In [1]:

```
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [2]:

```
documentA = 'Jupiter is the largest Planet'
documentB = 'Mars is the fourth planet from the sun'
```

In [3]:

```
bagOfWordsA = documentA.split(' ')
bagOfWordsB = documentB.split(' ')
```

In [4]:

```
uniqueWords =set (bagOfWordsA).union(set(bagOfWordsB))
```

In [5]:

```
numOfWordsA = dict.fromkeys(uniqueWords, 0)
```

In [6]:

```
for word in bagOfWordsA:numOfWordsA[word] += 1
numOfWordsB = dict.fromkeys(uniqueWords,0)
for word in bagOfWordsB:numOfWordsB[word] += 1
```

In [7]:

```
def computeTF(wordDict, bagOfWords):
    tfDict = {}
    bagOfWordsCount =len(bagOfWords)
    for word, count   in wordDict.items():
        tfDict[word] = count / float(bagOfWordsCount)
    return tfDict
tfA = computeTF(numOfWordsA,bagOfWordsA)
tfB =computeTF(numOfWordsB, bagOfWordsB)
tfA
```

Out[7]:

```
{'is': 0.2,
 'from': 0.0,
 'Planet': 0.2,
 'fourth': 0.0,
 'largest': 0.2,
 'planet': 0.0,
 'the': 0.2,
 'Jupiter': 0.2,
 'sun': 0.0,
 'Mars': 0.0}
```

In [12]:

```
def computeIDF(documents):
    import math
    N = len(documents)
    idfDict = dict.fromkeys(documents[0].keys(),0)
    for document in documents:
        for word, val in document.items():
            if val > 0:
                idfDict[word] += 1
    for word,  val in idfDict.items():
         idfDict[word] = math.log(N / float(val))
    return idfDict
idfs = computeIDF([numOfWordsA,numOfWordsB])
idfs
```

Out[12]:

```
{'is': 0.0,
 'from': 0.6931471805599453,
 'Planet': 0.6931471805599453,
 'fourth': 0.6931471805599453,
 'largest': 0.6931471805599453,
 'planet': 0.6931471805599453,
 'the': 0.0,
 'Jupiter': 0.6931471805599453,
 'sun': 0.6931471805599453,
 'Mars': 0.6931471805599453}
```

In [13]:

```
def computeTFIDF(tfBagOfWords, idfs):
    tfidf = {}
    for word, val in tfBagOfWords.items():
        tfidf[word] = val * idfs[word]
    return tfidf
tfidfA = computeTFIDF(tfA,idfs)
tfidfB = computeTFIDF(tfB,idfs)
df = pd.DataFrame([tfidfA,tfidfB])
df
```

Out[13]:

	is	from	Planet	fourth	largest	planet	the	Jupiter	sun	Mars
0	0.0	0.000000	0.138629	0.000000	0.138629	0.000000	0.0	0.138629	0.000000	0.000000
1	0.0	0.086643	0.000000	0.086643	0.000000	0.086643	0.0	0.000000	0.086643	0.086643

In [ ]: