Kodingan diambil dari: https://github.com/jin-zhe/boolean-retrieval-engine. Credits to: https://github.com/jin-zhe

Modifikasi **indeks.py** DAN **search.py** berupa:

Mengganti stemmer dari Porter menjadi Sastrawi, juga stopword dari NLTK menjadi stopword pada Sastrawi.

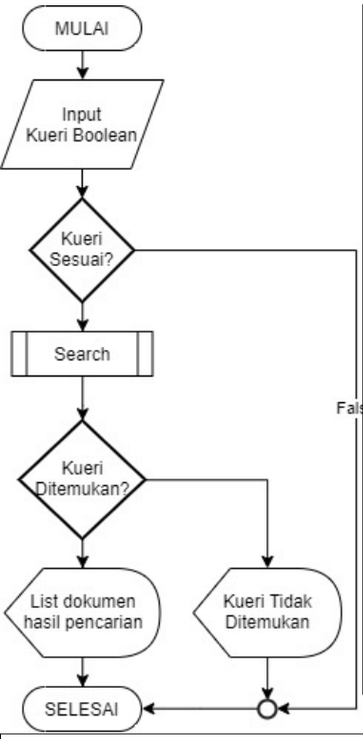
Menambah preprocessing sebelum di-stem.

Membuang beberapa fungsi pelengkap yang tidak perlu.

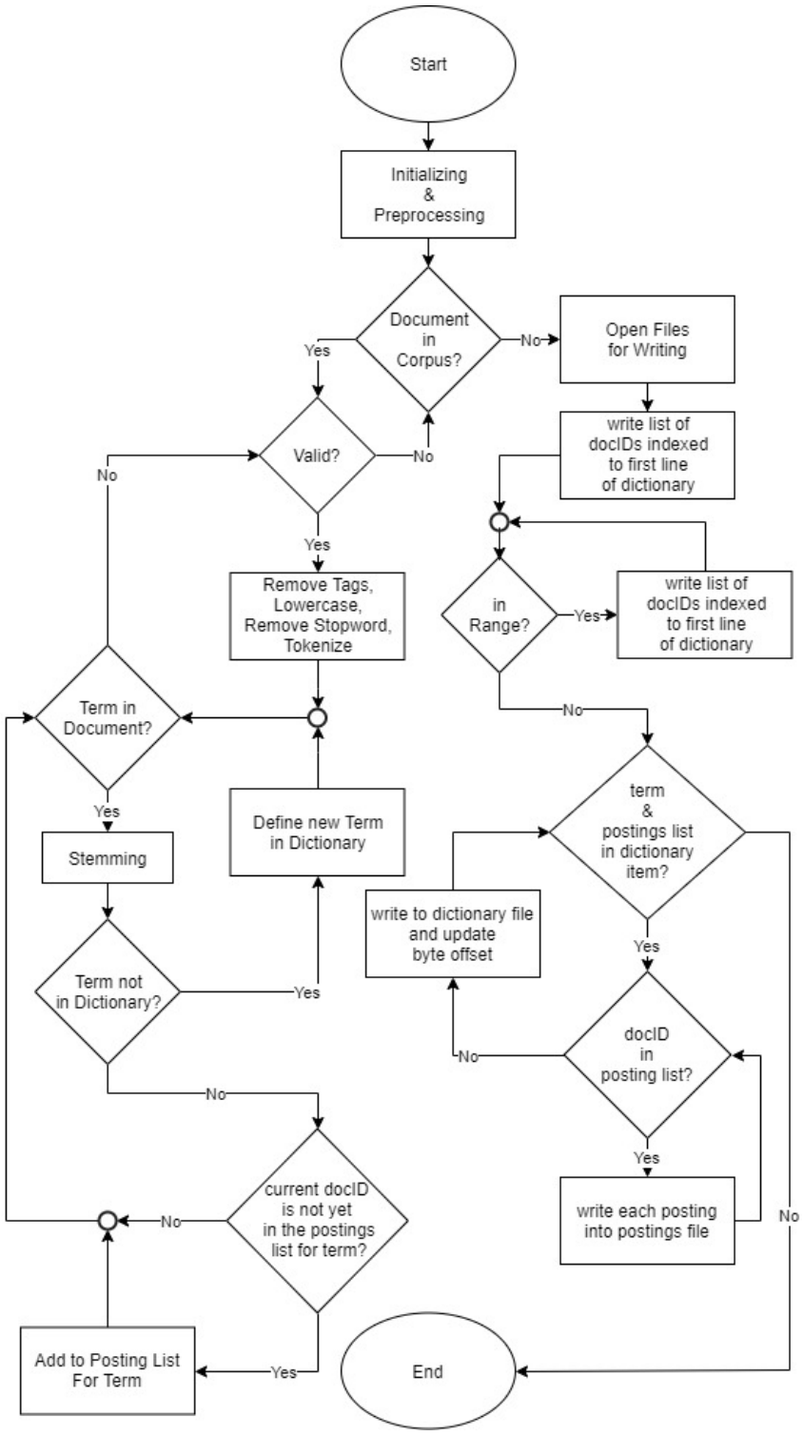
Mempermudah penggunaan/mengekseskusi file; mengubah fungsi menjadi baris-baris kode biasa agar dapat langsung dieksekusi tanpa perlu memanggil fungsi beserta parameternya, menambah inisialisasi variabel sebagai pengganti parameter pada fungsi.

Menambah contoh penggunaan program.

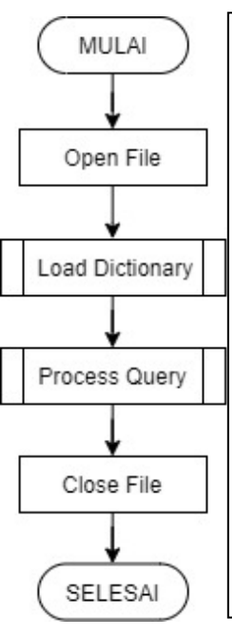
Semua file atau dokumentasi dapat diakses di: https://github.com/eliteraihan/boolean-retrieval-engine



[Search.py] Secara garis besar, alur program sama seperti pada umumnya, sistem menerima kueri Boolean, kueri diproses, dan menampilkan hasil proses, berupa **daftar dokumen** atau **dokumen tidak ditemukan**.



[Indeks.py] Akan tetapi, **pertama-tama** adalah membuat dictionary dan postings list terlebih dahulu sebelum sistem digunakan. Builder ini akan menelusuri dokumen-dokumen pada sebuah folder, dan akan membaca setiap dokumen yang **valid**. Yang masing-masing dokumen tersebut akan dilakukan tokenisasi dan stemming untuk kontennya, dan juga mengambil nomor dokumen untuk dijadikan **indeks/dictionary** dan **postings list**. File hasil dari pembentukan postings list tidak bisa dibaca oleh manusia, karena file tersebut berisi byte offset yang merepresentasikan bentuk postings list, berupa term t ada pada dokumen mana saja, akan tetapi dalam bentuk byte offset, sehingga penelusurannya akan direpresentasikan pada **file dictionary**, yang isinya adalah term t, frekuensi dokumen, dan posisi/offset pada postings list, baris pertama file berisi nomor-nomor dokumen yang sudah diindeks. Program ini hanya dijalankan sekali, atau ketika terjadi perubahan pada korpus sehingga indeks dan postings list perlu dibangun ulang.



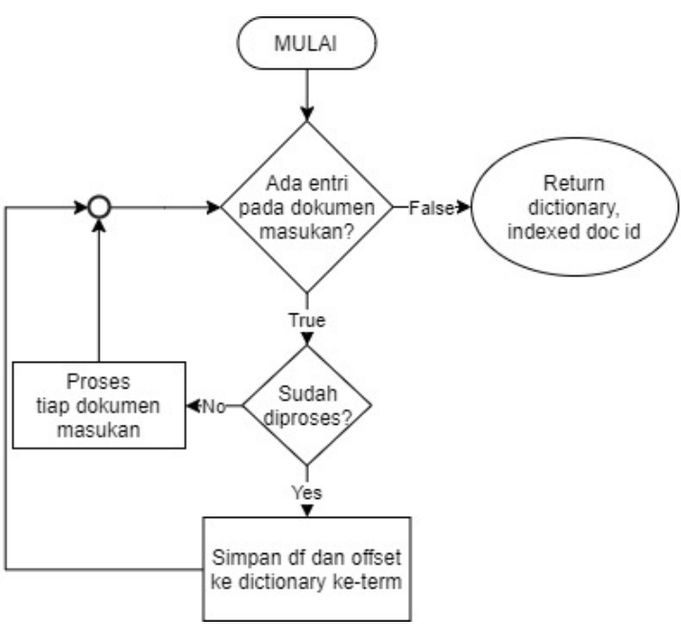
[| Search |]

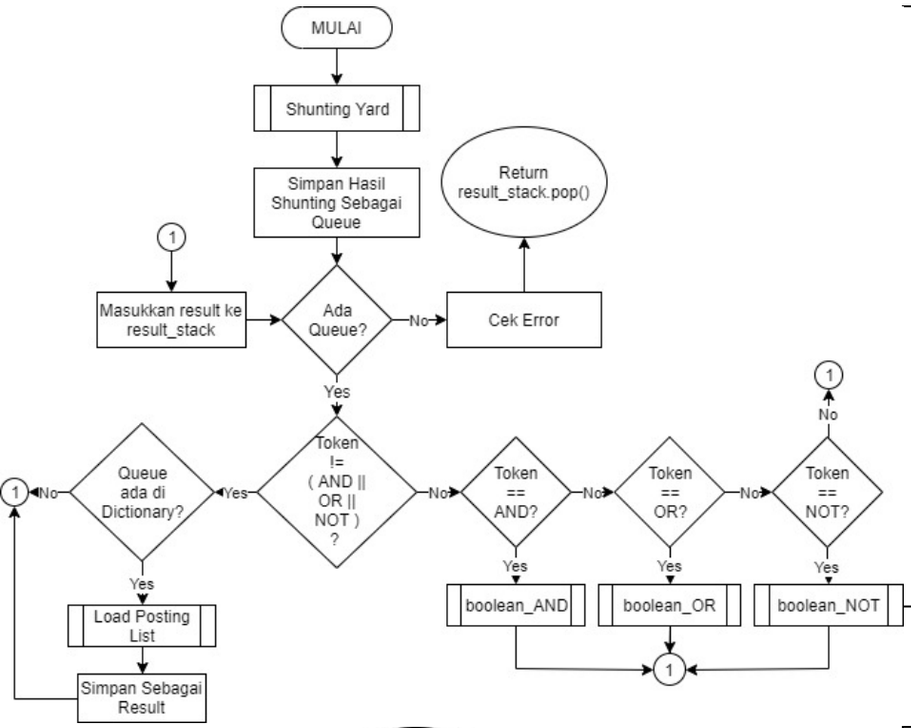
Inti dari sistem ini pertama-tama adalah **memuat dictionary** dan **posting list** yang sudah dibuat, kemudian **memproses kueri**, kemudian akan mengembalikan daftar dokumen atau dokumen tidak ditemukan.

[| Load Dictionary|]

Untuk setiap entri (pemisah '\n') pada **file dictionary**, baca baris pertama, dan selain baris pertama. Selain baris pertama, baca token (pemisah spasi): term, df, dan offset.

Kemudian mengembalikan **dictionary** dan **nomor-nomor dokumen yang diindeks**.

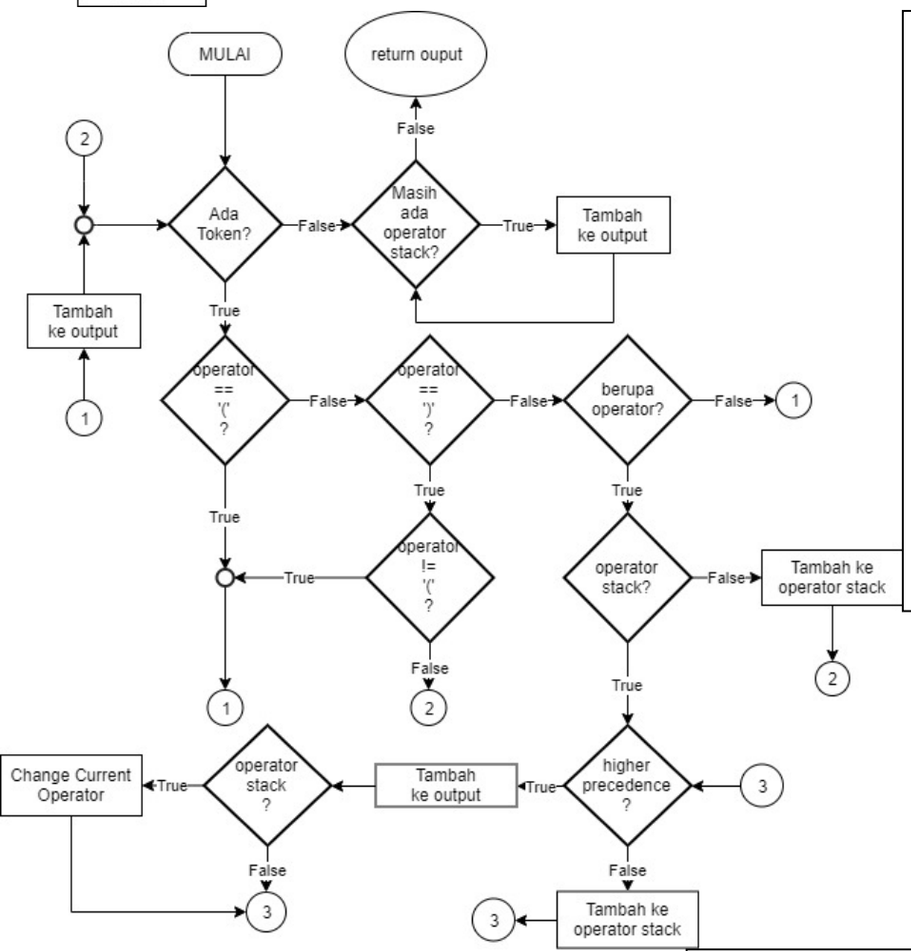




[| Process Query|]

Setelah kueri dilakukan algoritma shunting yard, setiap token pada queue akan dicek. Pertama membedakan antara **operator boolean** dengan **term**.

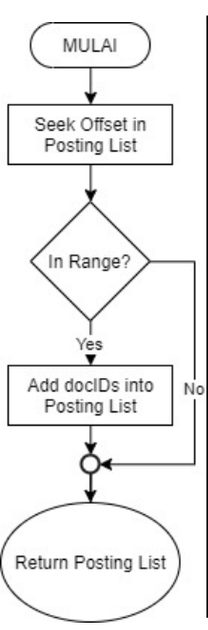
Apabila token berupa selain operator boolean, maka dianggap sebagai kueri term. Yang kemudian dicek ada/tidaknya pada dictionary.



[| Shunting Yard|]

precedence['NOT'] = 3
precedence['AND'] = 2
precedence['OR'] = 1
precedence['('] = 0
precedence[')'] = 0

Input disusun menjadi Reverse Polish notation (misal “3 + 4 × (2 − 1)” menjadi “3 4 2 1 − × +”). Disebut juga operator-precedence parsing.



[| Load Posting List|]

Fungsi ini dipanggil secara terus-menerus dengan parameter nilai offset, kemudian akan menelusuri pada offset sekian, akan ketemu term t beserta nomor dokumen yang mengandung term tersebut.

Mengembalikan nomor-nomor dokumen sebagai bagian dari hasil akhir.

[| boolean_AND |] [| boolean_OR |] [| boolean_NOT |]

Fungsi boolean_AND dan boolean_OR dijalankan dengan menerima 2 parameter, operand kiri dan kanan. Sedangkan boolean_NOT dengan 2 parameter, operand, dan nomor-nomor dokumen operand.

Operand di sini berupa term, yang mewakili nomor-nomor dokumen term.