classification

March 18, 2019

0.1 Klasifikasi Text - Tugas 3 Pengenalan Pola

0.2 Baskara - 16/398499/PA/17460

0.2.1 Import Library

In [1]:
```
import pandas as pd
import numpy as np
import nltk
import math
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
```

0.2.2 Data Preparation (Pembagian Menjadi Training & Test)

In [2]:
```
df = pd.read_csv('data_penyakit.csv', names=['tanda_gejala','diagnosis_penyakit'])
training = df[df['diagnosis_penyakit'].notnull()]
test = df[df['diagnosis_penyakit'].isnull()]
```

In [3]:
```
training.head()
```

Out[3]:

	tanda_gejala	diagnosis_penyakit
0	menggigil, demam, sakit kepala	Malaria (bentuk benigma)
1	Kaku kuduk, penurunan kesadaran, muntah proyek...	Meningitis + perdarahan subarachnoid
3	Pipi bengkak, nyeri saat mengunyah, nyeri testis	Parotitis
5	Sakit gigi, gigi sensitif pada makanan dingin ...	Karies dentis
6	Hidung tersumbat, bersin, batuk, sakit tenggor...	Common cold

In [4]:
```
test.head()
```

Out[4]:

	tanda_gejala	diagnosis_penyakit
2	Mata lengket, mata berair, pandangan sedikit k...	NaN
4	Gusi bengkak, gusi kemerahan, gusi berdarah	NaN
18	Batuk lebih dari tiga minggu, sesak napas atau...	NaN

```
19  Demam, menggigil, suhu tubuh meningkat, batuk ...              NaN
29  Demam, muntah, diare cair, ampas sedikit seper...              NaN
```

0.2.3 Tokenization & Stemming Data Training

```
In [5]: factory = StemmerFactory()
        stemmer = factory.create_stemmer()
        tokenizer = nltk.RegexpTokenizer(r'\w+')

        for index, row in training.iterrows():
            # Stemming
            stemmed = stemmer.stem(row[0])
            #Tokenization
            tokens = tokenizer.tokenize(row[0])
            #Case Folding
            words = [w.lower() for w in tokens]
            training.at[index, 'tanda_gejala'] = words
```

```
In [6]: training.head()
```

```
Out[6]:                                          tanda_gejala  \
        0                  [menggigil, demam, sakit, kepala]
        1  [kaku, kuduk, penurunan, kesadaran, muntah, pr...
        3  [pipi, bengkak, nyeri, saat, mengunyah, nyeri,...
        5  [sakit, gigi, gigi, sensitif, pada, makanan, d...
        6  [hidung, tersumbat, bersin, batuk, sakit, teng...

                         diagnosis_penyakit
        0              Malaria (bentuk benigma)
        1  Meningitis + perdarahan subarachnoid
        3                            Parotitis
        5                        Karies dentis
        6                          Common cold
```

0.2.4 Membuat Kolom Untuk Setiap Kata

```
In [7]: columnlist = []
        for index, row in training.iterrows():
            columnlist = np.concatenate((columnlist, row[0]))
        columnlist = np.unique(columnlist)
```

```
In [8]: for index in range(len(columnlist)):
            training.insert(2, str(columnlist[index]), 0)
```

```
In [9]: training.head()
```

```
Out[9]:                                          tanda_gejala  \
        0                  [menggigil, demam, sakit, kepala]
        1  [kaku, kuduk, penurunan, kesadaran, muntah, pr...
```

```
3  [pipi, bengkak, nyeri, saat, mengunyah, nyeri,...
5  [sakit, gigi, gigi, sensitif, pada, makanan, d...
6  [hidung, tersumbat, bersin, batuk, sakit, teng...

                    diagnosis_penyakit  yang  warna  wajah  vulva  volume  \
0              Malaria (bentuk benigma)     0      0      0      0       0
1  Meningitis + perdarahan subarachnoid     0      0      0      0       0
3                             Parotitis     0      0      0      0       0
5                         Karies dentis     0      0      0      0       0
6                           Common cold     0      0      0      0       0

   vesikuler  vesikul  vesikel  ...  ampas  amis  amandel  alis  aksila  \
0          0        0        0  ...      0     0        0     0       0
1          0        0        0  ...      0     0        0     0       0
3          0        0        0  ...      0     0        0     0       0
5          0        0        0  ...      0     0        0     0       0
6          0        0        0  ...      0     0        0     0       0

   akibat  akan  agak  ada  abdomen
0       0     0     0    0        0
1       0     0     0    0        0
3       0     0     0    0        0
5       0     0     0    0        0
6       0     0     0    0        0

[5 rows x 427 columns]
```

0.2.5 Menghitung Jumlah Frekuensi Setiap Kata

```
In [10]: for index, row in training.iterrows():
             for columnindex in range(len(columnlist)):
                 training.at[index, columnlist[columnindex]] = row[0].count(str(columnlist[colum

In [11]: training.head()

Out[11]:                                          tanda_gejala  \
         0                 [menggigil, demam, sakit, kepala]
         1  [kaku, kuduk, penurunan, kesadaran, muntah, pr...
         3  [pipi, bengkak, nyeri, saat, mengunyah, nyeri,...
         5  [sakit, gigi, gigi, sensitif, pada, makanan, d...
         6  [hidung, tersumbat, bersin, batuk, sakit, teng...

                           diagnosis_penyakit  yang  warna  wajah  vulva  volume  \
         0              Malaria (bentuk benigma)     0      0      0      0       0
         1  Meningitis + perdarahan subarachnoid     0      0      0      0       0
         3                             Parotitis     0      0      0      0       0
         5                         Karies dentis     0      0      0      0       0
         6                           Common cold     0      0      0      0       0
```

```
   vesikuler  vesikul  vesikel  ...  ampas  amis  amandel  alis  aksila  \
0          0        0        0  ...      0     0        0     0       0
1          0        0        0  ...      0     0        0     0       0
3          0        0        0  ...      0     0        0     0       0
5          0        0        0  ...      0     0        0     0       0
6          0        0        0  ...      0     0        0     0       0

   akibat  akan  agak  ada  abdomen
0       0     0     0    0        0
1       0     0     0    0        0
3       0     0     0    0        0
5       0     0     0    0        0
6       0     0     0    0        0

[5 rows x 427 columns]
```

0.2.6 Prepare Test Data

```
In [12]: test.insert(2, 'jarak', 0.0)
```

0.2.7 Preprocessing dan Penghitungan Jumlah Frekuensi Setiap Kata

```
In [13]: for index, row in test.iterrows():
             # Stemming
             stemmed = stemmer.stem(row[0])
             #Tokenization
             tokens = tokenizer.tokenize(row[0])
             #Case Folding
             words = [w.lower() for w in tokens]
             test.at[index, 'tanda_gejala'] = words
         for index in range(len(columnlist)):
             test.insert(3, str(columnlist[index]), 0)
         for index, row in test.iterrows():
             for columnindex in range(len(columnlist)):
                 test.at[index, columnlist[columnindex]] = row[0].count(columnlist[columnindex])
```

0.2.8 Penghitungan Jarak (Menggunakan Euclidean)

```
In [14]: for test_index, test_row in test.iterrows():
             distance = []
             for train_index, train_row in training.iterrows():
                 temp = 0
                 for columnindex in range(len(columnlist)):
                     temp = temp + (test_row[3+columnindex] - train_row[2+columnindex])**2
                 distance += [math.sqrt(temp)]
             test.at[test_index, 'jarak'] = (np.min(distance))
             test.at[test_index, 'diagnosis_penyakit'] = str(df['diagnosis_penyakit'][np.argmin(
```

0.2.9 Hasil Prediksi

In [15]:
```
test.iloc[:, : 3]
```

Out[15]:

	tanda_gejala	diagnosis_penyakit	jarak
2	[mata, lengket, mata, berair, pandangan, sedik...	Malaria (bentuk benigma)	4.242641
4	[gusi, bengkak, gusi, kemerahan, gusi, berdarah]	Malaria (bentuk benigma)	4.000000
18	[batuk, lebih, dari, tiga, minggu, sesak, napa...	Pneumonia	4.123106
19	[demam, menggigil, suhu, tubuh, meningkat, bat...	Malaria (bentuk benigma)	3.316625
29	[demam, muntah, diare, cair, ampas, sedikit, s...	Filariasis	2.449490
36	[nyeri, kolik, daerah, pinggang, malaise, mual...	Disentri	2.645751
76	[ruam, yang, gatal, terdri, dari, macula, maku...	Malaria (bentuk benigma)	4.242641

In [ ]:
```
```