

9. Data dan Maklumat

9.0 Data Raya

9.1 Penanganan Data

9.2 Analisis Data

9.3 Dapatan Semula Maklumat

9.4 Pemvisualan

- Tugas 16

9.0 Data Raya

9.0.0 Asas

Isyarat ialah sebarang bentuk data, tenaga, atau mesej yang membawa maklumat. Sesuatu ukuran

Tugasan 16

1. Buat latihan pencucian data menggunakan Python dengan **pandas** dan **numpy** seperti di laman <https://realpython.com/python-data-cleaning-numpy-pandas/> atau yang serupa.

2. Algoritma **perangkak web** adalah seperti berikut:

```
awalkan senarai-senarai
input url benih
masukkan ke senarai belum lihat
rangkak:
    pop url teratas yg belum lihat => urlsemasa
    capai urlsemasa
    uraikan kandungan urlsemasa, cari url dirujuk dalamnya
    bagi setiap url ini, jika tidak ada dalam senarai dah dilihat,
        tolak masuk ke senarai belum dilihat
    proses kandungan urlsemasa apa yg nak dibuat
    tolak masuk urlsemasa ke dalam senarai dah dilihat
    kalau dah cukup, berhenti
rangkak
```

Tulis aturcara utk merangkak web mencari kekerapan nama anda dan uji menggunakan laman benih yg mungkin menyebutnya. Anda mungkin ingin memuat naik (e.g. guna **pip**) pakej *Requests* (untuk capaian web) dan *Beautiful Soup* (e.g. bs4) (utk mengurai dokumen html. Masukkan cetakan keluar di tempat-tempat strategik utk mencerminkan perjalanan aturcara anda.

3. Eksperimen perlombongan data *twitter* (anda mungkin perlu daftar sebagai pengguna *twitter* utk ini) menggunakan perpustakaan Python iaitu **Tweepy**.
Anda perlu daftar apps di <https://apps.twitter.com/>.
Import Tweepy dgn maklumat kekunci rahsia anda, dsb.

Cipta objek API anda, misalnya:

```
# Cipta objek kuasa
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
# Set token dan rahsia capaian anda
auth.set_access_token(access_token, access_token_secret)
# Cipta objek API sementara menyerahkan maklumat kuasa
api = tweepy.API(auth)
```

Cuba, misalnya, mencari mesej2 yg ada perkataan “reformasi”, e.g.

```
# nak cari
query = "reformasi"
# kod bahasa (menurut ISO 639-1)
language = "ms"
# panggil fungsi garis masa pengguna
hasil = api.search(q=query, lang=language)
# utk setiap tweet ditarik
for tweet in hasil:
    # cetak teks
    print tweet.user.screen_name,"Tweet:",tweet.text
```