Nama: Oktaveian Aliansyah

NIM: 220411100099

Kelas: NLP(A)

library yang dibutuhkan

```
import re
import string
import pymupdf4llm
import fitz
import nltk
from nltk.tokenize import sent_tokenize, word_tokenize
from nltk.corpus import stopwords
import matplotlib.pyplot as plt
from nltk.probability import FreqDist
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
```

menggabungkan file pdf

```
def gabungkan_pdf(file_list, output_pdf):
    doc = fitz.open()
    for pdf in file_list:
        with fitz.open(pdf) as temp_doc:
            doc.insert_pdf(temp_doc)  # Menambahkan halaman dari file lain
    doc.save(output_pdf)  # Menyimpan hasil gabungan
    doc.close()

# Contoh penggunaan
daftar_pdf = ["../Doc/Doc1-22_099.pdf", "../Doc/Doc2-22_099.pdf"]
gabungkan_pdf(daftar_pdf, "../Doc/Doc3-22_099.pdf")
```

menampilan file pdf yang sudah digabungkan

```
md_text = pymupdf4llm.to_markdown("../Doc/Doc3-22_099.pdf")
print(md_text)
```

```
Processing ../Doc/Doc3-22_099.pdf...
[                                        ] (0/2==================[====================                    ] (1/2==================
Mata kuliah Data Mining selalu menjadi favoritku setiap Senin pagi. Sejak pertama kali
mengikuti kelas ini, aku langsung tertarik dengan cara data dapat diolah untuk
menemukan pola tersembunyi. Dosen yang mengajar sangat interaktif dan selalu
membawakan contoh nyata yang membuat materi lebih mudah dipahami.

Salah satu hal yang paling aku sukai dari kuliah ini adalah saat kami membahas
bagaimana algoritma Page Rank bekerja dalam menentukan urutan hasil pencarian di
Google. Dosen menjelaskan dengan sederhana, kemudian kami langsung mencoba
mengimplementasikannya menggunakan Python dan OpenMPI untuk pemrosesan
paralel. Aku semakin kagum melihat bagaimana data yang awalnya tampak acak bisa
diolah menjadi informasi yang berguna.

Selain itu, diskusi di kelas selalu seru. Kami sering diberikan studi kasus, misalnya
bagaimana Netflix merekomendasikan film atau bagaimana e-commerce memprediksi
barang yang akan dibeli pelanggan. Saat sesi praktikum, aku merasa seperti seorang
detektif yang sedang mencari pola di dalam lautan data.


-----

mat Islam di Indonesia akan menyambut bulan suci Ramadhan 1446 Hijriah pada akhir
pekan ini. Bulan penuh berkah ini merupakan momen penting bagi umat Islam untuk
menjalankan ibadah puasa, sebagaimana diperintahkan dalam Al-Qur'an, Surah AlBaqarah ayat 183.

Ramadhan dimulai setelah berakhirnya bulan Syakban. Di Indonesia, penetapan awal
Ramadan dilakukan melalui sidang isbat yang digelar oleh Kementerian Agama
(Kemenag). Sidang isbat ini mengacu pada Fatwa Majelis Ulama Indonesia (MUI) Nomor
2 Tahun 2004 tentang Penetapan Awal Ramadhan, Syawal, dan Dzulhijjah.
```

```
Selain pemerintah, organisasi Islam seperti Nahdlatul Ulama (NU) dan Muhammadiyah
juga menetapkan awal Ramadan berdasarkan metode masing-masing. Muhammadiyah
menggunakan metode hisab hakiki wujudul hilal, sementara NU dan pemerintah
menggunakan metode rukyatul hilal dan kriteria MABIMS (Menteri Agama Brunei,
Indonesia, Malaysia, dan Singapura).


-----
```

tokenisasi kalimat

```
token_kalimat = nltk.tokenize.sent_tokenize(md_text)
print(token_kalimat)
```

```
['Mata kuliah Data Mining selalu menjadi favoritku setiap Senin pagi.', 'Sejak pertama kali\nmengikuti kelas ini, aku langsung tertа
```

menghapus angka, tanda baca dan stopword

```
def text_prep(file):
    file = re.sub(r"\d", "", file)
    file = file.translate(str.maketrans('','',string.punctuation)).lower()
    tokens = word_tokenize(file)
    listStopword =  set(stopwords.words('indonesian'))
    file = ' '.join(kata for kata in tokens if kata not in listStopword)
    return file

file = md_text
tp = text_prep(file)
print(tp)
```

```
mata kuliah data mining favoritku senin pagi kali mengikuti kelas langsung tertarik data diolah menemukan pola tersembunyi dosen mer
```

mencari frekuensi kemunculan setiap kata dan memvisualisasikan frekuensi

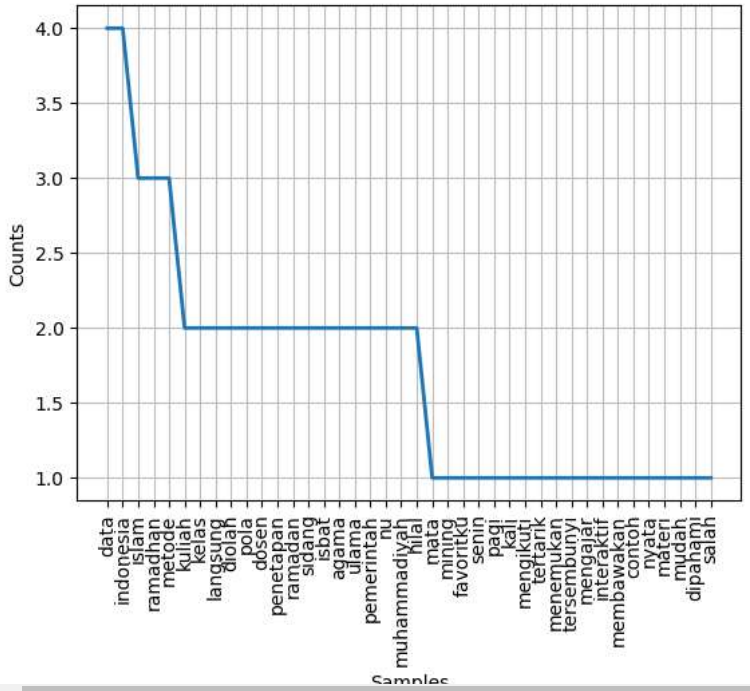
mencari frekuensi kata

```
# lakukan tokenize dulu
tokenize = nltk.tokenize.word_tokenize(tp)
# lalu cari frekuensi
frekuensi = nltk.FreqDist(tokenize)
print(frekuensi.most_common())
```

```
[('data', 4), ('indonesia', 4), ('islam', 3), ('ramadhan', 3), ('metode', 3), ('kuliah', 2), ('kelas', 2), ('langsung', 2), ('diolah
```

memvisualisasikan frekuensi

```
frekuensi.plot(40,cumulative=False)
plt.show()
```



stemming kata menggunakan sastrawi

```
stem = StemmerFactory()
stemer = stem.create_stemmer()

stemming = stemer.stem(tp)
print(stemming)
```

```
mata kuliah data mining favorit senin pagi kali ikut kelas langsung tarik data olah temu pola sembunyi dosen ajar interaktif bawa co
```