```
!ls
```

```
kaggle.json  sample_data
```

```
import os
```

```
!pip install -q kaggle
```

```
from google.colab import files
files.upload()
```

Choose files No file chosen Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

```
Saving kaggle.json to kaggle.json
{'kaggle.json': b'{"username":"sachinsirohi47","key":"f2fd0222aa9ce5c99d043fa94906a757"}'}
```

```
!mkdir ~/.kaggle
```

```
!cp kaggle.json ~/.kaggle/
```

```
!chmod 600 ~/.kaggle/kaggle.json
```

```
!kaggle datasets list
```

```
ref                                                         title                                          size  lastUpdated          downloadCount  voteCount  usabilityRating
----------------------------------------------------------  ---------------------------------------------  -----  -------------------  -------------  ---------  ---------------
nelgiriyewithana/top-spotify-songs-2023                     Most Streamed Spotify Songs 2023                47KB  2023-08-26 11:04:57           5739        184  1.0
nelgiriyewithana/global-youtube-statistics-2023             Global YouTube Statistics 2023                  60KB  2023-07-28 15:36:38          16463        546  1.0
joebeachcapital/students-performance                        Students Performance                             2KB  2023-08-31 00:50:11           1631         38  1.0
iamsouravbanerjee/airline-dataset                           Airline Dataset                                  4MB  2023-08-30 12:03:12           2451         67  1.0
aliessamali/ecommerce                                       E-commerce Dataset                               7MB  2023-08-13 18:57:42           1740         37  1.0
jjinho/wikipedia-20230701                                   Wikipedia Plaintext (2023-07-01)                12GB  2023-07-17 01:43:57            957         75  1.0
jjinho/wikipedia-2023-07-faiss-index                        Wikipedia 2023 07 faiss index                    5GB  2023-07-17 01:57:37            486         36  1.0
alitaqi000/world-university-rankings-2023                   World University Rankings 2023                  70KB  2023-08-31 14:35:38           1487         44  1.0
mohammadtalib786/retail-sales-dataset                       Retail Sales Dataset                            11KB  2023-08-22 18:33:09           2004         46  1.0
joebeachcapital/homicides                                   Homicides                                        1MB  2023-08-30 00:08:57            756         30  1.0
inductiveanks/top-1000-imdb-movies-dataset                  Top 1000 IMDb Movies Dataset                    91KB  2023-08-05 05:39:34           2090         40  0.9411765
uom190346a/water-quality-and-potability                     Water Quality and Potability                   251KB  2023-09-04 07:25:50            797         38  1.0
carlmcbrideellis/zzzs-lightweight-training-dataset-target   Zzzs: Lightweight training dataset + target    185MB  2023-09-11 07:21:51            115         31  1.0
khalidalam980/top-rated-movies-data-set                     Top Rated Movies Data Set                        1MB  2023-08-19 12:29:40           1096         34  1.0
nikhil1e9/comic-characters                                  DC-MARVEL Comic Characters                     394KB  2023-08-27 15:19:48            504         22  1.0
ranzeet013/migraine-dataset                                 Migraine Dataset                                 3KB  2023-09-04 12:00:47            539         24  0.8235294
nelgiriyewithana/global-weather-repository                  World Weather Repository ( Daily Updating )    198KB  2023-09-10 23:03:01           1264         53  1.0
jocelyndumlao/global-food-prices                            Global - Food Prices                           222KB  2023-08-29 06:35:55            908         33  1.0
judith007/my-1-epoch                                        Kaggle_LLM_Deberta                               1GB  2023-08-19 07:26:25             66         14  0.6875
sooyoungher/smoking-drinking-dataset                        Smoking and Drinking Dataset with body signal   27MB  2023-08-30 04:27:31           1174         46  1.0
```

```
import os


os.environ ['KAGGLE_CONFIG_DIR'] = '.'


!kaggle competitions download -c quora-insincere-questions-classification -f train.csv -p data
```

```
 Warning: Your Kaggle API key is readable by other users on this system! To fix this, you can run 'chmod 600 ./kaggle.json'
 Downloading train.csv.zip to data
  89% 49.0M/54.9M [00:00<00:00, 76.5MB/s]
 100% 54.9M/54.9M [00:00<00:00, 77.8MB/s]
```

```
IS_KAGGLE = 'KAGGLE_KERNEL_RUN_TYPE' in os.environ




if IS_KAGGLE:
    data_dir = '../input/quora-insincere-questions-classification'
    train_fname = data_dir + '/train.csv'
    test_fname = data_dir + '/test.csv'
    sample_fname = data_dir + '/sample_submission.csv'
else:
    os.environ['KAGGLE_CONFIG_DIR'] = '.'
    !kaggle competitions download -c quora-insincere-questions-classification -f train.csv -p data
    !kaggle competitions download -c quora-insincere-questions-classification -f test.csv -p data
    !kaggle competitions download -c quora-insincere-questions-classification -f sample_submission.csv -p data
    train_fname = 'data/train.csv.zip'
    test_fname = 'data/test.csv.zip'
    sample_fname = 'data/sample_submission.csv.zip'
```

```
 Warning: Your Kaggle API key is readable by other users on this system! To fix this, you can run 'chmod 600 ./kaggle.json'
 train.csv.zip: Skipping, found more recently modified local copy (use --force to force download)
 Warning: Your Kaggle API key is readable by other users on this system! To fix this, you can run 'chmod 600 ./kaggle.json'
 Downloading test.csv.zip to data
  32% 5.00M/15.8M [00:00<00:00, 38.9MB/s]
 100% 15.8M/15.8M [00:00<00:00, 81.3MB/s]
 Warning: Your Kaggle API key is readable by other users on this system! To fix this, you can run 'chmod 600 ./kaggle.json'
 Downloading sample_submission.csv.zip to data
 100% 4.09M/4.09M [00:00<00:00, 41.6MB/s]
 100% 4.09M/4.09M [00:00<00:00, 41.5MB/s]
```

Data Loading of Three DataSet

```
import pandas as pd
raw_df = pd.read_csv(train_fname)
raw_df
```

	qid	question_text	target
0	00002165364db923c7e6	How did Quebec nationalists see their province...	0
1	000032939017120e6e44	Do you have an adopted dog, how would you enco...	0
2	0000412ca6e4628ce2cf	Why does velocity affect time? Does velocity a...	0
3	000042bf85aa498cd78e	How did Otto von Guericke used the Magdeburg h...	0
4	0000455dfa3e01eae3af	Can I convert montra helicon D to a mountain b...	0
...	...	...	...
1306117	ffffcc4e2331aaf1e41e	What other technical skills do you need as a c...	0
1306118	ffffd431801e5a2f4861	Does MS in ECE have good job prospects in USA ...	0
1306119	ffffd48fb36b63db010c	Is foam insulation toxic?	0
1306120	ffffec519fa37cf60c78	How can one start a research project based on ...	0
1306121	ffffed09fedb5088744a	Who wins in a battle between a Wolverine and a...	0

1306122 rows × 3 columns

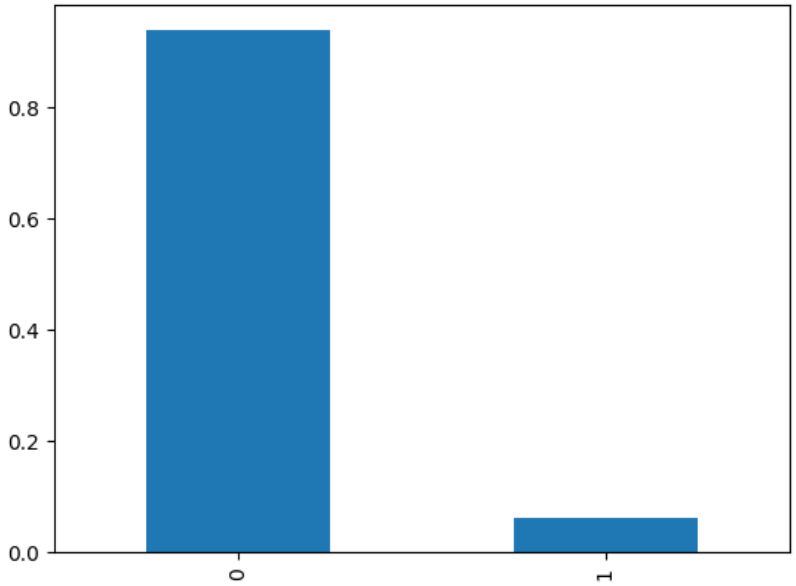
Read Train Dataset

```
sincere_df = raw_df[raw_df.target == 0]
sincere_df.question_text.values[:10]
insincere_df = raw_df[raw_df.target == 1]
insincere_df.question_text.values[:10]
raw_df.target.value_counts(normalize=True)
```

```
0    0.93813
1    0.06187
Name: target, dtype: float64
```

```
raw_df.target.value_counts(normalize=True).plot(kind='bar')
```

```
<Axes: >
```



Load Test Dataset

```
test_df = pd.read_csv(test_fname)
test_df
```

	qid	question_text
0	0000163e3ea7c7a74cd7	Why do so many women become so rude and arroga...
1	00002bd4fb5d505b9161	When should I apply for RV college of engineer...
2	00007756b4a147d2b0b3	What is it really like to be a nurse practitio...
3	000086e4b7e1c7146103	Who are entrepreneurs?
4	0000c4c3fbe8785a3090	Is education really making good people nowadays?
...	...	...
375801	ffff7fa746bd6d6197a9	How many countries listed in gold import in in...
375802	ffffa1be31c43046ab6b	Is there an alternative to dresses on formal p...
375803	ffffae173b6ca6bfa563	Where I can find best friendship quotes in Tel...
375804	fffffb1f7f1a008620287	What are the causes of refraction of light?
375805	fffff85473f4699474b0	Climate change is a worrying topic. How much t...

375806 rows × 2 columns

Load Test Sample

```
test_df = pd.read_csv(test_fname)
test_df
```

	qid	question_text
0	0000163e3ea7c7a74cd7	Why do so many women become so rude and arroga...
1	00002bd4fb5d505b9161	When should I apply for RV college of engineer...
2	00007756b4a147d2b0b3	What is it really like to be a nurse practitio...
3	000086e4b7e1c7146103	Who are entrepreneurs?
4	0000c4c3fbe8785a3090	Is education really making good people nowadays?
...	...	...
375801	ffff7fa746bd6d6197a9	How many countries listed in gold import in in...
375802	ffffa1be31c43046ab6b	Is there an alternative to dresses on formal p...
375803	ffffae173b6ca6bfa563	Where I can find best friendship quotes in Tel...
375804	fffffb1f7f1a008620287	What are the causes of refraction of light?
375805	fffff85473f4699474b0	Climate change is a worrying topic. How much t...

375806 rows × 2 columns

```
sub_df = pd.read_csv(sample_fname)
sub_df
```

	qid	prediction
0	0000163e3ea7c7a74cd7	0
1	00002bd4fb5d505b9161	0
2	00007756b4a147d2b0b3	0
3	000086e4b7e1c7146103	0
4	0000c4c3fbe8785a3090	0
...	...	...
375801	ffff7fa746bd6d6197a9	0
375802	ffffa1be31c43046ab6b	0
375803	ffffae173b6ca6bfa563	0
375804	fffffb1f7f1a008620287	0
375805	fffff85473f4699474b0	0

375806 rows × 2 columns

```
sub_df.prediction.value_counts()
```

```
0    375806
Name: prediction, dtype: int64
```

Create a Working Sample

Create a Working Sample

```
if IS_KAGGLE:
    SAMPLE_SIZE = len(raw_df)
else:
    SAMPLE_SIZE = 100_000


sample_df = raw_df.sample(SAMPLE_SIZE, random_state=42)
sample_df
```

	qid	question_text	target
443046	56d324bb1e2c29f43b12	What is the most effective classroom managemen...	0
947549	b9ad893dc78c577f8a63	Can I study abroad after 10th class from Bangl...	0
523769	6689ebaeeb65b209a412	How can I make friends as a college junior?	0
949821	ba1e2c4a0fef09671516	How do I download free APK Minecraft: Pocket E...	0
1030397	c9ea2b69bf0d74626f46	Like Kuvera, is "Groww" also a free online inv...	0
...	...	...	...
998930	c3c03a307a29c69971b4	How do I research list of reliable charcoal im...	0
66641	0d119aba95ee6684f506	What are petroleum products, and what is petro...	0
90024	11a46cd148a104b271cf	What are some services that will let you quick...	0
130113	1973e6e2111a0c93193a	What credit card processors do online marketpl...	0
1137	0037ed037520d82393c0	On which number system does a computer work?	0

100000 rows × 3 columns

Text Preprocessing Techniques

Outline:

Understand the bag of words model Tokenization Stop word removal Stemming Bag of Words Intuition Create a list of all the words across all the text documents You convert each document into vector counts of each word Limitations:

There may be too many words in the dataset

Some words may occur too frequently Some words may occur very rarely or only once A single word may have many forms (go, gone, going or bird vs. birds)

```
q0 = sincere_df.question_text.values[1]
q0
```

```
'Do you have an adopted dog, how would you encourage people to adopt and not shop?'
```

```
q1 = raw_df[raw_df.target == 1].question_text.values[0]
q1
```

```
'Has the United States become the largest dictatorship in the world?'
```

Tokenization

splitting a document into words and separators

```
import nltk
from nltk.tokenize import word_tokenize
```

```
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
True
```

```
q0
```

```
'Do you have an adopted dog, how would you encourage people to adopt and not shop?'
```

```
word_tokenize(q0)
```

```
['Do',
 'you',
 'have',
 'an',
 'adopted',
 'dog',
 ',',
 'how',
```

```
 'would',
 'you',
 'encourage',
 'people',
 'to',
 'adopt',
 'and',
 'not',
 'shop',
 '?']
```

```
word_tokenize(' this is (something) with, a lot of, punctuation;')
```

```
['this',
 'is',
 '(',
 'something',
 ')',
 'with',
 ',',
 'a',
 'lot',
 'of',
 ',',
 'punctuation',
 ';']
```

```
q1
```

```
'Has the United States become the largest dictatorship in the world?'
```

```
word_tokenize(q1)
```

```
['Has',
 'the',
 'United',
 'States',
 'become',
 'the',
 'largest',
 'dictatorship',
 'in',
 'the',
 'world',
 '?']
```

```
q0_tok = word_tokenize(q0)
q1_tok = word_tokenize(q1)
```

Stop Word Removal¶

Removing commonly occuring words

```
q1_tok
```

```
['Has',
 'the',
 'United',
 'States',
 'become',
 'the',
 'largest',
 'dictatorship',
 'in',
 'the',
 'world',
 '?']
```

```
from nltk.corpus import stopwords
nltk.download('stopwords')
english_stopwords = stopwords.words('english')
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

```
", ".join(english_stopwords)
```

```
'i, me, my, myself, we, our, ours, ourselves, you, you're, you've, you'll, you'd, your, yours, yourself, yourselves, he, him, h
is, himself, she, she's, her, hers, herself, it, it's, its, itself, they, them, their, theirs, themselves, what, which, who, wh
om, this, that, that'll, these, those, am, is, are, was, were, be, been, being, have, has, had, having, do, does, did, doing,
a, an, the, and, but, if, or, because, as, until, while, of, at, by, for, with, about, against, between, into, through, during,
before, after, above, below, to, from, up, down, in, out, on, off, over, under, again, further, then, once, here, there, when,
where, why, how, all, any, both, each, few, more, most, other, some, such, no, nor, not, only, own, same, so, than, too, very,
s, t, can, will, just, don, don't, should, should've, now, d, ll, m, o, re, ve, y, ain, aren, aren't, couldn, couldn't, didn, d
```

```
def remove_stopwords(tokens):
    return [word for word in tokens if word.lower() not in english_stopwords]
```

```
q0_tok
```

```
['Do',
 'you',
 'have',
 'an',
 'adopted',
 'dog',
 ',',
 'how',
```

```
 'would',
 'you',
 'encourage',
 'people',
 'to',
 'adopt',
 'and',
 'not',
 'shop',
 '?']
```

```
q0_stp = remove_stopwords(q0_tok)
```

```
q0_stp
```

```
['adopted', 'dog', ',', 'would', 'encourage', 'people', 'adopt', 'shop', '?']
```

```
q1_stp = remove_stopwords(q1_tok)
```

```
q1_tok
```

```
['Has',
 'the',
 'United',
 'States',
 'become',
 'the',
 'largest',
 'dictatorship',
 'in',
 'the',
 'world',
 '?']
```

```
q1_stp
```

```
['United', 'States', 'become', 'largest', 'dictatorship', 'world', '?']
```

Stemming

"go", "gone", "going" -> "go" "birds", "bird" -> "bird"

```
from nltk.stem.snowball import SnowballStemmer
stemmer = SnowballStemmer(language='english')
```

```
stemmer.stem('going')
```

```
'go'
```

```
stemmer.stem('supposedly')
```

```
'suppos'
```

```
q0_stm = [stemmer.stem(word) for word in q0_stp]
```

```
q0_stp
```

```
['adopted', 'dog', ',', 'would', 'encourage', 'people', 'adopt', 'shop', '?']
```

```
q0_stm
```

```
['adopt', 'dog', ',', 'would', 'encourag', 'peopl', 'adopt', 'shop', '?']
```

```
q1_stm = [stemmer.stem(word) for word in q1_stp]
```

```
q1_stp
```

```
['United', 'States', 'become', 'largest', 'dictatorship', 'world', '?']
```

```
q1_stm
```

```
['unit', 'state', 'becom', 'largest', 'dictatorship', 'world', '?']
```

Lemmatization

"love" -> "love" "loving" -> "love" "lovable" -> "love"

Implement Bag of Words

Outline:

1. Create a vocabulary using Count Vectorizer
2. Create a vocabulary using Count Vectorizer
3. Configure text preprocessing in Count Vectorizer

```
sample_df
```

	qid	question_text	target
443046	56d324bb1e2c29f43b12	What is the most effective classroom managemen...	0
947549	b9ad893dc78c577f8a63	Can I study abroad after 10th class from Bangl...	0
523769	6689ebaeeb65b209a412	How can I make friends as a college junior?	0
949821	ba1e2c4a0fef09671516	How do I download free APK Minecraft: Pocket E...	0
1030397	c9ea2b69bf0d74626f46	Like Kuvera, is "Groww" also a free online inv...	0
...	...	...	...
998930	c3c03a307a29c69971b4	How do I research list of reliable charcoal im...	0
66641	0d119aba95ee6684f506	What are petroleum products, and what is petro...	0
90024	11a46cd148a104b271cf	What are some services that will let you quick...	0
130113	1973e6e2111a0c93193a	What credit card processors do online marketpl...	0
1137	0037ed037520d82393c0	On which number system does a computer work?	0

100000 rows × 3 columns

```
small_df = sample_df[:5]
small_df
```

	qid	question_text	target
443046	56d324bb1e2c29f43b12	What is the most effective classroom managemen...	0
947549	b9ad893dc78c577f8a63	Can I study abroad after 10th class from Bangl...	0
523769	6689ebaeeb65b209a412	How can I make friends as a college junior?	0
949821	ba1e2c4a0fef09671516	How do I download free APK Minecraft: Pocket E...	0
1030397	c9ea2b69bf0d74626f46	Like Kuvera, is "Groww" also a free online inv...	0

```
small_df.question_text.values
```

```
array(['What is the most effective classroom management skill/technique to create a good learning environment?',
       'Can I study abroad after 10th class from Bangladesh?',
       'How can I make friends as a college junior?',
       'How do I download free APK Minecraft: Pocket Edition for iOS (iPhone)?',
       'Like Kuvera, is "Groww" also a free online investment platform where I can invest in direct mutual funds?'],
      dtype=object)
```

```
from sklearn.feature_extraction.text import CountVectorizer
small_vect = CountVectorizer()
small_vect.fit(small_df.question_text)
```

▾ CountVectorizer
CountVectorizer()

CountVectorizer() In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook. On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

```
small_vect.vocabulary_
```


```
{'what': 49,
 'is': 31,
 'the': 47,
 'most': 39,
 'effective': 16,
 'classroom': 9,
 'management': 37,
 'skill': 44,
 'technique': 46,
 'to': 48,
 'create': 11,
 'good': 23,
 'learning': 34,
 'environment': 17,
 'can': 7,
 'study': 45,
 'abroad': 1,
 'after': 2,
 '10th': 0,
 'class': 8,
 'from': 21,
 'bangladesh': 6,
 'how': 25,
 'make': 36,
 'friends': 20,
 'as': 5,
 'college': 10,
 'junior': 32,
 'do': 13,
 'download': 14,
 'free': 19,
 'apk': 4,
 'minecraft': 38,
 'pocket': 43,
 'edition': 15,
```

```
 'for': 18,
 'ios': 29,
 'iphone': 30,
 'like': 35,
 'kuvera': 33,
 'groww': 24,
 'also': 3,
 'online': 41,
 'investment': 28,
 'platform': 42,
 'where': 50,
 'invest': 27,
 'in': 26,
 'direct': 12,
 'mutual': 40,
 'funds': 22}
```

```
small_vect.get_feature_names_out()
```

```
array(['10th', 'abroad', 'after', 'also', 'apk', 'as', 'bangladesh',
       'can', 'class', 'classroom', 'college', 'create', 'direct', 'do',
       'download', 'edition', 'effective', 'environment', 'for', 'free',
       'friends', 'from', 'funds', 'good', 'groww', 'how', 'in', 'invest',
       'investment', 'ios', 'iphone', 'is', 'junior', 'kuvera',
       'learning', 'like', 'make', 'management', 'minecraft', 'most',
       'mutual', 'online', 'platform', 'pocket', 'skill', 'study',
       'technique', 'the', 'to', 'what', 'where'], dtype=object)
```

Start coding or generate with AI.

Transform documents into Vectors

```
vectors = small_vect.transform(small_df.question_text)
vectors.shape
```

```
(5, 51)
```

```
small_df.question_text.values[0]
```

```
'What is the most effective classroom management skill/technique to create a good learning environment?'
```

Double-click (or enter) to edit

```
vectors[0].toarray()
```

```
array([[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0,
        0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0,
        1, 0, 1, 1, 1, 1, 0]])
```

```
vectors.toarray()
```

```
array([[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0,
        0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0,
        1, 0, 1, 1, 1, 1, 0],
       [1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 1, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,
        0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 0,
        0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1,
        0, 0, 0, 0, 0, 0, 0],
       [0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0,
        1, 0, 1, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0,
        0, 0, 0, 0, 0, 0, 1]])
```

Learn More in this link

https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

```
Start coding or generate with AI.
```

ML Models for Text Classification

Outline:

- Create a training & validation set
- Train a logistic regression model
- Make predictions on training, validation & test data

Split into Training and Validation Set

```
sample_df
```

	qid	question_text	target
442046	50d224bb1e2e20f42b12	What is the most effective classroom managemen	0