Indonesia AI
AI for Everyone, AI for Indonesia

#2023JAGOTEKNOLOGIAI

AI Career Bootcamp

POS Tagging and Named-Entity Recognition

Indonesia AI (PT. Teknologi Artifisial Indonesia)

Pembukaan

Guide Book ini memuat beberapa informasi-informasi utama yang akan disampaikan oleh mentor di program AI Career Bootcamp yang bisa dijadikan pegangan para students untuk mempersiapkan diri sebelum sesi Live Class berlangsung.

Deskripsi

Text Processing adalah serangkaian teknik, metode dan algoritma yang digunakan untuk memanipulasi, menganalisis, dan memahami teks secara komputasional. Ini melibatkan pemecahan teks menjadi unit-unit yang lebih kecil seperti kata-kata, kalimat, atau entitas, serta menerapkan metode statistik, pembelajaran mesin, dan pendekatan linguistik untuk memperoleh informasi yang berguna dari teks tersebut.

POS Tagging dan NER adalah contoh spesifik dari teknik pemrosesan teks. POS tagging melibatkan penempatan tag gramatikal pada kata-kata dalam teks, sementara NER melibatkan identifikasi entitas bernama dalam teks. Keduanya membutuhkan pemahaman linguistik dan menggunakan pendekatan berbasis aturan atau berbasis statistik untuk melakukan tugas-tugas ini.

Dalam pengolahan bahasa alami dan analisis teks, POS Tagging dan NER adalah langkah-langkah penting untuk memahami struktur, makna, dan entitas dalam teks yang lebih besar. Mereka membantu dalam pengenalan konteks dan pengekstrakan informasi yang relevan, dan dengan

Mengenal POS Tagging

POS (Part-of-Speech) tagging, juga dikenal sebagai penandaan kata berdasarkan kelas kata, adalah proses untuk memberikan label atau tag kepada setiap kata dalam sebuah teks berdasarkan peran dan fungsi gramatikalnya dalam kalimat. Tujuan dari POS tagging adalah untuk mengidentifikasi kelas kata (seperti kata benda, kata kerja, kata sifat, dll.) serta informasi sintaksis lainnya (seperti kata ganti orang, bentuk waktu, kata penghubung, dll.) dalam teks.

POS tagging penting dalam pemrosesan bahasa alami (natural language processing) karena membantu dalam memahami struktur dan makna kalimat. Dengan mengetahui kelas kata setiap

kata, kita dapat memahami hubungan antara kata-kata dalam konteks kalimat dan menerapkan analisis lebih lanjut seperti analisis sintaksis, ekstraksi informasi, pemahaman teks, dan terjemahan mesin.

Proses POS tagging biasanya melibatkan penggunaan model statistik atau model berbasis aturan (rule-based). Model statistik menggunakan teknik pembelajaran mesin untuk mempelajari pola dari teks berlabel yang sudah ada, sedangkan model berbasis aturan mengandalkan aturan linguistik yang telah ditentukan sebelumnya. Model-model ini menggunakan kumpulan data pelatihan yang berisi teks-teks berlabel untuk mengenali pola dan mengasosiasikan kata-kata dengan kelas kata yang sesuai.

Contoh hasil POS tagging untuk sebuah kalimat sederhana adalah sebagai berikut:

"Saya sedang membaca buku di perpustakaan."

POS Tagging:

- Saya (Pronoun)
- sedang (Verb)
- membaca (Verb)
- buku (Noun)
- di (Preposition)
- perpustakaan (Noun)

Dalam contoh di atas, setiap kata diberikan tag yang sesuai dengan perannya dalam kalimat.

POS tagging memiliki banyak aplikasi dalam bidang pemrosesan bahasa alami, seperti pemahaman teks, ekstraksi informasi, analisis sentiment, pencarian informasi, dan terjemahan mesin. Dengan informasi POS tagging, komputer dapat memahami struktur bahasa manusia dengan lebih baik dan dapat digunakan untuk membangun berbagai sistem yang berhubungan dengan pemrosesan bahasa alami.

Mengenal NER

NER (Named Entity Recognition), juga dikenal sebagai pengenalan entitas bernama, adalah tugas dalam pemrosesan bahasa alami (natural language processing) yang bertujuan untuk mengidentifikasi dan mengklasifikasikan entitas bernama dalam teks, seperti nama orang, nama tempat, nama organisasi, tanggal, jumlah, dan sebagainya.

NER merupakan bagian penting dalam pemrosesan bahasa alami karena entitas bernama sering kali mengandung informasi penting dalam teks. Dengan mengenali dan mengklasifikasikan entitas bernama, kita dapat memahami konteks dan makna teks secara lebih baik serta menerapkan analisis lebih lanjut seperti ekstraksi informasi, pemahaman teks, dan analisis sentiment.

Proses NER melibatkan pendekatan berbasis aturan (rule-based) atau pendekatan berbasis statistik. Pendekatan berbasis aturan mengandalkan aturan linguistik dan heuristik untuk mengidentifikasi entitas bernama berdasarkan pola tertentu, seperti kata yang diawali dengan huruf kapital. Pendekatan berbasis statistik menggunakan teknik pembelajaran mesin untuk melatih model yang dapat mengenali entitas bernama berdasarkan contoh-contoh teks yang sudah diberi label.
Contoh hasil NER untuk sebuah kalimat sederhana adalah sebagai berikut:

> “Mark Zuckerberg adalah pendiri Facebook yang berasal dari Amerika Serikat.”

NER

- Mark Zuckerberg (Nama Orang)
- Facebook (Organisasi)
- Amerika Serikat (Lokasi)

Dalam contoh di atas, Named Entity seperti nama orang (Mark Zuckerberg), organisasi (Facebook), dan lokasi (Amerika Serikat) diidentifikasi dan diklasifikasikan sesuai dengan kategori yang sesuai.

NER memiliki berbagai aplikasi dalam pemrosesan bahasa alami, seperti ekstraksi informasi, pemodelan pengetahuan, analisis teks, analisis sosial media, dan banyak lagi. Dengan mengenali entitas bernama, komputer dapat memahami dan mengolah teks dengan lebih baik, serta menghasilkan informasi yang lebih kaya dan terstruktur.