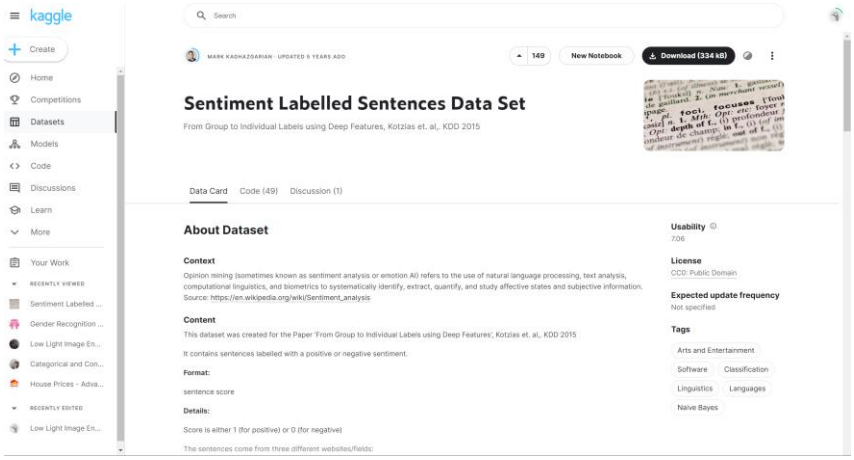
Soal

Untuk data yang ada pada link: https://www.kaggle.com/marklvl/sentiment-labelled-sentences-data-set, silahkan diolah dengan menggunakan text processing sehingga menghasilkan data bersih. Dikerjakan dengan jupyter notebook pada komputer ataupun google colaboratory, dan upload hasil pengerjaan dalam bentuk ipynb, rar, ataupun pdf (yang penting bisa terbaca oleh penilai).

Note : Untuk mengubah file ipynb ke pdf dapat lebih mudah dilakukan dengan menggunakan google colaboratory.

Link dataset

https://www.kaggle.com/datasets/marklvl/sentiment-labelled-sentences-data-set



Source Code

```python
import pandas as pd
import re
import string
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

# File path of the dataset CSV file
dataset_file = "C:/Users/LEGION 5 PRO/OneDrive/Documents/Semester 4/Temu Kembali Informasi/dataset 14/sentiment labelled sentences/datasets.csv"

# Define the regex pattern to remove punctuation and replace with whitespace
regex = re.compile('[%s]' % re.escape(string.punctuation.replace('.,', '')))

# Define stopwords
nltk.download('stopwords')
stop_words = set(stopwords.words('english'))

def preprocess_text(text):
    # Replace "1", "0", ",0,,,", and ",1,,,," with whitespace
    text = re.sub(r'\b[01]\b', ' ', text)
    text = re.sub(r',0,,,,,', ' ', text)
    text = re.sub(r',1,,,,,', ' ', text)

    # Tokenize the text
    tokens = word_tokenize(text)

    # Remove stopwords and punctuation
    tokens = [token for token in tokens if token.lower() not in stop_words and token not in string.punctuation]

    # Join the tokens back into a string
    preprocessed_text = ' '.join(tokens)

    return preprocessed_text

# Read the dataset from the CSV file
dataset_files = pd.read_csv(dataset_file).fillna('')

# Apply preprocessing to each column in the dataframe
for column in dataset_files.columns:
    dataset_files[column] = dataset_files[column].apply(preprocess_text)

# Save the preprocessed dataset as a new CSV file
output_file = "C:/Users/LEGION 5 PRO/OneDrive/Documents/Semester 4/Temu Kembali Informasi/dataset 14/sentiment labelled sentences/preprocessed_dataset.csv"
dataset_files.to_csv(output_file, index=False)

print("Datasets preprocessed and saved to:", output_file)
```

Penjelasan

Kode ini adalah contoh implementasi dari proses pra-pemrosesan teks pada dataset menggunakan library Pandas dan NLTK (Natural Language Toolkit).

Pada awalnya, library yang diperlukan diimpor, seperti Pandas untuk membaca dan menyimpan dataset dalam format CSV, re untuk melakukan operasi pemrosesan teks menggunakan ekspresi reguler, string untuk mendefinisikan kumpulan tanda baca yang akan dihapus dari teks, nltk (Natural Language Toolkit) untuk melakukan pemrosesan teks seperti tokenisasi, dan nltk.corpus untuk mengakses kamus stopwords.

Kemudian, path file dataset ditentukan untuk mengakses file CSV yang berisi dataset. Selanjutnya, pola regex didefinisikan untuk menghapus tanda baca dari teks dan menggantikannya dengan spasi, kecuali tanda baca '.' dan ','.

Pada tahap selanjutnya, kamus stopwords dalam bahasa Inggris diunduh menggunakan nltk.download('stopwords'). Stopwords adalah kata-kata yang umumnya tidak memberikan informasi penting dalam pemrosesan teks.

Setelah itu, sebuah fungsi bernama preprocess_text dibuat untuk melakukan proses pra-pemrosesan teks. Fungsi ini melakukan beberapa langkah, seperti menggantikan angka "1" dan "0" dengan spasi, menggantikan ",0,,,," dengan spasi, dan menggantikan ",1,,,," dengan spasi. Kemudian, teks di-tokenisasi menggunakan word_tokenize, yaitu membaginya menjadi token-token kata. Stopwords dan tanda baca dihapus dari token-token tersebut. Akhirnya, token-token digabungkan kembali menjadi string teks.

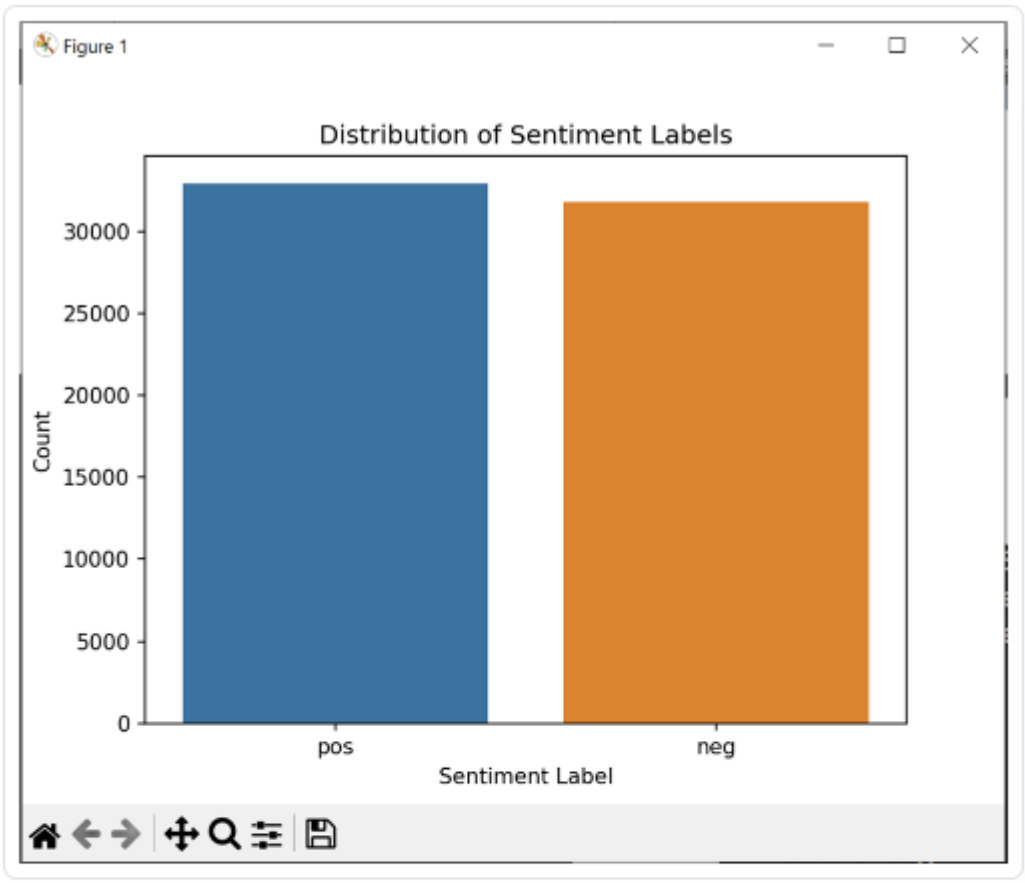
Dataset dibaca dari file CSV menggunakan Pandas, dan nilai-nilai kosong dalam dataset diisi dengan string kosong. Selanjutnya, proses pra-pemrosesan dilakukan pada setiap kolom dalam dataframe. Looping digunakan untuk mengiterasi melalui setiap kolom dataset, dan fungsi preprocess_text dipanggil untuk setiap nilai dalam kolom dataset. Hasilnya disimpan dalam kolom yang sama.

Dataset yang telah diproses kemudian disimpan sebagai file CSV baru. Path file output ditentukan, dan dataframe yang telah diproses disimpan ke file CSV baru dengan menggunakan metode to_csv dari Pandas. Indeks tidak disertakan dalam file output.

Terakhir, sebuah pesan konfirmasi dicetak untuk menginformasikan bahwa dataset telah diproses dan disimpan dalam file CSV baru. Dengan demikian, kode ini melakukan proses pra-pemrosesan teks pada dataset dengan menghapus stopwords dan tanda baca, dan menyimpan dataset yang telah diproses dalam file CSV baru.

Output

Bukti Upload Portofolio



Sentiment analysis in movie_review dataset

Course: Basic Text Processing

Menunggu Approval

Sertifikat

2023/221/19659

CERTIFICATE

OF APPRECIATION

Sertifikat ini diberikan kepada :

HEYDAR EMIR ALVARO

Telah Menyelesaikan Pembelajaran Free Course Mengenai Materi

Basic Text Processing

June 12, 2023

M. Octaviano Pratama S.Kom., M.Kom

President Director Of BISA AI ACADEMY