In [1]:
```
import numpy as np
```

In [2]:
```
doc=['hello my name is chandan how r u ','Hello ,win money ,win from me','Call me hello,call me tomorrow','Welcome India']
```

In [3]:
```
doc
```

Out[3]:
```
['hello my name is chandan how r u ',
 'Hello ,win money ,win from me',
 'Call me hello,call me tomorrow',
 'Welcome India']
```

In [4]:
```
small_doc=[]
for i in doc:
    small_doc.append(i.lower())
print(small_doc)
```

```
['hello my name is chandan how r u ', 'hello ,win money ,win from me', 'call me hello,call me tomorrow', 'welcome india']
```

In [5]:
```
#remove punctuation
doc_pun=[]
import string
for i in small_doc:
    doc_pun.append(i.translate(str.maketrans('','',string.punctuation)))
print(doc_pun)
```

```
['hello my name is chandan how r u ', 'hello win money win from me', 'call me hellocall me tomorrow', 'welcome india']
```

In [6]:
```
# every token is splitted as individual entry
doc_new=[]
for i in doc_pun:
    doc_new.append(i.split(' '))
print(doc_new)
```

```
[['hello', 'my', 'name', 'is', 'chandan', 'how', 'r', 'u', ''], ['hello', 'win', 'money', 'win', 'from', 'me'], ['call', 'me', 'hellocall', 'me', 't
omorrow'], ['welcome', 'india']]
```

In [7]:

```
# checking each sample and count token in particular sample
word_list=[]
import pprint #used for text
from collections import Counter
for i in doc_new:
    word_list.append(Counter(i))
pprint.pprint(word_list)
```

```
[Counter({'hello': 1,
          'my': 1,
          'name': 1,
          'is': 1,
          'chandan': 1,
          'how': 1,
          'r': 1,
          'u': 1,
          '': 1}),
 Counter({'win': 2, 'hello': 1, 'money': 1, 'from': 1, 'me': 1}),
 Counter({'me': 2, 'call': 1, 'hellocall': 1, 'tomorrow': 1}),
 Counter({'welcome': 1, 'india': 1})]
```

In [8]:

```
from sklearn.feature_extraction.text import CountVectorizer
```

In [9]:

```
count_vect=CountVectorizer()
count_vect.fit(doc)
```

Out[9]:

```
CountVectorizer(analyzer='word', binary=False, decode_error='strict',
                dtype=<class 'numpy.int64'>, encoding='utf-8', input='content',
                lowercase=True, max_df=1.0, max_features=None, min_df=1,
                ngram_range=(1, 1), preprocessor=None, stop_words=None,
                strip_accents=None, token_pattern='(?u)\\b\\w\\w+\\b',
                tokenizer=None, vocabulary=None)
```

In [10]:

```
# to get the feature names
count_vect.get_feature_names()
```

Out[10]:

```
['call',
 'chandan',
 'from',
 'hello',
 'how',
 'india',
 'is',
 'me',
 'money',
 'my',
 'name',
 'tomorrow',
 'welcome',
 'win']
```

In [11]:

```
mydoc_array=count_vect.transform(doc).toarray()
mydoc_array
```

Out[11]:

```
array([[0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0],
       [0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 2],
       [2, 0, 0, 1, 0, 0, 0, 2, 0, 0, 0, 1, 0, 0],
       [0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0]], dtype=int64)
```