

Modul 1 Praktikum Natural Language Processing Pengenalannya Pengolahan Data Teks menggunakan Python

Deskripsi Singkat

Modul ini akan membahas pengenalan pengolahan data teks menggunakan bahasa Python. Modul ini terdiri dari 5 materi yang akan membahas pengumpulan data menggunakan API, membaca data dari berbagai format file yang ada seperti PDF, Word, JSON, dan HTML.

Tujuan

1. Dapat mengumpulkan data menggunakan melalui API yang tersedia di internet.
2. Dapat membaca data dari berbagai format file yang ada seperti PDF, Word, JSON, dan HTML.

Materi 1 - Pengumpulan Data

Studi Kasus

Anda ingin mengumpulkan data menggunakan API Twitter

Langkah-langkah

1. Mendapatkan Bearer Token pada halaman Portal Developer Twitter
<https://developer.x.com/en/portal/>

2. Menginstall library tweepy

```
pip install tweepy
```

3. Import library yang dibutuhkan

```
1 import tweepy
```

4. Memasukkan bearer token sebagai akses token ke API Twitter

```
1 consumer_key = "your_consumer_key"
2 consumer_secret = "your_consumer_secret"
3 access_token = "your_access_token"
4 access_token_secret = "your_access_token_secret"
5 bearer_token = 'your_bearer_token'
```

5. Membuat client ke API Twitter

```
1 client = tweepy.Client(bearer_token=bearer_token, consumer_key
    =consumer_key, consumer_secret=consumer_secret,
    access_token=access_token, access_token_secret=
    access_token_secret)
```

6. Membuat query untuk mengambil data tweet berdasarkan kata kunci

```
1 query = 'covid-19'
2 result = client.search_recent_tweets(query=query, max_results
    =20)
```

Materi 2 - Membaca data dari PDF

Studi Kasus

Anda ingin membaca data dari file PDF

Langkah-langkah

1. Menginstall library PyPDF2

```
pip install PyPDF2
```

2. Import library yang dibutuhkan

```
1 import PyPDF2
2 from PyPDF2 import PdfReader
```

3. Membaca file PDF

```
1 # Membuka file PDF
2 pdf = open("pertemuan_1.pdf", "rb")
3
4 # Membuat objek pembaca PDF
5 pdf_reader = PyPDF2.PdfReader(pdf)
6
7 # Mengecek jumlah halaman dalam file PDF
8 print(pdf_reader.pages)
9
10 # Mendapatkan object sebuah halaman
11 page = pdf_reader.pages[0]
12
13 # Mengekstrak teks dari halaman
14 print(page.extract_text())
15
16 # Menutup file PDF setelah selesai
17 pdf.close()
```

Materi 3 - Membaca data dari File Word (Microsoft Word)

Studi Kasus

Anda ingin membaca data dari file Word

Langkah-langkah

1. Menginstall library python-docx

```
pip install docx # Library untuk python versi lama
pip install python-docx # Library terbaru
```

2. Import library yang dibutuhkan

```
1 import docx
```

3. Membaca text dari file Word

```
1 # Membaca text dari file Word
2 document = docx.Document("file.docx")
3
4 text = ""
5
6 for para in document.paragraphs:
7     text += para.text
8
9 # Cetak hasil text dari file Word
10 print(text)
```

Materi 4 - Membaca data dari JSON

Studi Kasus

Anda ingin membaca data dari file JSON dengan fetch data dari API

Langkah-langkah

1. Import library yang dibutuhkan

```
1 import requests
2 import json
```

2. Melakukan fetch data dari API

```
1 # Fetch data dari API
2 r = requests.get("https://jsonplaceholder.typicode.com/todos")
3 res = r.json()
4
5 # Cetak hasil fetch data dari API
6 print(res)
```

3. Melakukan ekstraksi data dari JSON

```
1 #extract contents
2 content = res[0]['title']
3 print(content)
```

atau jika data yang ingin diambil berada di dalam file JSON

1. Membaca data dari file JSON

```
1 # Membaca data dari file JSON
2 with open('data.json') as f:
3     data = json.load(f)
```

2. Melakukan ekstraksi data dari JSON

```
1 #extract contents
2 content = data[0]['title']
3 print(content)
```

Materi 5 - Membaca data dari HTML

Studi Kasus

Anda ingin membaca data dari file HTML

Langkah-langkah

1. Menginstall library BeautifulSoup

```
pip install beautifulsoup4
```

2. Import library yang dibutuhkan

```
1 from bs4 import BeautifulSoup
2 import urllib.request as urllib2
```

3. Mengambil file HTML melalui URL

```
1 response = urllib2.urlopen('https://en.wikipedia.org/wiki/
    Natural_language_processing')
2 html_doc = response.read()
```

4. Melakukan parsing file HTML

```
1 #Parsing
2 soup = BeautifulSoup(html_doc, 'html.parser')
3 # Formating the parsed html file
4 strhtm = soup.prettify()
5
6 # Mengambil data
7 print(soup.title)
8 print(soup.title.string)
9 print(soup.a.string)
10 print(soup.b.string)
11
12 # Mengambil semua tag tertentu
13 for x in soup.find_all('a'): print(x.string)
```

Tugas

1. Buatlah program yang dapat mengambil data dari file PDF, Word, JSON, dan HTML
2. Buatlah program yang dapat mengambil data dari API Twitter