LEARNING PROGRESS REVIEW

WEEK 10 - BY CHARLIE'S ANGELS

WEB SCRAPING

MACHINE LEARNING

DigitalSkola

Web Scraping

Web Scraping adalah metode otomatis untuk mendapatkan data dalam jumlah besar dari sebuah situs web. Sebagian besar data tersebut merupakan data yang tidak terstruktur dalam format HTML, kemudian diubah menjadi data terstruktur dalam spreadsheet atau database sehingga dapat digunakan dalam berbagai aplikasi.

Manfaat Web Scraping

- Mendapatkan Leads
- Mendalami Kebutuhan Konsumen dari Kompetitor
- Mencari Informasi
- Analisa kompetitor dan pasar
- Memantau Berita dan Konten

Teknik-Teknik Web Scraping

Menyalin Data Manual

Cara ini merupakan cara paling sederhana dalam web scraping adalah dengan menyalin data laman secara manual. Teknik ini perlu mengambil dan menyimpan informasi yang diperlukan satu per satu.

Menggunakan Regular Expression

Regular expression merupakan baris kode yang digunakan dalam algoritma pencarian untuk menemukan data spesifik dari sebuah file. Dalam konteks ini, file yang dimaksud adalah file yang dapat digunakan untuk menunjang sebuah website.

Parsing HTML

Biasanya teknik ini dilakukan melalui JavaScript serta menargetkan halaman HTML linear dan bercabang. Metode ini lebih efisien dalam mengidentifikasi script HTML dari website yang kemudian digunakan untuk mengekstraksi teks, tautan, dan data.

Menggunakan Text Pattern Matching

Teknik ini menggunakan UNIX grep command dan bahasa program seperti Python dengan mencocokan ekspresi regular.

Teknik-Teknik Web Scraping

Parsing DOM

DOM adalah singkatan dari Document Object Model, representasi struktur sebuah halaman website yang ditulis dengan XML dan HTML. Ketika melakukan parsing HTML, DOM dari halaman yang ingin diekstrak dimuat terlebih dahulu. Ini dapat digunakan apabila ingin mengetahui cara kerja internal halaman web.

Menggunakan XPath

XPath adalah bahasa query yang digunakan untuk memilih node dari struktur file XML dan HTML. Implementasinya tidak jauh berbeda dengan analisa DOM, dengan tujuan mencari data dari struktur file penunjang halaman.

Menggunakan Google Sheet

Selain google sheet, diperlukan juga browser yang memiliki fitur inspect element. Setelah itu, cukup mengcopy expression XPath dari elemen halaman website yang datanya ingin disalin ke dalam command IMPORT XML yang ada di google sheet.

Library Python untuk Web Scraping

LIBRARY REQUESTS (HTTP FOR HUMANS)

Library Requests merupakan library python yang paling dasar untuk web scraping. Request HTML dibuat untuk mengambil data dari halaman website dengan mengirimkan request ke server situs web. LIbrary Requests pada Python digunakan untuk membuat berbagai jenis permintaan HTTP seperti POST, GET, dan lain sebagainya.

LIBRARY LXML DAN BEAUTIFUL SOUP

Library lxml berfokus pada parsing HTML yang mengambil data dari halaman web. Fitur unik dari perpustakaan ini melibatkan kinerja tinggi, produksi HTML, dan penguraian XML yang lebih cepat. Beautiful Soup adalah library Python yang paling banyak digunakan untuk project web scraping. Salah satu alasan paling dasar mengapa menggunakan library ini adalah karena relatif lebih mudah untuk pemula.

LIBRARY SCRAPY

Library ini digunakan untuk project web scraping besar. Keuntungan menggunakan library ini adalah library ini bersifat asynchronous, dokumentasi yang lebih baik, bisa menggunakan berbagai plugin, bisa dikombinasikan dengan pipeline dan middlewares buatan.

Kendala Web Scraping

Tidak ada teknik web scraping yang 100% efektif	Data yang didapat tidak selalu rapi
Pemahaman tentang struktur halaman website tetap menjadi kewajiban	Akses ke suatu website dapat diblokir

Tidak semua website mudah diekstrak datanya

CLEANING DATA

- Data celaning merupakan suatu prosedur untuk memastikan kebenaran, konsistensi, dan kegunaan suatu data yang ada dalam dataset
- Data cleaning termasuk bagian **Transform** dalam proses **ETL**
- Berikut beberapa hal yang biasanya dilakukan pada proses cleaning
  - Drop atau menghapus kolom yang tidak relevan
  - Rename kolom
  - Merubah format kolom
  - Mengganti/menghitung nilai yang hilang

STORING DATA TO DATABASE

- Proses penyimpanan data ke database termasuk dalam proses **Load** pada ETL
- Load data dapat juga dilakukan kedalam bentuk selain database. Seperti file csv/parquet/etc
- Dalam proses web scraping dengan python, diperlukan SQLAlchemy untuk storing data to database

ETL PROCESS IN WEB SCRAPING

Dalam proses end to end web scraping, berikut rincian proses ETL yang terjadi :

- Scraping data from web: termasuk kedalam proses extract data
- Cleaning dataframe: data yang telah di scrape diubah menjadi dataframe dan dilakukan cleaning untuk merapihkan column, proses ini masuk ke proses transform
- Storing data to database: termasuk kedalam proses load data. Dataframe yang telah di-*cleaning* disimpan ke dalam database atau bisa juga diubah ke bentuk csv

Machine Learning

Kecerdasan buatan atau yg sering kita dengar dengan istilah Artificial Intelligence (AI) sebenarnya terdiri dari beberapa cabang, salah satunya Machine Learning (ML) atau pembelajaran mesin. Teknologi Machine Learning merupakan salah satu cabang AI yang sangat menarik perhatian karena Machine Learning merupakan mesin yang belajar selayaknya manusia

Teknologi machine learning (ML) adalah mesin yang dikembangkan untuk bisa belajar dengan sendirinya tanpa arahan dari penggunanya. Pembelajaran mesin dikembangkan berdasarkan disiplin ilmu lainnya seperti statistika, matematika dan data mining sehingga mesin dapat belajar dengan menganalisa data tanpa perlu di program ulang atau diperintah.

Machine Learning

Beberapa contoh program machine learning yang telah digunakan dalam kehidupan sehari-hari:

- Pendeteksi Spam
- Pendeteksi Wajah
- Rekomendasi Produk
- Asisten Virtual
- Diagnosa Medis
- Pendeteksi Penipuan Kartu Kredit
- Pengenal Digit
- Perdagangan Saham
- Segmentasi Pelanggan
- Mobil yang bisa Mengendarai Sendiri

Teknik-Teknik Belajar Machine Learning

Supervised Learning

Teknik supervised learning merupakan teknik yang bisa diterapkan pada pembelajaran mesin yang bisa menerima informasi yang sudah ada pada data dengan memberikan label tertentu. Diharapkan teknik ini bisa memberikan target terhadap output yang dilakukan dengan membandingkan pengalaman belajar di masa lalu.

Unsupervised Learning

Teknik unsupervised learning merupakan teknik yang bisa diterapkan pada machine learning yang digunakan pada data yang tidak memiliki informasi yang bisa diterapkan secara langsung. Diharapkan teknik ini dapat membantu menemukan struktur atau pola tersembunyi pada data yang tidak memiliki label.

Teknik-Teknik Belajar Machine Learning

Semi-supervised learning

Semi-supervised learning menawarkan jalan alternatif antara supervised dan unsupervised learning. Selama pelatihan, semi-supervised learning menggunakan kumpulan data berlabel yang lebih kecil untuk memandu klasifikasi dan ekstraksi fitur dari kumpulan data yang lebih besar dan tidak berlabel

Reinforcement Learning

Teknik supervised learning merupakan teknik yang bisa diterapkan pada pembelajaran mesin dengan menggunakan agent yang terus menerus belajar di dalam sebuah environment

Cara kerja Machine Learning

pada dasarnya prinsip cara kerja pembelajaran mesin masih sama

- pengumpulan data
- eksplorasi data
- pemilihan model atau teknik,(regresi linear, regresi logistik, neural network, dll)
- memberikan pelatihan terhadap model yang dipilih
- mengevaluasi hasil dari ML
- Prediksi

Bias and Variance Tradeoff-Measurements

BIAS

ADALAH KEADAAN DIMANA MODEL PELATIHAN DATA YANG DIBUAT TIDAK MEWAKILKAN KESELURUHAN DATA YANG AKAN DIGUNAKAN NANTINYA. SEHINGGA MENGHASILKAN PERFORMA YANG BURUK DALAM PELATIHAN DATA. UNDERFITTING TERJADI KARENA MODEL MASIH MEMPELAJARI STRUKTUR DARI DATA. HASILNYA, TREE BEKERJA DENGAN BURUK PADA MASA PELATIHAN DAN TES. SEBAGAIMANA BANYAKNYA NODE DALAM POHON KEPUTUSAN MENINGKAT, TREE MEMILIKI GALAT PELATIHAN DAN TES YANG LEBIH KECIL. PADA SAAT TREE BERUKURAN SANGAT BESAR, TINGKAT TERJADINYA GALAT TES MULAI MENINGKAT WALAUPUN TINGKAT GALAT PELATIHANNYA TERUS MENURUN.

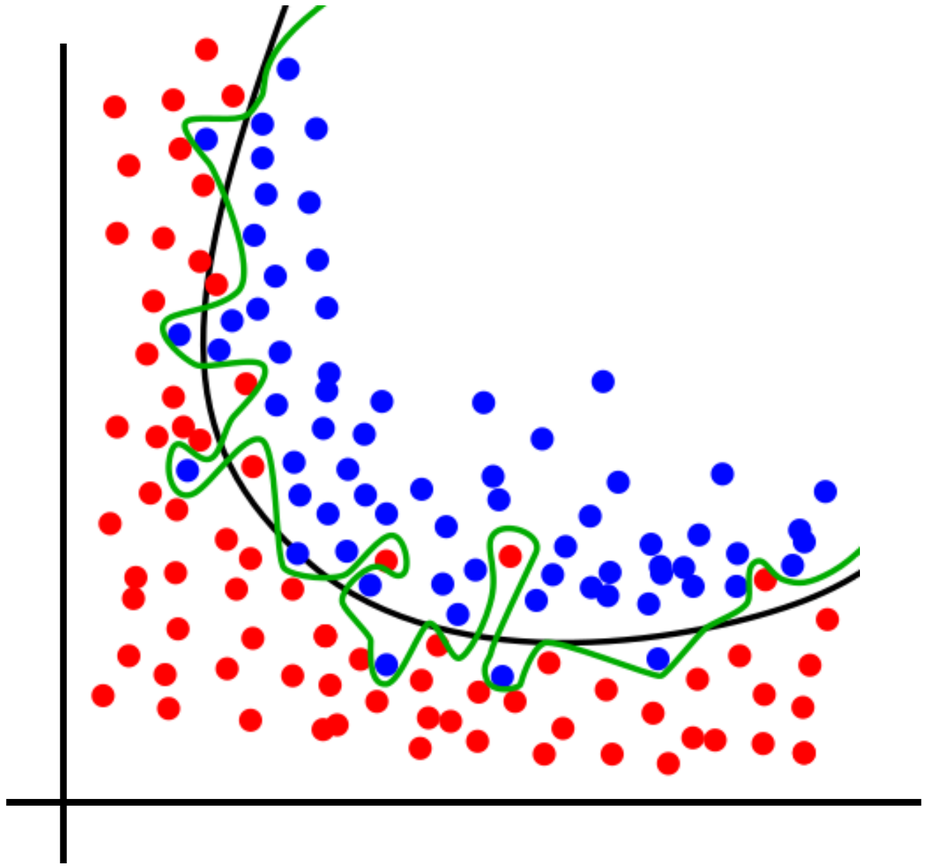
VARIANCE

ADALAH SUATU KEADAAN DIMANA DATA YANG DIGUNAKAN UNTUK PELATIHAN ITU ADALAH YANG "TERBAIK". SEHINGGA APABILA DILAKUKAN TES DENGAN MENGGUNAKAN DATA YANG BERBEDA DAPAT MENGURANGI AKURASI (HASIL YANG DIBUAT TIDAK SESUAI YANG DIHARAPKAN). OVERFITTING DAPAT TERJADI KETIKA BEBERAPA BATASAN DIDASARKAN PADA SIFAT KHUSUS YANG TIDAK MEMBUAT PERBEDAAN PADA DATA. SELAIN ITU DUPLIKASI DATA MINOR YANG BERLEBIHAN JUGA DAPAT MENGAKIBATKAN TERJADINYA OVERFITTING

Contoh Dari Grafik Bias and Variance

Garis hitam adalah contoh bias yang menggambarkan keseluruhan data.

Garis Hijau adalah contoh Variance yang menggambarkan Detail dari data.



TERIMA KASIH

Sampai jumpa di
Learning Progress Review
kami berikutnya!

HILDA
MEIRANITA
PRASTIKA DEWI

NUR
INDRASARI

REZHA
SULVIAN

THASHA DINYA
AINSHA

DigitalSkola