

Dasar-Dasar Pemrograman 1 Gasal 2021/2022

Tugas Pemrograman 02

Versi 1

Deadline: 15 Oktober 2021
23:55 Waktu SCELE



FAKULTAS
ILMU
KOMPUTER

Versi Dokumen

- Versi 1 (05-10-2021): Rilis Pertama

Komponen Penilaian

Sesuai dengan ketentuan yang ada di [Rubrik Penilaian DDP1](#).

Plagiarisme tidak akan ditoleransi. Kamu diperbolehkan berdiskusi dengan teman terkait ide implementasi tugas ini. Harap menuliskan kolaborator jika berdiskusi dengan mahasiswa lain. Perlu diingat bahwa **implementasi kode dilakukan secara individu**. Tim pengajar akan melakukan **code similarity checking** pada **implementasi kode mahasiswa**. Jika terbukti melakukan kecurangan/plagiarisme akan mendapat sanksi berupa **nilai 0 pada TP ini**.

Word Clouds

A Word Cloud of CommencementSpeechByGates2014.txt



Tujuan

Tugas ini akan membantumu dalam memahami *functions*, *strings*, *lists*, *text processing* dan *file I/O*

Problem

Salah satu hal yang sering dilihat pada halaman-halaman *website* saat ini adalah *word clouds* atau *tag clouds* dari sebuah dokumen. *Word cloud* adalah representasi visual dari kemunculan suatu kata (*word*). Semakin banyak frekuensi kemunculan kata tersebut pada dokumen maka semakin besar ukuran *font* untuk kata tersebut.

Dalam tugas ini, kamu diminta untuk menganalisis suatu dokumen *text* yang diberikan oleh pengguna dan membuat *word cloud* dari dokumen tersebut.

Ketentuan

Anda akan disediakan dengan beberapa *file* pendukung dalam menyelesaikan tugas berikut:

- *Text documents* untuk diproses: `CommencementSpeechByGates2014.txt` dan `JokoWidodoSpeechAPEC2014.txt`
- *File* yang berisi *stop words*: `stopwords.txt`
- Fungsi yang digunakan untuk membuat HTML file yang bisa dibuka oleh browser: `html_functions.py`

File-file tersebut dapat diakses di [sini](#).

Stop Words

Tidak semua kata kita masukkan ke dalam *word clouds*, contohnya seperti 'a', 'the', 'was', dan lain-lain. Semua kata tersebut sering muncul di dokumen dan kurang memberikan makna. Kata-kata ini tersedia dalam file *stop words*, setiap baris berisi satu kata.

Tidak boleh ada kata dalam *stopwords* yang masuk dalam *word cloud*.

Functions

Ada tiga buah fungsi dan contoh pemakaiannya di `html_functions.py`. Gunakan fungsi-fungsi tersebut pada program kamu. Program tersebut berisi:

- `make_HTML_word(word, count, high, low)`: fungsi ini mengambil `word` dan membungkusnya dalam **HTML font tag** dengan ukuran font tertentu dan warna yang random, lalu mengembalikan string tersebut yang `font_size`-nya berada di antara `html_big` dan `html_little` (dua variabel lokal dalam fungsi yang bisa diubah sesuka hati). Parameternya adalah:
 - `word` (string), kata yang akan dibungkus dengan tag HTML
 - `count`, berapa kali kemunculan kata tersebut
 - `high`, jumlah kemunculan dari kata terbanyak dalam dokumen
 - `low`, jumlah kemunculan dari kata paling sedikit dalam dokumen

Fungsi ini mengembalikan sebuah **HTML_formatted string**.

Contoh, jika parameter `word` berisi 'africa', fungsi mengembalikan string:

```
'<span style="color: rgb(182, 125, 40); font-size:14px;">africa</span>'
```

Keterangan: nilai-nilai pada `rgb()` akan di-*random* dan `font_size` akan diperhitungkan dari parameter `count`, `high`, dan `low`.

- `make_HTML_box(body)`: fungsi ini menerima sebuah string `body` yang berisi semua kata yang sudah dibungkus dari `make_HTML_word()` dan menempatkannya ke dalam HTML box untuk ditampilkan di *website*. Fungsi ini mengembalikan string kode HTML box.
- `print_HTML_file(body, title)`: menerima `body` (string) yang dikembalikan oleh `make_HTML_box`, membungkusnya dengan HTML *tags*, dan memasukkannya ke sebuah file HTML. String `title` juga akan digunakan dalam *file* HTML tadi. Selain itu, `title` juga digunakan sebagai nama file output dengan format `.html`. File outputnya disimpan di directory yang sama.

Spesifikasi program

Buatlah `word_clouds.py` dengan spesifikasi:

- Meminta file input yang akan diproses
- **Mencetak 56 kata teratas** berdasarkan jumlah kemunculan terbanyak (pasangan kata dan jumlah kemunculan), dalam format 14 baris dan 4 kolom. Kata-kata tersebut **diurutkan dari yang paling sering muncul sampai paling jarang** (untuk tahap selanjutnya diurutkan sesuai abjad).
- **Membuat file HTML** (dengan format `.html`) menggunakan fungsi-fungsi dari `html_functions.py` untuk membuat *word cloud* berisi 56 kata teratas. Kata-kata yang ada dalam *word cloud* **diurutkan sesuai abjad** agar terlihat lebih menarik. Kamu bisa melihat *word cloud* tersebut pada browser favoritmu.

Hints

Ada beberapa masalah yang harus dipecahkan. Pikirkan tiap masalah tersebut sebelum mulai membuat program.

1. Kamu harus mengakses isi file dan memisahkan setiap katanya
2. Setelah dipisahkan, kamu harus menghilangkan semua *stop words* yang disediakan di `stopwords.txt`. Jangan lupa untuk menghilangkan semua *punctuation* (tanda baca) dari setiap kata. Hanya karena suatu kata memiliki *punctuation* di akhir kata tersebut, bukan berarti merupakan sebuah kata yang berbeda.

Misalnya, jangan sampai kata “**married**” dan “**married,**” dilihat sebagai kata yang berbeda. (hint: meng-*import* modul **string** dan menggunakan **string.punctuation** akan mempermudah proses penghilangan tanda baca)

3. Setelah itu, hitung jumlah kemunculan setiap kata. Gunakan struktur data list atau dictionary.
4. Setelah list-nya selesai, urutkan list tersebut menggunakan sort atau sorted.
5. Gunakan fungsi yang telah diberikan untuk mengubah kata dan menghitungnya menjadi HTML page.
(hint: lihat contoh file `html_functions.py`)

Sample Output:

1. Berikut adalah hasil program yang dicetak pada IDLE

```
Program untuk membuat word cloud dari text file
```

```
-----
```

```
hasilnya disimpan sebagai file html,  
yang bisa ditampilkan di browser.
```

```
Silakan masukan nama file: CommencementSpeechByGates2014.txt
```

```
CommencementSpeechByGates2014.txt :
```

```
56 kata diurutkan berdasarkan jumlah kemunculan dalam pasangan  
(jumlah:kata)
```

26:people	17:optimism	12:world	12:melinda
11:stanford	11:make	9:bill	8:poor
8:innovation	8:computers	7:leave	7:knew
7:kids	6:years	6:workers	6:time
6:talk	6:sex	6:problems	6:empathy
6:cure	6:change	6:center	6:aids
5:women	5:wanted	5:trip	5:technology
5:soweto	5:south	5:poverty	5:patients
5:made	5:lives	5:hospital	5:future
5:foundation	5:day	5:children	4:united
4:things	4:tb	4:suffering	4:states
4:software	4:power	4:pessimists	4:microsoft
4:luck	4:home	4:helped	4:heart
4:gates	4:empower	4:community	4:africa

```
Tekan Enter untuk keluar ...
```

2. Lalu, hasil file HTML yang dibuat: [CommencementSpeechByGates2014.txt.html](#)

```
<html> <head>
<title>A Word Cloud of CommencementSpeechByGates2014.txt</title>
</head>

<body>
<h1>A Word Cloud of CommencementSpeechByGates2014.txt</h1>
<div style="
width: 560px;
background-color:
rgb(250,250,250); border:
1px grey solid;
text-align: center" ><span style="color: rgb(182, 125, 40); font-
size:14px;">africa</span> <span style="color: rgb(201, 22, 200); font-
size:21px;">aids</span> <span style="color: rgb(200, 35, 215); font-
size:32px;">bill</span> <span style="color: rgb(37, 177, 97); font-
size:21px;">center</span> <span style="color: rgb(153, 117, 175); font-
size:21px;">change</span> <span style="color: rgb(170, 85, 107); font-
size:17px;">children</span> <span style="color: rgb(93, 26, 68); font-
size:14px;">community</span> <span style="color: rgb(3, 15, 138); font-
size:28px;">computers</span> <span style="color: rgb(111, 32, 140); font-
size:21px;">cure</span> <span style="color: rgb(131, 110, 173); font-
size:17px;">day</span> <span style="color: rgb(234, 102, 19); font-
size:21px;">empathy</span> <span style="color: rgb(57, 177, 156); font-
size:14px;">empower</span> <span style="color: rgb(233, 74, 91); font-
size:17px;">foundation</span> <span style="color: rgb(59, 172, 102); font-
size:17px;">future</span> <span style="color: rgb(34, 232, 145); font-
size:14px;">gates</span> <span style="color: rgb(79, 255, 8); font-
size:14px;">heart</span> <span style="color: rgb(166, 125, 88); font-
size:14px;">helped</span> <span style="color: rgb(165, 7, 249); font-
size:14px;">home</span> <span style="color: rgb(194, 84, 0); font-
size:17px;">hospital</span> <span style="color: rgb(9, 91, 228); font-
size:28px;">innovation</span> <span style="color: rgb(76, 0, 178); font-
size:25px;">kids</span> <span style="color: rgb(162, 27, 197); font-
size:25px;">knew</span> <span style="color: rgb(39, 122, 167); font-
size:25px;">leave</span> <span style="color: rgb(35, 102, 3); font-
size:17px;">lives</span> <span style="color: rgb(226, 91, 65); font-
size:14px;">luck</span> <span style="color: rgb(169, 155, 7); font-
size:17px;">made</span> <span style="color: rgb(26, 192, 91); font-
size:40px;">make</span> <span style="color: rgb(231, 11, 190); font-
size:43px;">melinda</span> <span style="color: rgb(243, 167, 105); font-
size:14px;">microsoft</span> <span style="color: rgb(85, 50, 90); font-
size:62px;">optimism</span> <span style="color: rgb(131, 109, 33); font-
size:17px;">patients</span> <span style="color: rgb(76, 73, 136); font-
size:96px;">people</span> <span style="color: rgb(114, 147, 98); font-
size:14px;">pessimists</span> <span style="color: rgb(166, 33, 73); font-
size:28px;">poor</span> <span style="color: rgb(33, 29, 89); font-
size:17px;">poverty</span> <span style="color: rgb(80, 103, 61); font-
size:14px;">power</span> <span style="color: rgb(113, 164, 17); font-
size:21px;">problems</span> <span style="color: rgb(51, 170, 96); font-
size:21px;">sex</span> <span style="color: rgb(119, 220, 96); font-
size:14px;">software</span> <span style="color: rgb(31, 140, 9); font-
size:17px;">south</span> <span style="color: rgb(153, 61, 179); font-
size:17px;">soweto</span> <span style="color: rgb(28, 241, 20); font-
size:40px;">stanford</span> <span style="color: rgb(13, 249, 105); font-
size:14px;">states</span> <span style="color: rgb(25, 200, 145); font-
size:14px;">suffering</span> <span style="color: rgb(3, 157, 102); font-
size:21px;">talk</span> <span style="color: rgb(90, 37, 203); font-
size:14px;">tb</span> <span style="color: rgb(230, 17, 200); font-
size:17px;">technology</span> <span style="color: rgb(62, 183, 235); font-
size:14px;">things</span> <span style="color: rgb(120, 63, 192); font-
size:21px;">time</span> <span style="color: rgb(3, 83, 249); font-
size:17px;">trip</span> <span style="color: rgb(208, 51, 54); font-
```

```
size:14px;">united</span> <span style="color: rgb(53, 159, 84); font-  
size:17px;">wanted</span> <span style="color: rgb(44, 103, 241); font-  
size:17px;">women</span> <span style="color: rgb(57, 247, 49); font-  
size:21px;">workers</span> <span style="color: rgb(48, 231, 147); font-  
size:43px;">world</span> <span style="color: rgb(188, 234, 169); font-  
size:21px;">women</span>  
  
</body> </html>
```

Jika dibuka di browser, maka [CommencementSpeechByGates2014.txt.html](#) akan terlihat seperti ini (catatan: perhatikan bahwa kata-kata tersebut terurut secara alfabetis)

A Word Cloud of CommencementSpeechByGates2014.txt



Deliverables

Kumpulkan berkas **word_clouds.py** dan **html_functions.py** yang telah di-zip dengan format penamaan seperti berikut.

[KodeAsdos]_[Kelas]_[NPM]_[NamaLengkap]_TP02.zip

Contoh:

ORI_B_1234567890_AhmadHaroriZakiIchsan_TP02.zip