Modul 1 Praktikum Natural Language Processing
Pengenalan 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

```
import tweepy
```

4. Memasukkan bearer token sebagai akses token ke API Twitter

```
consumer_key = "your_consumer_key"
consumer_secret = "your_consumer_secret"
access_token = "your_access_token"
access_token_secret = "your_access_token_secret"
bearer_token = 'your_bearer_token'
```

5. Membuat client ke API Twitter

```
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

```
query = 'covid-19'
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

```
import PyPDF2
from PyPDF2 import PdfReader
```

3. Membaca file PDF

```
# Membuka file PDF
pdf = open("pertemuan_1.pdf", "rb")

# Membuat objek pembaca PDF
pdf_reader = PyPDF2.PdfReader(pdf)

# Mengecek jumlah halaman dalam file PDF
print(pdf_reader.pages)

# Mendapatkan object sebuah halaman
page = pdf_reader.pages[0]

# Mengekstrak teks dari halaman
print(page.extract_text())

# Menutup file PDF setelah selesai
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

```
import docx
```

3. Membaca text dari file Word

```
# Membaca text dari file Word
document = docx.Document("file.docx")

text = ""

for para in document.paragraphs:
    text += para.text

# Cetak hasil text dari file Word
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

```
import requests
import json
```

2. Melakukan fetch data dari API

```
# Fetch data dari API
r = requests.get("https://jsonplaceholder.typicode.com/todos")
res = r.json()

# Cetak hasil fetch data dari API
print(res)
```

3. Melakukan ekstraksi data dari JSON

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

atau jika data yang ingin diambil berada di dalam file JSON

1. Membaca data dari file JSON

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

2. Melakukan ekstraksi data dari JSON

```
#extract contents
content = data[0]['title']
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

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

3. Mengambil file HTML melalui URL

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

4. Melakukan parsing file HTML

```
#Parsing
soup = BeautifulSoup(html_doc, 'html.parser')
# Formating the parsed html file
strhtm = soup.prettify()

# Mengambil data
print(soup.title)
print(soup.title.string)
print(soup.a.string)
print(soup.b.string)

# Mengambil semua tag tertentu
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