



Minggu ke-14

# Praktikum Text Mining

Ali Ridho Barakbah

Knowledge Engineering Research Group  
Department of Information and Computer Engineering  
Politeknik Elektronika Negeri Surabaya

# Eksperimen dengan Data News

---

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

# Membaca File

```
f = open("news.txt", "r")
text=f.read()
f.close()

print("\nText:\n-----\n", text)
```

Text:

-----

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

# Preprocessing

```
import re
import string

f = open("news.txt", "r")
text=f.read()
f.close()

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

text = text.lower()
print("\nHuruf kecil semua:\n-----\n",text)

text = re.sub(r"\d+", "", text)
print("\nAngka hilang:\n-----\n",text)

text = text.translate(str.maketrans("", "", string.punctuation))
print("\nTanda Baca hilang:\n-----\n",text)

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

Huruf kecil semua:

-----  
sejak sebulan terakhir harga bawang putih dan bombay yang sempat melonjak tinggi akhirnya kembali turun dan stabil di rata-rata rp. 20.000 per kg. namun, berkah harga murah yang dinikmati masyarakat kembali terusik dengan mulai naiknya harga bawang putih di pasar. pemerhati pertanian, syaiful bahari, menjelaskan masalah kenaikan harga komoditi yang terkait dengan impor seperti bawang putih, bombay dan gula, selama ini lebih banyak disebabkan oleh kebijakan restriksi atau pembatasan yang diberlakukan oleh pemerintah sendiri. untuk kasus bawang putih dan bombay, lanjut syaiful, ketika relaksasi diberlakukan terbukti harga turun drastis. bombay dari rp 150.000 per kilo gram menjadi rp 17.000 sampai rp. 20.000 per kilo gram. sehingga kedua komoditi ini menyumbang deflasi.

Angka hilang:

-----  
sejak sebulan terakhir harga bawang putih dan bombay yang sempat melonjak tinggi akhirnya kembali turun dan stabil di rata-rata rp. . per kg. namun, berkah harga murah yang dinikmati masyarakat kembali terusik dengan mulai naiknya harga bawang putih di pasar. pemerhati pertanian, syaiful bahari, menjelaskan masalah kenaikan harga komoditi yang terkait dengan impor seperti bawang putih, bombay dan gula, selama ini lebih banyak disebabkan oleh kebijakan restriksi atau pembatasan yang diberlakukan oleh pemerintah sendiri. untuk kasus bawang putih dan bombay, lanjut syaiful, ketika relaksasi diberlakukan terbukti harga turun drastis. bombay dari rp . per kilo gram menjadi rp . sampai rp. . per kilo gram. sehingga kedua komoditi ini menyumbang deflasi.

Tanda Baca hilang:

-----  
sejak sebulan terakhir harga bawang putih dan bombay yang sempat melonjak tinggi akhirnya kembali turun dan stabil di ratarata rp per kg namun berkah harga murah yang dinikmati masyarakat kembali terusik dengan mulai naiknya harga bawang putih di pasar pemerhati pertanian syaiful bahari menjelaskan masalah kenaikan harga komoditi yang terkait dengan impor seperti bawang putih bombay dan gula selama ini lebih banyak disebabkan oleh kebijakan restriksi atau pembatasan yang diberlakukan oleh pemerintah sendiri untuk kasus bawang putih dan bombay lanjut syaiful ketika relaksasi diberlakukan terbukti harga turun drastis bombay dari rp per kilo gram menjadi rp sampai rp per kilo gram sehingga kedua komoditi ini menyumbang deflasi

Karakter Kosong hilang:

-----  
sejak sebulan terakhir harga bawang putih dan bombay yang sempat melonjak tinggi akhirnya kembali turun dan stabil di ratarata rp per kg namun berkah harga murah yang dinikmati masyarakat kembali terusik dengan mulai naiknya harga bawang putih di pasar pemerhati pertanian syaiful bahari menjelaskan masalah kenaikan harga komoditi yang terkait dengan impor seperti bawang putih bombay dan gula selama ini lebih banyak disebabkan oleh kebijakan restriksi atau pembatasan yang diberlakukan oleh pemerintah sendiri untuk kasus bawang putih dan bombay lanjut syaiful ketika relaksasi diberlakukan terbukti harga turun drastis bombay dari rp per kilo gram menjadi rp sampai rp per kilo gram sehingga kedua komoditi ini menyumbang deflasi

# Tokenizing

```
import re
import string
from nltk.tokenize import word_tokenize

f = open("news.txt", "r")
text=f.read()
f.close()

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

text = text.lower()
print("\nHuruf kecil semua:\n-----\n",text)

text = re.sub(r"\d+", "", text)
print("\nAngka hilang:\n-----\n",text)

text = text.translate(str.maketrans("", "", string.punctuation))
print("\nTanda Baca hilang:\n-----\n",text)

text = text.strip()
print("\nKarakter Kosong hilang:\n-----\n",text)

tokens = word_tokenize(text)
print("\nTokenizing:\n-----\n", tokens)
```

Tokenizing:

```
['sejak', 'sebulan', 'terakhir', 'harga', 'bawang', 'putih', 'dan',
'bombay', 'yang', 'sempat', 'melonjak', 'tinggi', 'akhirnya', 'kembali',
'turun', 'dan', 'stabil', 'di', 'ratarata', 'rp', 'per', 'kg', 'namun',
'berkah', 'harga', 'murah', 'yang', 'dinikmati', 'masyarakat', 'kembali',
'terusik', 'dengan', 'mulai', 'naiknya', 'harga', 'bawang', 'putih', 'di',
'pasar', 'pemerhati', 'pertanian', 'syaiful', 'bahari', 'menjelaskan',
'masalah', 'kenaikan', 'harga', 'komoditi', 'yang', 'terkait', 'dengan',
'impor', 'seperti', 'bawang', 'putih', 'bombay', 'dan', 'gula', 'selama',
'ini', 'lebih', 'banyak', 'disebabkan', 'oleh', 'kebijakan', 'restriksi',
'atau', 'pembatasan', 'yang', 'diberlakukan', 'oleh', 'pemerintah',
'sendiri', 'untuk', 'kasus', 'bawang', 'putih', 'dan', 'bombay', 'lanjut',
'syaiful', 'ketika', 'relaksasi', 'diberlakukan', 'terbukti', 'harga',
'turun', 'drastis', 'bombay', 'dari', 'rp', 'per', 'kilo', 'gram',
'menjadi', 'rp', 'sampai', 'rp', 'per', 'kilo', 'gram', 'sehingga',
'kedua', 'komoditi', 'ini', 'menyumbang', 'deflasi']
```

NLTK (<http://www.nltk.org/>):

Buka comment prompt dan tuliskan:  
python

```
>>> import nltk
>>> nltk.download()
```

# Filtering – dengan Library Sastrawi

```
import re
import string
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory

f = open("news.txt", "r")
text=f.read()
f.close()

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

text = text.lower()
print("\nHuruf kecil semua:\n-----\n",text)

text = re.sub(r"\d+", "", text)
print("\nAngka hilang:\n-----\n",text)

text = text.translate(str.maketrans("", "", string.punctuation))
print("\nTanda Baca hilang:\n-----\n",text)

text = text.strip()
print("\nKarakter Kosong hilang:\n-----\n",text)

tokens = word_tokenize(text)
print("\nTokenizing:\n-----\n", tokens)

# Filtering dengan Sastrawi -----
factory = StopWordRemoverFactory()
stopword = factory.create_stop_word_remover()
text = stopword.remove(text)
print("\nSetelah filtering:\n-----\n", text)
```

Setelah filtering:

-----  
sejak sebulan terakhir harga bawang putih bombay sempat melonjak tinggi akhirnya turun stabil rata-rata Rp per kg berkah harga murah dinikmati masyarakat terusik mulai naiknya harga bawang putih pasar pemerhati pertanian Syaiful Bahari menjelaskan masalah kenaikan harga komoditi terkait impor bawang putih bombay gula selama lebih banyak disebabkan kebijakan restriksi pembatasan diberlakukan pemerintah sendiri kasus bawang putih bombay lanjut Syaiful Relaksasi diberlakukan terbukti harga turun drastis bombay Rp per kilo gram menjadi Rp Rp per kilo gram kedua komoditi menyumbang deflasi

Sastrawi  
(<https://pypi.org/project/Sastrawi/>):

- pip install Sastrawi

# Stemming – dengan Library Sastrawi

```
import re
import string
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory

f = open("news.txt", "r")
text=f.read()
f.close()

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

text = text.lower()
print("\nHuruf kecil semua:\n-----\n",text)

text = re.sub(r"\d+", "", text)
print("\nAngka hilang:\n-----\n",text)

text = text.translate(str.maketrans("", "", string.punctuation))
print("\nTanda Baca hilang:\n-----\n",text)

text = text.strip()
print("\nKarakter Kosong hilang:\n-----\n",text)

# Filtering dengan Sastrawi -----
factory = StopWordRemoverFactory()
stopword = factory.create_stop_word_remover()
text = stopword.remove(text)
print("\nSetelah filtering:\n-----\n", text)

# Stemming dengan Sastrawi -----
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  
turun stabil rata-rata rp per kg berkah harga murah nikmat masyarakat usik  
mulai naik harga bawang putih pasar perhati tani syaiful bahari jelas  
masalah naik harga komoditi kait impor bawang putih bombay gula lama lebih  
banyak sebab bijak restriksi batas laku perintah sendiri kasus bawang putih  
bombay lanjut syaiful relaksasi laku bukti harga turun drastis bombay rp  
per kilo gram jadi rp rp per kilo gram dua komoditi sumbang deflasi



# Analyzing

```
import nltk
from nltk.tokenize import word_tokenize
from nltk.probability import FreqDist
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
import re
import string
import matplotlib.pyplot as plt
from nltk.corpus import stopwords
from Sastrawi.StopWordRemover.StopWordRemoverFactory import
StopWordRemoverFactory

f = open("news.txt", "r")
text=f.read()

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

text = text.lower()
print("\nHuruf kecil semua:\n-----\n",text)

text = re.sub(r"\d+", "", text)
print("\nAngka hilang:\n-----\n",text)

text = text.translate(str.maketrans("", "",string.punctuation))
print("\nTanda Baca hilang:\n-----\n",text)

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

```
# Filtering dengan Sastrawi -----
factory = StopWordRemoverFactory()
stopword = factory.create_stop_word_remover()
text = stopword.remove(text)
print("\nSetelah filtering:\n-----\n", text)

# Stemming dengan Sastrawi -----
factory = StemmerFactory()
stemmer = factory.create_stemmer()
text = stemmer.stem(text)
print("\nOutput stemming:\n-----\n", text)

tokens = word_tokenize(text)
print("\nTokenizing:\n-----\n", tokens)

tf = FreqDist(tokens)
print("\nTerm Frequency:\n-----\n", tf.most_common())

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

print("\nKeseluruhan keywords:\n-----\n")
for word, frequency in tf.most_common():
    print(word, ":", frequency)

tf.plot(cumulative=False)
plt.show()
```



#### Term Frequency:

```
[('harga', 5), ('bawang', 4), ('putih', 4), ('bombay', 4), ('rp', 4),
('per', 3), ('akhir', 2), ('turun', 2), ('naik', 2), ('syaiful', 2),
('komoditi', 2), ('laku', 2), ('kilo', 2), ('gram', 2), ('sejak', 1),
('bulan', 1), ('sempat', 1), ('lonjak', 1), ('tinggi', 1), ('stabil', 1),
('ratarata', 1), ('kg', 1), ('berkah', 1), ('murah', 1), ('nikmat', 1),
('masyarakat', 1), ('usik', 1), ('mulai', 1), ('pasar', 1), ('perhati', 1),
('tani', 1), ('bahari', 1), ('jelas', 1), ('masalah', 1), ('kait', 1),
('impor', 1), ('gula', 1), ('lama', 1), ('lebih', 1), ('banyak', 1),
('sebab', 1), ('bijak', 1), ('restriksi', 1), ('batas', 1), ('perintah',
1), ('sendiri', 1), ('kasus', 1), ('lanjut', 1), ('relaksasi', 1),
('bukti', 1), ('drastis', 1), ('jadi', 1), ('dua', 1), ('sumbang', 1),
('deflasi', 1)]
```

#### Keyword yang paling banyak muncul:

```
-----
harga = 5
```

#### Keseluruhan keywords:

```
-----
harga : 5
bawang : 4
putih : 4
bombay : 4
rp : 4
per : 3
akhir : 2
turun : 2
naik : 2
syaiful : 2
komoditi : 2
laku : 2
kilo : 2
gram : 2
sejak : 1
bulan : 1
sempat : 1
lonjak : 1
tinggi : 1
stabil : 1
ratarata : 1
kg : 1
berkah : 1
murah : 1
nikmat : 1
masyarakat : 1
usik : 1
mulai : 1
```

# Text Mining – (dengan stemming memakai Porter)

```
import nltk
from nltk.tokenize import word_tokenize
from nltk.probability import FreqDist
from nltk.stem import PorterStemmer
import re
import string
import matplotlib.pyplot as plt
from nltk.corpus import stopwords

f = open("news.txt", "r")
text=f.read()

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

text = text.lower()
print("\nHuruf kecil semua:\n-----\n",text)

text = re.sub(r"\d+", "", text)
print("\nAngka hilang:\n-----\n",text)

text = text.translate(str.maketrans("", "", string.punctuation))
print("\nTanda Baca hilang:\n-----\n",text)

text = text.strip()
print("\nKarakter Kosong hilang:\n-----\n",text)

tokens = word_tokenize(text)
print("\nTokenizing:\n-----\n", tokens)
```

```
# Filtering dengan Porter -----
listStopword = set(stopwords.words('indonesian'))
tmpstr = []
for t in tokens:
    if t not in listStopword:
        tmpstr.append(t)
tokens=tmpstr
print("\nSetelah filtering --> ", tokens)

# Stemming dengan Porter -----
tmpstr = []
ps = PorterStemmer()
for k in tokens:
    tmpstr.append(ps.stem(k))
tokens=tmpstr
print("\nOutput stemming:\n", tokens)

tf = FreqDist(tokens)
print("\nTerm Frequency:\n-----\n", tf.most_common())

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

print("\nKeseluruhan keywords:\n-----\n")
for word, frequency in tf.most_common():
    print(word, ":", frequency)

tf.plot(cumulative=False)
plt.show()
```

Output stemming:

```
['sebulan', 'harga', 'bawang', 'putih', 'bombay', 'melonjak', 'turun',  
'stabil', 'ratarata', 'rp', 'kg', 'berkah', 'harga', 'murah', 'dinikmati',  
'masyarakat', 'terusik', 'naiknya', 'harga', 'bawang', 'putih', 'pasar',  
'pemerhati', 'pertanian', 'syaiful', 'bahari', 'kenaikan', 'harga',  
'komod', 'terkait', 'impor', 'bawang', 'putih', 'bombay', 'gula',  
'disebabkan', 'kebijakan', 'restriksi', 'pembatasan', 'diberlakukan',  
'pemerintah', 'bawang', 'putih', 'bombay', 'syaiful', 'relaksasi',  
'diberlakukan', 'terbukti', 'harga', 'turun', 'drasti', 'bombay', 'rp',  
'kilo', 'gram', 'rp', 'rp', 'kilo', 'gram', 'komod', 'menyumbang',  
'deflasi']
```

Term Frequency:

```
[('harga', 5), ('bawang', 4), ('putih', 4), ('bombay', 4), ('rp', 4),  
( 'turun', 2), ('syaiful', 2), ('komod', 2), ('diberlakukan', 2), ('kilo',  
2), ('gram', 2), ('sebulan', 1), ('melonjak', 1), ('stabil', 1),  
( 'ratarata', 1), ('kg', 1), ('berkah', 1), ('murah', 1), ('dinikmati', 1),  
( 'masyarakat', 1), ('terusik', 1), ('naiknya', 1), ('pasar', 1),  
( 'pemerhati', 1), ('pertanian', 1), ('bahari', 1), ('kenaikan', 1),  
( 'terkait', 1), ('impor', 1), ('gula', 1), ('disebabkan', 1), ('kebijakan',  
1), ('restriksi', 1), ('pembatasan', 1), ('pemerintah', 1), ('relaksasi',  
1), ('terbukti', 1), ('drasti', 1), ('menyumbang', 1), ('deflasi', 1)]
```

Keyword yang paling banyak muncul:

-----  
harga = 5

Keseluruhan keywords:

-----  
harga : 5  
bawang : 4  
putih : 4  
bombay : 4  
rp : 4  
turun : 2  
syaiful : 2  
komod : 2  
diberlakukan : 2  
kilo : 2  
gram : 2  
sebulan : 1  
melonjak : 1  
stabil : 1  
ratarata : 1  
kg : 1  
berkah : 1  
murah : 1  
dinikmati : 1  
masyarakat : 1