LAPORAN PRAKTIKUM

DATA MAINING

Disusun Oleh :

Nama : Fahmi Adi Setiawan

NIM : 22230010

Mata Kuliah : Prak. Data Maining

Program Studi Sistem Informasi

Fakultas Sains dan Teknologi

Universitas Respati Yogyakarta

2025/2026

Kodingan dan hasil Runing

```
import pandas as pd
import re
from collections import defaultdict
from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory

# Load data CSV
df = pd.read_csv('kelompok2_22230010.csv')
tweets = df['Isi Tweet'].dropna().tolist()

# Preprocessing & hapus stopword Indonesia
factory = StopWordRemoverFactory()
stop_words = set(factory.get_stop_words())

def preprocess(text):
    text = re.sub(r"http\S+", "", text)                 # hapus link
    text = re.sub(r"[^a-zA-Z\s]", "", text)             # hapus simbol
    text = text.lower()
    tokens = text.split()
    return ' '.join([w for w in tokens if w not in stop_words])

tweets_cleaned = [preprocess(t) for t in tweets]

# Simulasi STC (Suffix Tree Clustering) sederhana
def extract_phrases(text, min_words=2, max_words=4):
    tokens = text.split()
    phrases = []
    for size in range(min_words, max_words + 1):
        for i in range(len(tokens) - size + 1):
            phrase = ' '.join(tokens[i:i+size])
            phrases.append(phrase)
    return phrases

# Buat indeks frasa dokumen
phrase_to_docs = defaultdict(set)
for idx, tweet in enumerate(tweets_cleaned):
    words = extract_phrases(tweet)
    for phrase in words:
        phrase_to_docs[phrase].add(idx)

# Ambang batas jumlah dokumen
min_docs_per_cluster = 5  # Ambang batas/threshold
common_phrases = {
    phrase: docs for phrase, docs in phrase_to_docs.items()
    if len(docs) >= min_docs_per_cluster
}

# Kelompokkan dokumen berdasarkan frasa
clusters = defaultdict(set)
for phrase, docs in common_phrases.items():
    clusters[f"Cluster: '{phrase}'"] = docs

# Tampilkan hasil clustering
if not clusters:
    print("Tidak ada cluster yang memenuhi ambang batas !!!!!")
else:
    for i, (cluster_name, doc_ids) in enumerate(clusters.items(), start=1):
        print(f"\n{i}. {cluster_name} (total: {len(doc_ids)} tweet)")
        for doc_id in doc_ids:
            print(f"- {tweets[doc_id]}")
```

1. Cluster: 'rt doktertifa' (total: 14 tweet)
- RT @DokterTifa: Dugaan Ijazah Palsu sudah sampai di Media Internasional.

Jokowi sejak sekarang harus menghitung langkah, sebab sebetulnya…
- RT @DokterTifa: Dugaan Ijazah Palsu sudah sampai di Media Internasional.

Jokowi sejak sekarang harus menghitung langkah, sebab sebetulnya…
- RT @DokterTifa: Sekarang giliran Prof Amin Rais yang dilaporkan buntut isu ijazah palsu.

Lama-lama Jokowi jadi sasaran tukang palak para…
- RT @DokterTifa: Dugaan Ijazah Palsu sudah sampai di Media Internasional.

Jokowi sejak sekarang harus menghitung langkah, sebab sebetulnya…
- RT @DokterTifa: Dugaan Ijazah Palsu sudah sampai di Media Internasional.