**Penggalian Sosial Media**

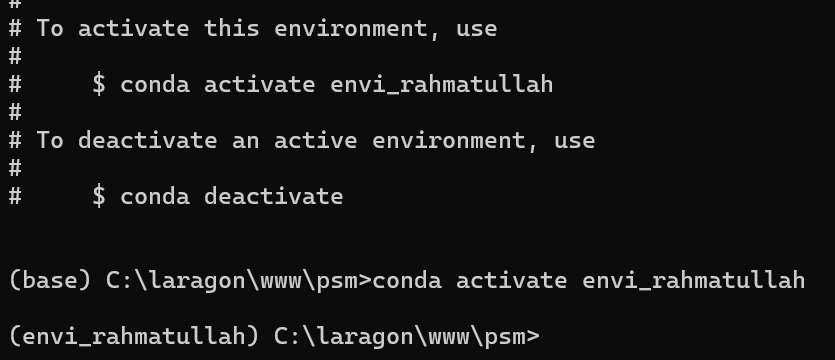
**Tugas 2**

Nama : Rahmatullah

Nim : 10221027

1. Buatkan Environtment Conda dengan nama envi\_nama-mahasiswa

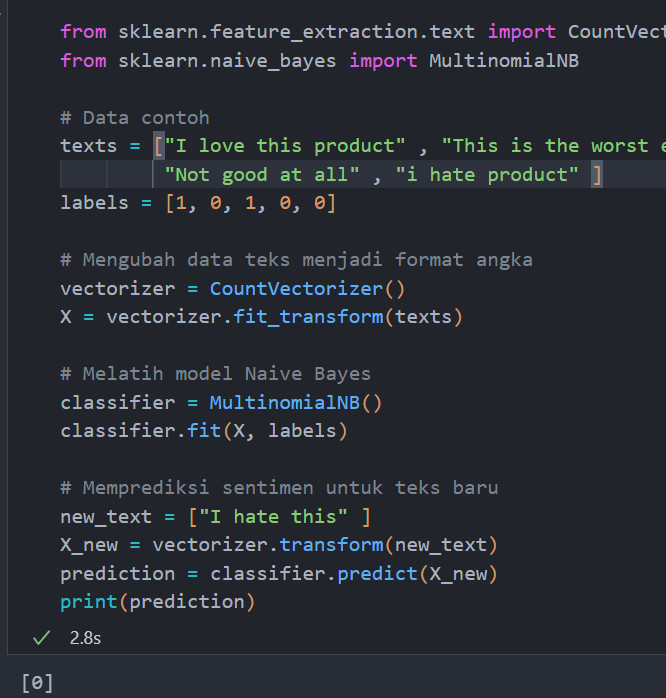
* Tunjukkan dengan screnshot conda activate envi\_nama-mahasiswa



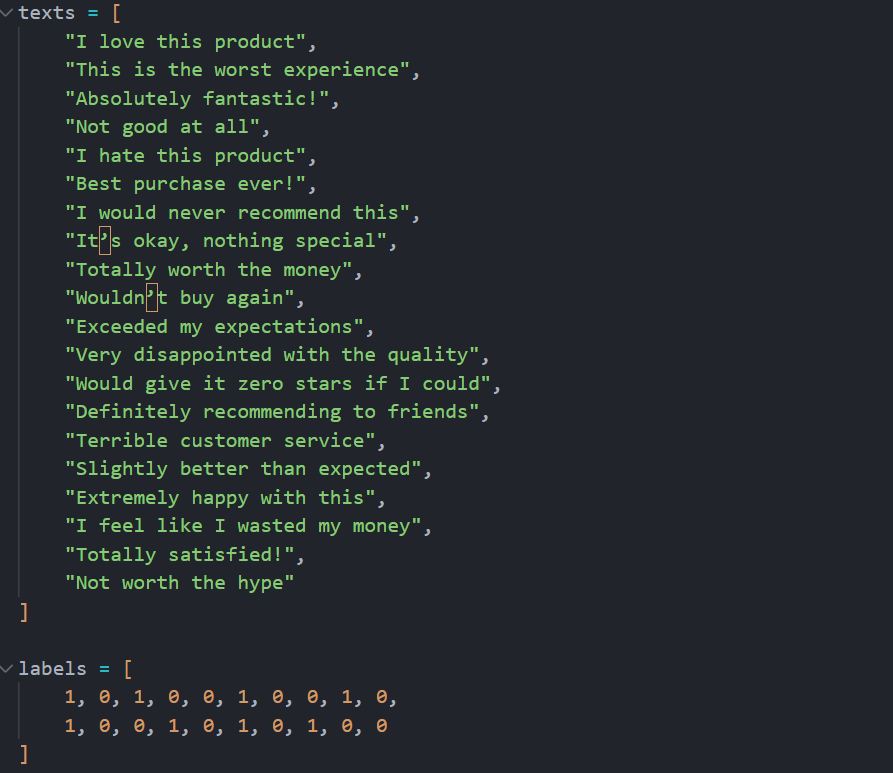
1. Modifikasi Dataset

* Tambahkan beberapa kalimat baru ke dalam dataset (texts) dan label (labels).
* Pastikan untuk menambahkan contoh teks dengan sentimen positif dan negatif.
* Latih ulang model dan uji apakah prediksi untuk teks baru berubah.

Sebelum ditambah :



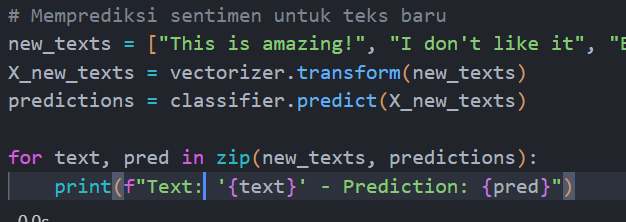
Setelah ditambah:

A screen shot of a computer code

AI-generated content may be incorrect.  
setelah dataset ditambah, dapat dilihat bahwa hasilnya tetap sama. Model masih bisa memprediksi kalimat “I hate this” merupakan sentiment negative.

1. Prediksi untuk Beberapa Teks Baru

* Uji model dengan beberapa teks baru
* Transformasikan teks baru menggunakan vectorizer.transform() dan prediksi sentimen untuk setiap teks.



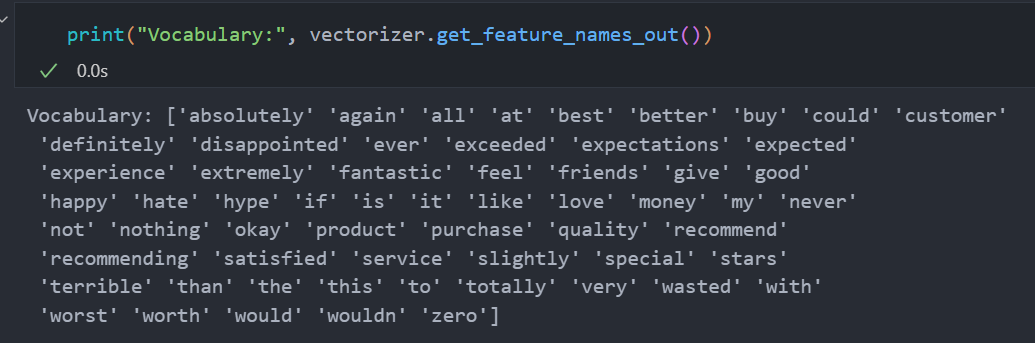
A black background with white text

AI-generated content may be incorrect.

Dari hasil yang didapat, model memprediksi 3 text baru dengan dengan banar dari 4 text baru. Model gagal memprediksi text “This is amazing!” yang seharusnya positif.

1. Analisis Vocabulary

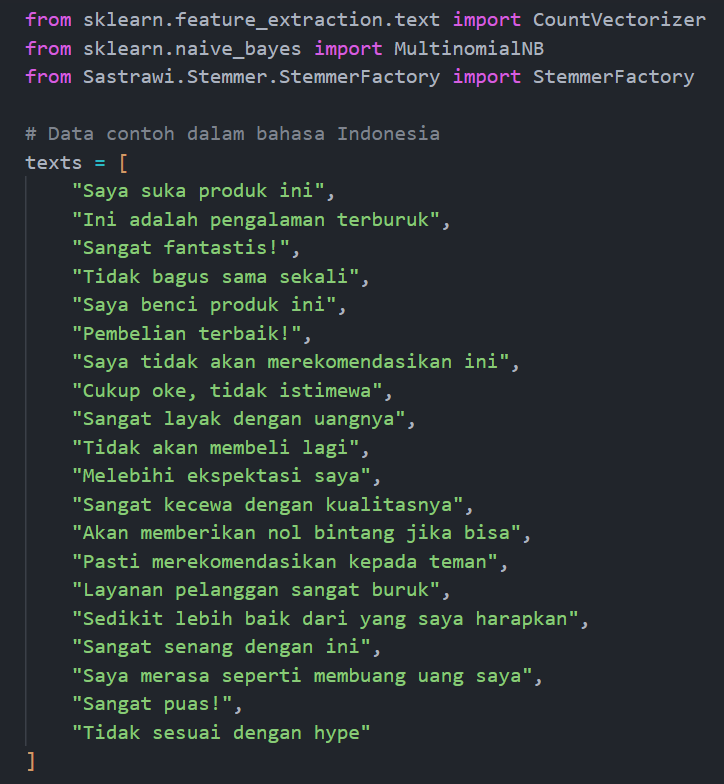
* Tampilkan vocabulary yang dihasilkan oleh CountVectorizer.
* print("Vocabulary:", vectorizer.get\_feature\_names\_out())
* Jelaskan bagaimana vocabulary ini digunakan dalam pembentukan matriks fitur.



Vocabulary dalam vectorizer seperti CountVectorizer, adalah daftar kata unik atau token yang ditemukan dalam seluruh korpus (kumpulan teks) yang digunakan untuk membangun matriks fitur. Setiap kata dalam vocabulary ini diberi indeks unik, dan vectorizer menggunakan vocabulary ini untuk mengubah teks mentah menjadi format numerik. Pertama dengan membangun vocabulary dari semua kata yang ada di dalam teks, kemudian setiap teks diubah menjadi vektor fitur berdasarkan kemunculan kata-kata dalam vocabulary. Setiap teks diubah menjadi vektor di mana setiap elemen mewakili jumlah kemunculan kata tertentu yang ada di vocabulary. Hasil akhirnya adalah matriks di mana setiap baris mewakili satu teks dan setiap kolom mewakili kata dalam vocabulary, dengan nilai dalam matriks menunjukkan frekuensi atau bobot kata tersebut dalam teks yang bersangkutan.

1. Modifikasi Dataset

* Tambahkan beberapa kalimat baru berbahasa indonesia ke dalam dataset (texts) dan label (labels).
* Pastikan untuk menambahkan contoh teks dengan sentimen positif dan negatif.
* Gunakan library sastrawi
* Latih ulang model dan uji apakah prediksi untuk teks baru berubah.

 A screen shot of a computer program

AI-generated content may be incorrect.

A screen shot of a computer program

AI-generated content may be incorrect.  
Model memprediksi 3 kalimat secara benar dari 4 kalimat yang diuji, Dimana prediksi kalimat yang salah adalah “Ini luar biasa” yang seharusnya sentiment positive, model memprediksi kalimat tersebut sebagai sentiment negative.