

Text Cleansing REST API

Mohamad Ikhsan Nudin
Binar Academy Gold Challenge - Data Science Batch 1

Text Cleansing REST API adalah sebuah service API yang berfungsi untuk melakukan cleansing terhadap data text menggunakan metode regex. adapun fitur yang tersedia antara lain:

- ❖ Auth yang berfungsi sebagai keamanan dan identifikasi user.
- ❖ Input data berupa text dalam format json atau file csv.
- ❖ Data input dan data hasil akan disimpan kedalam database.
- ❖ Response API dengan format json.

Struktur Folder



BINAR

Project ini menggunakan arsitektur Model View Controller atau yang dapat disingkat MVC, yaitu sebuah pola arsitektur dalam membuat sebuah aplikasi dengan cara memisahkan kode menjadi tiga bagian agar mempermudah pengembangan dan debugging menjadi lebih mudah.

```
app.py
config.py

controllers
└── api
    ├── AuthController.py
    ├── DataController.py
    └── UserController.py

data
├── databases
│   └── binarapi_db.db
├── data_set
│   ├── data.csv
│   ├── new_kamusalay.csv
│   └── stopwordbahasa.csv
└── files

models
├── Text.py
├── User.py
└── __init__.py

routes
├── api
│   ├── AuthRoute.py
│   ├── DataRoute.py
│   ├── UserRoute.py
│   └── __init__.py
└── web
    └── __init__.py

utils
├── auth_handler.py
├── doc.yaml
├── text.py
└── __init__.py

views
```

Metode Cleansing

1. Menghilangkan gambar, merubah string menjadi lowercase dan menghilangkan URL.

2. Mengubah kata tidak baku menjadi kata baku (gue -> saya) & menghilangkan new line agar text hanya dalam 1 baris saja.

3. Agar mempermudah melakukan analisis perlu menghilangkan kata - kata yang dirasa tidak perlu seperti halnya rt, user, secure dll.

```
def unnecessary_char_remover(text):  
    new_text = re.sub(r'pic.twitter.com.[\w]+', '', text)  
    new_text = new_text.lower()  
    new_text = re.sub('((www\.[^\s]+)|(https?://[^\s]+)|(http?://[^\s]+))', ' ', new_text)  
  
    new_text = re.sub('gue', 'saya', new_text)  
    new_text = re.sub('\n', ' ', new_text)  
  
    to_delete = ['hypertext', 'transfer', 'protocol', 'over', 'secure', 'socket', 'layer', 'dtype', 'tweet', 'name', 'object' ...  
  
    for word in to_delete:  
        new_text = re.sub(word, '', new_text)  
        new_text = re.sub(word.upper(), ' ', new_text)  
  
    retweet_user = ['rt ', ' rt ', ' user ']  
  
    for word in retweet_user:  
        new_text = re.sub(word, ' ', new_text)  
        new_text = re.sub(word.upper(), ' ', new_text)  
  
    new_text = re.sub(' +', ' ', new_text)  
  
    result = {'original' : text, 'result' : new_text}  
    return result
```

Pada function ini menghasilkan tipe data dictionary berupa key original untuk text input dan key result untuk hasil dari cleaning pada tahap pertama

Metode Cleansing

```
def remove_nonalphanumeric(text):  
    new_text = re.sub('[^0-9a-zA-Z]+', ' ', text)  
    result = {'original' : text, 'result' : new_text}  
    return result
```

Menghilangkan karakter selain alphanumerik

Hasil

BINAR-BOOTCAMP

Public

Text Cleansing REST API adalah sebuah service API yang berfungsi untuk melakukan cleansing terhadap data text menggunakan metode regex.

Python

Updated 2 hours ago

repository proyek ini dapat di akses di link berikut :
<https://github.com/Xsanjaya/BINAR-BOOTCAMP.git>

auth Operations about user

GET /auth/login Logs user into the system

POST /auth/register Create user

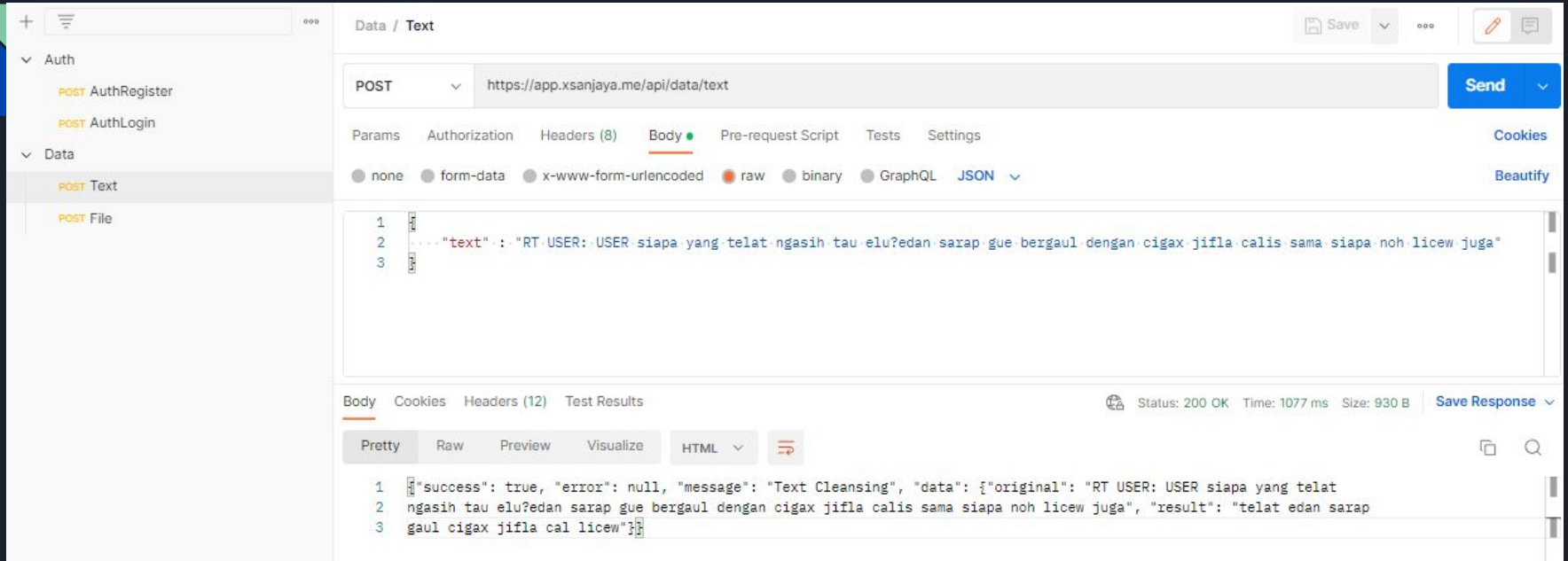
data Everything about data

POST /data/file uploads an file csv

POST /data/text Text Cleansing

tampilan dokumentasi api
menggunakan swagger UI

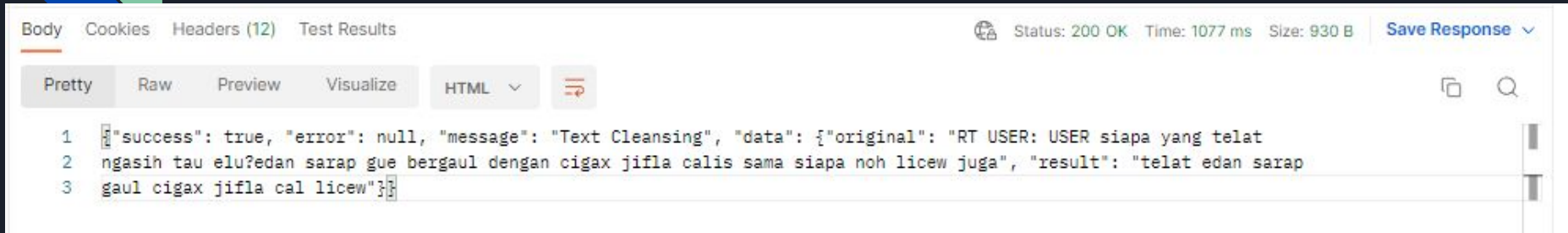
<https://app.xsanjaya.me/api/docs>



<https://www.postman.com/xsanjaya/workspace/public/request/16689317-7f9dabde-5608-4a29-8c8e-b4760fa5ee69>

Selain dengan Swagger UI, dokumentasi API juga dapat diakses menggunakan postman pada link diatas

Response endpoint untuk input text



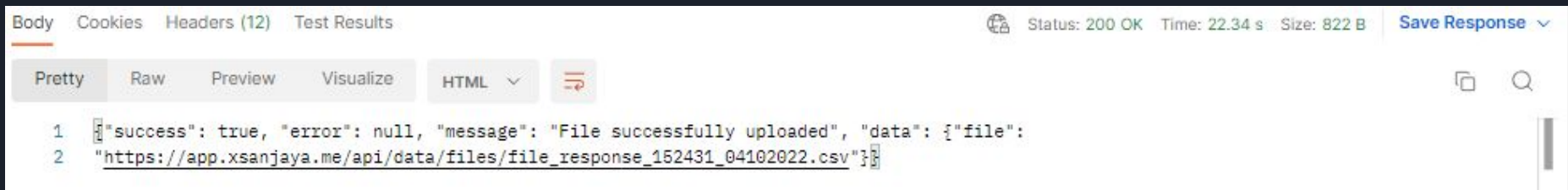
Body Cookies Headers (12) Test Results

Status: 200 OK Time: 1077 ms Size: 930 B Save Response

Pretty Raw Preview Visualize HTML

```
1 {"success": true, "error": null, "message": "Text Cleansing", "data": {"original": "RT USER: USER siapa yang telat  
2 ngasih tau elu?edan sarap gue bergaul dengan cigax jifla calis sama siapa noh licew juga", "result": "telat edan sarap  
3 gaul cigax jifla cal licew"}}
```

Response endpoint untuk input text



Body Cookies Headers (12) Test Results

Status: 200 OK Time: 22.34 s Size: 822 B Save Response

Pretty Raw Preview Visualize HTML

```
1 {"success": true, "error": null, "message": "File successfully uploaded", "data": {"file":  
2 "https://app.xsanjaya.me/api/data/files/file_response_162431_04102022.csv"}}
```


- ❖ Auth terdapat 2 Endpoint yaitu register & login.
- ❖ Input data bisa berupa text dalam format json atau file csv.
- ❖ Response API dengan format json.