

INDEKS *FILE* (TEKS) MENGGUNAKAN SWISH-E

by: [M. SALSABILA JAMIL](#)

SWISH-e adalah sebuah *tool* yang digunakan untuk meng-indeks teks dalam berbagai format, seperti PDF, html, txt, XML, PostScript, dll. Selain meng-indeks, *tool* ini juga dapat mencari dokumen berdasarkan *query* yang di-*input*-kan. *Tool* ini memiliki banyak fitur/kemampuan yang dimilikinya, contohnya:

- Dapat menentukan *stopword list* yang hendak digunakan
- Dapat Menentukan direktori (berisi file) yang ingin di-indeks
- Dapat Menentukan jenis file yang di-indeks (PDF?, XML?, email?, txt?, html?, dll)
- Bisa melakukan *stemming* (dalam beberapa bahasa)
- *Fast*
- Dapat meng-indeks berita dari suatu *website* (*crawling*)
- Dapat Mencari dokumen yang mengandung *query* yang dimasukkan
- Dan masih banyak lagi -> [Baca Lebih Lanjut](#)

Berikut dijelaskan cara menggunakan *tool* ini.

Langkah pertama yaitu **DOWNLOAD FILE BERISI TEKS** (disini penulis menggunakan TEKS BERITA)

1. Penulis membuat skrip untuk mengunduh berita pada website:

- **KOMPAS**: `src/spider/spider_kompas.go`
- **TEMPO**: `src/spider/spider_tempo.go`
- **ANTARA**:
 - `src/spider/spider_antara.go`
 - `src/spider/spider_antara_oto.go`

2. Tema-tema berita yang di-*download*:

- *Entertainment*/Hiburan
- *Sport*/Olahraga
- Tekno
- Otomotif

3. Bagian-bagian *HTML* yang diambil:

- *Title*/Judul
- Tanggal Terbit
- *Tag* (Opsional)
- *Content*/Isi

4. Berita yang terkumpul Dapat dilihat pada *folder*:

- `kompas/`
- `tempo/`
- `antara/`

5. Selanjutnya dilakukan proses **PREPROCESSING**:

- Menghapus `\n\r`
- Menghapus tanda baca yang dianggap ***TIDAK PENTING**

- Mengubah *spasi* lebih dari satu menjadi satu *spasi*
- *Lowercase every letter*
- Hasil disimpan pada folder:
 - **ALL/**
 - Secara keseluruhan berjumlah: **~55K**

Langkah kedua, **STEMMING**:

1. Pengertian: Mengubah kata ke bentuk dasarnya, contoh:
 - berjalan -> jalan
 - memancing -> pancing
 - menggunakan -> guna
2. Tujuan:
Sebisa mungkin mengurangi kata yang nantinya akan di-indeks, sehingga jumlah memori yang digunakan semakin sedikit
3. Proses **STEMMING** menggunakan pustaka/modul yang dikembangkan oleh [RadhiFadlillah](#) dalam bahasa **GO**, dimana RadhiFadlillah mengambil referensi dari [andylibrian](#) yang mengembangkannya dalam bahasa PHP.
4. Pustaka (kopi sebahagian dari [README.md RadhiFadlillah](#))

Algoritma

1. Algoritma Nazief dan Adriani
2. Asian J. 2007. **Effective Techniques for Indonesian Text Retrieval**. PhD thesis School of Computer Science and Information Technology RMIT University Australia. ([PDF](#) dan [Amazon](#))
3. Arifin, A.Z., I.P.A.K. Mahendra dan H.T. Ciptaningtyas. 2009. **Enhanced Confix Stripping Stemmer and Ants Algorithm for Classifying News Document in Indonesian Language**, Proceeding of International Conference on Information & Communication Technology and Systems (ICTS). ([PDF](#))
4. A. D. Tahitoe, D. Purwitasari. 2010. **Implementasi Modifikasi Enhanced Confix Stripping Stemmer Untuk Bahasa Indonesia dengan Metode Corpus Based Stemming**, Institut Teknologi Sepuluh Nopember (ITS) – Surabaya, 60111, Indonesia. ([PDF](#))
5. Tambahan aturan *stemming* dari [kontributor Sastrawi](#).

Kamus Kata Dasar

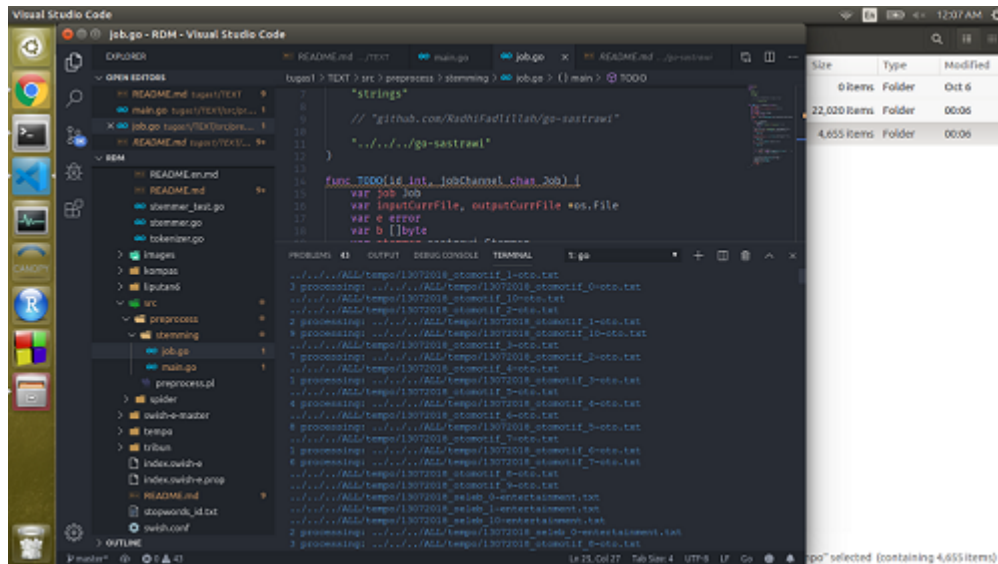
Proses stemming oleh Sastrawi sangat bergantung pada kamus kata dasar. Sastrawi menggunakan kamus kata dasar dari [kateglo.com](#) dengan sedikit perubahan.

Lisensi

Sebagaimana [Sastrawi](#) untuk PHP, [Go-Sastrawi](#) untuk GO, proyek ini juga disebar dengan lisensi **MIT**. Untuk lisensi kamus kata dasar dari Kateglo adalah **CC-BY-NC-SA 3.0**.

5. Proses **STEMMING** dilakukan menggunakan skrip yang terdapat di folder [src/preprocess/stemming/main.go](#)

6. Cuplikan ketika proses *STEMMING* berlangsung:

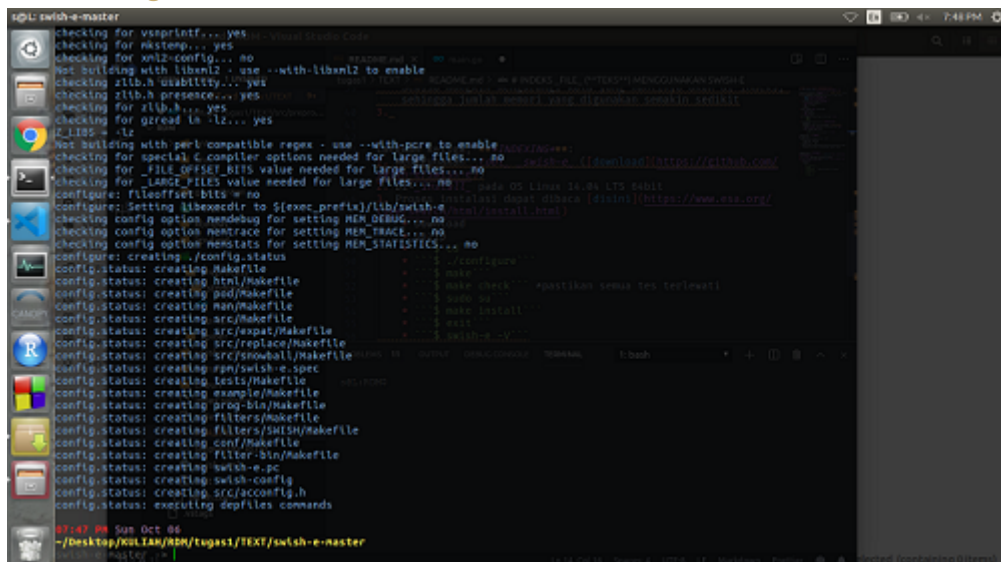


7. Hasil *STEMMING* ditaruh di folder:

- ALL_2/kompas
- ALL_2/tempo
- ALL_2/antara

Langkah ketiga, **INDEXING**:

1. Menggunakan *tool swish-e* ([download](#))
2. Di-*install* pada OS Linux 16.04 LTS 64bit
3. Proses instalasi dapat dibaca [disini](#)
 - Download
 - Unzip
 - \$ cd
 - \$./configure



- \$ make

[illegible]

- \$ make check *pastikan semua tes berhasil

```
nake[1]: Nothing to be done for 'check'.
nake[1]: Leaving directory '/home/s/Desktop/KULIAH/RDN/tugas1/TEXT/swish-e-master/jan'
Making check in src
nake[1]: Entering directory '/home/s/Desktop/KULIAH/RDN/tugas1/TEXT/swish-e-master/src'
Making check in expat
nake[1]: Entering directory '/home/s/Desktop/KULIAH/RDN/tugas1/TEXT/swish-e-master/src/expat'
nake[1]: Nothing to be done for 'check'.
nake[1]: Leaving directory '/home/s/Desktop/KULIAH/RDN/tugas1/TEXT/swish-e-master/src/expat'
Making check in replace
nake[1]: Entering directory '/home/s/Desktop/KULIAH/RDN/tugas1/TEXT/swish-e-master/src/replace'
nake[1]: Nothing to be done for 'check'.
nake[1]: Leaving directory '/home/s/Desktop/KULIAH/RDN/tugas1/TEXT/swish-e-master/src/replace'
Making check in snowball
nake[1]: Entering directory '/home/s/Desktop/KULIAH/RDN/tugas1/TEXT/swish-e-master/src/snowball'
nake[1]: Nothing to be done for 'check'.
nake[1]: Leaving directory '/home/s/Desktop/KULIAH/RDN/tugas1/TEXT/swish-e-master/src/snowball'
nake[1]: Entering directory '/home/s/Desktop/KULIAH/RDN/tugas1/TEXT/swish-e-master/src'
nake[1]: Leaving directory '/home/s/Desktop/KULIAH/RDN/tugas1/TEXT/swish-e-master/src'
nake[1]: Leaving directory '/home/s/Desktop/KULIAH/RDN/tugas1/TEXT/swish-e-master/src'
Making check in tests
nake[1]: Entering directory '/home/s/Desktop/KULIAH/RDN/tugas1/TEXT/swish-e-master/tests'
nake[1]: check_tests
PASS: check_index
PASS: check_search
PASS: check_metasearch
PASS: check_fuzzy
=====
All 4 tests passed
nake[1]: Leaving directory '/home/s/Desktop/KULIAH/RDN/tugas1/TEXT/swish-e-master/tests'
nake[1]: Leaving directory '/home/s/Desktop/KULIAH/RDN/tugas1/TEXT/swish-e-master/tests'
Making check in pod
nake[1]: Entering directory '/home/s/Desktop/KULIAH/RDN/tugas1/TEXT/swish-e-master/pod'
nake[1]: Nothing to be done for 'check'.
nake[1]: Leaving directory '/home/s/Desktop/KULIAH/RDN/tugas1/TEXT/swish-e-master/pod'
nake[1]: Entering directory '/home/s/Desktop/KULIAH/RDN/tugas1/TEXT/swish-e-master'
nake[1]: Nothing to be done for 'check-an'.
nake[1]: Leaving directory '/home/s/Desktop/KULIAH/RDN/tugas1/TEXT/swish-e-master'

B7:49 PM Sun Oct 26
~/Desktop/KULIAH/RDN/tugas1/TEXT/swish-e-master
git
```

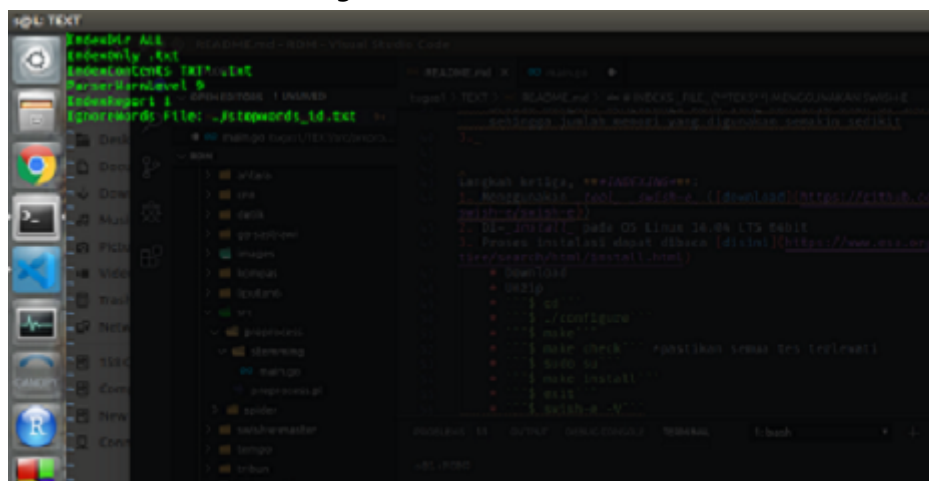
- `$ sudo su`
- `$ make install`

```
root@kali: /home/s/Desktop/KULIAH/RDM/tugas1/TEXT/swish-e-master
test -z "/usr/local/include" || mkdir -p -- "/usr/local/include"
/usr/bin/install -c -m 644 "swish-e.h" "/usr/local/include/swish-e.h"
make[3]: Leaving directory /home/s/Desktop/KULIAH/RDM/tugas1/TEXT/swish-e-master/src/
make[2]: Leaving directory /home/s/Desktop/KULIAH/RDM/tugas1/TEXT/swish-e-master/src/
make[1]: Leaving directory /home/s/Desktop/KULIAH/RDM/tugas1/TEXT/swish-e-master/src/
making install in tests
make[1]: Entering directory /home/s/Desktop/KULIAH/RDM/tugas1/TEXT/swish-e-master/tests/
make[2]: Entering directory /home/s/Desktop/KULIAH/RDM/tugas1/TEXT/swish-e-master/tests/
make[2]: Nothing to be done for "install-exec-am".
make[2]: Nothing to be done for "install-data-am".
make[2]: Leaving directory /home/s/Desktop/KULIAH/RDM/tugas1/TEXT/swish-e-master/tests/
make[1]: Leaving directory /home/s/Desktop/KULIAH/RDM/tugas1/TEXT/swish-e-master/tests/
making install in pod
make[1]: Entering directory /home/s/Desktop/KULIAH/RDM/tugas1/TEXT/swish-e-master/pod/
make[2]: Entering directory /home/s/Desktop/KULIAH/RDM/tugas1/TEXT/swish-e-master/pod/
make[2]: Nothing to be done for "install-exec-am".
test -z "/usr/local/share/doc/swish-e-pod" || mkdir -p -- "/usr/local/share/doc/swish-e-pod"
/usr/bin/install -c -m 644 "CHANGES.pod" /usr/local/share/doc/swish-e-pod/CHANGES.pod
/usr/bin/install -c -m 644 "INSTALL.pod" /usr/local/share/doc/swish-e-pod/INSTALL.pod
/usr/bin/install -c -m 644 "README.pod" /usr/local/share/doc/swish-e-pod/README.pod
/usr/bin/install -c -m 644 "SWISH-FAQ.pod" /usr/local/share/doc/swish-e-pod/SWISH-FAQ.pod
/usr/bin/install -c -m 644 "SWISH-BUGS.pod" /usr/local/share/doc/swish-e-pod/SWISH-BUGS.pod
/usr/bin/install -c -m 644 "SWISH-CONFIG.pod" /usr/local/share/doc/swish-e-pod/SWISH-CONFIG.pod
/usr/bin/install -c -m 644 "swish-e.pod" /usr/local/share/doc/swish-e-pod/swish-e.pod
/usr/bin/install -c -m 644 "SWISH-FAQ.pod" /usr/local/share/doc/swish-e-pod/SWISH-FAQ.pod
/usr/bin/install -c -m 644 "SWISH-BUGS.pod" /usr/local/share/doc/swish-e-pod/SWISH-BUGS.pod
/usr/bin/install -c -m 644 "SWISH-CONFIG.pod" /usr/local/share/doc/swish-e-pod/SWISH-CONFIG.pod
/usr/bin/install -c -m 644 "swish-e.pod" /usr/local/share/doc/swish-e-pod/swish-e.pod
/usr/bin/install -c -m 644 "SWISH-FAQ.pod" /usr/local/share/doc/swish-e-pod/SWISH-FAQ.pod
/usr/bin/install -c -m 644 "SWISH-BUGS.pod" /usr/local/share/doc/swish-e-pod/SWISH-BUGS.pod
/usr/bin/install -c -m 644 "SWISH-CONFIG.pod" /usr/local/share/doc/swish-e-pod/SWISH-CONFIG.pod
/usr/bin/install -c -m 644 "swish-e.pod" /usr/local/share/doc/swish-e-pod/swish-e.pod
make[1]: Leaving directory /home/s/Desktop/KULIAH/RDM/tugas1/TEXT/swish-e-master/pod/
make[1]: Leaving directory /home/s/Desktop/KULIAH/RDM/tugas1/TEXT/swish-e-master/pod/
make[1]: Entering directory /home/s/Desktop/KULIAH/RDM/tugas1/TEXT/swish-e-master/
make[2]: Entering directory /home/s/Desktop/KULIAH/RDM/tugas1/TEXT/swish-e-master/
test -z "/usr/local/bin" || mkdir -p -- "/usr/local/bin"
/usr/bin/install -c "swish-config" /usr/local/bin/swish-config
test -z "/usr/local/share/doc/swish-e" || mkdir -p -- "/usr/local/share/doc/swish-e"
/usr/bin/install -c -m 644 "INSTALL" /usr/local/share/doc/swish-e/INSTALL
/usr/bin/install -c -m 644 "README" /usr/local/share/doc/swish-e/README
/usr/bin/install -c -m 644 "README.cvs" /usr/local/share/doc/swish-e/README.cvs
test -z "/usr/local/lib/pkgconfig" || mkdir -p -- "/usr/local/lib/pkgconfig"
/usr/bin/install -c -m 644 "swish-e.pc" /usr/local/lib/pkgconfig/swish-e.pc
make[1]: Leaving directory /home/s/Desktop/KULIAH/RDM/tugas1/TEXT/swish-e-master/
make[1]: Leaving directory /home/s/Desktop/KULIAH/RDM/tugas1/TEXT/swish-e-master/
root@kali: /home/s/Desktop/KULIAH/RDM/tugas1/TEXT/swish-e-master
```

- o \$ exit

-

- Versi tidak ada *stemming*



-

5 / 8

dapat di-download melalui link [ini](https://sites.google.com/site/kevinbouge/stopwords-lists). Data *Stopwords* ini didapat pada *website* sites.google.com/site/kevinbouge/stopwords-lists.

- no stem

```

07:53 PM Sun Oct 06
~/Desktop/KULIAH/RM/tugas1/TEXT
$ ./swish-e -c swish.conf
Indexing Data Source: "File-System"
Indexing "All"
Removing very common words...
Writing main index...
Sorting words ...
Sorting 186,758 words alphabetically
Writing header ...
Writing index entries ...
Writing word text: Complete
Writing word hash: Complete
Writing word data: Complete
186,758 unique words indexed.
4 properties sorted.
56,295 files indexed. 99,057,441 total bytes. 186,758 total words.
Elapsed time: 00:07:28 CPU time: 00:00:25
Indexing done!
  
```

- with stem

```

12:14 AM Mon Oct 07
~/Desktop/KULIAH/RM/tugas1/TEXT
$ ./swish-e -c swish.conf -f index_after_stemmed.swish-e.conf
Indexing Data Source: "File-System"
Indexing "All"
Removing very common words...
Writing main index...
Sorting words ...
Sorting 162,450 words alphabetically
Writing header ...
Writing index entries ...
Writing word text: Complete
Writing word hash: Complete
Writing word data: Complete
162,450 unique words indexed.
4 properties sorted.
56,295 files indexed. 88,943,833 total bytes. 162,450 total words.
Elapsed time: 00:09:48 CPU time: 00:00:20
Indexing done!
  
```

6. Dapat dilihat *file* yang diindeks berjumlah **56295 files**
7. Waktu peng-indeks-an: **~10 menit**
8. Jumlah *Unique words*:
 - no stem: **~186K**
 - with stem: **~162K**
9. Terlihat perbedaan mencolok antara **no stem** dan **with stem** dalam hal jumlah *Unique words*
10. Uji Coba *search* sebuah kata:

- no stem

```

12:55 AM Mon Oct 07
~/Desktop/KULIAH/RM/tugas1/TEXT
TEXT -> swish-e -f index.swish-e -m 10 -w jakarta
# SMISH Format: 2.5.0
# Search words: jakarta
# Removed stopwords:
# Number of hits: 37282
# Search time: 0.087 seconds
# Run time: 0.016 seconds
1000 ALL/tempo/04012018_sport_8-sport.txt "04012018_sport_8-sport.txt" 8357
932 ALL/tempo/12032018_sport_8-sport.txt "12032018_sport_8-sport.txt" 2871
825 ALL/tempo/10042018_sport_2-sport.txt "10042018_sport_2-sport.txt" 3579
934 ALL/tempo/13122018_sport_3-sport.txt "13122018_sport_3-sport.txt" 3593
904 ALL/tempo/20032018_sport_5-sport.txt "20032018_sport_5-sport.txt" 2263
781 ALL/kompas/02012018_1-oto.txt "02012018_1-oto.txt" 2785
788 ALL/tempo/01042018_sport_1-sport.txt "01042018_sport_1-sport.txt" 1383
788 ALL/jantara/09112018_0_0-entertainment.txt "09112018_0_0-entertainment.txt" 3291
755 ALL/jantara/11032018_0_3-entertainment.txt "11032018_0_3-entertainment.txt" 3208
755 ALL/kompas/400_1-sport.txt "400_1-sport.txt" 3329

```

- with stem

```

12:57 AM Mon Oct 07
~/Desktop/KULIAH/RM/tugas1/TEXT
TEXT -> swish-e -f index_after_stemmed.swish-e -m 10 -w jakarta
# SMISH Format: 2.5.0
# Search words: jakarta
# Removed stopwords:
# Number of hits: 37281
# Search time: 0.089 seconds
# Run time: 0.016 seconds
1000 ALL_2/tempo/04012018_sport_8-sport.txt "04012018_sport_8-sport.txt" 7972
932 ALL_2/tempo/12032018_sport_8-sport.txt "12032018_sport_8-sport.txt" 2785
825 ALL_2/tempo/10042018_sport_2-sport.txt "10042018_sport_2-sport.txt" 3446
934 ALL_2/tempo/13122018_sport_3-sport.txt "13122018_sport_3-sport.txt" 3228
904 ALL_2/tempo/20032018_sport_5-sport.txt "20032018_sport_5-sport.txt" 2164
781 ALL_2/kompas/02012018_1-oto.txt "02012018_1-oto.txt" 1681
788 ALL_2/jantara/09112018_0_0-entertainment.txt "09112018_0_0-entertainment.txt" 2964
755 ALL_2/jantara/11032018_0_3-entertainment.txt "11032018_0_3-entertainment.txt" 2186
755 ALL_2/tempo/02042018_sport_11-sport.txt "02042018_sport_11-sport.txt" 1832

```

- Perbedaannya tidak terlalu jauh dari hal jumlah dokumen yang berhasil ditemukan menggunakan kata tersebut.

11. Jika ingin mencari dokumen dengan jumlah kata lebih dari satu, maka gunakan tanda petik dua ("[query]"), contohnya `swish-e -f [index.swish-e] -m 10 -w "jakarta bandung surabaya"`. *Query* tersebut akan mencari dokumen yang mengandung kata **jakarta AND bandung AND surabaya**. Jika dalam *query* tersebut ada kata yang terdapat dalam daftar *stopword*, maka akan dihapus/dilewatkan/tidak di proses. Argument dari parameter `-f` adalah nama file hasil *indexing* (default: `index.swish-e`).

12. Arti dari Output (format):

- Kolom-1: Ranking
- Kolom-2: Lokasi dokumen/file
- Kolom-3: Judul>Nama dokumen/file
- Kolom-4: Size/Ukuran dokumen (dalam *bytes*) **BACA LEBIH LANJUT**

---Download---

Semua file yang digunakan dalam projek ini dapat di unduh pada link berikut [UNDUH FILE](#)

github

Versi enak dilihat -> [link](#)