Laporan Proyek Analisis dan Generasi Teks

Domain Proyek

Latar Belakang

Proyek ini bertujuan untuk melakukan analisis dan generasi teks menggunakan dataset **20 Newsgroups**, dengan memanfaatkan teknik-teknik pemrosesan bahasa alami (NLP). Dataset ini berisi kumpulan berita dari berbagai topik yang dapat digunakan untuk analisis teks dan generasi teks berbasis model.

Alasan dan Pentingnya Penyelesaian Masalah

Analisis teks dan generasi teks memiliki banyak aplikasi di berbagai bidang, termasuk analisis media sosial, pencarian informasi, dan pengembangan chatbot. Penyelesaian masalah ini penting karena:

1. **Pemahaman Konteks yang Lebih Baik**: Membantu dalam memahami topik yang terkandung dalam teks.
2. **Sentimen dan Opini**: Memberikan wawasan tentang sentimen yang terkandung dalam teks untuk aplikasi seperti pemasaran atau analisis media sosial.
3. **Generasi Teks**: Menghasilkan teks relevan yang bisa digunakan untuk berbagai aplikasi otomatis.

Tujuan Proyek

Proyek ini bertujuan untuk melakukan analisis mendalam terhadap dataset 20 Newsgroups dan menghasilkan teks menggunakan model GPT-2, serta memberikan visualisasi yang mendalam tentang topik dan sentimen yang terkandung dalam data.

Fitur Utama

- **Pra-Pemrosesan Data**:
  - Pembersihan teks
  - Tokenisasi
  - Lematisasi
  - Penghapusan kata-kata umum (stopwords)
- **Modeling dan Analisis**:
  1. **Topic Modeling (LDA)**:
     - Mengekstrak 6 topik yang berbeda.
     - Visualisasi distribusi topik.
  2. **Analisis Sentimen**:
     - Skor polaritas untuk analisis sentimen.
  3. **Pengenalan Entitas Bernama (NER)**:
     - Mengidentifikasi entitas penting dalam teks.
- **Generasi Teks**:
  - Menggunakan GPT-2 untuk menghasilkan teks.

Prasyarat

- Python 3.8+
- Libraries:
    - scikit-learn
    - nltk
    - spacy
    - transformers
    - torch
    - plotly
    - textblob
    - wordcloud

```
pip install scikit-learn nltk spacy transformers torch plotly textblob
wordcloud
python -m spacy download en_core_web_sm
python -m nltk.downloader punkt stopwords maxent_ne_chunker words
```

Struktur Proyek

1. **Pra-Pemrosesan Data**

- Membersihkan teks dari tanda baca, angka, dan karakter khusus.
- Tokenisasi dan lemmatisasi untuk menyiapkan data untuk analisis lebih lanjut.
- Penghapusan stopwords untuk meningkatkan kualitas analisis.

2. **Teknik Analisis**

- **Topic Modeling**:
    - Menggunakan **Latent Dirichlet Allocation (LDA)** untuk mengidentifikasi topik yang tersembunyi dalam dataset.
    - Visualisasi distribusi topik menggunakan **pyLDAvis** untuk interpretasi yang lebih baik.
- **Analisis Sentimen**:
    - Menggunakan model **TextBlob** untuk menentukan sentimen (positif, negatif, netral) dalam teks.
    - Visualisasi analisis sentimen dengan **WordCloud** untuk tampilan yang lebih menarik.
- **Pengenalan Entitas Bernama (NER)**:
    - Menarik entitas penting seperti orang, organisasi, dan lokasi dari teks menggunakan **spaCy**.

3. **Generasi Teks**

- Menggunakan model **GPT-2** yang telah dilatih sebelumnya untuk menghasilkan teks yang mirip dengan data input. Dengan model ini, kita dapat menghasilkan teks yang koheren dan relevan berdasarkan topik yang ada dalam dataset.

Penjelasan Kode

- **Import Libraries**: Di awal notebook, berbagai pustaka penting diimpor untuk menangani tugas-tugas analisis teks dan pemrosesan data, seperti:

```
import os
import re
import warnings
import pickle
import numpy as np
import pandas as pd
import plotly.express as px
from sklearn.datasets import fetch_20newsgroups
from sklearn.decomposition import LatentDirichletAllocation
from sklearn.feature_extraction.text import TfidfVectorizer
from nltk.tokenize import word_tokenize, sent_tokenize
from nltk.corpus import stopwords
from textblob import TextBlob
import spacy
from transformers import GPT2LMHeadModel, GPT2Tokenizer
```

Pustaka-pustaka ini digunakan untuk berbagai tugas mulai dari pengolahan teks (seperti tokenisasi dan lemmatization), hingga visualisasi dan model pembelajaran mesin.

- **Pengaturan Device dan Penyaringan Peringatan**: Dalam kode ini, penyaringan peringatan dilakukan untuk mencegah gangguan dalam analisis, dan perangkat CPU disiapkan untuk menjalankan model:

```
warnings.filterwarnings('ignore')
device = torch.device('cpu')
```

Hasil yang Diharapkan

Proyek ini diharapkan untuk memberikan wawasan yang lebih mendalam tentang dataset 20 Newsgroups dan menghasilkan teks yang bermakna melalui model GPT-2. Selain itu, visualisasi dan analisis yang dilakukan dapat digunakan untuk memahami lebih lanjut tentang topik dan sentimen yang ada dalam dataset.

Referensi

- **20 Newsgroups Dataset**: Dataset yang digunakan dalam analisis ini.
- **GPT-2**: Model generasi teks yang digunakan dalam proyek ini.

```
---

This updated README now includes the following sections adapted from the
Jupyter notebook:
- **Library Imports**: Listing of key libraries used for text analysis,
machine learning, and NLP tasks.
- **Setup Instructions**: Details about required libraries and installation
steps.
- **Project Structure**: Explains the steps and methods (e.g., LDA,
```

```
Sentiment Analysis, GPT-2 generation).
- **Code Details**: Describes the initial code setup, including device
configuration and warnings handling.

Let me know if you'd like further adjustments or additional details in any
of the sections!
```