TEXT PROCESSING MENGGUNAKAN STOP WORD FILTERING
Soal No. 4

UAS
SISTEM CERDAS

Oleh:
DEWI AROFAH
NIM 18.52.0013

PROGRAM STUDI S1 – TEKNOLOGI INFORMASI

KEMENTERIAN RISTEK DAN PENDIDIKAN TINGGI
SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER
STMIK PPKIA PRADNYA PARAMITA
MALANG
2021

1. Masukkan Dataset dalam bentuk corpus yang disajikan per kalimat. Sebagai contoh topik yang digunakan yaitu Omicron.

```
corpus = [
    'CDC has been collaborating with global public health and industry partners to learn about Omicron',
    'The Omicron variant likely will spread more easily than the original SARS-CoV-2 virus and how easily Omicron spreads compar
    'CDC expects that anyone with Omicron infection can spread the virus to others, even if they are vaccinated or don't have sy
]

corpus
```

```
['CDC has been collaborating with global public health and industry partners to learn about Omicron',
 'The Omicron variant likely will spread more easily than the original SARS-CoV-2 virus and how easily Omicron spreads compared
to Delta remains unknown',
 'CDC expects that anyone with Omicron infection can spread the virus to others, even if they are vaccinated or don't have symp
toms']
```

2. Selanjutnya yaitu melakukan Bag of Words model dengan memanfaatkan CountVectorizer.

```
from sklearn.feature_extraction.text import CountVectorizer

vectorizer = CountVectorizer()
vectorizer_X = vectorizer.fit_transform(corpus).todense()
vectorizer_X
```

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

3. Untuk melihat daftar kata yang ada dalam text, bisa menggunakan syntax berikut

```
vectorizer.get_feature_names()
```

```
Out[3]: ['about',
         'and',
         'anyone',
         'are',
         'been',
         'can',
         'cdc',
         'collaborating',
         'compared',
         'cov',
         'delta',
         'don',
         'easily',
         'even',
         'expects',
         'global',
         'has',
         'have',
         'health',
         'how',
         'if',
         'industry',
         'infection',
         'learn',
         'likely',
         'more',
         'omicron',
         'or',
         'original',
         'others',
         'partners',
         'public',
         'remains',
         'sars',
         'spread',
         'spreads',
         'symptoms',
         'than',
         'that',
         'the',
         'they',
         'to',
         'unknown',
         'vaccinated',
         'variant',
         'virus',
         'will',
         'with']
```

4. Euclidean Distance digunakan untuk mengukur jarak antar dokumen.

```
In [4]: from sklearn.metrics.pairwise import euclidean_distances

for i in range(len(vectorizer_X)) :
    for j in range(i, len(vectorizer_X)):
        if i == j :
            continue
        jarak = euclidean_distances(vectorizer_X[i], vectorizer_X[j])
        print(f'Jarak dokumen {i+1} dan {j+1}: {jarak}')
```

```
Jarak dokumen 1 dan 2: [[6.08276253]]
Jarak dokumen 1 dan 3: [[5.38516481]]
Jarak dokumen 2 dan 3: [[6.164414]]
```

5. Stop Word Filtering pada text. Proses ini menghilangkan beberapa kata seperti determiners, aux verb dan preposition

```
In [5]: from sklearn.feature_extraction.text import CountVectorizer

vectorizer = CountVectorizer(stop_words='english')
vectorizer_X = vectorizer.fit_transform(corpus).todense()
vectorizer_X
```

```
Out[5]: matrix([[1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0,
                 0, 0, 0, 0, 0, 0],
                [0, 0, 1, 1, 1, 0, 2, 0, 0, 0, 0, 0, 0, 1, 2, 1, 0, 0, 1, 1, 1,
                 1, 0, 1, 0, 1, 1],
                [1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1,
                 0, 1, 0, 1, 0, 1]], dtype=int64)
```

6. Tampilkan lagi menggununakan get_feature_names. Dapat dilihat bahwa terdapat pengurangan beberapa kata yang tadinya berjumlah 48 kata, sekarang hanya tersisa 27 kata.

```
In [6]: vectorizer.get_feature_names()
```

```
F:\Anaconda\lib\site-packages\sklearn
ture_names is deprecated in 1.0 and w
  warnings.warn(msg, category=FutureW
```

```
Out[6]: ['cdc',
 'collaborating',
 'compared',
 'cov',
 'delta',
 'don',
 'easily',
 'expects',
 'global',
 'health',
 'industry',
 'infection',
 'learn',
 'likely',
 'omicron',
 'original',
 'partners',
 'public',
 'remains',
 'sars',
 'spread',
 'spreads',
 'symptoms',
 'unknown',
 'vaccinated',
 'variant',
 'virus']
```