Nama : Oktaveian Aliansyah

NIM : 220411100099

Lowercase

```
kalimat = "Akhir pekan kemarin Perdana Menteri Israel Benjamin Netanyahu menunda pembebasan 602 tahanan Palestina, karena menuduh Hamas

lower_case = kalimat.lower()
print(lower_case)
```

```
akhir pekan kemarin perdana menteri israel benjamin netanyahu menunda pembebasan 602 tahanan palestina, karena menuduh hamas melakuk
```

dari pergantian data diatas didapati perubahan mengganti huruf kapital menjadi lower case.

Remove number

```
import re

kalimat = "Akhir pekan kemarin Perdana Menteri Israel Benjamin Netanyahu menunda pembebasan 602 tahanan Palestina, karena menuduh Hamas
hasil = re.sub(r"\d+", "", kalimat)
print(hasil)
```

```
Akhir pekan kemarin Perdana Menteri Israel Benjamin Netanyahu menunda pembebasan  tahanan Palestina, karena menuduh Hamas melakukan
```

hasil yang di dapati dari code diatas adalah penghapusan angka yang ada didalam kalimat

Removing white space

```
kalimat = " \t ini kalimat contoh\t "
hasil = kalimat.strip()
print(hasil)
```

```
ini kalimat contoh
```

Separating Sentences with Split () Method

```
kalimat = "Akhir pekan kemarin Perdana Menteri Israel Benjamin Netanyahu menunda pembebasan 602 tahanan Palestina, karena menuduh Hamas
pisah = kalimat.split()
print(pisah)
```

```
, 'Hamas', 'melakukan', 'pelanggaran', 'gencatan', 'senjata.', 'Hamas', 'Ogah', 'Lanjut', 'Gencatan', 'Senjata', 'sampai', 'Israel',
```

hasil output dari code diatas adalah memisah kalimat menjadi setiap kalimat

Tokenizing: Word Tokenizing Using NLTK Module

```
import nltk
from nltk.tokenize import word_tokenize

nltk.download('punkt_tab')
nltk.download('punkt')

kalimat = "Akhir pekan kemarin Perdana Menteri Israel Benjamin Netanyahu menunda pembebasan 602 tahanan Palestina, karena menuduh Hamas

tokens = nltk.tokenize.word_tokenize(kalimat)
print(tokens)
```

```
', 'Hamas', 'melakukan', 'pelanggaran', 'gencatan', 'senjata', '.', 'Hamas', 'Ogah', 'Lanjut', 'Gencatan', 'Senjata', 'sampai', 'Isr
```

didapati hasil seperti split, namun ada perbedaan seperti stopword juga dipisah

Tokenizing with Case Folding

```
from nltk.tokenize import word_tokenize
import string

kalimat = "Akhir pekan kemarin Perdana Menteri Israel Benjamin Netanyahu menunda pembebasan 602 tahanan Palestina, karena menuduh Hamas
kalimat = kalimat.translate(str.maketrans('','',string.punctuation)).lower()

tokens = nltk.tokenize.word_tokenize(kalimat)
print(tokens)
```

```
['akhir', 'pekan', 'kemarin', 'perdana', 'menteri', 'israel', 'benjamin', 'netanyahu', 'menunda', 'pembebasan', '602', 'tahanan', 'p
```

output yang didapatkan berupa gabungan antara tokenize dan case folding

Frequency Distribution

```
ıan Palestina, karena menuduh Hamas melakukan pelanggaran gencatan senjata. Hamas Ogah Lanjut Gencatan Senjata sampai Israel Bebaskan 602
```

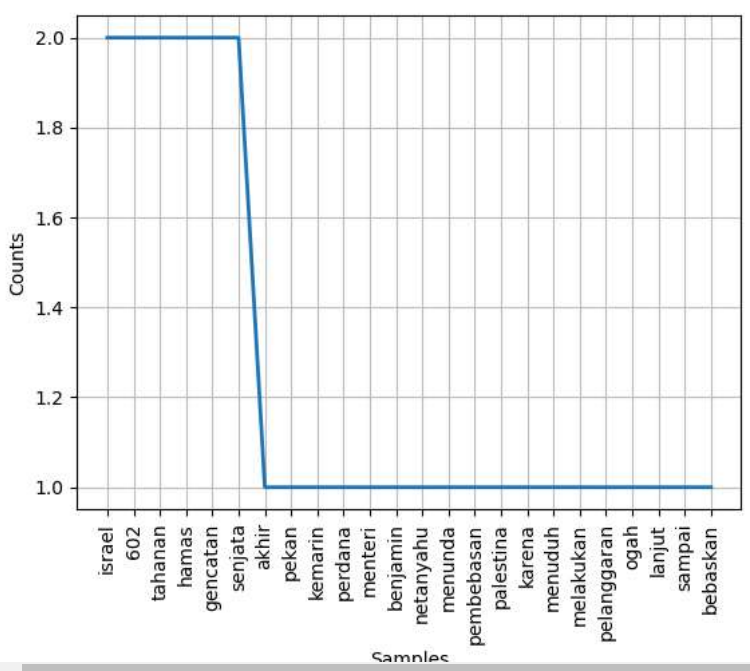
```
[('israel', 2), ('602', 2), ('tahanan', 2), ('hamas', 2), ('gencatan', 2), ('senjata', 2), ('akhir', 1), ('pekan', 1), ('kemarin', 1
```

didapati frekuensi kemunculan dari setiap kata

Frequency Distribution Visualization with Matplotlib

```
import matplotlib.pyplot as plt

kemunculan.plot(30,cumulative=False)
plt.show()
```



menampilkan visualisasi data menggunakan library matplotlib dari frekuensi kata yang di tokenize

Tokenizing: Sentences Tokenizing Using NLTK Module

```
from nltk.tokenize import sent_tokenize

kalimat = "Akhir pekan kemarin Perdana Menteri Israel Benjamin Netanyahu menunda pembebasan 602 tahanan Palestina, karena menuduh Hamas

tokens = nltk.tokenize.sent_tokenize(kalimat)
print(tokens)
```

```
['Akhir pekan kemarin Perdana Menteri Israel Benjamin Netanyahu menunda pembebasan 602 tahanan Palestina, karena menuduh Hamas melak
```

hasilnya berupa tokenize per-kalimat yang dipisah dengan tanda baca titik

Filtering using NLTK

```
ıan Palestina, karena menuduh Hamas melakukan pelanggaran gencatan senjata. Hamas Ogah Lanjut Gencatan Senjata sampai Israel Bebaskan 602
```

```
['pekan', 'kemarin', 'perdana', 'menteri', 'israel', 'benjamin', 'netanyahu', 'menunda', 'pembebasan', '602', 'tahanan', 'palestina
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

hasil berupa tokenize dan penghapusan stopword

Filtering using Sastrawi: Stopword List

```
!pip install sastrawi
```

```
Collecting sastrawi
  Downloading Sastrawi-1.0.1-py2.py3-none-any.whl.metadata (909 bytes)
Downloading Sastrawi-1.0.1-py2.py3-none-any.whl (209 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 209.7/209.7 kB 8.0 MB/s eta 0:00:00
Installing collected packages: sastrawi
Successfully installed sastrawi-1.0.1
```

```
from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory

factory = StopWordRemoverFactory()
stopwords = factory.get_stop_words()
print(stopwords)
```

```
['yang', 'untuk', 'pada', 'ke', 'para', 'namun', 'menurut', 'antara', 'dia', 'dua', 'ia', 'seperti', 'jika', 'jika', 'sehingga', 'ke
```

melihat daftar stopword yang ada di library sastrawi

Filtering using Sastrawi

```
from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory
from nltk.tokenize import word_tokenize

factory = StopWordRemoverFactory()
stopword = factory.create_stop_word_remover()

kalimat = "Akhir pekan kemarin Perdana Menteri Israel Benjamin Netanyahu menunda pembebasan 602 tahanan Palestina, karena menuduh Hamas
kalimat = kalimat.translate(str.maketrans('','',string.punctuation)).lower()

stop = stopword.remove(kalimat)
tokens = nltk.tokenize.word_tokenize(stop)

print(tokens)
```

```
['akhir', 'pekan', 'kemarin', 'perdana', 'menteri', 'israel', 'benjamin', 'netanyahu', 'menunda', 'pembebasan', '602', 'tahanan', 'p
```

hasil berupa penghapusan stopword di dalam kalimat berita

Add Custom Stopword

```
from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory, StopWordRemover, ArrayDictionary
from nltk.tokenize import word_tokenize

# ambil stopword bawaan
stop_factory = StopWordRemoverFactory().get_stop_words()
more_stopword = ['tahanan', 'senjata']

kalimat = "Akhir pekan kemarin Perdana Menteri Israel Benjamin Netanyahu menunda pembebasan 602 tahanan Palestina, karena menuduh Hamas
kalimat = kalimat.translate(str.maketrans('','',string.punctuation)).lower()

# menggabungkan stopword
data = stop_factory + more_stopword

dictionary = ArrayDictionary(data)
str = StopWordRemover(dictionary)
tokens = nltk.tokenize.word_tokenize(str.remove(kalimat))

print(tokens)
```

```
, 'palestina', 'menuduh', 'hamas', 'melakukan', 'pelanggaran', 'gencatan', 'hamas', 'ogah', 'lanjut', 'gencatan', 'sampai', 'israel'
```

hasil berupa penghapusan kata 'tanahan' dan 'senjata'

Stemming : Porter Stemming Algorithm using NLTK

```
from nltk.stem import PorterStemmer
ps = PorterStemmer()

kata = ["program", "programs", "programer", "programing", "programers"]

for k in kata:
    print(k, " : ", ps.stem(k))
```

```
striker  :  striker
programs  :  program
programer  :  program
programing  :  program
programers  :  program
```

Stemming Bahasa Indonesia using Sastrawi

```
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
factory = StemmerFactory()
stemmer = factory.create_stemmer()

kalimat = "Akhir pekan kemarin Perdana Menteri Israel Benjamin Netanyahu menunda pembebasan 602 tahanan Palestina, karena menuduh Hamas m

hasil = stemmer.stem(kalimat)
print(hasil)
```

```
akhir pekan kemarin perdana menteri israel benjamin netanyahu tunda bebas 602 tahan palestina karena tuduh hamas laku langgar genca
```