



CSCE604135 • Perolehan Informasi
Semester Gasal 2023/2024
Fakultas Ilmu Komputer, Universitas Indonesia

Tugas Eksplorasi Pemrograman 1: Text Preprocessing
Tenggat Waktu: Rabu, 13 September 2023, 23.55 WIB

Ketentuan:

1. Anda diberikan sebuah Google Colab *notebook* berisi contoh beserta dengan soal yang anda harus kerjakan. Data yang dibutuhkan dapat diunduh dengan code yang sudah tersedia dalam *notebook* tersebut.
2. Untuk pengumpulan, silahkan *share* Google Colab *notebook* yang sudah berisi jawaban anda. Langkah-langkah pengumpulan ini akan dijelaskan lebih detil selanjutnya pada dokumen ini. Tulis *share link* notebook anda ke dalam sebuah file dengan format penamaan **TEP1_NPM.url**, dan kumpulkan file tersebut melalui submisi di SCeLe.
Contoh penamaan file: TEP1_2206137126.url
3. Kumpulkan dokumen tersebut pada submisi yang telah disediakan di SCeLe sebelum **Rabu, 13 September 2023, 23.55 WIB**. Keterlambatan pengumpulan akan dikenakan pinalti.
4. Anda tidak diperbolehkan untuk mengubah isi dari Google Colab Notebook yang telah anda submit setelah tenggat waktu berakhir. Bila *last modified time notebook* anda melebihi tenggat waktu, akan dikenakan pinalti sesuai aturan keterlambatan.
5. Tugas ini dirancang sebagai **tugas mandiri**. Plagiarisme tidak diperkenankan dalam bentuk apapun. Adapun kolaborasi berupa diskusi (tanpa menyalin maupun mengambil jawaban orang lain) dan literasi masih diperbolehkan dengan mencantumkan kolaborator dan sumber.
6. Anda boleh berkonsultasi terkait tugas ini dengan 2 asisten dosen berikut. Asisten dosen diperbolehkan membantu anda dengan memberikan petunjuk.

a. Gibran Brahmana

Email: gibranbrahmana@gmail.com

Whatsapp: +6281219578792

Line: gibranbrahmana

b. Adrianus Saga Ekakristi

Email: adrianus.saga21@ui.ac.id / saga.ekakristi@gmail.com

Whatsapp: +6285216616696

Line: saga_ekakristi

Petunjuk Pengerjaan Tugas

Pada tugas eksplorasi pemrograman ini, anda diberikan sebuah Google Colab *notebook* berisi contoh *text preprocessing* yang umum dilakukan sebagai langkah awal dalam sistem Information Retrieval, beserta dengan soal yang perlu anda kerjakan. Berikut adalah *link* menuju *notebook* tersebut:

Link: <https://colab.research.google.com/drive/1M3fFRUfrlxE6adFFz7ToH1qcXdv37s4?usp=sharing>

Silahkan *copy notebook* tersebut ke Google Drive anda masing-masing dengan format nama *notebook* **TEP1_NPM.ipynb**

Contoh penamaan *notebook*: TEP1_2206137126.ipynb

Notebook ini terdiri dari 2 bagian. Pertama, anda akan diberikan contoh *code text preprocessing* yang diterapkan untuk data dalam bahasa Indonesia, yaitu tokenisasi, lematisasi, stemming, dan *stop words removal*. Pada bagian kedua, anda akan diminta untuk menulis *code* untuk melakukan pemrosesan serupa untuk diterapkan pada data berbahasa Inggris.

Bagian 1 terdiri dari pengambilan data dan empat contoh *preprocessing*, yaitu tokenisasi (1.1), lematisasi (1.2), *stemming* (1.3), dan *stop words removal* (1.4). Keempat pemrosesan ini kemudian dirangkai sebagai *pipeline* dan diterapkan kepada dataset bahasa Indonesia. Silahkan anda jalankan dan pelajari *code* dan tampilan *input-output* yang sudah disediakan.

Pada Bagian 2, anda diminta untuk melakukan eksplorasi terhadap *text preprocessing* serupa dengan Bagian 1, namun diterapkan ke data dalam bahasa Inggris. *Code* untuk penarikan data sudah disediakan dan dapat langsung anda gunakan. Berikut adalah beberapa tugas yang perlu anda kerjakan beserta dengan ketentuannya:

- Bagian 2.1: Anda diminta melakukan tokenisasi *text* menjadi sekumpulan token sama seperti contoh pada Bagian 1.1 dengan ketentuan harus menggunakan regex. Tampilkan penerapan tokenisasi ini kepada contoh salah satu artikel (*example_passage_en*).
- Bagian 2.2: Anda diminta melakukan lematisasi untuk token-token yang sudah diproses sebelumnya dengan ketentuan harus menggunakan *library* [Spacy](#). Tampilkan penerapan lematisasi ini terhadap hasil tokenisasi kepada contoh sebelumnya. Silahkan lakukan eksplorasi terhadap cara penggunaan *tools* ini, seperti pembentukan *instance* Doc, pengambilan lemma dari kata, dsb. Beberapa tips untuk bagian ini dapat anda baca dalam *notebook*.

- Bagian 2.3: Anda diminta untuk melakukan *stemming* terhadap kumpulan token dari pemrosesan sebelumnya dengan ketentuan harus menggunakan library [NLTK](#). NLTK memiliki beberapa jenis *stemmer* untuk bahasa Inggris, dan anda boleh memilih salah satunya. Tampilkan penerapan *stemming* ini terhadap hasil lematisasi pada contoh sebelumnya.
- Bagian 2.4: Anda diminta untuk mencari kumpulan *stop words* yang telah disediakan oleh library NLTK. Silahkan lakukan eksplorasi bagaimana cara mendapatkan kumpulan *stop words* tersebut dan tampilkan ke *output* untuk anda pelajari. Anda juga diminta untuk mengimplementasikan penghapusan *stop words* dari token-token yang telah diproses sebelumnya menggunakan kumpulan *stop words* tersebut.
- Bagian 2.5: Menggunakan *function-function* yang telah anda buat dari eksplorasi diatas, silahkan rangkai tokenisasi, lematisasi, dan *stemming* untuk diterapkan terhadap seluruh data dalam dataset bahasa Inggris yang sudah disediakan, serupa seperti Bagian 1. Khusus untuk tahap ini, anda **tidak** perlu melakukan *stop words* removal sebagai persiapan untuk soal selanjutnya.
- Bagian 2.6: Anda diminta untuk mengumpulkan seluruh token unik beserta dengan frekuensi kemunculannya dalam dataset ini, atau dengan kata lain anda diminta melakukan *word counting*. Hasil akhir yang diharapkan berbentuk *list of tuple* (*string* token, *integer* count).
 - Sebagai contoh, bila terdapat artikel A, B, dan C dalam dataset, dimana masing-masing artikel berisi:
 - Artikel A: "computer", "science"
 - Artikel B: "computer", "program"
 - Artikel C: "program", "execution"
 - Maka, diharapkan output *list of tuple* (*string*, *integer*) berupa:
 - (computer, 2)
 - (program, 2)
 - (science, 1)
 - (execution, 1)

Beberapa tips untuk bagian ini dapat anda baca dalam *notebook*.

- Bagian 2.7: Setelah anda mendapatkan kumpulan token-frekuensi dari proses sebelumnya, gunakan data tersebut untuk membentuk *stop words* anda sendiri. Ambil top 200 token yang memiliki kemunculan terbanyak.

Kemudian, tampilkan *stop words* yang sudah anda kumpulkan dengan *stop words* dari NLTK. Bandingkan keduanya. Apa yang anda temukan dari komparasi tersebut? Token jenis seperti apa yang hanya ditemukan dari hasil pengumpulan anda?

- Bagian 2.8: Selanjutnya, mirip dengan *word counting*, anda diminta untuk mengumpulkan karakter unik dari seluruh token-token dalam dataset beserta dengan frekuensi kemunculannya (*character counting*).
 - Sebagai contoh, bila terdapat artikel A dan B dalam dataset, dimana masing-masing artikel berisi:
 - Artikel A: "computer", "science"
 - Artikel B: "computer", "program"
 - Maka, diharapkan output *list of tuple (string, integer)* berupa:
 - ('c', 4)
 - ('e', 4)
 - ('r', 4)
 - dst
 - Ketentuan:
 - Untuk simplifikasi, anda hanya perlu menghitung karakter alfabet (a-z). Silahkan hapus atau abaikan karakter lain (e.g. tanda baca).
 - Untuk simplifikasi, anda dapat mengubah huruf kapital dalam seluruh token menjadi huruf kecil.
- Bagian 2.9: Kumpulkan juga character count untuk dataset bahasa Indonesia pada Bagian 1. Tampilkan kedua data tersebut dan bandingkan. Apa karakter yang paling dominan dalam bahasa Indonesia? Apa karakter yang paling dominan dalam bahasa Inggris?

Catatan: Contoh, soal, dan jumlah data dalam tugas ini dibuat sedemikian sehingga durasi eksekusi program dari awal hingga akhir dapat selesai dalam 45 menit atau lebih cepat. Bila anda mengalami isu eksekusi program yang lambat, silahkan cek kembali *throughput* (jumlah data terproses per detik) pada setiap langkah dan implementasi *code* anda.

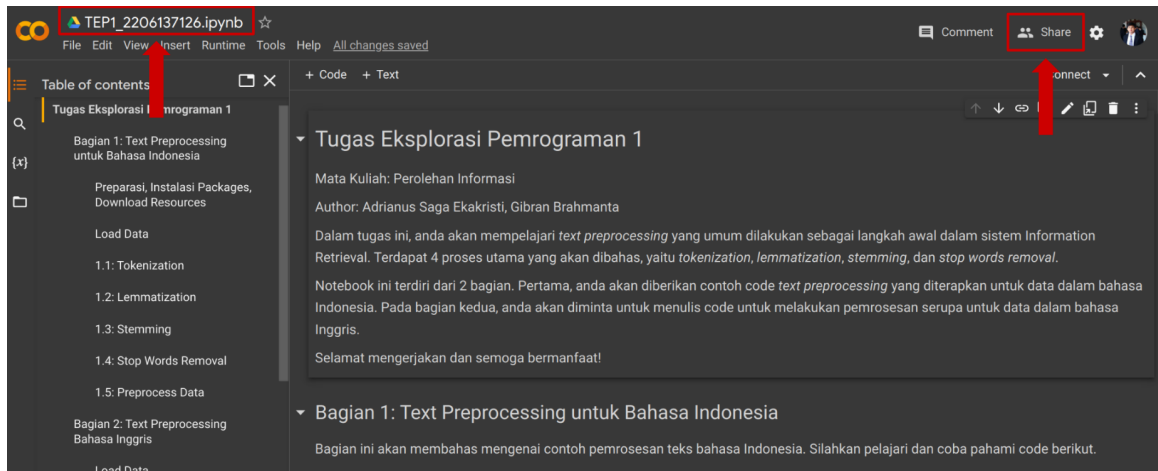
Catatan Revisi:

- 2023-09-08 15:45: Pengurangan jumlah data pada Template Bagian 2. Silahkan menyesuaikan dengan template baru.

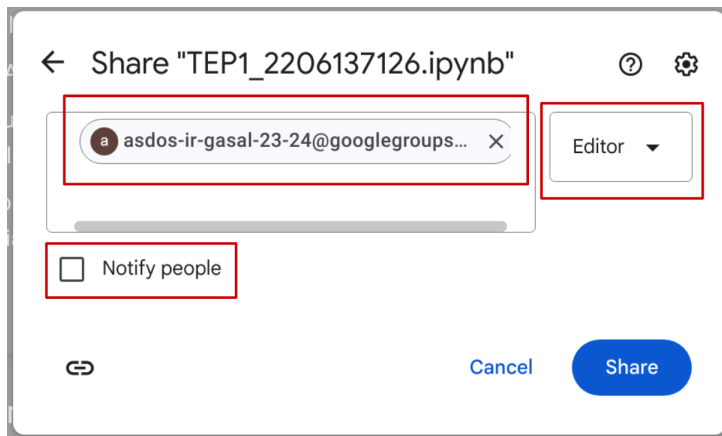
Pengumpulan

Berikut adalah cara melakukan sharing *notebook* untuk keperluan pengumpulan tugas:

1. Pastikan nama *notebook* anda sudah sesuai format yang sudah dijelaskan sebelumnya. Kemudian, buka *section Share* pada bagian atas kanan *notebook* anda.



2. Isi alamat email asdos-ir-gasal-23-24@googlegroups.com dalam tujuan *share*. Pilih akses *Editor*. *Disable* pilihan “*Notify People*”. Tekan tombol *Share*.




3. Buka kembali *section Share* tersebut. Pastikan kembali alamat Google Group asdos telah terdaftar sebagai *Editor*. Tekan tombol ‘*Copy link*’ untuk mendapatkan link untuk mengakses *notebook*


anda yang perlu anda kumpulkan ke *submission* SCell.

Share "TEP1_2206137126.ipynb" ? ⚙


Add people and groups

People with access


 **Saga Ekakristi (you)**
saga.ekakristi@gmail.com Owner

 **Asdos IR Gasal 2023-2024**
asdos-ir-gasal-23-24@googlegroups.com Editor ▼

General access

 **Restricted** ▼

Only people with access can open with the link

 **Copy link**

Done

Penilaian

| Komponen | Bagian | Proporsi |
|--|---------------|----------|
| Tokenisasi | 2.1 | 15% |
| Lematisasi | 2.2 | 15% |
| Stemming | 2.3 | 15% |
| Stop Words | 2.4 | 15% |
| Word Counting & Komparasi Stop Words | 2.5, 2.6, 2.7 | 20% |
| Character Counting & Komparasi Distribusi Karakter | 2.8, 2.9 | 20% |

Selamat mengerjakan!