By Nitish Adhikari

Email id :nitishbuzzpro@gmail.com (mailto:nitishbuzzpro@gmail.com) , +91-9650740295

Topic Modeling Project

A dataset of over 400,000 quora questions that have no labeled cateogry, and attempting to find 20 cateogries to assign these questions to.

Import pandas and read in the quora_questions.csv file.

In [1]:
```
import pandas as pd
```

In [2]:
```
df = pd.read_csv('quora_questions.csv')
```

In [3]:
```
df.head()
```

Out[3]:

	Question
0	What is the step by step guide to invest in sh...
1	What is the story of Kohinoor (Koh-i-Noor) Dia...
2	How can I increase the speed of my internet co...
3	Why am I mentally very lonely? How can I solve...
4	Which one dissolve in water quikly sugar, salt...

Preprocessing

Use TF-IDF Vectorization to create a vectorized document term matrix.

In [4]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [7]:
```
tfidf = TfidfVectorizer(max_df=0.95, min_df=2, stop_words='english')
```

In [11]:
```
dtm = tfidf.fit_transform(df ['Question'])
```

In [12]:
```
dtm.shape
```

Out[12]:
```
(404289, 38669)
```

Non-negative Matrix Factorization

Using Scikit-Learn create an instance of NMF with 20 expected components.

In [13]:
```
from sklearn.decomposition import NMF
```

In [19]:
```
nmf_model = NMF(n_components=20, random_state=42)
```

In [20]:
```
nmf_model.fit(dtm)
```

```
C:\Users\DELL PC\AppData\Local\Programs\Python\Python310\lib\site-packages\sklearn\decomposition\_nmf.py:1692: Conver
genceWarning: Maximum number of iterations 200 reached. Increase it to improve convergence.
  warnings.warn(
```

Out[20]:

NMF

```
NMF(n_components=20, random_state=42)
```

In [21]:
```
nmf_model.components_.shape
```

Out[21]:
```
(20, 38669)
```

Print out the top 15 most common words for each of the 20 topics.

In [24]:
```
for index,topic in enumerate(nmf_model.components_):
    print(f'Topic {index}')
    print([tfidf.get_feature_names()[i] for i in topic.argsort()[-15:]])
    print('\n')
```

```
Topic 0
['thing', 'read', 'place', 'visit', 'places', 'phone', 'buy', 'laptop', 'movie', 'ways', '2016', 'books', 'book', 'mo
vies', 'best']


Topic 1
['majors', 'recruit', 'sex', 'looking', 'differ', 'use', 'exist', 'really', 'compare', 'cost', 'long', 'feel', 'wor
k', 'mean', 'does']


Topic 2
['add', 'answered', 'needing', 'post', 'easily', 'improvement', 'delete', 'asked', 'google', 'answers', 'answer', 'as
k', 'question', 'questions', 'quora']


Topic 3
['using', 'website', 'investment', 'friends', 'black', 'internet', 'free', 'home', 'easy', 'youtube', 'ways', 'earn',
'online', 'make', 'money']


Topic 4
['balance', 'earth', 'day', 'death', 'changed', 'live', 'want', 'change', 'moment', 'real', 'important', 'thing', 'me
aning', 'purpose', 'life']


Topic 5
['reservation', 'engineering', 'minister', 'president', 'company', 'china', 'business', 'country', 'olympics', 'avail
able', 'job', 'spotify', 'war', 'pakistan', 'india']


Topic 6
['beginners', 'online', 'english', 'book', 'did', 'hacking', 'want', 'python', 'languages', 'java', 'learning', 'star
t', 'language', 'programming', 'learn']


Topic 7
['happen', 'presidency', 'think', 'presidential', '2016', 'vote', 'better', 'election', 'did', 'win', 'hillary', 'pre
sident', 'clinton', 'donald', 'trump']


Topic 8
```

['russia', 'business', 'win', 'coming', 'countries', 'place', 'pakistan', 'happen', 'end', 'country', 'iii', 'start', 'did', 'war', 'world']

Topic 9
['indian', 'companies', 'don', 'guy', 'men', 'culture', 'women', 'work', 'girls', 'live', 'girl', 'look', 'sex', 'feel', 'like']

Topic 10
['ca', 'departments', 'positions', 'movies', 'songs', 'business', 'read', 'start', 'job', 'work', 'engineering', 'ways', 'bad', 'books', 'good']

Topic 11
['money', 'modi', 'currency', 'economy', 'think', 'government', 'ban', 'banning', 'black', 'indian', 'rupee', 'rs', '1000', 'notes', '500']

Topic 12
['blowing', 'resolutions', 'resolution', 'mind', 'likes', 'girl', '2017', 'year', 'don', 'employees', 'going', 'day', 'things', 'new', 'know']

Topic 13
['aspects', 'fluent', 'skill', 'spoken', 'ways', 'language', 'fluently', 'speak', 'communication', 'pronunciation', 'speaking', 'writing', 'skills', 'improve', 'english']

Topic 14
['diet', 'help', 'healthy', 'exercise', 'month', 'pounds', 'reduce', 'quickly', 'loss', 'fast', 'fat', 'ways', 'gain', 'lose', 'weight']

Topic 15
['having', 'feel', 'long', 'spend', 'did', 'person', 'machine', 'movies', 'favorite', 'job', 'home', 'sex', 'possible', 'travel', 'time']

Topic 16
['marriage', 'make', 'did', 'girlfriend', 'feel', 'tell', 'forget', 'really', 'friend', 'true', 'know', 'person', 'girl', 'fall', 'love']

```
Topic 17
['easy', 'hack', 'prepare', 'quickest', 'facebook', 'increase', 'painless', 'instagram', 'account', 'best', 'commit',
'fastest', 'suicide', 'easiest', 'way']


Topic 18
['web', 'java', 'scripting', 'phone', 'mechanical', 'better', 'job', 'use', 'account', 'data', 'software', 'science',
'computer', 'engineering', 'difference']


Topic 19
['earth', 'blowing', 'stop', 'use', 'easily', 'mind', 'google', 'flat', 'questions', 'hate', 'believe', 'ask', 'don',
'think', 'people']
```

Add a new column to the original quora dataframe that labels each question into one of the 20 topic categories.

In [25]:
```
df.head()
```

Out[25]:

	Question
0	What is the step by step guide to invest in sh...
1	What is the story of Kohinoor (Koh-i-Noor) Dia...
2	How can I increase the speed of my internet co...
3	Why am I mentally very lonely? How can I solve...
4	Which one dissolve in water quikly sugar, salt...

In [31]:
```
topic_result = nmf_model.transform(dtm) #document-topic matrix
topic_result
```

Out[31]:
```
array([[2.74927015e-04, 5.88014577e-05, 6.17412189e-06, ...,
        6.97429434e-04, 2.13458466e-04, 0.00000000e+00],
       [1.95705655e-04, 8.80743307e-05, 0.00000000e+00, ...,
        0.00000000e+00, 5.51003101e-05, 1.05546944e-05],
       [1.77372166e-04, 6.43938862e-04, 1.60463804e-03, ...,
        3.02446249e-03, 1.05890709e-03, 1.23898603e-03],
       ...,
       [0.00000000e+00, 1.61570029e-05, 5.23565129e-06, ...,
        0.00000000e+00, 2.76224751e-06, 0.00000000e+00],
       [5.34282407e-04, 1.01028959e-03, 0.00000000e+00, ...,
        1.28754707e-04, 7.76842889e-04, 0.00000000e+00],
       [0.00000000e+00, 0.00000000e+00, 0.00000000e+00, ...,
        0.00000000e+00, 0.00000000e+00, 1.25204924e-04]])
```

In [32]:
```
topic_result.shape #document-topic matrix
```

Out[32]:
```
(404289, 20)
```

In [37]:
```
topic_result.argmax(axis=1)
```

Out[37]:
```
array([ 5, 16, 17, ..., 11, 11,  9], dtype=int64)
```

In [38]:
```
df['Topic Category'] = topic_result.argmax(axis=1)
```

In [39]:
```
df #updated dataframe with 'Topic Category' feature
```

Out[39]:

	Question	Topic Category
0	What is the step by step guide to invest in sh...	5
1	What is the story of Kohinoor (Koh-i-Noor) Dia...	16
2	How can I increase the speed of my internet co...	17
3	Why am I mentally very lonely? How can I solve...	11
4	Which one dissolve in water quikly sugar, salt...	14
...	...	...
404284	How many keywords are there in the Racket prog...	6
404285	Do you believe there is life after death?	4
404286	What is one coin?	11
404287	What is the approx annual cost of living while...	11
404288	What is like to have sex with cousin?	9

404289 rows × 2 columns

Done!