Week 5 - AI Mentorship Program

NLP and Advanced Topic

Google Developer Group
Jogjakarta

Outline

1. Progress Mini Project
2. NLP
3. Demo NLP

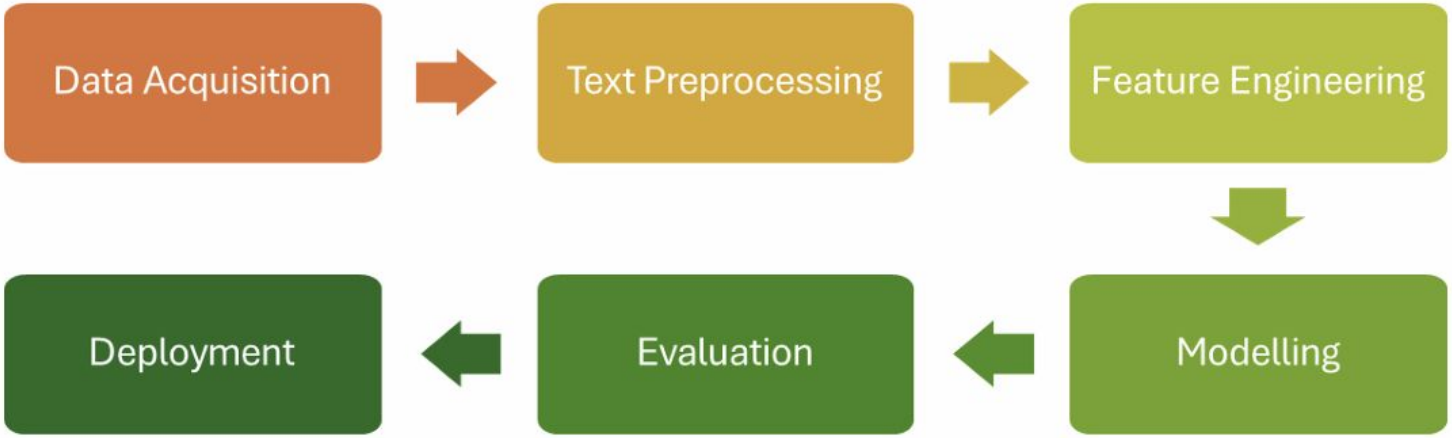
Progress Mini Project

NLP

Google Developer Group
Jogjakarta

- NLP adalah cabang kecerdasan buatan yang berfokus pada interaksi antara komputer dan bahasa manusia. Contohnya termasuk:
- Tokenisasi: Memecah teks menjadi kata atau karakter.
- Bag-of-Words (BoW): Merepresentasikan teks dalam bentuk vektor kata.
- TF-IDF: Mengukur seberapa penting sebuah kata dalam dokumen tertentu dibandingkan keseluruhan korpus.

NLP Pipeline



Google Developer Group
Jogjakarta

NLP Pipeline

Google Developer Group Jogjakarta

1. Data Acquisition

Ini adalah langkah pertama dalam alur NLP di mana data teks mentah dikumpulkan dari sumber seperti situs web, API, database, atau konten buatan pengguna.

Tujuannya adalah untuk mengumpulkan data yang relevan khusus untuk masalah yang sedang dipecahkan.

Ini sering melibatkan penanganan berbagai format (JSON, CSV, atau teks biasa) dan memastikan penggunaan etis dan kepatuhan terhadap undang-undang privasi data.

2. Text Preprocessing

Pada langkah ini, data mentah yang dikumpulkan dibersihkan dan disiapkan untuk dianalisis.

Ini termasuk menghilangkan kebisingan (misalnya, tanda baca, tag HTML, dan kata berhenti), tokenisasi (memisahkan teks menjadi unit yang lebih kecil seperti kata atau kalimat), dan normalisasi (misalnya, mengonversi ke huruf kecil, lemmatisasi, dan stemming).

Prapemrosesan memastikan bahwa teks terstruktur, konsisten, dan siap untuk ekstraksi fitur.

3. Feature Engineering

Feature engineering Mengubah teks yang telah diproses sebelumnya menjadi representasi numerik yang dapat ditafsirkan oleh model pembelajaran mesin.

Tekniknya meliputi Bag-of-Words, TF-IDF, penyematan kata (misalnya, Word2Vec atau GloVe), dan N-gram.

Fitur tambahan, seperti skor sentimen atau entitas bernama, juga dapat disertakan untuk meningkatkan representasi konteks dan semantik teks.

4. Modelling

Langkah ini melibatkan pemilihan, pelatihan, dan penyempurnaan model pembelajaran mesin atau pembelajaran mendalam untuk melakukan tugas NLP yang diinginkan, seperti analisis sentimen, klasifikasi teks, atau terjemahan mesin.

Algoritme seperti Naive Bayes, Mesin Vektor Pendukung, atau model transformator canggih seperti BERT digunakan tergantung pada kompleksitas dan persyaratan tugas.

5. Evaluation

Setelah melatih model, penting untuk mengevaluasi kinerjanya menggunakan metrik seperti akurasi, presisi, penarikan, dan skor F1.

Evaluasi memastikan model menggeneralisasi dengan baik ke data yang tidak terlihat dan membantu mengidentifikasi area yang perlu ditingkatkan.

Matriks validasi silang dan kebingungan sering digunakan selama fase ini.

6. Deployment

Pada langkah terakhir, model terlatih disebarkan ke lingkungan produksi untuk melayani aplikasi dunia nyata. Ini melibatkan pembuatan API, menyematkan model ke dalam aplikasi, atau menghostingnya di platform cloud.

Pemantauan dan pembaruan rutin diperlukan untuk mempertahankan kinerja model dan beradaptasi dengan data yang berkembang.

Tokenisasi

- Tokenisasi adalah proses membagi teks menjadi unit-unit kecil seperti kata atau karakter.

Google Developer Group
Jogjakarta

```
import nltk
from nltk.tokenize import
word_tokenize
nltk.download('punkt')

text = "Natural Language Processing
is amazing!"
tokens = word_tokenize(text)
print(tokens)
```

Bag of Words

Bag-of-Words (BoW) : Representasi teks dalam bentuk vektor berdasarkan frekuensi kata.

Google Developer Group
Jogjakarta

```
from sklearn.feature_extraction.text
import CountVectorizer
corpus = ["Saya suka belajar NLP",
"NLP sangat menarik"]
vectorizer = CountVectorizer()
X = vectorizer.fit_transform(corpus)
print(X.toarray())
print(vectorizer.get_feature_names_ou
t())
```

TF-IDF

Google Developer Group
Jogjakarta

- TF-IDF (Term Frequency - Inverse Document Frequency) adalah Mengukur pentingnya sebuah kata dalam suatu dokumen relatif terhadap seluruh korpus.

```
from sklearn.feature_extraction.text
import TfidfVectorizer
vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(corpus)
print(X.toarray())
```

Transfer learning

Transfer learning memungkinkan kita menggunakan model yang telah dilatih sebelumnya untuk tugas baru dengan fine-tuning. Ini mempercepat pelatihan dan meningkatkan akurasi karena model sudah memahami fitur dasar dari data.

Google Developer Group
Jogjakarta

```
# Load model ResNet yang telah dilatih
model = models.resnet18(pretrained=True)
model.eval()

# Transformasi gambar agar sesuai dengan input model
transform = transforms.Compose([
    transforms.Resize((224, 224)),
    transforms.ToTensor(),
])
image = Image.open("example.jpg")
image = transform(image).unsqueeze(0)
```

Optimasi Model dan Interpretasi Hasil

Google Developer Group Jogjakarta

Hyperparameter tuning adalah proses mencari kombinasi parameter terbaik untuk meningkatkan performa model machine learning. Contohnya GridSearchCV, yang mencoba berbagai kombinasi parameter seperti learning rate, jumlah layer, atau jenis kernel dalam SVM. Tujuannya adalah menemukan konfigurasi optimal tanpa melakukan overfitting atau underfitting.

```
from sklearn.model_selection import GridSearchCV
from sklearn.svm import SVC

parameters = {'C': [0.1, 1, 10], 'kernel': ['linear', 'rbf']}
svc = SVC()
grid_search = GridSearchCV(svc, parameters, cv=5)
grid_search.fit(X, [1, 0])
print(grid_search.best_params_)
```

Demo

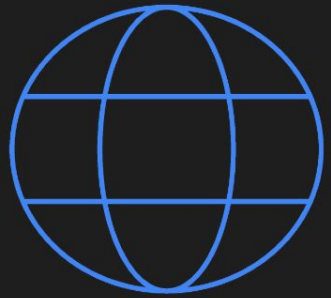
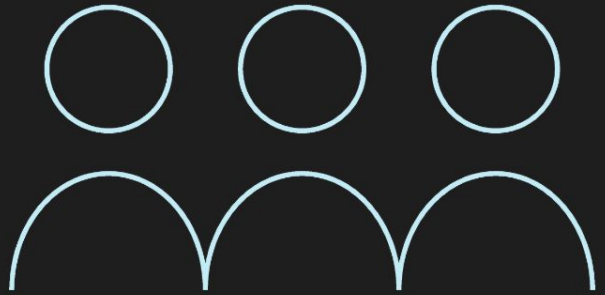
Google Developer Group
Jogjakarta

- Scrap berita sebagai data
- Lakukan pemprosesan terhadap data tersebut
- Buatlah model clustering

Summary

Google Developer Group
Jogjakarta

1. NLP memungkinkan analisis teks dengan berbagai metode seperti tokenisasi, BoW, dan TF-IDF.
2. Transfer learning memanfaatkan model pretrained untuk meningkatkan performa tanpa perlu melatih dari awal.
3. Optimasi model sangat penting untuk meningkatkan akurasi prediksi.
4. Implementasi langsung dilakukan untuk klasifikasi teks dan gambar dengan model yang telah tersedia.

Thank You

See You Next Week!

Google Developer Group
Jogjakarta