In [5]:

```
from sklearn.feature_extraction.text import CountVectorizer
vect = CountVectorizer(binary = True)
```

In [6]:

```
corpus=["I have a german shepard","German shepard is from german","germans love gossipin
```

In [7]:

```
vect.fit(corpus)
```

Out[7]:

```
CountVectorizer(binary=True)
```

In [17]:

```
vocab=vect.vocabulary_
```

In [19]:

```
for key in sorted(vocab.keys()):
    print("{}:{}" .format(key,vocab[key]))
```

```
from:0
german:1
germans:2
gossiping:3
have:4
is:5
love:6
shepard:7
```

In [21]:

```
print(vect.transform(["Germany has german shepard"]).toarray())
```

```
[[0 1 0 0 0 0 0 1]]
```

In [25]:

```
from sklearn.metrics.pairwise import cosine_similarity
similarity=cosine_similarity(vect.transform(["German has German shepard,german has capit
```

In [26]:

```
print(similarity)
```

```
[[1.]]
```

In [ ]: