In [2]:

```
import pandas as pd
df=pd.read_csv('~/Desktop/subreddit.csv')
```

In [3]:

```
df.head()
```

Out[3]:

	title	score	id	url	comms_num	c
0	Every little bit counts	9187	5pcees	https://i.reddituploads.com/fa91b455743943a6bf...	1035	1.4850
1	This is how many people are at the Women's Mar...	5794	5pbdhm	https://i.redd.it/4r0jhwibd3by.jpg	1103	1.4850
2	Just out for a little jog today before the hur...	2876	129vlt	http://i.imgur.com/bHtwi.jpg	334	1.3515
3	We really live in a beautiful place.	2225	5e5nhz	https://i.redd.it/hfdeeuyqo0zx.jpg	46	1.4797
4	Just watched this guy house a 2 liter bottle o...	1906	340x5o	http://imgur.com/B0baNHN	164	1.4301

In [6]:

```
X = df['body']
y = df['title']
```

In [7]:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_sta
```

In [9]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
tvec = TfidfVectorizer(stop_words='english', max_features=1000, ngram_range=[1,2] )
tvec
```

Out[9]:

```
TfidfVectorizer(analyzer='word', binary=False, decode_error='strict',
                dtype=<class 'numpy.float64'>, encoding='utf-8',
                input='content', lowercase=True, max_df=1.0, max_featu
res=1000,
                min_df=1, ngram_range=[1, 2], norm='l2', preprocessor=
None,
                smooth_idf=True, stop_words='english', strip_accents=N
one,
                sublinear_tf=False, token_pattern='(?u)\\b\\w\\w+\\b',
                tokenizer=None, use_idf=True, vocabulary=None)
```

In [10]:

```
X_train_tvec = tvec.fit_transform(X_train.values.astype('U'))
```

In [11]:

```
X_test_tvec = tvec.transform(X_test.values.astype('U'))
```