Lab3. Computing Document Similarity using VSM

In [ ]:

```
Name:Swetha.K
Roll No:235229143
```

In [3]:

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

docs = [
    "good movie", "not a good movie", "did not like",
    "i like it", "good one"
]

# using default tokenizer in TfidfVectorizer
tfidf = TfidfVectorizer(min_df=2, max_df=0.5, ngram_range=(1, 2))
features = tfidf.fit_transform(docs)
print(features)

# Pretty printing
df = pd.DataFrame(
    features.todense(),
    columns=tfidf.get_feature_names_out())  # Use get_feature_names_out() insteac
print(df)
```

```
  (0, 0)	0.7071067811865476
  (0, 2)	0.7071067811865476
  (1, 3)	0.5773502691896257
  (1, 0)	0.5773502691896257
  (1, 2)	0.5773502691896257
  (2, 1)	0.7071067811865476
  (2, 3)	0.7071067811865476
  (3, 1)	1.0
   good movie      like     movie       not
0    0.707107  0.000000  0.707107  0.000000
1    0.577350  0.000000  0.577350  0.577350
2    0.000000  0.707107  0.000000  0.707107
3    0.000000  1.000000  0.000000  0.000000
4    0.000000  0.000000  0.000000  0.000000
```

In [ ]: