Topic Modeling Assessment Project

Welcome to your Topic Modeling Assessment! For this project you will be working with a dataset of over 400,000 quora questions that have no labeled cateogry, and attempting to find 20 cateogries to assign these questions to. The .csv file of these text questions can be found underneath the Topic-Modeling folder.

Remember you can always check the solutions notebook and video lecture for any questions.

Task: Import pandas and read in the quora_questions.csv file.

In [8]:

In [52]:

In [53]:

Out[53]:

	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

Task: Use TF-IDF Vectorization to create a vectorized document term matrix. You may want to explore the max_df and min_df parameters.

In [40]:

In [41]:

In [42]:

In [43]:

Out[43]:
```
<404289x38669 sparse matrix of type '<class 'numpy.float64'>'
        with 2002912 stored elements in Compressed Sparse Row format>
```

Non-negative Matrix Factorization

TASK: Using Scikit-Learn create an instance of NMF with 20 expected components. (Use random_state=42)..

In [44]:

In [48]:

In [49]:

Out[49]:
```
NMF(alpha=0.0, beta_loss='frobenius', init=None, l1_ratio=0.0, max_iter=200,
  n_components=20, random_state=42, shuffle=False, solver='cd', tol=0.0001,
  verbose=0)
```

TASK: Print our the top 15 most common words for each of the 20 topics.

In [50]:

```
THE TOP 15 WORDS FOR TOPIC #0
['thing', 'read', 'place', 'visit', 'places', 'phone', 'buy', 'laptop', 'movie', 'ways', '2016', 'books', 'book', '
movies', 'best']


THE TOP 15 WORDS FOR TOPIC #1
['majors', 'recruit', 'sex', 'looking', 'differ', 'use', 'exist', 'really', 'compare', 'cost', 'long', 'feel', 'wor
k', 'mean', 'does']


THE TOP 15 WORDS FOR TOPIC #2
['add', 'answered', 'needing', 'post', 'easily', 'improvement', 'delete', 'asked', 'google', 'answers', 'answer', '
ask', 'question', 'questions', 'quora']


THE TOP 15 WORDS FOR TOPIC #3
['using', 'website', 'investment', 'friends', 'black', 'internet', 'free', 'home', 'easy', 'youtube', 'ways', 'earn
', 'online', 'make', 'money']


THE TOP 15 WORDS FOR TOPIC #4
['balance', 'earth', 'day', 'death', 'changed', 'live', 'want', 'change', 'moment', 'real', 'important', 'thing', '
meaning', 'purpose', 'life']
```

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

In [54]:

Out[54]:

	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 [55]:

In [56]:

Out[56]:

	Question	Topic
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
5	Astrology: I am a Capricorn Sun Cap moon and c...	1
6	Should I buy tiago?	0
7	How can I be a good geologist?	10
8	When do you use シ instead of し?	19
9	Motorola (company): Can I hack my Charter Moto...	17