Text Mining		
Praktikan	**Aslab**	
Nama: xxxx Nim: xxxx	Annur Hangga Prihadi	065001800028
	Faiz Kumara	065001800003

PRAKTIKUM 5

DATA SAINS DAN ANALITIK

Topik pertemuan praktikum ke-lima adalah mengolah teks untuk kegiatan analisa bahan data berupa teks.

Source Code:
https://github.com/hanggaa/PrakDSDA/blob/main/Prak%205%20Text%20Processing.ipynb

Latihan 1

1. Memasang library yang dibutuhkan

```
In [1]: import sys
        !{sys.executable} -m pip install nltk

Requirement already satisfied: nltk in c:\users\hangg\anaconda3\lib\site-packages (3.6.1)
Requirement already satisfied: click in c:\users\hangg\anaconda3\lib\site-packages (from nlt
Requirement already satisfied: tqdm in c:\users\hangg\anaconda3\lib\site-packages (from nltk
Requirement already satisfied: joblib in c:\users\hangg\anaconda3\lib\site-packages (from nl
Requirement already satisfied: regex in c:\users\hangg\anaconda3\lib\site-packages (from nlt

In [2]: import re
        import string
        import nltk
        nltk.download('all')
        from nltk.tokenize import word_tokenize

[nltk_data] Downloading collection 'all'
[nltk_data]    |
```

2. Membaca file

```
In [3]: f = open("C:/Users/hangg/Downloads/Random Aslab/DSDA/Material/news.txt","r") #Pastikan sesuai dengan directory file
        text = f.read()
        f.close()

        print("Bacaan:\n----------------\n", text)

Bacaan:
----------------
 Sejak sebulan terakhir harga bawang putih dan bombay yang sempat melonjak tinggi akhirnya kembali turun dan stabil
a Rp. 20.000 per kg. Namun, berkah harga murah yang dinikmati masyarakat kembali terusik dengan mulai naiknya harga
h di pasar. Pemerhati Pertanian, Syaiful Bahari, menjelaskan masalah kenaikan harga komoditi yang terkait dengan im
bawang putih, bombay dan gula, selama ini lebih banyak disebabkan oleh kebijakan restriksi atau pembatasan yang dib
eh pemerintah sendiri. Untuk kasus bawang putih dan bombay, lanjut Syaiful, ketika relaksasi diberlakukan terbukti
drastis. Bombay dari Rp 150.000 per kilo gram menjadi Rp 17.000 sampai Rp. 20.000 per kilo gram. Sehingga kedua kom
nyumbang deflasi.
```

3. Mengubah huruf menjadi huruf kecil

```
In [4]: text = text.lower()
        print("Huruf kecil semua:\n----------------\n", text)

Huruf kecil semua:
----------------
 sejak sebulan terakhir harga bawang putih dan bombay yang sempat
a rp. 20.000 per kg. namun, berkah harga murah yang dinikmati mas
h di pasar. pemerhati pertanian, syaiful bahari, menjelaskan masa
bawang putih, bombay dan gula, selama ini lebih banyak disebabkan
eh pemerintah sendiri. untuk kasus bawang putih dan bombay, lanju
```

4. **Menghapus angka pada paragraf**

```
In [5]: text = re.sub(r"\d+", "", text)
        print("Menghilangkan angka:\n----------------\n", text)
```

```
Menghilangkan angka:
----------------
 sejak sebulan terakhir harga bawang putih dan bombay yang semp
a rp. . per kg. namun, berkah harga murah yang dinikmati masyar
pasar. pemerhati pertanian, syaiful bahari, menjelaskan masalah
g putih, bombay dan gula, selama ini lebih banyak disebabkan ol
merintah sendiri. untuk kasus bawang putih dan bombay, lanjut s
is. bombay dari rp . per kilo gram menjadi rp . sampai rp. . pe
```

5. **Menghapus tanda baca**

```
In [6]: text = text.translate(str.maketrans("","",string.punctuation))
        print("Menghilangkan Tanda Baca:\n----------------\n",text)
```

```
Menghilangkan Tanda Baca:
----------------
 sejak sebulan terakhir harga bawang putih dan bombay yang sempat melon
rp  per kg namun berkah harga murah yang dinikmati masyarakat kembali t
pemerhati pertanian syaiful bahari menjelaskan masalah kenaikan harga k
ombay dan gula selama ini lebih banyak disebabkan oleh kebijakan restri
ndiri untuk kasus bawang putih dan bombay lanjut syaiful ketika relaksa
rp  per kilo gram menjadi rp  sampai rp  per kilo gram sehingga kedua k
```

6. **Menghapus karakter kosong**

```
In [7]: text = text.strip()
        print("\nKarakter Kosong hilang:\n----------------\n",text)
```

```
Karakter Kosong hilang:
----------------
 sejak sebulan terakhir harga bawang putih dan bombay yang sempat
rp  per kg namun berkah harga murah yang dinikmati masyarakat kem
pemerhati pertanian syaiful bahari menjelaskan masalah kenaikan h
ombay dan gula selama ini lebih banyak disebabkan oleh kebijakan
ndiri untuk kasus bawang putih dan bombay lanjut syaiful ketika r
rp  per kilo gram menjadi rp  sampai rp  per kilo gram sehingga k
```

7. **Proses tokenizing**

```
In [8]: tokens = word_tokenize(text)
        print("Tokenizing:\n----------------\n", tokens)
```

```
Tokenizing:
----------------
 ['sejak', 'sebulan', 'terakhir', 'harga', 'bawang', 'putih', 'dan'
a', 'kembali', 'turun', 'dan', 'stabil', 'di', 'ratarata', 'rp', 'p
ikmati', 'masyarakat', 'kembali', 'terusik', 'dengan', 'mulai', 'na
i', 'pertanian', 'syaiful', 'bahari', 'menjelaskan', 'masalah', 'ke
'impor', 'seperti', 'bawang', 'putih', 'bombay', 'dan', 'gula', 'se
jakan', 'restriksi', 'atau', 'pembatasan', 'yang', 'diberlakukan',
'putih', 'dan', 'bombay', 'lanjut', 'syaiful', 'ketika', 'relaksasi
'bombay', 'dari', 'rp', 'per', 'kilo', 'gram', 'menjadi', 'rp', 'sa
moditi', 'ini', 'menyumbang', 'deflasi']
```

8. Memasang library Sastrawi

```
In [16]: !{sys.executable} -m pip install Sastrawi
         from nltk.corpus import stopwords
         from nltk.probability import FreqDist
         from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory
         from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
         import matplotlib.pyplot as plt

         Requirement already satisfied: Sastrawi in c:\users\hang\anaconda3\lib\site-packag
```

9. Proses filter menggunakan Sastrawi

```
In [17]: factory = StopWordRemoverFactory()
         stopword = factory.create_stop_word_remover()
         text = stopword.remove(text)
         print("\nSetelah filtering:\n----------------\n", text)

         Setelah filtering:
         ----------------
          sejak bulan akhir harga bawang putih bombay sempat lonjak tinggi akhir
         at masyarakat usik mulai naik harga bawang putih pasar perhati tani syai
         bawang putih bombay gula lama lebih banyak bijak restriksi batas laku pe
         relaksasi laku bukti harga turun drastis bombay rp per kilo gram jadi rp
```

10. Proses Stemming Sastrawi

```
In [18]: factory = StemmerFactory()
         stemmer = factory.create_stemmer()
         text = stemmer.stem(text)
         print("\nOutput stemming:\n----------------\n", text)

         Output stemming:
         ----------------
          sejak bulan akhir harga bawang putih bombay sempat lonjak tinggi akhir
         at masyarakat usik mulai naik harga bawang putih pasar perhati tani sya
         bawang putih bombay gula lama lebih banyak bijak restriksi batas laku p
         relaksasi laku bukti harga turun drastis bombay rp per kilo gram jadi r
```

11. Frekuensi huruf yang muncul menggunakan Sastrawi

```
In [19]: tf = FreqDist(tokens)
         print("\nTerm Frequency:\n----------------\n", tf.most_common())

         Term Frequency:
         ----------------
          [('harga', 5), ('bawang', 4), ('putih', 4), ('dan', 4), ('bombay', 4
         urun', 2), ('di', 2), ('dengan', 2), ('syaiful', 2), ('komoditi', 2),
         2), ('gram', 2), ('sejak', 1), ('sebulan', 1), ('terakhir', 1), ('ser
         ('stabil', 1), ('ratarata', 1), ('kg', 1), ('namun', 1), ('berkah', 1
         rusik', 1), ('mulai', 1), ('naiknya', 1), ('pasar', 1), ('pemerhati',
         ('masalah', 1), ('kenaikan', 1), ('terkait', 1), ('impor', 1), ('sepe
         ak', 1), ('disebabkan', 1), ('kebijakan', 1), ('restriksi', 1), ('ata
         1), ('untuk', 1), ('kasus', 1), ('lanjut', 1), ('ketika', 1), ('relak
         ('menjadi', 1), ('sampai', 1), ('sehingga', 1), ('kedua', 1), ('menyu
```

12. Kata yang sering muncul menggunakan Sastrawi

```
In [20]: word, frequency= tf.most_common()[0]
         print("\nKeyword yang paling banyak muncul:\n----------------\n", wo

         Keyword yang paling banyak muncul:
         ----------------
          harga = 5
```

13. Keseluruhan kata yang muncul

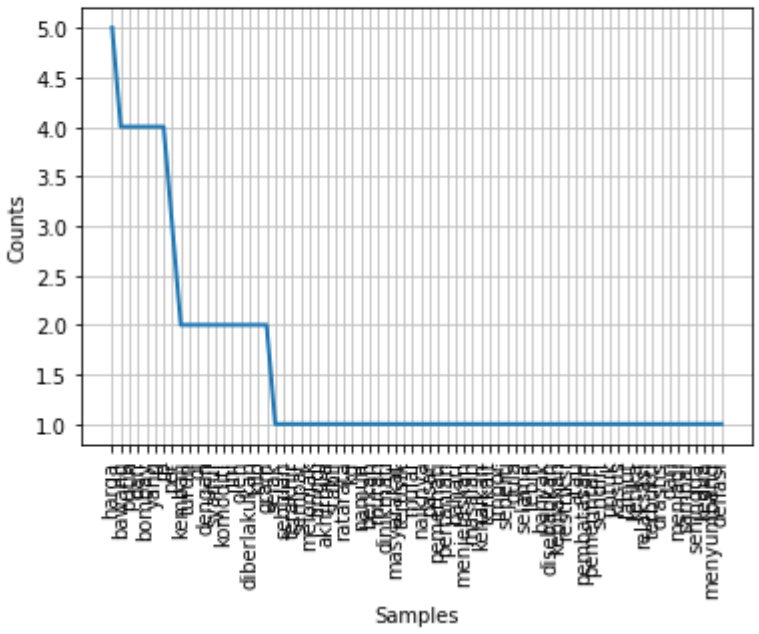
```
In [21]: print("\nKeseluruhan keywords:\n----------------\n")
         for word, frequency in tf.most_common():
             print(word, ":", frequency)

          Keseluruhan keywords:
          ----------------

          harga : 5
          bawang : 4
          putih : 4
          dan : 4
          bombay : 4
```

14. Plot data kata yang sering muncul

```
In [22]: tf.plot(cumulative=False)
         plt.show()
```



15. Memasang library Porter

```
In [23]: from nltk.stem import PorterStemmer
```

16. Proses filter menggunakan Porter

```
In [26]: listStopword = set(stopwords.words('indonesian'))
         tmpstr = []
         for t in tokens:
             if t not in listStopword:
                 tmpstr.append(t)
                 tokens=tmpstr
         print("\nSetelah filtering\n----------------\n", tokens)

         Setelah filtering
         ----------------
          ['sebulan', 'harga', 'bawang', 'putih', 'bombay', 'melonjak', 'tu
         urah', 'dinikmati', 'masyarakat', 'terusik', 'naiknya', 'harga', 'l
         l', 'bahari', 'kenaikan', 'harga', 'komoditi', 'terkait', 'impor',
         n', 'restriksi', 'pembatasan', 'diberlakukan', 'pemerintah', 'bawa
         n', 'terbukti', 'harga', 'turun', 'drastis', 'bombay', 'rp', 'kilo
         g', 'deflasi']
```

17. Proses Stemming menggunakan Porter

```
In [28]: tmpstr = []
         ps = PorterStemmer()
         for k in tokens:
             tmpstr.append(ps.stem(k))
         tokens=tmpstr
         print("\nOutput stemming:\n----------------\n", tokens)

         Output stemming:
         ----------------
          ['sebulan', 'harga', 'bawang', 'putih', 'bombay', 'melonjak', 'tu
         urah', 'dinikmati', 'masyarakat', 'terusik', 'naiknya', 'harga', '
         l', 'bahari', 'kenaikan', 'harga', 'komod', 'terkait', 'impor', 'b
         'restriksi', 'pembatasan', 'diberlakukan', 'pemerintah', 'bawang',
         'terbukti', 'harga', 'turun', 'drasti', 'bombay', 'rp', 'kilo', 'g
         lasi']
```

18. Frekuensi kata yang muncul menggunakan Porter

```
In [29]: tf = FreqDist(tokens)
         print("\nTerm Frequency:\n----------------\n", tf.most_common())

         Term Frequency:
         ----------------
          [('harga', 5), ('bawang', 4), ('putih', 4), ('bombay', 4), ('rp'
         an', 2), ('kilo', 2), ('gram', 2), ('sebulan', 1), ('melonjak', 1
         ('murah', 1), ('dinikmati', 1), ('masyarakat', 1), ('terusik', 1)
         n', 1), ('bahari', 1), ('kenaikan', 1), ('terkait', 1), ('impor',
         iksi', 1), ('pembatasan', 1), ('pemerintah', 1), ('relaksasi', 1)
         i', 1)]
```

19. Kata yang sering muncul menggunakan Porter

```
In [30]: word, frequency=tf.most_common()[0]
         print("\nKeyword yang paling banyak muncul:\n----------------\n", word, "=", frequency , "\n")
```

```
Keyword yang paling banyak muncul:
----------------
 harga = 5
```

20. Keseluruhan kata yang muncul

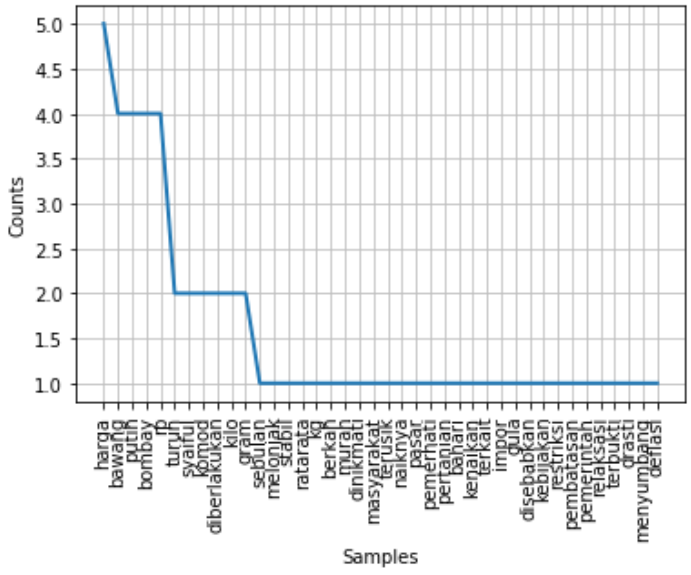
```
In [31]: print("\nKeseluruhan keywords:\n-----------------\n")
         for word, frequency in tf.most_common():
             print(word, ":", frequency)
```

```
Keseluruhan keywords:
-----------------

harga : 5
bawang : 4
putih : 4
bombay : 4
rp : 4
turun : 2
syaiful : 2
komod : 2
```

21. Plot data yang sering muncul

```
In [32]: tf.plot(cumulative=False)
         plt.show()
```



Latihan 2

1. Diberikan 2 bahan berita (news2 dan news3) silahkan lakukan olah teks mining seperti diatas (Silahkan pilih salah satu bahan berita)

Lampiran Screenshot hasil

Input screenshot disini