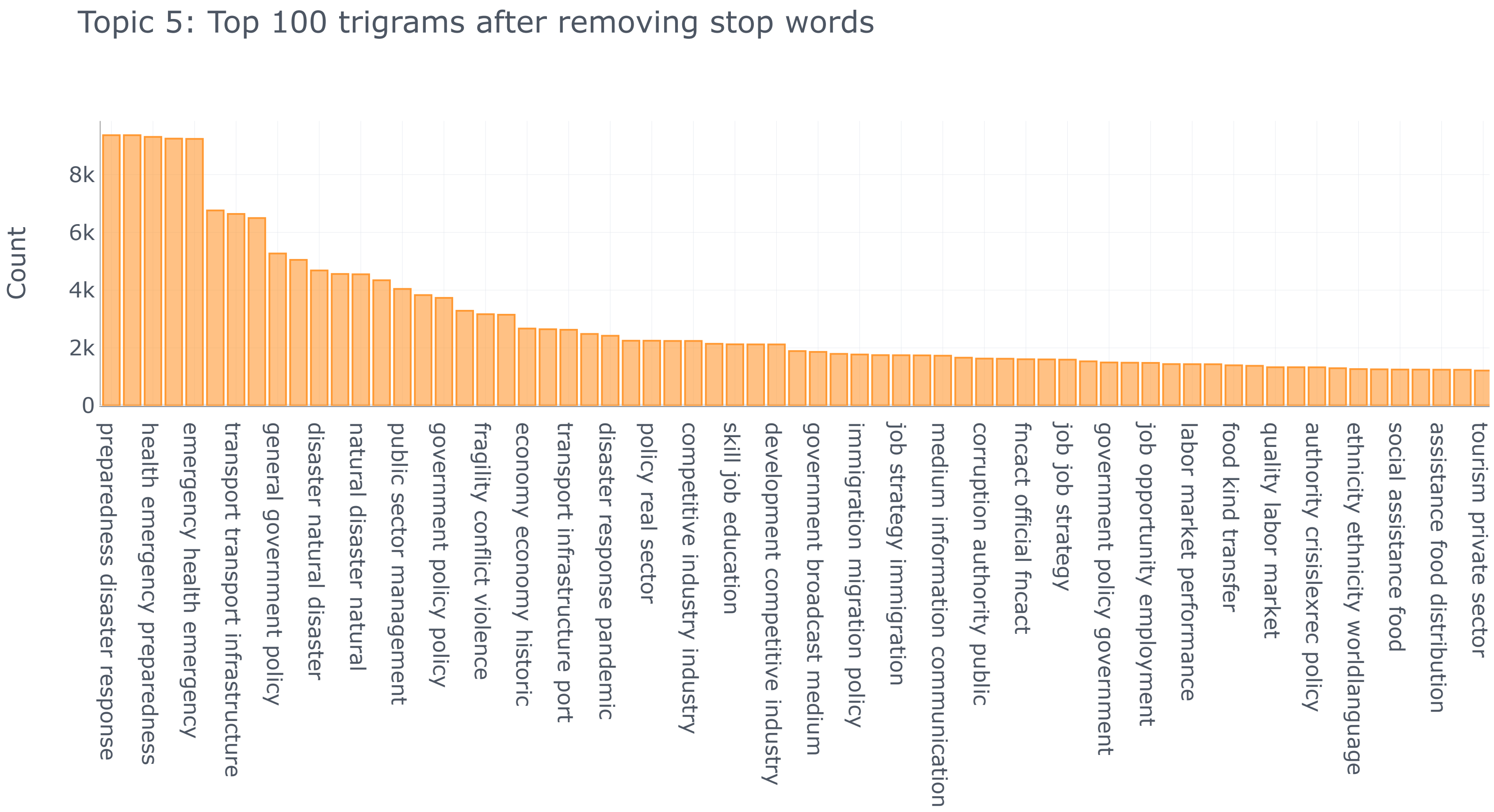
```
plot_model(lda, plot='frequency', topic_num = 'Topic 5')
```



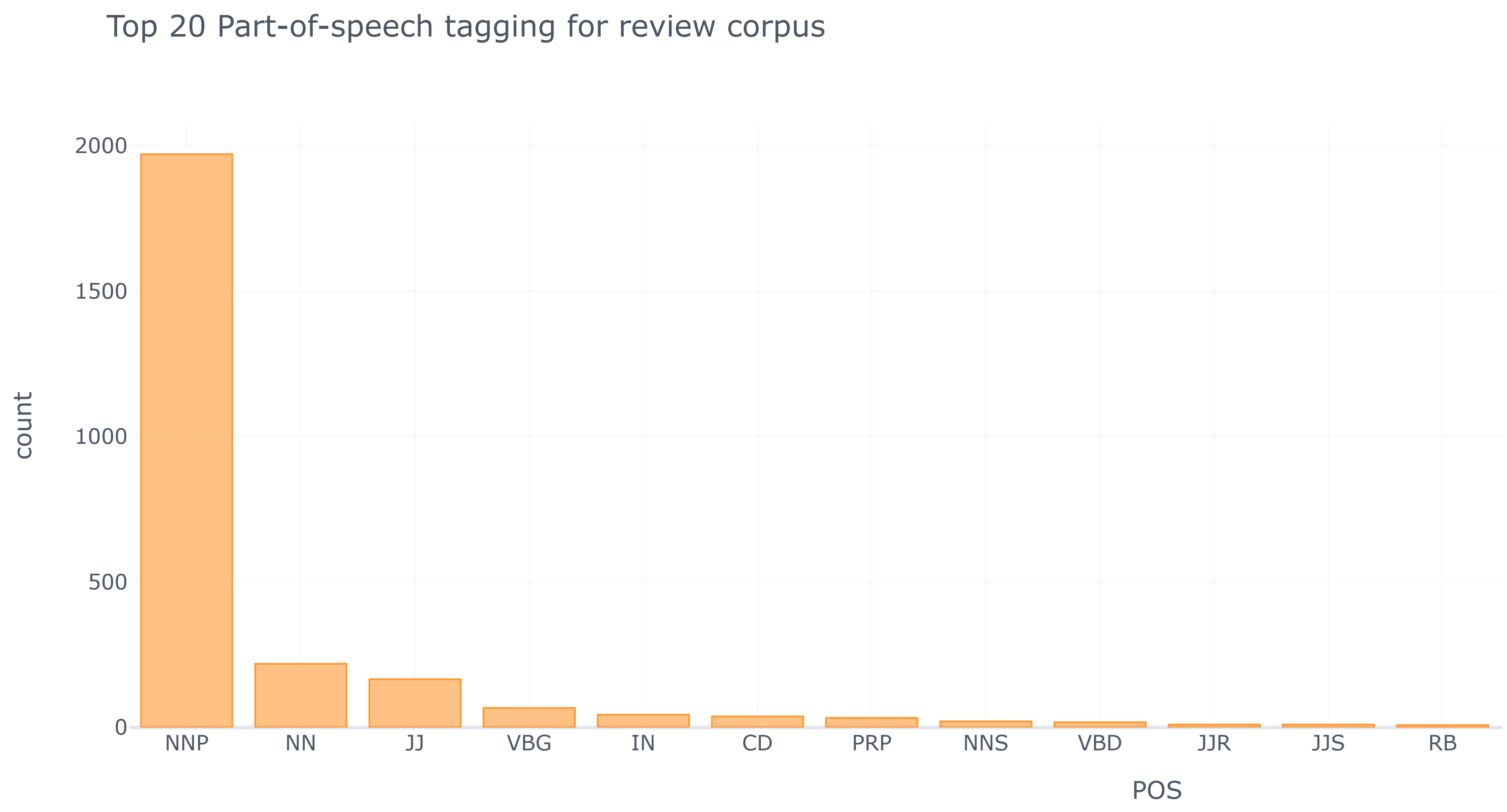
In [24]:
```
plot_model(lda, plot='bigram', topic_num = 'Topic 5')
```



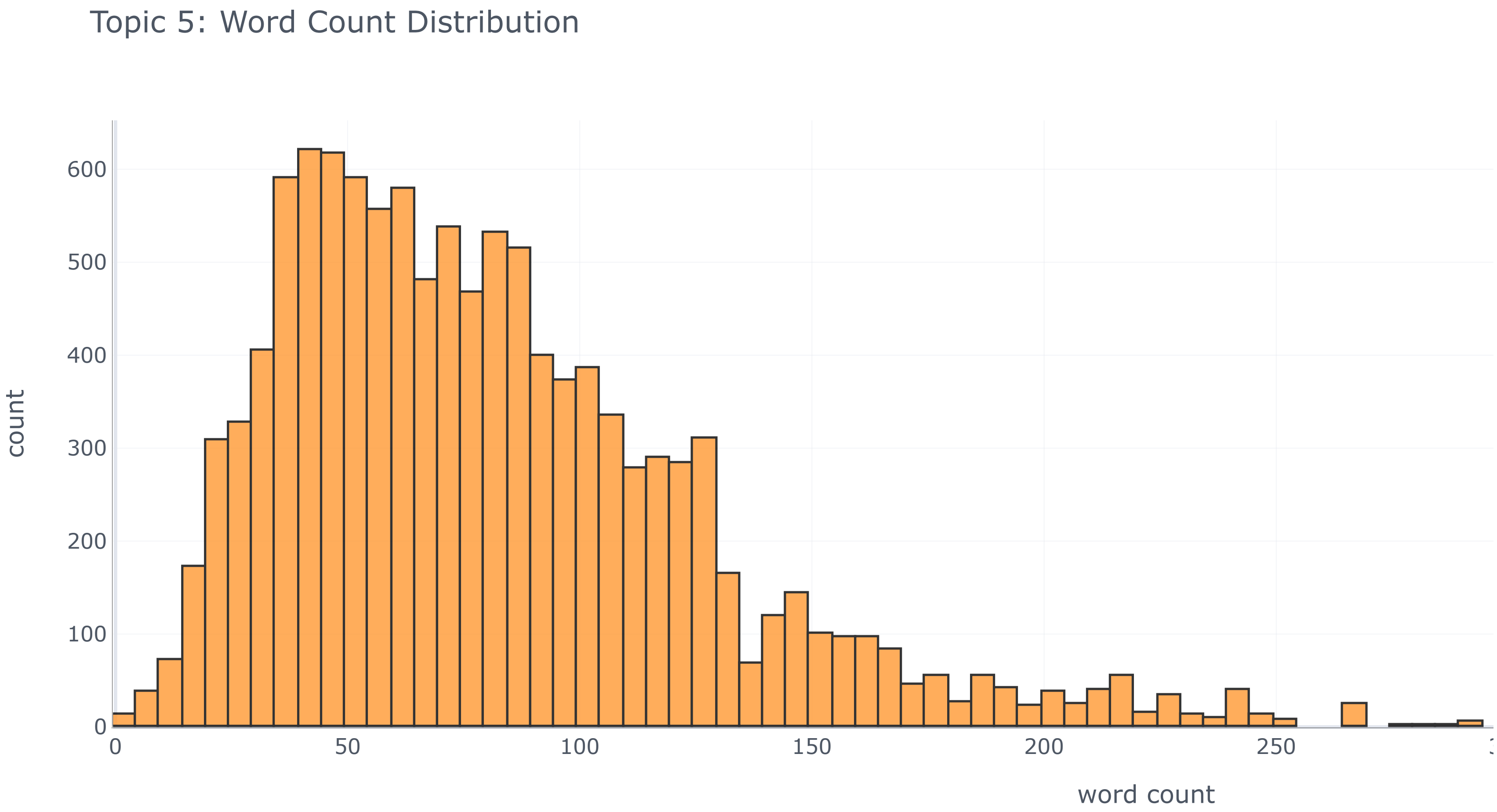
```
In [25]: plot_model(lda, plot='trigram', topic_num = 'Topic 5')
```



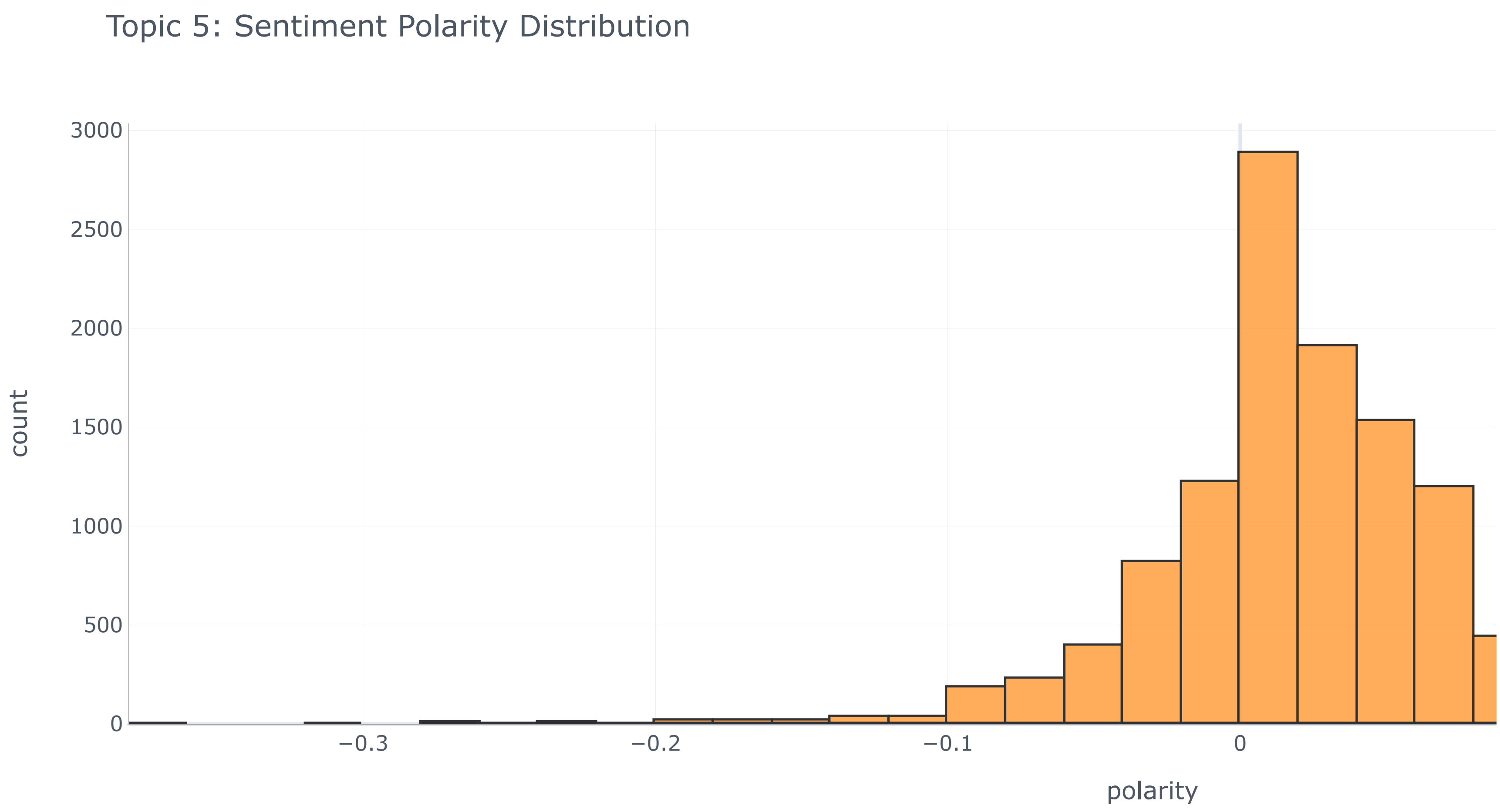
In [26]:
```
plot_model(lda, plot="pos")
```



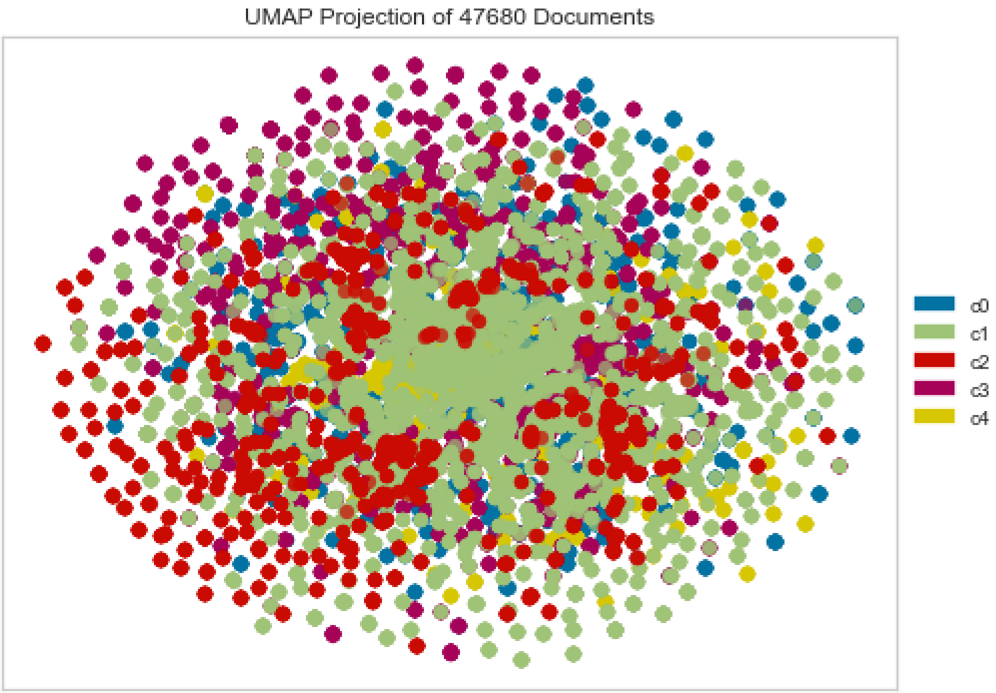
In [27]:
```
plot_model(lda, plot='distribution', topic_num = 'Topic 5')
```



```
In [28]: plot_model(lda, plot='sentiment', topic_num = 'Topic 5')
```



```
In [32]: #plot_model(lda, plot='tsne')
In [30]: plot_model(lda, plot='umap')
```



```
In [34]: evaluate_model(lda)
```

```
interactive(children=(ToggleButtons(description='Plot Type:', icons=('',), options=(('Frequency Plot', 'freque…
```

```
In [ ]:
```