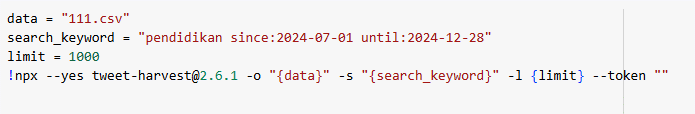
**LAPORAN TOPIC MODELLING LDA**

**Proses Pembuatan**

Untuk mengimplementasikan topic modelling LDA, pertama-tama, dilakukan scraping data 1000 tweet di X tentang pendidikan dalam rentang waktu Juni 2024 – Desember 2024 dengan menggunakan tweet harvest. Berikut adalah coding yang digunakan.







Namun karena per putaran scrape data hanya diperoleh 100 data per hari. Maka dilakukan iterasi beberapa kali hingga terkumpul 1000 data.

Selanjutnya, dilakukan preprocessing data dengan membersihkan data dari kata-kata kurang penting yang sering muncul seperti “ga, tdk, ya, gua”. Kata-kata juga dibuat ke bentuk paling dasarnya dengan menghapus imbuhan.

import pandas as pd

import re

from nltk.corpus import stopwords

from nltk.tokenize import word\_tokenize

from nltk.stem import WordNetLemmatizer

from sklearn.feature\_extraction.text import CountVectorizer

import gensim

from gensim import corpora

from gensim.models import LdaModel

from sklearn.feature\_extraction.text import TfidfVectorizer

from sklearn.decomposition import LatentDirichletAllocation

import nltk

nltk.download('punkt')

nltk.download('stopwords')

nltk.download('wordnet')

stop\_words = stopwords.words('indonesian')

custom\_stop\_words = ['yang', 'ga', 'https', 'co', 'yg','t', 'ya', 'nya', 'gue','gw', 'gua', 'aja', 'kalo', 'ga', 'pemerintahnaikkangajiguru', 'amp','sih']

stop\_words.extend(custom\_stop\_words)

wtk = nltk.tokenize.RegexpTokenizer(r'\w+')

wnl = WordNetLemmatizer()

from google.colab import files

uploaded = files.upload()

file\_name = list(uploaded.keys())[0]

app\_reviews\_df = pd.read\_csv(file\_name)

print(app\_reviews\_df.head())

reviews = app\_reviews\_df['description']

reviews.dropna(inplace=True)

reviews.reset\_index(drop=True, inplace=True)

def preprocess\_text(text):

    tokens = wtk.tokenize(text.lower())

    filtered\_tokens = [word for word in tokens if word not in stop\_words]

    lemmatized\_tokens = [wnl.lemmatize(word) for word in filtered\_tokens]

    return " ".join(lemmatized\_tokens)

processed\_reviews = reviews.apply(preprocess\_text)

print(processed\_reviews.head())

Kemudian, program menjalankan algoritma LDA untuk menghasilkan 5 buah topik yang paling sering tentang keseluruhan tweet.

from gensim import corpora

from gensim.models import LdaModel

processed\_reviews\_list = processed\_reviews.tolist()

dictionary = corpora.Dictionary([review.split() for review in processed\_reviews\_list])

corpus = [dictionary.doc2bow(review.split()) for review in processed\_reviews\_list]

num\_topics = 5

lda\_model = LdaModel(corpus=corpus, id2word=dictionary, num\_topics=num\_topics, passes=10, random\_state=42)

topics = lda\_model.show\_topics(num\_topics=num\_topics, num\_words=7, formatted=False)

topics = sorted(topics, key=lambda x: x[0])  # Mengurutkan topik berdasarkan indeks

for i, (\_, word\_list) in enumerate(topics):

    topic\_words = ', '.join([word for word, \_ in word\_list])

    print(f"Topic {i}: {topic\_words}")

**Hasil dan Pembahasan**

Berikut adalah topik yang dihasilkan dari ke-1000 tweet.

Topic 0: pendidikan, guru, indonesia, kenaikan, kesehatan, berkualitas, gaji

Topic 1: pendidikan, wayang, indonesia, budaya, batik, orang, pelatihan

Topic 2: pendidikan, anak, makanbergizibangunnegeri, guru, orang, sekolah, pemerintah

Topic 3: pendidikan, orang, kerja, dunia, kesehatan, sma, sekolah

Topic 4: pendidikan, orang, banget, susah, butuh, menteri, tu

Berikut adalah analisis untuk masing-masing topik berdasarkan kata kunci.

*Topic 0: pendidikan, guru, indonesia, kenaikan, kesehatan, berkualitas, gaji*

Fokus utama dari topik ini adalah isu kesejahteraan guru di Indonesia, termasuk gaji, kesehatan, dan kualitas pendidikan. Berarti, ada perhatian terhadap kesejahteraan guru dan bagaimana hal ini memengaruhi kualitas pendidikan di Indonesia. Selain itu, kata "kenaikan" kemungkinan mengacu pada harapan atau diskusi terkait kenaikan gaji guru.

*Topic 1: pendidikan, wayang, indonesia, budaya, batik, orang, pelatihan*

Fokus utama dari topik ini adalah pendidikan berbasis budaya dan pelestarian budaya lokal Indonesia. Selain itu, kata "wayang" dan "batik" menyoroti pentingnya memperkenalkan budaya lokal dalam pendidikan. Berikutnya, pelatihan mungkin berhubungan dengan upaya memperkenalkan budaya melalui pendidikan formal atau non-formal.

*Topic 2: pendidikan, anak, makanbergizibangunnegeri, guru, orang, sekolah, pemerintah*

Fokus utama dari topik ini adalah pendidikan anak, nutrisi, dan peran pemerintah dalam mendukungnya. Isu nutrisi ("makanbergizibangunnegeri") menunjukkan perhatian terhadap pentingnya gizi dalam mendukung keberhasilan pendidikan anak. Ada pula penekanan pada guru dan sekolah sebagai faktor penting dalam membangun negeri.

*Topic 3: pendidikan, orang, kerja, dunia, kesehatan, sma, sekolah*

Fokus utama dari topik ini adalah pendidikan yang relevan dengan dunia kerja dan kesehatan. Ada hubungan antara pendidikan (terutama di tingkat SMA) dan persiapan kerja. Kesehatan mungkin menjadi faktor yang memengaruhi kesiapan siswa dalam menghadapi dunia kerja.

*Topic 4: pendidikan, orang, banget, susah, butuh, menteri, tu*

Fokus utama dari topik ini adalah keluhan atau tantangan dalam sistem pendidikan. Kata seperti "susah", "butuh", dan "banget" menunjukkan adanya keluhan terkait akses atau kualitas pendidikan. "Menteri" mungkin mengacu pada ekspektasi terhadap kinerja Menteri Pendidikan dalam menyelesaikan masalah pendidikan.

**Proses Lengkap: Scraping Data, Insert ke Supabase, Hingga Deploy ke Vercel**

**Langkah 1: Scraping Data**

1. **Load Data**
   * File CSV dengan data awal (all.csv) dimuat menggunakan pandas: pd.read\_csv(file\_path).
   * Data diperiksa untuk memastikan kolom description tersedia.
   * Data dengan nilai NaN dalam kolom description dihapus.
2. **Hasil Awal**
   * Menampilkan 5 baris pertama dari dataset awal untuk validasi.

**Langkah 2: Pembersihan Data**

1. **Preprocessing Data**
   * Fungsi clean\_text dibuat untuk:
     + Mengubah teks ke huruf kecil.
     + Menghapus URL, angka, dan karakter tidak relevan.
     + Melakukan tokenisasi menggunakan WordTokenizer.
     + Menghapus stop words.
     + Melakukan lemmatization.
2. **Aplikasi Pembersihan**
   * Kolom description diproses dengan fungsi clean\_text.
   * Kolom baru cleaned\_description ditambahkan ke dataset.
3. **Simpan Hasil**
   * Data yang sudah dibersihkan disimpan ke file CSV baru: all\_cleaned.csv.

**Langkah 3: Insert Data ke Supabase**

1. **Persiapan Kolom**
   * Pastikan dataset memiliki semua kolom yang dibutuhkan:

conversation\_id\_str, created\_at, favorite\_count, description,

id\_str, image\_url, in\_reply\_to\_screen\_name, lang, location,

quote\_count, reply\_count, retweet\_count, tweet\_url, user\_id\_str

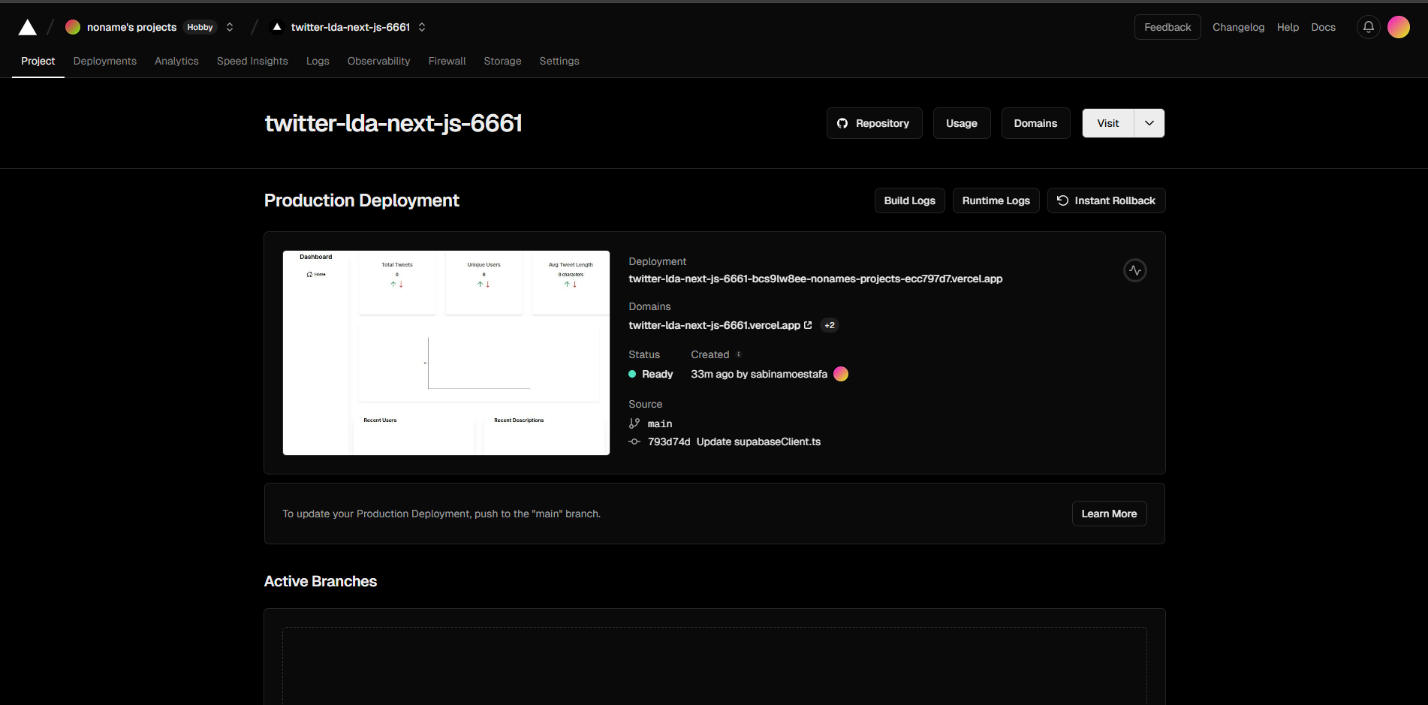
* + Kolom created\_at dikonversi ke format string.

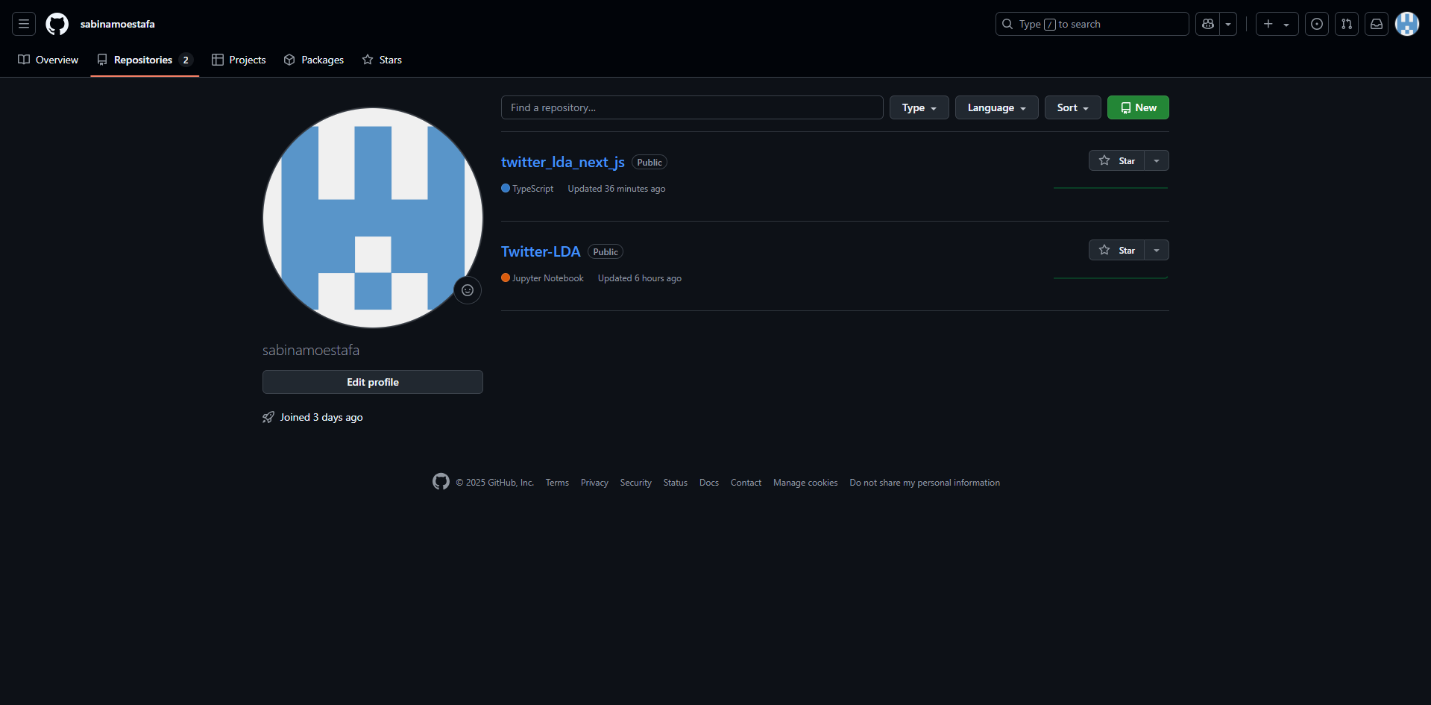
1. **Persiapan Data untuk Insert**
   * Dataset diubah menjadi dictionary (records) menggunakan to\_dict(orient="records").
   * Kolom description diisi dengan data dari cleaned\_description.
2. **Insert ke Supabase**
   * Data bersih dimasukkan ke tabel twitter\_data di Supabase menggunakan metode insert.
3. **Validasi Insert**
   * Setelah setiap insert, validasi dilakukan untuk memastikan status berhasil.
   * Kesalahan ditangani dengan mencetak pesan error.

**Langkah 4: Visualisasi di Next.js**

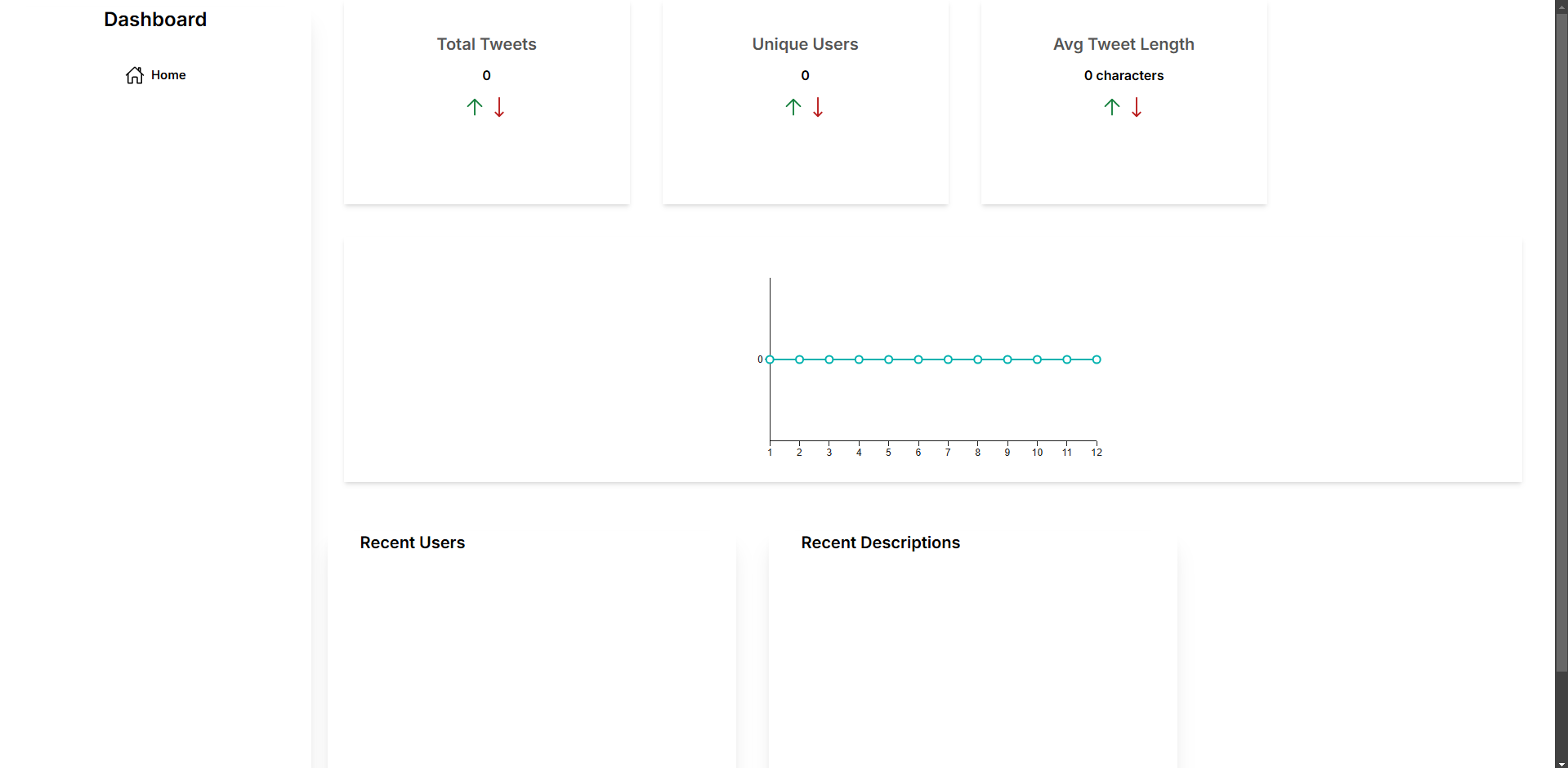
1. **Membuat Komponen Visualisasi**
   * Gunakan Next.js untuk membuat aplikasi frontend.
   * Data dari Supabase di-fetch melalui API Supabase.
   * Komponen visualisasi (contoh: grafik atau tabel) dibuat menggunakan library seperti Chart.js atau Recharts.
2. **Integrasi dengan Supabase**
   * Tambahkan kredensial Supabase ke proyek Next.js untuk memungkinkan koneksi database.
   * Data diambil dari Supabase dan ditampilkan secara dinamis di halaman.

**Langkah 5: Deploy ke Vercel**

1. **Persiapan GitHub**
   * Proyek Next.js diunggah ke repositori GitHub.
   * Pastikan file konfigurasi (seperti .env) berisi kredensial Supabase.
2. **Login ke Vercel**
   * Login ke Vercel menggunakan akun GitHub.
3. **Pilih Repositori**
   * Pilih repositori GitHub yang berisi proyek Next.js.
   * Klik "Deploy" untuk memulai proses deployment.
4. **Proses Deployment**
   * Tunggu hingga proses deployment selesai.
   * Setelah selesai, aplikasi dapat diakses melalui URL publik yang diberikan oleh Vercel.



**Hasil Akhir**

* Aplikasi Next.js yang terhubung ke Supabase berhasil di-deploy ke Vercel.
* Data dari Supabase divisualisasikan secara dinamis.
* URL publik dapat diakses oleh pengguna untuk melihat hasil visualisasi.

**Kesimpulan**

Semua topik berpusat pada pendidikan, namun dengan fokus berbeda: kesejahteraan guru, pelestarian budaya, nutrisi anak, relevansi pendidikan dengan dunia kerja, dan tantangan sistem pendidikan.

Link Project nya : <https://colab.research.google.com/drive/1OcRx2R5bYCbN6c9Zvd1-2ybYMtCQyapI?usp=sharing>