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
   ```

2. Import kaggle.json

   Code:

   ```
   import os
   !mkdir -p ~/.kaggle
   !cp kaggle.json ~/.kaggle/
   !chmod 600 ~/.kaggle/kaggle.json
   ```

3. 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)
   ```

4. Connect to drive

   Code:

   ```
   from google.colab import drive
   drive.mount('/content/drive')
   ```

5. Import datasets.crv

   (i) First datasets

   Code:

   ```
   import pandas as pd
   df = pd.read_csv('/content/drive/MyDrive/mixed_data.csv')
   print(df.head())
   ```

   (ii) Second datasets

Code:

```
import pandas as pd
df = pd.read_csv('/content/drive/MyDrive/search_query_data.csv')
print(df.head())
```

6. 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())
```

7. 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()
```

8. Save datasets file

   Code:

```
filtered_df.to_csv('Language( Malay & English )_dataset.csv', index=False)
```