In [1]:

```
import pandas as pd
import urllib.request
urllib.request.urlretrieve("https://raw.githubusercontent.com/franciscadias/data/master/abcnews-date-text.csv", f
data = pd.read_csv('abcnews-date-text.csv', error_bad_lines=False)

print(len(data))
```

```
1082168
```

In [2]:

```
print(data.head(10))
```

```
   publish_date                                      headline_text
0      20030219  aba decides against community broadcasting lic...
1      20030219     act fire witnesses must be aware of defamation
2      20030219     a g calls for infrastructure protection summit
3      20030219           air nz staff in aust strike for pay rise
4      20030219      air nz strike to affect australian travellers
5      20030219                  ambitious olsson wins triple jump
6      20030219         antic delighted with record breaking barca
7      20030219  aussie qualifier stosur wastes four memphis match
8      20030219       aust addresses un security council over iraq
9      20030219         australia is locked into war timetable opp
```

In [3]:

```
text = data[['headline_text']]

import nltk

# word tokenization
text['headline_text'] = text.apply(lambda row: nltk.word_tokenize(row['headline_text']), axis=1)

# stop words removal
from nltk.corpus import stopwords

stop = stopwords.words('english')
text['headline_text'] = text['headline_text'].apply(lambda x: [word for word in x if word not in (stop)])

print(text.head(10))
```

```
<ipython-input-3-fcc9fa4d5631>:6: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#retur
ning-a-view-versus-a-copy
  text['headline_text'] = text.apply(lambda row: nltk.word_tokenize(row['headline_text']), axis=1)
```

```
                                       headline_text
0   [aba, decides, community, broadcasting, licence]
1    [act, fire, witnesses, must, aware, defamation]
2     [g, calls, infrastructure, protection, summit]
3          [air, nz, staff, aust, strike, pay, rise]
4  [air, nz, strike, affect, australian, travellers]
5            [ambitious, olsson, wins, triple, jump]
6        [antic, delighted, record, breaking, barca]
7  [aussie, qualifier, stosur, wastes, four, memp...
8     [aust, addresses, un, security, council, iraq]
9           [australia, locked, war, timetable, opp]
```

```
<ipython-input-3-fcc9fa4d5631>:12: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#retur
ning-a-view-versus-a-copy
  text['headline_text'] = text['headline_text'].apply(lambda x: [word for word in x if word not in (stop)])
```

In [4]:

```
from nltk.stem import WordNetLemmatizer

text['headline_text'] = text['headline_text'].apply(lambda x: [WordNetLemmatizer().lemmatize(word, pos='v') for w

tokenized_doc = text['headline_text'].apply(lambda x: [word for word in x if len(word) > 3])

print(tokenized_doc[:10])
```

```
<ipython-input-4-6e6314925ca3>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
```

```
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#retur
ning-a-view-versus-a-copy
  text['headline_text'] = text['headline_text'].apply(lambda x: [WordNetLemmatizer().lemmatize(word, pos='v') for
word in x])
```

```
0              [decide, community, broadcast, licence]
1             [fire, witness, must, aware, defamation]
2           [call, infrastructure, protection, summit]
3                          [staff, aust, strike, rise]
4             [strike, affect, australian, travellers]
5                  [ambitious, olsson, triple, jump]
6               [antic, delight, record, break, barca]
7    [aussie, qualifier, stosur, waste, four, memph...
8             [aust, address, security, council, iraq]
9                         [australia, lock, timetable]
Name: headline_text, dtype: object
```

In [5]:

```
# 역토큰화 (토큰화 작업을 되돌림)
detokenized_doc = []
for i in range(len(text)):
    t = ' '.join(tokenized_doc[i])
    detokenized_doc.append(t)

text['headline_text'] = detokenized_doc # 다시 text['headline_text']에 재저장
```

```
<ipython-input-5-93d63d960f46>:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#retur
ning-a-view-versus-a-copy
  text['headline_text'] = detokenized_doc # 다시 text['headline_text']에 재저장
```

In [6]:

```
#3,000 개

from sklearn.feature_extraction.text import TfidfVectorizer
# 상위 3,000개의 단어를 보존
vectorizer3 = TfidfVectorizer(stop_words='english', max_features= 3000)

T = vectorizer3.fit_transform(text['headline_text'])
T.shape # TF-IDF 3000 행렬의 크기 확인
```

Out[6]:

```
(1082168, 3000)
```

In [7]:

```
from sklearn.decomposition import LatentDirichletAllocation

lda_model=LatentDirichletAllocation(n_components=10,learning_method='online',random_state=777,max_iter=1)
```

In [8]:

```
lda_top3=lda_model.fit_transform(T)
terms3 = vectorizer3.get_feature_names() # 단어 집합. 3,000개의 단어가 저장됨.
```

In [9]:

```
def get_topics(components, feature_names, n=10):
    for idx, topic in enumerate(components):
        print("Topic %d:" % (idx+1), [(feature_names[i], topic[i].round(2)) for i in topic.argsort()[:-n - 1:-1]]
```

In [10]:

```
get_topics(lda_model.components_,terms3)
```

```
Topic 1: [('melbourne', 6123.81), ('south', 5489.87), ('open', 4393.29), ('accuse', 3279.73), ('fall', 2963.08),
('budget', 2931.44), ('work', 2881.09), ('health', 2875.98), ('victoria', 2811.34), ('city', 2505.54)]
Topic 2: [('police', 9651.47), ('rise', 3298.73), ('death', 3196.13), ('flood', 2822.28), ('indigenous', 2726.02)
, ('beat', 2541.34), ('centre', 2487.28), ('search', 2441.13), ('talk', 2401.51), ('royal', 2358.97)]
Topic 3: [('change', 4761.16), ('year', 4627.02), ('hour', 3432.46), ('time', 3186.76), ('dead', 2870.28), ('clai
m', 2801.94), ('lead', 2775.57), ('federal', 2626.23), ('release', 2483.45), ('help', 2223.02)]
Topic 4: [('court', 5986.97), ('woman', 4548.92), ('face', 4114.78), ('test', 3981.04), ('price', 3124.0), ('figh
t', 3095.02), ('trial', 2914.05), ('child', 2909.81), ('farm', 2760.55), ('guilty', 2753.62)]
Topic 5: [('attack', 5430.16), ('canberra', 4965.14), ('interview', 4439.72), ('coast', 4334.15), ('tasmanian', 3
993.18), ('gold', 3089.53), ('service', 3030.97), ('final', 2861.95), ('lose', 2860.8), ('community', 2626.11)]
Topic 6: [('trump', 9080.34), ('government', 7189.86), ('world', 5247.35), ('country', 4737.61), ('home', 4601.0)
, ('school', 4508.37), ('warn', 4104.63), ('drug', 3560.57), ('people', 3431.21), ('national', 3341.36)]
Topic 7: [('australian', 8772.92), ('queensland', 6274.04), ('perth', 5235.21), ('house', 4980.23), ('market', 47
06.06), ('north', 4260.99), ('brisbane', 4034.79), ('tasmania', 3808.37), ('miss', 3664.85), ('west', 3315.57)]
Topic 8: [('sydney', 6729.77), ('report', 4519.95), ('live', 4355.77), ('donald', 3721.05), ('plan', 3415.89), ('
power', 3252.55), ('minister', 3126.5), ('leave', 3104.18), ('news', 2847.05), ('return', 2711.12)]
```

```
Topic 9: [('election', 6091.44), ('adelaide', 5572.41), ('rural', 4813.94), ('2016', 4543.96), ('make', 4408.25),
('crash', 4356.76), ('state', 4089.57), ('shoot', 3653.04), ('hospital', 3605.45), ('women', 3456.99)]
Topic 10: [('australia', 7824.37), ('charge', 6917.13), ('kill', 4673.94), ('years', 4167.26), ('jail', 3755.43),
('china', 3679.36), ('life', 3435.01), ('league', 3274.94), ('arrest', 3194.5), ('murder', 3160.97)]
```

In [11]:

```
#5,000

vectorizer5 = TfidfVectorizer(stop_words='english', max_features= 5000)

Y = vectorizer5.fit_transform(text['headline_text'])
Y.shape # TF-IDF 3000 행렬의 크기 확인
```

Out[11]:

```
(1082168, 5000)
```

In [12]:

```
lda_top5=lda_model.fit_transform(Y)
terms5 = vectorizer5.get_feature_names() # 단어 집합. 5,000개의 단어가 저장됨.
```

In [13]:

```
get_topics(lda_model.components_,terms5)
```

```
Topic 1: [('australian', 8122.21), ('government', 6740.12), ('attack', 5057.59), ('world', 4832.76), ('country',
4584.93), ('tasmania', 3572.61), ('miss', 3393.17), ('labor', 3082.25), ('minister', 2901.5), ('news', 2749.02)]
Topic 2: [('trump', 8500.1), ('adelaide', 5189.49), ('home', 4299.32), ('make', 3991.47), ('tasmanian', 3737.44),
('bank', 2567.5), ('return', 2484.5), ('federal', 2482.57), ('victorian', 2468.07), ('centre', 2317.92)]
Topic 3: [('court', 5553.79), ('market', 4440.27), ('report', 4192.57), ('face', 3817.26), ('brisbane', 3747.09),
('hour', 3410.47), ('share', 3116.5), ('rise', 3111.54), ('plan', 3054.87), ('trial', 2657.95)]
Topic 4: [('election', 5713.22), ('school', 4185.79), ('open', 4070.83), ('years', 3841.02), ('jail', 3442.16), (
'women', 3215.75), ('people', 3198.9), ('life', 3149.46), ('arrest', 2978.23), ('police', 2968.07)]
Topic 5: [('melbourne', 5722.01), ('change', 4449.62), ('year', 4296.24), ('live', 4140.6), ('crash', 4100.45), (
'donald', 3504.51), ('hospital', 3318.38), ('national', 3161.71), ('time', 2902.23), ('leave', 2832.15)]
Topic 6: [('perth', 4882.88), ('kill', 4367.42), ('2016', 4280.73), ('interview', 3816.45), ('china', 3453.33), (
'accuse', 3063.33), ('child', 2973.11), ('beat', 2371.49), ('near', 2322.92), ('push', 2134.25)]
Topic 7: [('house', 4597.83), ('north', 3987.88), ('test', 3722.37), ('shoot', 3400.21), ('west', 3121.76), ('pow
er', 3008.69), ('family', 2969.32), ('budget', 2776.09), ('drum', 2771.06), ('help', 2747.22)]
Topic 8: [('australia', 7419.87), ('queensland', 5826.46), ('police', 5084.35), ('woman', 4252.05), ('coast', 408
7.4), ('warn', 3807.55), ('death', 3546.59), ('drug', 3262.8), ('2015', 3110.45), ('league', 3108.2)]
Topic 9: [('south', 5141.07), ('canberra', 4653.86), ('turnbull', 3159.39), ('lose', 2643.41), ('farm', 2542.89),
('city', 2346.79), ('search', 2235.9), ('violence', 2212.56), ('hold', 2200.38), ('port', 2081.42)]
Topic 10: [('charge', 6435.89), ('sydney', 4755.74), ('rural', 4616.43), ('state', 3835.01), ('indigenous', 3171.
72), ('murder', 3130.99), ('fall', 2771.6), ('need', 2699.67), ('high', 2658.73), ('claim', 2595.18)]
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js