1. Install Necessary Libraries (data preprocessing, text processing, and machine learning)

**Code:**

!pip install pandas spacy scikit-learn joblib

!python -m spacy download en\_core\_web\_sm

1. Import kaggle.json

**Code:**

import os

!mkdir -p ~/.kaggle

!cp kaggle.json ~/.kaggle/

!chmod 600 ~/.kaggle/kaggle.json

1. Download datasets from kaggle

**Code:**

import kagglehub

# Download latest version

path = kagglehub.dataset\_download("shippii/multilingual-search-queries-mixed-datasets")

print("Path to dataset files:", path)

1. Connect to drive

**Code:**

from google.colab import drive

drive.mount('/content/drive')

1. Import datasets.crv
2. First datasets

**Code:**

import pandas as pd

df = pd.read\_csv('/content/drive/MyDrive/mixed\_data.csv')

print(df.head())

1. Second datasets

**Code:**

import pandas as pd

df = pd.read\_csv('/content/drive/MyDrive/search\_query\_data.csv')

print(df.head())

1. Filter datasets (English and Malay)

**Code:**

# Filter for Malay ('ms') and English ('en') rows

filtered\_df = df[df['lan\_code'].isin(['ms', 'en'])]

# Check the filtered data

print(filtered\_df['lan\_code'].value\_counts())

1. Text Preprocessing

**Code:**

filtered\_df = filtered\_df.copy()

filtered\_df['query'] = filtered\_df['query'].str.replace('[^a-zA-Z]', ' ', regex=True)

filtered\_df['query'] = filtered\_df['query'].str.lower()

1. Save datasets file

**Code:**

filtered\_df.to\_csv('Language( Malay & English )\_dataset.csv', index=False)