**Analisis Teks Twitter Ujaran Kebencian**

1. Pendahuluan

Teks di twitter sangat digemari dan bangak dibaca oleh kaum muda. Hal ini tentu berdampak pada pola pikir dan perilaku pembacanya, apabila tidak diantisipasi dengan tepat. Salah satu cara untuk mengantisipasi teks yang memiliki ujaran kebencian adalah dengan menganalisisnya. Teks twitter ujaran kebencian ini sangat tepat jika dianalisis menggunakan python, yang terdiri dari: a) input user berupa teks dan file dengan endpoint, b) server API dibuat dengan flask dan swagger UI, c) penyimpanan data dalam SQLite menggunakan modul SQLite, d) API yang bisa menghasilkan output berupa teks ygng sudah di-cleansing, e) API yang bisa memproses text cleansing

Bahasa pemrograman python merupakan bahasa *Open Source*, dapat digunakan tanpa lisensi, dan dikembangkan semampu yang bisa dilakukan. Sebenarnya bahasa pemrograman python ini mudah dipelajari karena penulisan sintaks yang lebih fleksibel. Selain itu, juga memiliki efisiensi tinggi untuk struktur data level tinggi, pemrograman berorientasi objek lebih sederhana tetapi efektif, dapat bekerja pada multi platform, dan digabungkan dengan bahasa pemrograman lain untuk menghasilkan aplikasi yang diinginkan.

Python dikenal sebagai bahasa pemograman interpreter, karena python dieksekusi dengan sebuah interpreter. Terdapat dua cara untuk menggunakan interpreter, yaitu dengan mode baris perintah dan modus script. Pada mode baris, perintah diketikkan pada shell atau command line dan python langsung menampilkan hasilnya. Bila menggunakan shell, semua definisi yang telah dibuat baik fungsi atau variabel akan dihapus. Cara lain adalah dengan menyimpan perintah –perintah python dalam satu file, yang disebut selanjutnya sebagai script. Kita dapat mengetikkan perintah-perintah python dengan menggunakan text editor seperti notepad. Lalu menyimpannya dengan akhiran ".py". kemudian menjalankannya dengan python.

Berdasarkan uraian di atas, penting untuk menganalisis teks twitter ujaran kebencian menggunakan python untuk mengetahui seberapa banyak kata-kata ujaran kebencian yang terdapat pada data teks yang dianalisis dan meminimalisir dampak negatifnya pada masyarakat.

1. Metode

Metode dalam penelitian ini menggunakan EDA python yaitu proses menyelidiki data set untuk menemukan pola dan anomaly, serta membentuk hipotesis berdasarkan pemahaman tentang data teks twitter ujaran kebencian berikut ini:

1. Proses *code function cleansing* dalam analisis ini adalah cara melakukan *data cleaning*. Untuk pembersihan data yang benar-benar menyeluruh, mengikuti langkah-langkah berikut:
   1. Mendeteksi error

Langkah awal yang harus dilakukan adalah memantau notifikasi *error* atau *corrupt*. Dicatat titik yang paling sering terjadi *error*. Dengan begitu, deteksi *error* bisa dilakukan lebih cepat.

* 1. Hapus duplikat data atau data yang tidak perlu

Jika sudah menemukan data yang *error*, *corrupt,* atau inkonsisten, langsung dilakukan tindakan, baik itu perbaikan atau penghapusan. Ini juga berlaku untuk duplikat dan data yang dirasa tidak perlu. Untuk mencegah terjadinya duplikasi data saat menggabungkan beberapa *data sources*, lakukan standarisasi proses. Dengan begitu, saat terdapat data ganda, sistem secara otomatis akan membuangnya.

* 1. Perbaiki kesalahan struktur

Menemukan penamaan yang aneh, *typo,* atau penggunaan simbol aneh saat sedang melakukan transfer data? Bisa jadi ada kesalahan struktur pada *dataset*. Kesalahan struktur tersebut biasanya disebabkan oleh adanya inkonsistensi data.

* 1. Filter outlier yang tidak diinginkan

Dalam proses pengolahan data, terkadang muncul data yang sekilas tampak tidak sesuai atau terpaut jauh dengan data lain. Inilah yang disebut dengan *outlier* atau pencilan.

Boleh saja menghapus *outlier* yang ditemukan, tentu saja dengan alasan jelas. Sebab, penyaringan *outlier* memang bisa membantu performa data yang sedang dikerjakan. Meski begitu, perlu diingat bahwa kemunculan *outlier* bukan berarti teori yang sedang dikerjakan salah. Justru sebaliknya, adanya *outlier* bisa menjadi indikator untuk menentukan validitas data.

* 1. Tangani data yang hilang

Data yang hilang harus ditangani dengan serius. Sebab, sering kali algoritma tidak akan menerima nilai yang hilang. Untuk menanganinya, bisa mengikuti cara berikut:

* Cara pertama, masukkan nilai yang hilang berdasarkan observasi lain. Cara ini sangat riskan karena mengandalkan asumsi, yang mana bisa mengancam integritas data.
* Cara kedua, buang observasi dengan nilai yang hilang. Namun, langkah ini bisa membuat kehilangnya informasi penting.
* Cara terakhir, mengubah cara data digunakan agar nilai yang kosong dapat dinavigasikan dengan efektif.
  1. Validasi dan lakukan QA

Langkah terakhir *data cleaning* adalah validasi dan QA (*quality assurance*). Harus bisa memastikan bahwa data diterima dan memang masuk akal. Selain itu, data juga harus sesuai dengan aturan yang ada. Dengan teori kerja yang diterapkan, apakah terbukti atau justru tidak. Jika tidak, apakah kualitas data berpengaruh? Pertanyaan-pertanyaan tersebut harus terjawab sebagai validasi data.

1. Pemilihan metode statistika dan EDA, dapat digunakan di bidang apapun, salah satunya adalah Data Science. Data Science adalah ilmu yang saat ini banyak diaplikasikan hampir di semua industri. Hal ini terjadi karena produksi data di dunia yang meningkat tajam dari waktu ke waktu, bahkan dalam satu hari dunia dapat memproduksi jutaan data. Salah satu profesi yang berkaitan dengan pengolahan data menggunakan metode statistika adalah Data Scientist.

Langkah pertama dalam menganalisis data adalah mempelajari karakteristik data tersebut. Analisis data eksploratif sangat penting untuk menunjang statistisi dalam menelaah dan menemukan karakteristik data yang selanjutnya dapat berguna dalam pemilihan model statistika yang tepat. Dua alat penting dalam EDA untuk menunjukkan karakteristik data adalah ringkasan numerik dan diagram kotak (boxplot). Ringkasan numerik menunjukkan dimana data memusat dan bagaimana sebaran datanya, sedangkan boxplot digunakan untuk mengetahui bagaimana distribusi data, dimana distribusi data ini sangat dibutuhkan untuk menentukan motode analisis statistika apa yang tepat untuk data tersebut. Untuk lebih jelasnya dapat dicermati temuan berikut:

* 1. Input user berupa teks dan file dengan 2 endpoint.
  2. server API dibuat dengan Flask dan Swagger UI
  3. Penyimpanan data dalam SQLite menggunakan modul SQLite
  4. API yang bisa menghasilkan output berupa teks ygng sudah di-cleansing
  5. API yang bisa memproses text cleansing

1. Proses visualisasi

Exploratory Data Analysis (EDA) atau dikenal pula dengan analisis data eksploratif merupakan pendekatan analisis untuk suatu data guna membuat gambaran keseluruhan (summary) data sehingga mudah untuk dipahami. Metode analisis ini menyediakan berbagai alat untuk meringkas dan memperoleh wawasan tentang sekumpulan data dengan cepat menggunakan graﬁk sebagai bentuk visualisasi data, tanpa menggunakan model statistik, atau formulasi hipotesis, yang dapat dilihat pada grafik di bawah ini:

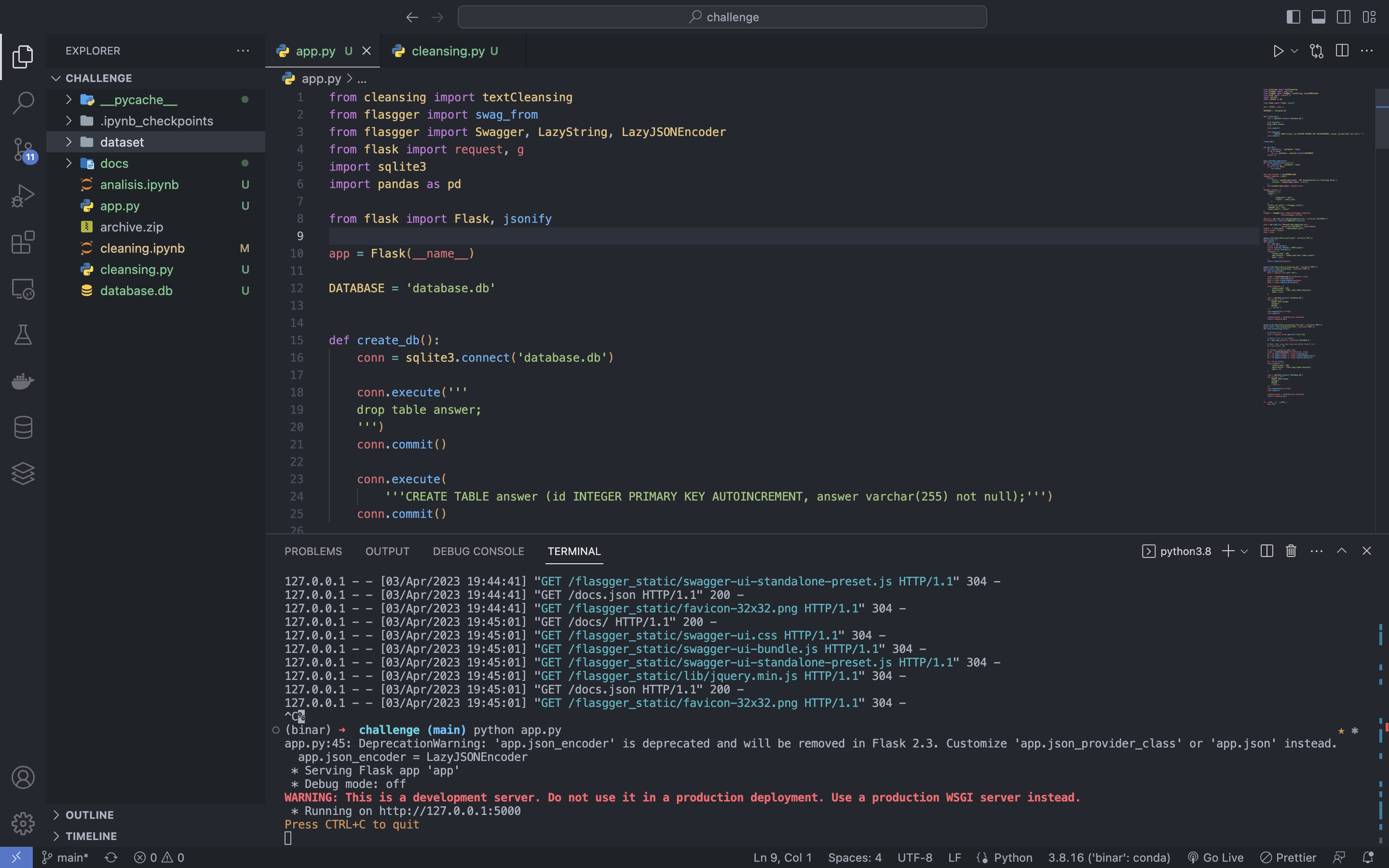
1. Hasil dan Kesimpulan
2. Hasil

Gambar 1 input user berupa teks dan file dengan 2 endpoint

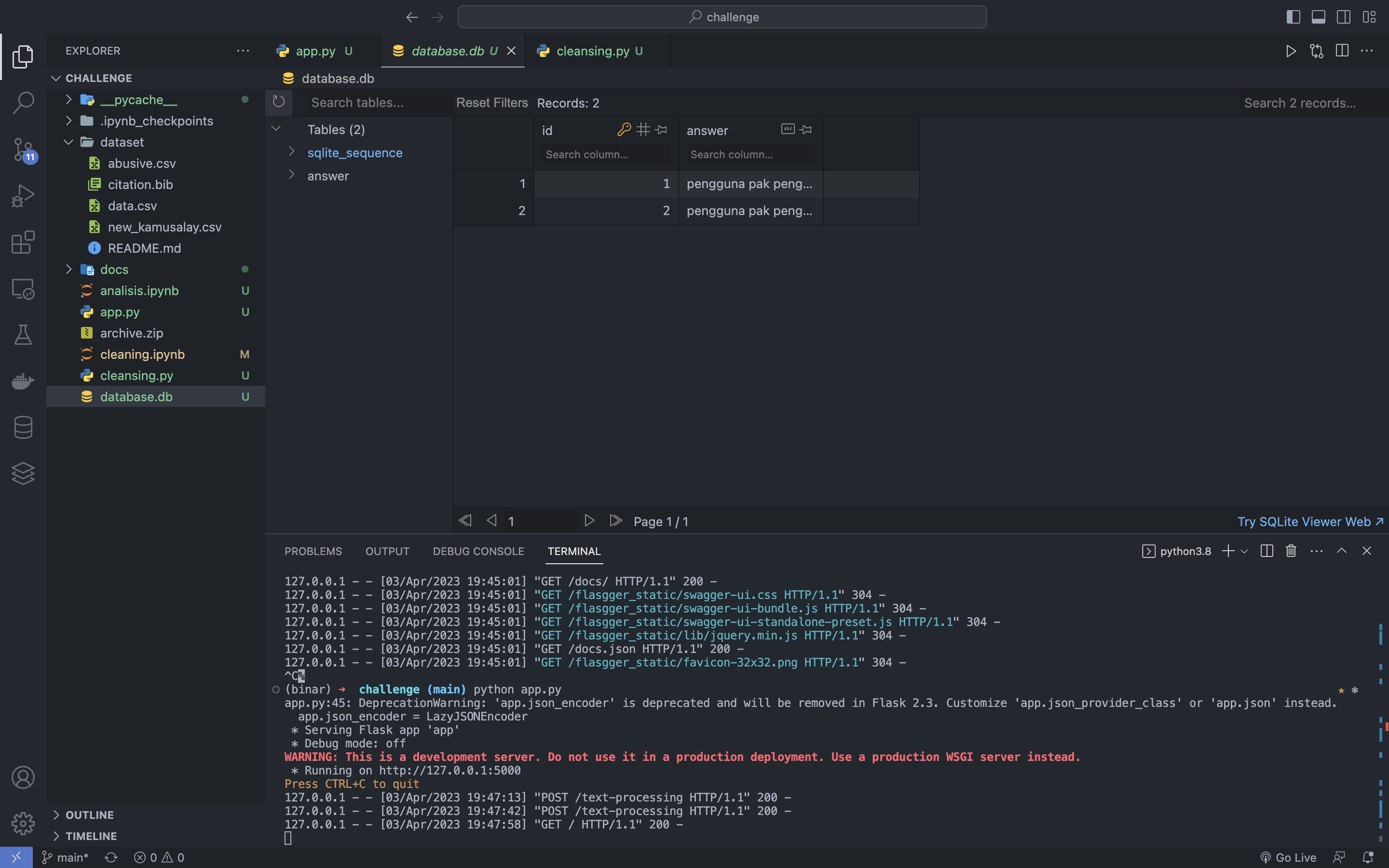
Graphical user interface, application

Description automatically generated

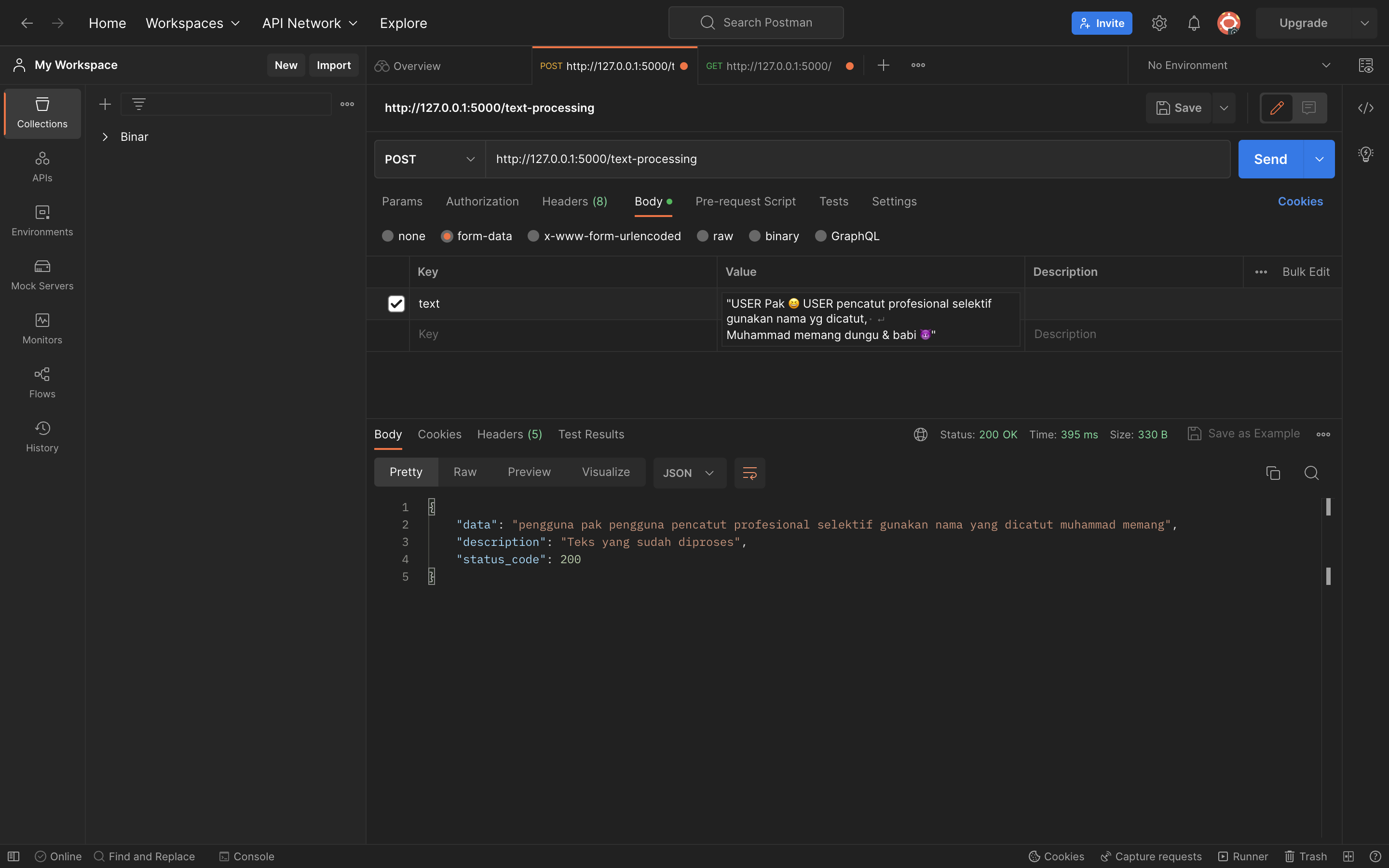
Gambar 2 server API dibuat dengan flask dan swagger UI



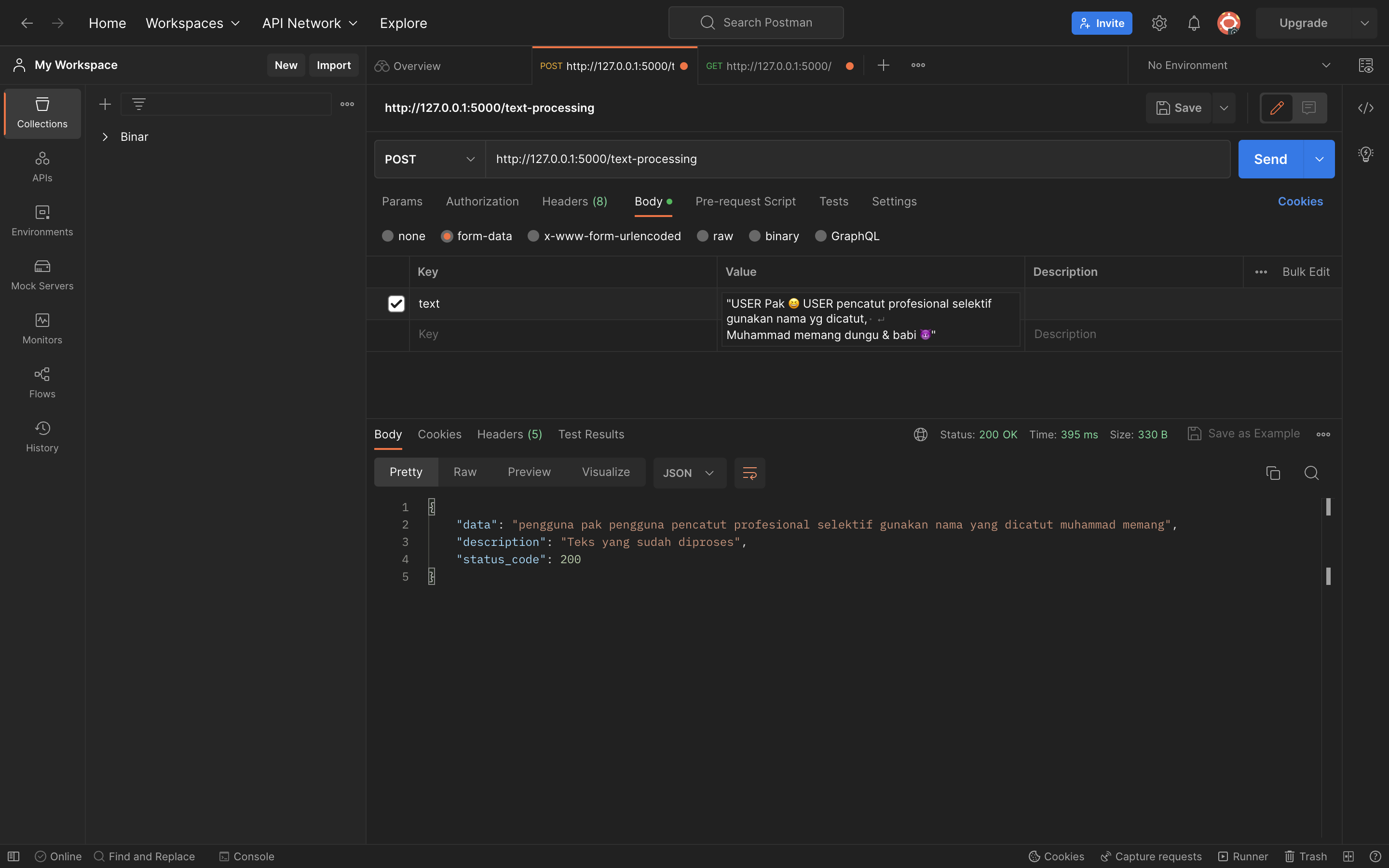
Gambar 3 penyimpanan data dalam SQLite menggunakan modul SQLite



Gambar 4 API yang bisa menghasilkan output berupa teks yang sudah di-cleansing



Gambar 5 API yang bisa memproses text cleansing



1. Kesimpulan