

## ✓ Responsi 2 Praktikum Natural Language Processing

Nama : Gavrilla Claudia

NIM : 21110004

Kelas : S1SD02A

```
import pandas as pd
import numpy as np
import warnings
warnings.filterwarnings("ignore")
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.svm import LinearSVC
from string import punctuation
import re
import nltk
nltk.download('punkt')
nltk.download('wordnet')

[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
True
```

```
faq =pd.read_csv('/content/dataset_mentalhealth.csv')
faq
```

Question_ID	Questions	Jawaban	
0	1590140 Apa yang dimaksud dengan penyakit mental?	Penyakit mental adalah kondisi kesehatan yang ...	
1	2110618 Siapa yang terpengaruh oleh penyakit mental?	Diperkirakan bahwa penyakit mental mempengaru... 	
2	6361820 Apa penyebab penyakit mental?	Diperkirakan bahwa penyakit mental mempengaru... 	
3	9434130 Apa sajakah tanda-tanda peringatan penyakit me...	Gejala gangguan kesehatan mental bervariasi te... 	
4	7657263 Apakah penderita penyakit jiwa bisa sembuh?	Ketika penyembuhan dari penyakit mental, ident... 	
...	...	...	...
93	4373204 Bagaimana saya tahu kalau saya minum terlalu b...	Menyortir jika Anda minum terlalu banyak bisa ... 	
94	7807643 Jika ganja berbahaya, mengapa kita melegalkannya?	Asap ganja, misalnya, mengandung racun penye... 	

```
faq_quest = faq[['Question_ID', 'Questions']]
faq_anw = faq[['Question_ID', 'Jawaban']]
```

## ✓ Text Pre-processing

```
def to_lower(text):
    return text.lower()

def remove_number(text):
    output = ''.join(c for c in text if not c.isdigit())
    return output

def remove_punct(text):
    return "".join(c for c in text if c not in punctuation)

def to_strip(text):
    return " ".join([c for c in text.split() if len(c)>2])
```

```

def remove_char(text):
    text = re.sub(r'[^a-zA-Z\s]', '', text, re.I|re.A)
    return text

def remove_duplicate(text):
    text = re.sub("(.)\\1{2,}", "\\1", text)
    return text

import nltk
nltk.download('stopwords')

[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
True

import nltk
from nltk.corpus import stopwords
stopwords.words('indonesian')

def remove_stopwords(text):
    stop_words= stopwords.words('indonesian')

    return ' '.join(c for c in nltk.word_tokenize(text) if c not in stop_words)

from nltk.stem import WordNetLemmatizer

wordnet_lemma = WordNetLemmatizer()

def lemma(text):
    lemmatize_words = [wordnet_lemma.lemmatize(word) for sent in nltk.sent_tokenize(text) for word in nltk.word_tokenize(sent)]
    return ' '.join(lemmatize_words)

from sklearn.preprocessing import LabelEncoder
label = LabelEncoder()
faq['JawabanEncode'] = label.fit_transform(faq['Jawaban'])
faq

```

	Question_ID	Questions	Jawaban	JawabanEncode	
0	1590140	Apa yang dimaksud dengan penyakit mental?	Penyakit mental adalah kondisi kesehatan yang ...	72	
1	2110618	Siapa yang terpengaruh oleh penyakit mental?	Diperkirakan bahwa penyakit mental mempengaru... ...	23	
2	6361820	Apa penyebab penyakit mental?	Diperkirakan bahwa penyakit mental mempengaru... ...	24	
3	9434130	Apa sajakah tanda-tanda peringatan penyakit me...	Gejala gangguan kesehatan mental bervariasi te...	33	
4	7657263	Apakah penderita penyakit jiwa bisa sembuh?	Ketika penyembuhan dari penyakit mental, ident...	50	
...	...	...	...	...	
93	4373204	Bagaimana saya tahu kalau saya minum terlalu b...	Menyortir jika Anda minum terlalu banyak bisa ...	61	

```

faq_quest['prep1']= faq_quest['Questions'].apply(to_lower)
faq_quest['prep2']= faq_quest['prep1'].apply(remove_number)
faq_quest['prep3']= faq_quest['prep2'].apply(remove_punct)
faq_quest['prep4']= faq_quest['prep3'].apply(to_strip)
faq_quest['prep5']= faq_quest['prep4'].apply(remove_char)
faq_quest['prep6']= faq_quest['prep5'].apply(remove_duplicate)
faq_quest['prep7']= faq_quest['prep6'].apply(remove_stopwords)
faq_quest['lemma']= faq_quest['prep7'].apply(lemma)
faq_quest.head(10)

```

Question_ID	Questions	prep1	prep2	prep3	prep4	prep5	prep6	prep7
0	1590140	Apa yang dimaksud dengan penyakit mental?	apa yang dimaksud dengan penyakit mental?	apa yang dimaksud dengan penyakit mental?	apa dimaksud dengan penyakit mental?			
1	2110618	Siapa yang terpengaruh oleh penyakit mental?	siapa yang terpengaruh oleh penyakit mental?	siapa yang terpengaruh oleh penyakit mental?	siapa yang terpengaruh oleh penyakit mental?			
2	6361820	Apa penyebab penyakit mental?	apa penyebab penyakit mental?	apa penyebab penyakit mental?	apa penyebab penyakit mental?			
3	9434130	Apa sajakah tanda-tanda peringatan penyakit me...	apa sajakah tanda-tanda peringatan penyakit me...	apa sajakah tanda-tanda peringatan penyakit me...	apa sajakah tanda-tanda peringatan penyakit mental	apa sajakah tanda-tanda peringatan penyakit mental	apa sajakah tanda-tanda peringatan penyakit mental	apa sajakah tanda-tanda peringatan penyakit mental?
4	7657263	Apakah penderita penyakit jiwa bisa sembuh?	apakah penderita penyakit jiwa bisa sembuh?	apakah penderita penyakit jiwa bisa sembuh?	apakah penderita penyakit jiwa bisa sembuh?			
5	1619387	Apa yang harus saya lakukan jika saya mengenal...	apa yang harus saya lakukan jika saya mengenal...	apa yang harus saya lakukan jika saya mengenal...	apa yang harus saya lakukan jika saya mengenal...	apa yang harus saya lakukan jika saya mengenal...	apa yang harus saya lakukan jika saya mengenal...	apa yang harus saya lakukan jika saya mengenal...
6	1030153	Bagaimana saya bisa menemukan ahli kesehatan m...	bagaimana saya bisa menemukan ahli kesehatan m...	bagaimana saya bisa menemukan ahli kesehatan m...	bagaimana saya bisa menemukan ahli kesehatan m...			
7	8022026	Pilihan pengobatan apa yang tersedia?	pilihan pengobatan apa yang tersedia?	pilihan pengobatan apa yang tersedia?	pilihan pengobatan apa yang tersedia?			
8	1155199	Jika saya terlibat dalam pengobatan, apa yang ...	jika saya terlibat dalam pengobatan, apa yang ...	jika saya terlibat dalam pengobatan, apa yang ...	jika saya terlibat dalam pengobatan, apa yang ...	jika saya terlibat dalam pengobatan, apa yang ...	jika saya terlibat dalam pengobatan, apa yang ...	jika saya terlibat dalam pengobatan, apa yang ...

```

faq_answ['prep1']= faq_answ['Jawaban'].apply(to_lower)
faq_answ['prep2']= faq_answ['prep1'].apply(remove_number)
faq_answ['prep3']= faq_answ['prep2'].apply(remove_punct)
faq_answ['prep4']= faq_answ['prep3'].apply(to_strip)
faq_answ['prep5']= faq_answ['prep4'].apply(remove_char)
faq_answ['prep6']= faq_answ['prep5'].apply(remove_duplicate)
faq_answ['prep7']= faq_answ['prep6'].apply(remove_stopwords)
faq_answ['Lemma']= faq_answ['prep7'].apply(lemma)
faq_answ.head(10)

```

Question_ID	Jawaban	prep1	prep2	prep3
0 1590140	Penyakit mental adalah kondisi kesehatan yang			
1 2110618	Diperkirakan bahwa penyakit mental mempengaru...			
2 6361820	Diperkirakan bahwa penyakit mental mempengaru...			
3 9434130	Gejala gangguan kesehatan mental bervariasi te...			
4 7657263	Ketika penyembuhan dari penyakit mental, ident...	ketika penyembuhan dari penyakit mental, ident...	ketika penyembuhan dari penyakit mental, ident...	ketika penyembuhan dari penyakit mental identi...
5 1619387	Meskipun situs web ini tidak dapat mengantik...			
6 1030153	Merasa nyaman dengan profesional yang Anda ata...			
7 8022026	Sama seperti ada berbagai jenis obat untuk pen...	sama seperti ada berbagai jenis obat untuk pen...	sama seperti ada berbagai jenis obat untuk pen...	sama seperti ada berbagai jenis obat untuk pen...
8 1155199	Karena memulai perawatan adalah langkah besar ...			
9 7760466	Ada banyak jenis profesional kesehatan mental....	ada banyak jenis profesional kesehatan mental....	ada banyak jenis profesional kesehatan mental....	ada banyak jenis profesional kesehatan mental ...

Pada proses text preprocessing, dilakukan lowercase, remove number, remove punctuation, remove duplicate, stopword removal, dan lemmatization.

## ✓ Analisis Eksplorasi Data

```

import pandas as pd
from nltk.tokenize import word_tokenize
from nltk.probability import FreqDist

# Gabungkan semua kata dari kolom "Question"
all_questions = ' '.join(faq_quest['lemma'])

# Tokenisasi teks menjadi kata-kata
words = word_tokenize(all_questions)

# Hitung frekuensi kata
freq_dist = FreqDist(words)

# Tampilkan frekuensi kata
print(freq_dist.most_common(10)) # Menampilkan 10 kata paling umum

```

```
[('mental', 20), ('kesehatan', 17), ('gangguan', 12), ('menemukan', 12), ('perbedaan', 12), ('bantuan', 12), ('lakukan', 11), ('pen  

◀ ▶  

#Menghitung frekuensi kata-kata pada kolom "Questions"  

def dictionary(check):  

    check = check.str.extractall('([a-zA_Z]+)')  

    check.columns = ['check']  

    b = check.reset_index(drop=True)  

    check = b['check'].value_counts()  

    dictionary = pd.DataFrame({'word': check.index, 'freq': check.values})  

    dictionary.index = dictionary['word']  

    dictionary.drop('word', axis = 1, inplace=True)  

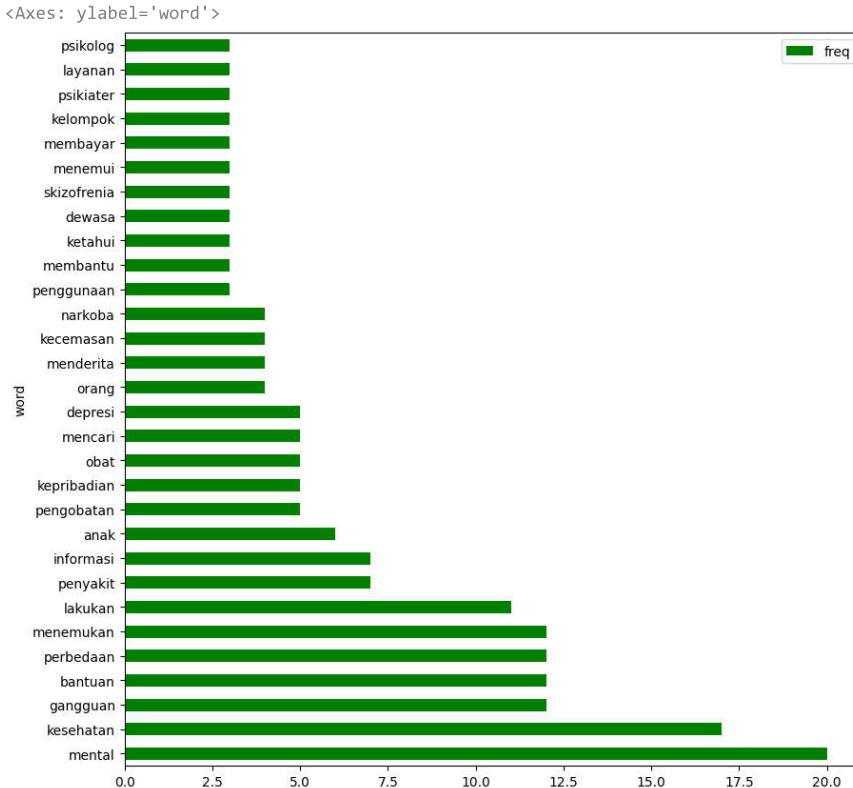
    dictionary.sort_values('freq', inplace= True, ascending= False)  

    return dictionary
```

```
dictionary_clean = dictionary(faq_quest['lemma'])  

dictionary_clean[:30].plot(kind = 'barh',figsize = (10,10), color='green')
```



```
from wordcloud import WordCloud  

import matplotlib.pyplot as plt  

# Menggabungkan semua kata dalam kolom 'Questions' menjadi satu teks  

text = ' '.join(faq_quest['Questions'])  

# Membuat word cloud  

wordcloud = WordCloud(width = 800, height = 800,  

                      background_color ='white',  

                      stopwords = stopwords.words('indonesian'),  

                      min_font_size = 10).generate(text)  

# Menampilkan word cloud  

plt.figure(figsize = (8, 8), facecolor = None)  

plt.imshow(wordcloud)  

plt.axis("off")  

...  

...  

...
```

```
plt.tight_layout(pad = 0)
```

```
plt.show()
```



Berdasarkan digram batang di atas, dapat dilihat bahwa frekuensi kata yang paling sering muncul pada pertanyaan yang berkaitan dengan kesehatan mental adalah kata "mental" dengan frekuensi kata sebesar 20 dan kata "kesehatan" dengan frekuensi kata 17.

```
import pandas as pd
from nltk.tokenize import word_tokenize
from nltk.probability import FreqDist

# Gabungkan semua kata dari kolom "Question"
all_questions = ' '.join(faq_anw['lemma'])

# Tokenisasi teks menjadi kata-kata
words = word_tokenize(all_questions)

# Hitung frekuensi kata
freq_dist = FreqDist(words)

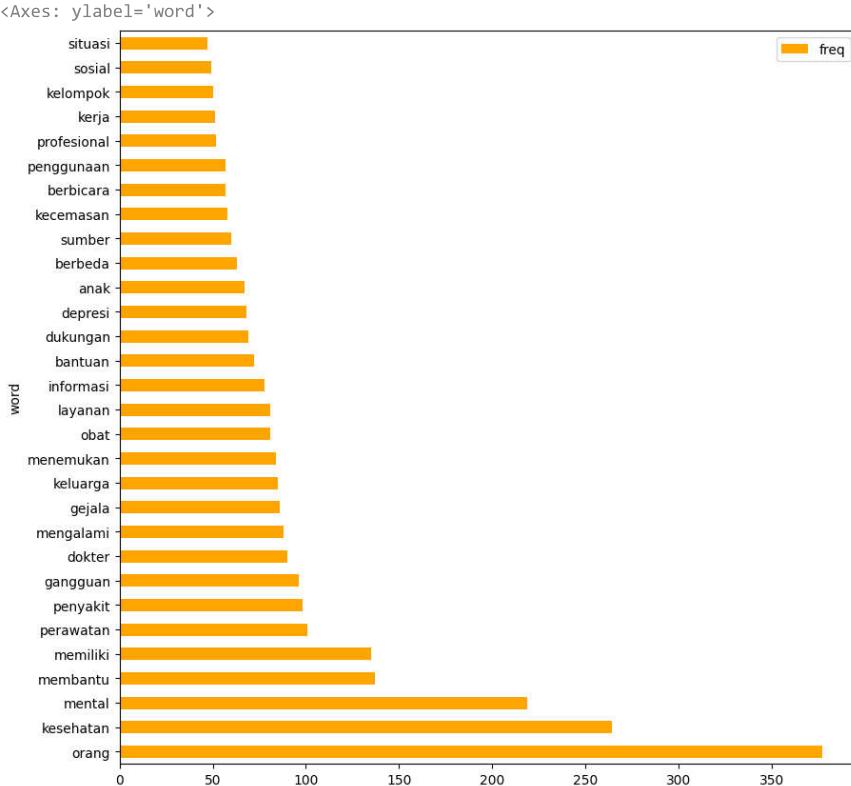
# Tampilkan frekuensi kata
print(freq_dist.most_common(10)) # Menampilkan 10 kata paling umum
```

```
#Menghitung frekuensi kata-kata pada kolom "Jawaban"
def dictionary(check):
    check = check.str.extractall('([a-zA_Z]+)')
    check.columns = ['check']
    b = check.reset_index(drop=True)
    check = b['check'].value_counts()

    dictionary = pd.DataFrame({'word': check.index, 'freq': check.values})
    dictionary.index = dictionary['word']
    dictionary.drop('word', axis = 1, inplace=True)
    dictionary.sort_values('freq', inplace= True, ascending= False)

    return dictionary

dictionary_clean = dictionary(faq_anw['lemma'])
dictionary_clean[:30].plot(kind = 'barh',figsize = (10,10),color='orange')
```



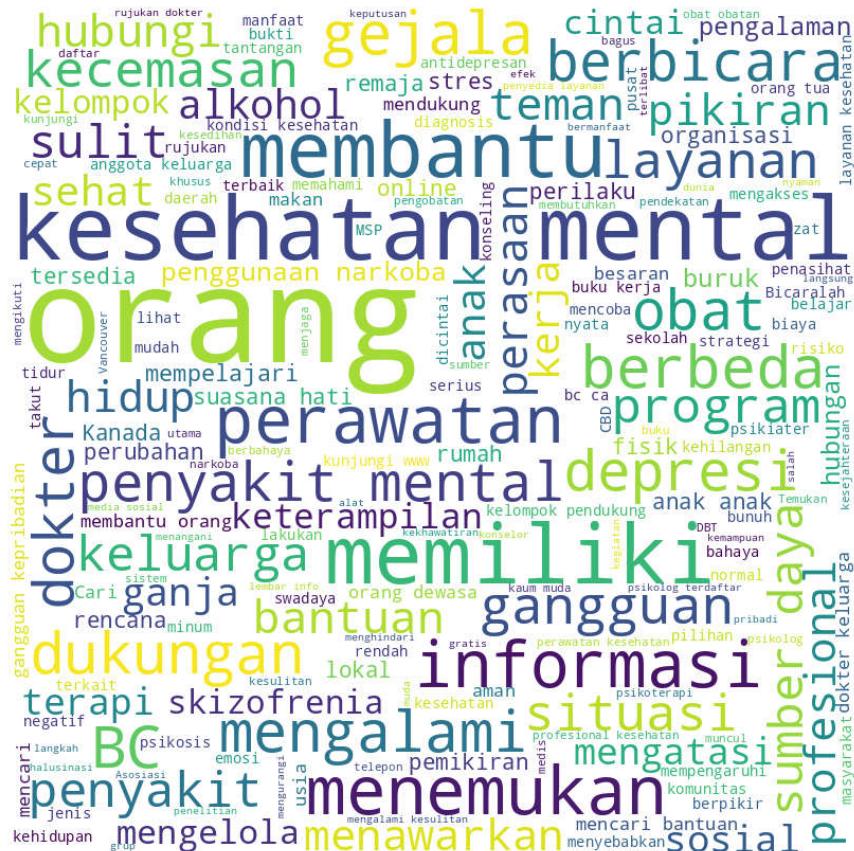
```
from wordcloud import WordCloud
import matplotlib.pyplot as plt

# Menggabungkan semua kata dalam kolom 'Jawaban' menjadi satu teks
text = ' '.join(faq_anw['Jawaban'])

# Membuat word cloud
wordcloud = WordCloud(width = 800, height = 800,
                      background_color ='white',
                      stopwords = stopwords.words('indonesian'),
                      min_font_size = 10).generate(text)

# Menampilkan word cloud
plt.figure(figsize = (8, 8), facecolor = None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad = 0)

plt.show()
```



Berdasarkan diagram batang di atas, dapat dilihat bahwa frekuensi kata yang paling sering muncul pada jawaban yang berkaitan dengan kesehatan mental adalah kata "orang" dengan frekuensi kata sebesar 377, kata "kesehatan" dengan frekuensi kata 264, dan kata "mental" dengan frekuensi kata 219.

## Normalisasi/Vektorisasi

```
from sklearn.model_selection import train_test_split  
from sklearn.feature_extraction.text import TfidfVectorizer  
from sklearn.svm import LinearSVC
```

```
text = faq['Questions']
y= faq['JawabanEncode'].values
```

```
tfidf = TfidfVectorizer(use_idf=True, analyzer='word', stop_words='english', token_pattern=r'\b[\^\d\W]+\b', ngram_range=(1,2))
X_train = tfidf.fit_transform(text)
print(X_train)
```

(0, 506)	0.29766662771136054
(0, 170)	0.3998547064187586
(0, 187)	0.36678085172921204
(0, 727)	0.36678085172921204
(0, 45)	0.21621646894305466
(0, 446)	0.21203246659316707
(0, 504)	0.2867744503381645
(0, 167)	0.3433145783784615
(0, 186)	0.36678085172921204
(0, 725)	0.1906313709496605
(0, 36)	0.14062033956802386
(1, 486)	0.3414532789086613
(1, 686)	0.3414532789086613
(1, 737)	0.3414532789086613
(1, 639)	0.3414532789086613
(1, 485)	0.3414532789086613

```

(1, 685)      0.3414532789086613
(1, 638)      0.3414532789086613
(1, 506)      0.25419044573476496
(1, 446)      0.18106372087441705
(1, 504)      0.24488914298946848
(1, 725)      0.16278839695698333
(2, 509)      0.49051145702302673
(2, 40)       0.44993895808100537
(2, 507)      0.44993895808100537
:           :
(95, 18)      0.2459979689026959
(95, 14)      0.21328947396573344
(95, 709)     0.16874381501354382
(95, 64)       0.13888701381380952
(95, 577)     0.0886998500546177
(96, 464)     0.29833042133121224
(96, 115)     0.29833042133121224
(96, 141)     0.29833042133121224
(96, 235)     0.29833042133121224
(96, 649)     0.29833042133121224
(96, 44)      0.29833042133121224
(96, 127)     0.29833042133121224
(96, 463)     0.29833042133121224
(96, 114)     0.29833042133121224
(96, 648)     0.29833042133121224
(96, 234)     0.2736541155476652
(96, 138)     0.15522746697748319
(96, 36)      0.10491642208440592
(97, 676)     0.4196695874272384
(97, 117)     0.4196695874272384
(97, 37)      0.4196695874272384
(97, 724)     0.4196695874272384
(97, 116)     0.4196695874272384
(97, 670)     0.31241755777073815
(97, 36)      0.14758880899250104

```

Berdasarkan output vektorisasi TF-IDF tersebut, dapat disimpulkan bahwa :

(0, 506) 0.2976666277113606

(0, 170) 0.39985470641875864

(0, 187) 0.36678085172921204

kata dengan indeks 170 memiliki bobot TF-IDF tertinggi di antara ketiga kata tersebut, yang menunjukkan bahwa kata tersebut lebih spesifik atau penting dalam konteks dokumen ke-0. Sementara itu, kata dengan indeks 506 memiliki bobot yang lebih rendah, menunjukkan bahwa kata tersebut mungkin lebih umum dan lebih banyak muncul dalam korpus secara keseluruhan.

## ▼ Modeling

```
lsvc = LinearSVC(random_state = 2021)
lsvc.fit(X_train, y)
```

```
    ▾ LinearSVC
    LinearSVC(random_state=2021)
```

```
search_test = [
    "Apa yang dimaksud dengan penyakit mental?"
]
```

```
search_engine = tfidf.transform(search_test)
result = lsvc.predict(search_engine)
```

```
for question in result:
    faq_data = faq.loc[faq.isin([question]).any(axis=1)]
    print("Jawaban: ", faq_data['Jawaban'].values)
```

```
Jawaban: ['Penyakit mental adalah kondisi kesehatan yang mengganggu pikiran, emosi, hubungan, dan fungsi seseorang. Mereka dikaitk
```

## ▼ Uji/Testing

```

import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.svm import LinearSVC

# Inisialisasi TF-IDF Vectorizer
tfidf_vectorizer = TfidfVectorizer()

# Inisialisasi model LinearSVC
lsvc_model = LinearSVC()

# Transformasi teks pertanyaan menjadi vektor TF-IDF
X = tfidf_vectorizer.fit_transform(faq['Questions']) # Assuming 'Pertanyaan' is the column name for questions
lsvc_model.fit(X, faq['Jawaban']) # Assuming 'Jawaban' is the column name for answers

# Fungsi chatbot
def chatbot():
    # List untuk menyimpan pertanyaan dan jawaban
    pertanyaan = []
    jawaban = []

    while True:
        print("Welcome to the Mental Health FAQ Chatbot!")
        print("*****")
        # Meminta pengguna untuk memasukkan pertanyaan
        user_input = input("Pertanyaan : ")

        # Keluar dari loop jika pengguna mengetik 'quit'
        if user_input.lower() == 'quit':
            break

        # Transformasi input pengguna menjadi vektor TF-IDF
        user_input_vectorized = tfidf_vectorizer.transform([user_input])

        # Melakukan prediksi jawaban
        predicted_answer = lsvc_model.predict(user_input_vectorized)

        # Menampilkan jawaban
        print("Jawaban:", predicted_answer[0])

        # Menambahkan pertanyaan dan jawaban ke dalam list
        pertanyaan.append(user_input)
        jawaban.append(predicted_answer[0])

        # Menambahkan spasi
        print()

    # Membuat DataFrame dari pertanyaan dan jawaban
    df = pd.DataFrame({'Pertanyaan': pertanyaan, 'Jawaban': jawaban})

    # Menyimpan DataFrame ke dalam file CSV
    df.to_csv('chatbot_output.csv', index=False)

# Memanggil fungsi chatbot
chatbot()

Welcome to the Mental Health FAQ Chatbot!
*****
Pertanyaan : Apa perbedaan kesehatan mental dan penyakit mental?
Jawaban: 'Kesehatan mental' dan 'penyakit mental' semakin banyak digunakan seolah -olah memaksudkan hal yang sama, tetapi tidak. Ketika kita berbicara tentang kesehatan mental, kita berbicara tentang kesejahteraan mental kita: emosi kita, pikiran dan perasaan. Penyakit mental adalah penyakit yang memengaruhi cara orang berpikir, merasakan, berperilaku, atau berinteraksi dengan orang lain. Kesehatan tidak seperti saklar hidup/mati. Ada berbagai tingkat kesehatan. Orang-orang beralih pada kontinum mulai dari kesehatan yang sama seperti seseorang yang merasa tidak sehat mungkin tidak memiliki penyakit serius, orang mungkin memiliki kesehatan mental yang buruk tetapi tidak ada penyakit mental, sangat mungkin untuk memiliki penyakit mental. Dengan dukungan dan alat yang tepat, siapa pun dapat hidup dengan baik - tetapi mereka mendefinisikan dengan baik - dan mereka yang tidak.
Dengan dukungan dan alat yang tepat, siapa pun dapat hidup dengan baik - tetapi mereka mendefinisikan dengan baik - dan mereka yang tidak.

Welcome to the Mental Health FAQ Chatbot!
*****
Pertanyaan : Apa penyebab gangguan kesehatan mental terjadi?
Jawaban: Tantangan atau masalah dengan kesehatan mental Anda dapat muncul dari masalah psikologis, biologis, dan sosial, serta pribadi.

Welcome to the Mental Health FAQ Chatbot!
*****
Pertanyaan : Siapa saja yang bisa mengalami gangguan penyakit mental?
Jawaban: Diperkirakan bahwa penyakit mental mempengaruhi 1 dari 5 orang dewasa di Amerika, dan bahwa 1 dari 24 orang dewasa memiliki penyakit mental. Meskipun penyakit mental dapat mempengaruhi siapa pun, kondisi tertentu mungkin lebih umum pada populasi yang berbeda. Misalnya, selain itu, segala usia rentan, tetapi yang muda dan yang tua sangat rentan. Penyakit mental biasanya mengejutkan individu dalam suasana hati, kepribadian, kebiasaan pribadi, dan pengasuh harus menyadari fakta ini, dan memperhatikan perubahan dalam suasana hati, kepribadian, kebiasaan pribadi.

Welcome to the Mental Health FAQ Chatbot!
*****
Pertanyaan : Apa tanda - tanda seseorang mengalami gangguan kesehatan mental?
Jawaban: Gejala gangguan kesehatan mental bervariasi tergantung pada jenis dan tingkat keparahan kondisi tersebut. Berikut ini adalah beberapa gejala yang mungkin dialami oleh orang dengan gangguan kesehatan mental:
Pada orang dewasa:
Bingung berpikir
```

Kesedihan atau mudah marah yang tahan lama  
Tinggi dan rendah dalam suasana hati  
Ketakutan yang berlebihan, mengkhawatirkan, atau kecemasan  
Penarikan sosial  
Perubahan dramatis dalam kebiasaan makan atau tidur  
Perasaan marah yang kuat  
Delusi atau halusinasi (melihat atau mendengar hal -hal yang tidak benar -benar ada)  
Meningkatkan ketidakmampuan untuk mengatasi masalah dan kegiatan sehari -hari  
Pikiran bunuh diri  
Penolakan masalah yang jelas  
Banyak masalah fisik yang tidak dapat dijelaskan  
Penyalahgunaan narkoba dan/atau alkohol  
Pada anak yang lebih besar dan pra-remaja:  
Penyalahgunaan narkoba dan/atau alkohol  
Ketidakmampuan untuk mengatasi masalah dan aktivitas sehari -hari  
Perubahan kebiasaan tidur dan/atau makan  
Keluhan Masalah Fisik yang Berlebihan  
Menentang otoritas, melewatkkan sekolah, mencuri, atau merusak properti  
Ketakutan yang intens untuk menambah berat badan  
Suasana hati negatif yang tahan lama, seringkali bersama dengan nafsu makan yang buruk dan pikiran mati  
Ledakan kemarahan yang sering  
Pada anak kecil:  
Perubahan kinerja sekolah  
Nilai yang buruk meskipun ada upaya kuat  
Kekhawatiran atau kecemasan yang berlebihan  
Hiperaktif  
Mimpi buruk vano pisih

**Berdasarkan output jawaban yang diberikan oleh chatbot terkait pertanyaan yang diberikan oleh user yang berkaitan dengan kesehatan mental telah dijawab dengan baik oleh chatbot.**