

# 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:  
Sebisanya 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

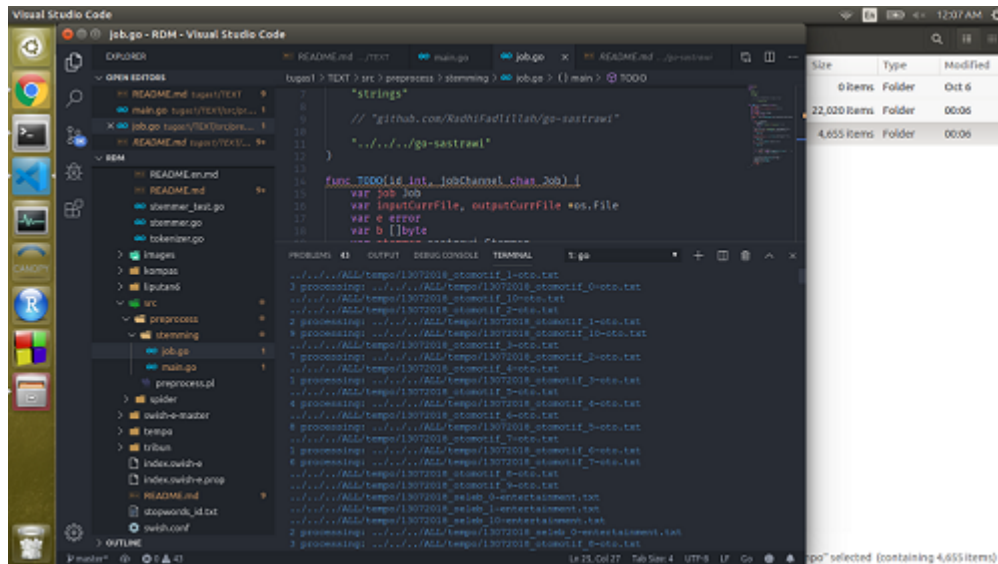
Proses stemming oleh Sastrawi sangat bergantung pada kamus kata dasar. Sastrawi menggunakan kamus kata dasar dari [kateglo.com](#) dengan sedikit perubahan.

### Lisensi

Sebagaimana [Sastrawi](#) untuk PHP, [Go-Sastrawi](#) untuk GO, projek 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:



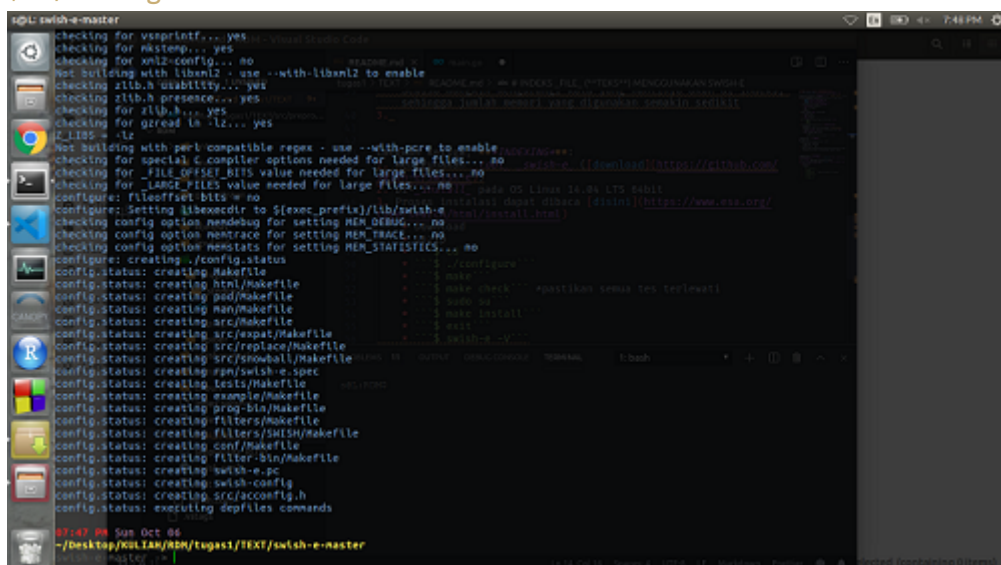
## 7. Hasil *STEMMING* ditaruh di folder:

- 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



- [illegible]

- ```

kali@kali:~$ sudo swish-e-master
make[1]: Nothing to be done for 'check'.
make[1]: Leaving directory '/home/s/Desktop/KULIAH/RDM/tugasi/TEXT/swish-e-master/jman'
Making check in src
make[1]: Entering directory '/home/s/Desktop/KULIAH/RDM/tugasi/TEXT/swish-e-master/src'
make[1]: Entering directory '/home/s/Desktop/KULIAH/RDM/tugasi/TEXT/swish-e-master/src/compat'
make[2]: Nothing to be done for 'check'.
make[2]: Leaving directory '/home/s/Desktop/KULIAH/RDM/tugasi/TEXT/swish-e-master/src/compat'
Making check in replace
make[2]: Entering directory '/home/s/Desktop/KULIAH/RDM/tugasi/TEXT/swish-e-master/src/replace'
make[2]: Nothing to be done for 'check'.
make[2]: Leaving directory '/home/s/Desktop/KULIAH/RDM/tugasi/TEXT/swish-e-master/src/replace'
Making check in snowball
make[2]: Entering directory '/home/s/Desktop/KULIAH/RDM/tugasi/TEXT/swish-e-master/src/snowball'
make[2]: Nothing to be done for 'check'.
make[2]: Leaving directory '/home/s/Desktop/KULIAH/RDM/tugasi/TEXT/swish-e-master/src/snowball'
make[1]: Entering directory '/home/s/Desktop/KULIAH/RDM/tugasi/TEXT/swish-e-master/src'
make[1]: Leaving directory '/home/s/Desktop/KULIAH/RDM/tugasi/TEXT/swish-e-master/src'
Making check in tests
make[1]: Entering directory '/home/s/Desktop/KULIAH/RDM/tugasi/TEXT/swish-e-master/tests'
make[1]: check-tests
PASS: check_index
PASS: check_search
PASS: check_metasearch
PASS: check_fuzzy
=====
All 4 tests passed
make[1]: Leaving directory '/home/s/Desktop/KULIAH/RDM/tugasi/TEXT/swish-e-master/tests'
make[1]: Leaving directory '/home/s/Desktop/KULIAH/RDM/tugasi/TEXT/swish-e-master/tests'
Making check in pod
make[1]: Entering directory '/home/s/Desktop/KULIAH/RDM/tugasi/TEXT/swish-e-master/pod'
make[1]: Nothing to be done for 'check'.
make[1]: Leaving directory '/home/s/Desktop/KULIAH/RDM/tugasi/TEXT/swish-e-master/pod'
make[1]: Entering directory '/home/s/Desktop/KULIAH/RDM/tugasi/TEXT/swish-e-master'
make[1]: Nothing to be done for 'check-an'.
make[1]: Leaving directory '/home/s/Desktop/KULIAH/RDM/tugasi/TEXT/swish-e-master'

07:49 PM Sun Oct 26
~/Desktop/KULIAH/RDM/tugasi/TEXT/swish-e-master
git

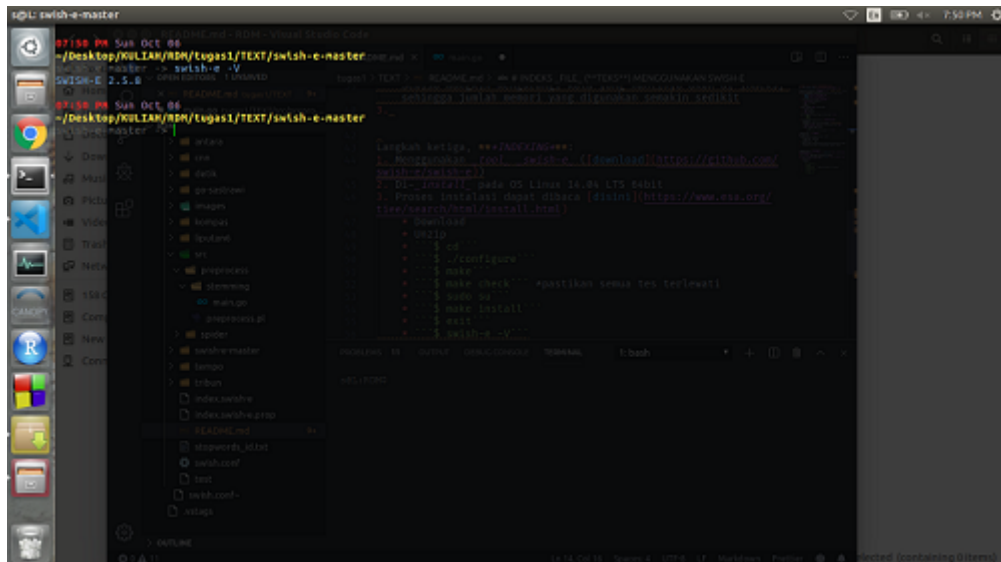
```

- ```
root@kali: /home/s/Desktop/KULIAH/RDR/tugas1/TEXT/switch-e-master
test -z "/usr/local/include" || mkdir -p -- "/usr/local/include"
/usr/bin/install -c -m 444 "/usr/local/include/swish-e.h" "/usr/local/include/swish-e.h"
make[1]: Leaving directory /home/s/Desktop/KULIAH/RDR/tugas1/TEXT/switch-e-master/src/
make[1]: Leaving directory /home/s/Desktop/KULIAH/RDR/tugas1/TEXT/switch-e-master/src/
make[1]: Leaving directory /home/s/Desktop/KULIAH/RDR/tugas1/TEXT/switch-e-master/src/
make[1]: Entering directory /home/s/Desktop/KULIAH/RDR/tugas1/TEXT/switch-e-master/test/
make[1]: Entering directory /home/s/Desktop/KULIAH/RDR/tugas1/TEXT/switch-e-master/test/
make[1]: Nothing to be done for 'install-exec-am'.
make[1]: Nothing to be done for 'install-data-am'.
make[1]: Leaving directory /home/s/Desktop/KULIAH/RDR/tugas1/TEXT/switch-e-master/test/
make[1]: Leaving directory /home/s/Desktop/KULIAH/RDR/tugas1/TEXT/switch-e-master/test/
make[1]: Leaving directory /home/s/Desktop/KULIAH/RDR/tugas1/TEXT/switch-e-master/test/
make[1]: Entering directory /home/s/Desktop/KULIAH/RDR/tugas1/TEXT/switch-e-master/pod/
make[1]: Entering directory /home/s/Desktop/KULIAH/RDR/tugas1/TEXT/switch-e-master/pod/
make[1]: 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 444 "CHANGES.pod" "/usr/local/share/doc/swish-e-pod/CHANGES.pod"
/usr/bin/install -c -m 444 "INSTALL.pod" "/usr/local/share/doc/swish-e-pod/INSTALL.pod"
/usr/bin/install -c -m 444 "README.pod" "/usr/local/share/doc/swish-e-pod/README.pod"
/usr/bin/install -c -m 444 "SWISH-3.pod" "/usr/local/share/doc/swish-e-pod/SWISH-3.pod"
/usr/bin/install -c -m 444 "SWISH-BUGS.pod" "/usr/local/share/doc/swish-e-pod/SWISH-BUGS.pod"
/usr/bin/install -c -m 444 "SWISH-CONFIG.pod" "/usr/local/share/doc/swish-e-pod/SWISH-CONFIG.pod"
/usr/bin/install -c -m 444 "swish-e.pod" "/usr/local/share/doc/swish-e-pod/swish-e.pod"
/usr/bin/install -c -m 444 "SWISH-FAQ.pod" "/usr/local/share/doc/swish-e-pod/SWISH-FAQ.pod"
/usr/bin/install -c -m 444 "SWISH-FAQ.pod" "/usr/local/share/doc/swish-e-pod/SWISH-FAQ.pod"
/usr/bin/install -c -m 444 "SWISH-RUN.pod" "/usr/local/share/doc/swish-e-pod/SWISH-RUN.pod"
/usr/bin/install -c -m 444 "SWISH-RUN.pod" "/usr/local/share/doc/swish-e-pod/SWISH-RUN.pod"
/usr/bin/install -c -m 444 "SWISH-SEARCH.pod" "/usr/local/share/doc/swish-e-pod/SWISH-SEARCH.pod"
make[1]: Leaving directory /home/s/Desktop/KULIAH/RDR/tugas1/TEXT/switch-e-master/pod/
make[1]: Leaving directory /home/s/Desktop/KULIAH/RDR/tugas1/TEXT/switch-e-master/pod/
make[1]: Entering directory /home/s/Desktop/KULIAH/RDR/tugas1/TEXT/switch-e-master/
make[1]: Entering directory /home/s/Desktop/KULIAH/RDR/tugas1/TEXT/switch-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 444 "INSTALL" "/usr/local/share/doc/swish-e/INSTALL"
/usr/bin/install -c -m 444 "README" "/usr/local/share/doc/swish-e/README"
/usr/bin/install -c -m 444 "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 444 "swish-e.pc" "/usr/local/lib/pkgconfig/swish-e.pc"
make[1]: Leaving directory /home/s/Desktop/KULIAH/RDR/tugas1/TEXT/switch-e-master/
make[1]: Leaving directory /home/s/Desktop/KULIAH/RDR/tugas1/TEXT/switch-e-master/
root@kali: /home/s/Desktop/KULIAH/RDR/tugas1/TEXT/switch-e-master
```

- 4 / 9

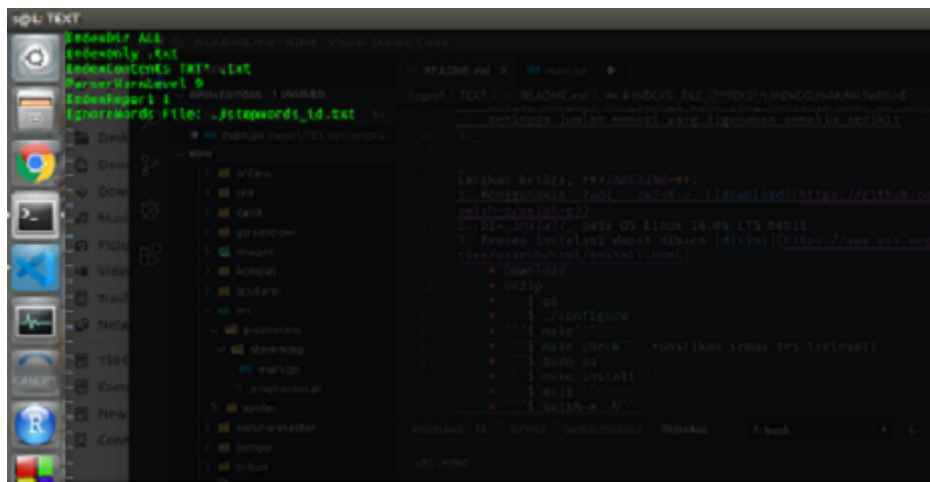


- `$ swish-e -V` SWISH-E 2.5.8

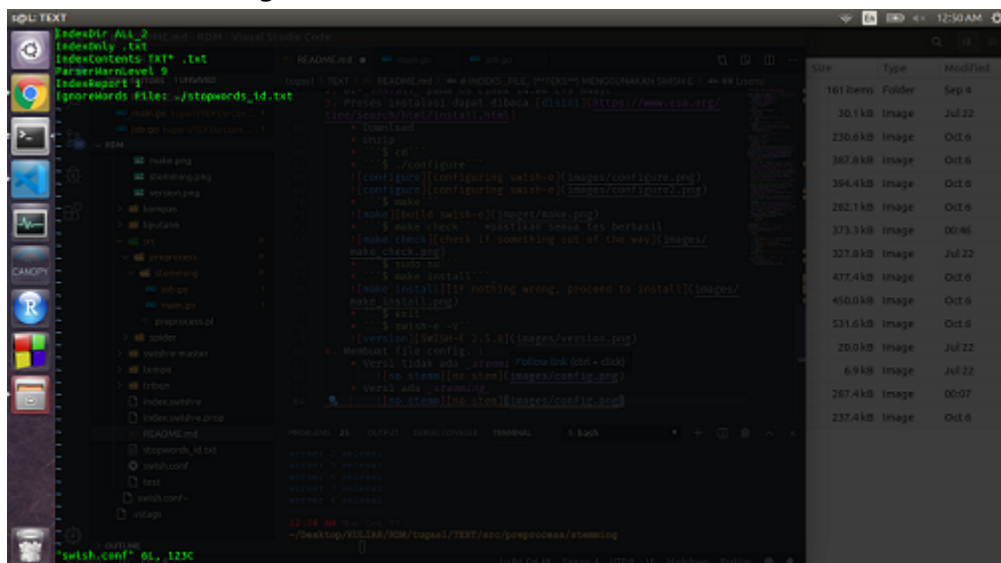


#### 4. Membuat file config. : [BACA LEBIH LANJUT](#)

- Versi tidak ada *stemming*



- Versi ada *stemming*



#### 5. Konfigurasi proses peng-indeks-an tercantum dalam file config yang telah dibuat: **swish.conf**. \*Perlu diperhatikan adanya pemangkasan kata umum (*common words*)

(*stopwords*) menggunakan data *stopwords* bahasa indonesia (**stopwords\_id.txt**) yang dapat di-download melalui link [ini](https://sites.google.com/site/kevinbougue/stopwords-lists). Data *Stopwords* ini didapat pada *website* [sites.google.com/site/kevinbougue/stopwords-lists](https://sites.google.com/site/kevinbougue/stopwords-lists).

```
IndexFile ./indeks.swish-e
IndexDir ALL_2
IndexOnly .txt
IndexContents TXT* .txt
ParserWarnLevel 9
IndexReport 1
IgnoreWords File: ./stopwords_id.txt
```

swish.conf

- no stem  
IndexDir ALL
- with stem  
IndexDir ALL\_2

6. Dapat dilihat *file* yang diindeks berjumlah **56295 files**

7. Waktu peng-indeks-an: **~10 menit**

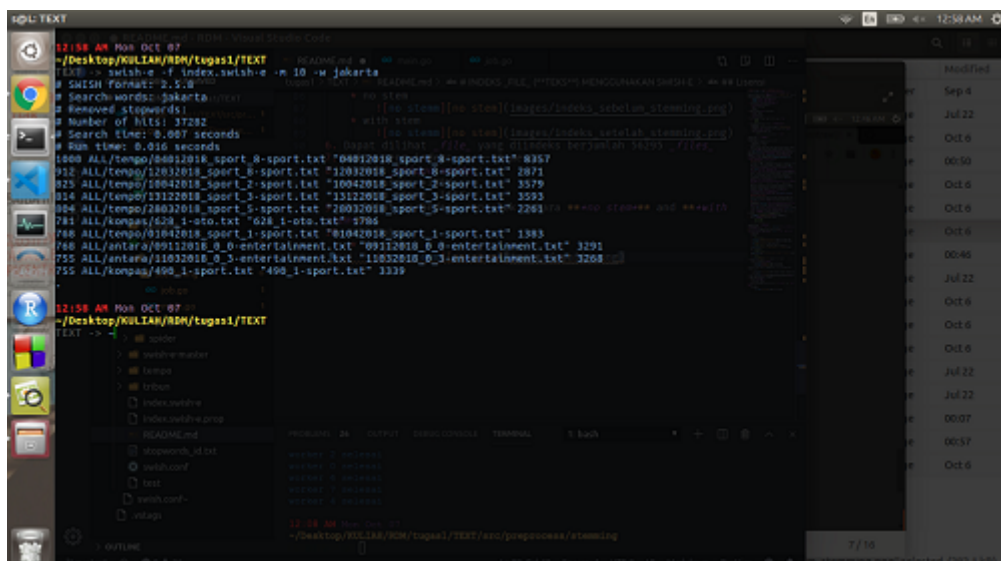
8. Jumlah *Unique words*:

- 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



- 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
# SWISH Format: 2.5.0
# Search words: jakarta
# Removed stopwords:
# Number of hits: 37281
# Search time: 0.009 seconds
# Run time: 0.016 seconds
1408 ALL_2/tempo/04012018_sport_8-sport.txt "04012018_sport_8-sport.txt" 7972
912 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" 3448
834 ALL_2/tempo/13122018_sport_3-sport.txt "13122018_sport_3-sport.txt" 3228
864 ALL_2/tempo/28032018_sport_1-sport.txt "28032018_sport_1-sport.txt" 2164
741 ALL_2/tempo/02042018_1-oto.txt "02042018_1-oto.txt" 1681
748 ALL_2/antara/09112018_0-0-entertainment.txt "09112018_0-0-entertainment.txt" 2984
768 ALL_2/tempo/01042018_sport_1-sport.txt "01042018_sport_1-sport.txt" 1305
755 ALL_2/antara/02112018_1-5-entertainment.txt "02112018_1-5-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`).

- Query: "jakarta bandung surabaya" atau **jakarta AND bandung AND surabaya**  
Hit: 390

```

03:07 AM Sat Oct 12
~/Desktop/KULIAH/RM/tugas1/TEXT
TEXT: swish-e -f index_after_stemmed.swish-e -m 10 -w "jakarta AND bandung AND surabaya"
# SWISH Format: 2.5.0
# Search words: jakarta AND bandung AND surabaya
# Removed stopwords:
# Number of hits: 390
# Search time: 0.015 seconds
# Run time: 0.015 seconds
1408 ALL_2/tempo/12032018_sport_8-sport.txt "12032018_sport_8-sport.txt" 2785
940 ALL_2/tempo/10042018_sport_2-sport.txt "10042018_sport_2-sport.txt" 3448
836 ALL_2/tempo/28032018_sport_5-sport.txt "28032018_sport_5-sport.txt" 2164
748 ALL_2/tempo/03042018_sport_11-sport.txt "03042018_sport_11-sport.txt" 1580
756 ALL_2/tempo/02042018_sport_11-sport.txt "02042018_sport_11-sport.txt" 1832
756 ALL_2/tempo/01042018_sport_4-sport.txt "01042018_sport_4-sport.txt" 1658
741 ALL_2/tempo/13102018_tekno_4-tekno.txt "13102018_tekno_4-tekno.txt" 3557
729 ALL_2/tempo/09032017_sport_5-sport.txt "09032017_sport_5-sport.txt" 2820
943 ALL_2/tempo/13122018_sport_3-sport.txt "13122018_sport_3-sport.txt" 3228
943 ALL_2/tempo/09032017_sport_8-sport.txt "09032017_sport_8-sport.txt" 2408

```

- Query: "pandji pragiwaksono dan raditya dika"  
Hit: 8

```

03:07 AM Sat Oct 12
~/Desktop/KULIAH/RM/tugas1/TEXT
TEXT --
./Desktop/KULIAH/RM/tugas1/TEXT
# SWISH Format: 2.5.0
# Search words: pandji pragiwaksono dan raditya dika
# Removed stopwords: dan
# Number of hits: 0
# Search time: 0.001 seconds
# Run time: 0.070 seconds
1000 ALL_2/kompas/565_5-entertainment.txt "565_5-entertainment.txt" 1235
982 ALL_2/kompas/545_8-entertainment.txt "545_8-entertainment.txt" 1541
982 ALL_2/kompas/313_9-entertainment.txt "313_9-entertainment.txt" 1455
948 ALL_2/kompas/545_8-entertainment.txt "545_8-entertainment.txt" 1356
710 ALL_2/kompas/583_6-entertainment.txt "583_6-entertainment.txt" 1276
986 ALL_2/kompas/513_7-entertainment.txt "513_7-entertainment.txt" 1876
475 ALL_2/kompas/568_11-entertainment.txt "568_11-entertainment.txt" 1310
475 ALL_2/kompas/568_8-entertainment.txt "568_8-entertainment.txt" 1781

```

Dapat di-intepret sebagai: "cari dokumen yang mengandung kata **pandji AND pragiwaksono AND dan AND raditya AND dika**"

Kata 'dan' tidak diproses karena termasuk dalam daftar *stopwords*

- Query: "raditya dika AND NOT pandji pragiwaksono" dan "raditya dika OR pandji pragiwaksono"

Hit: 0 dan 71

Dapat di-intepret sebagai:

- "cari dokumen yang mengandung kata **raditya AND dika tapi tidak ada kata pandji AND pragiwaksono**"
- "cari dokumen yang mengandung kata **raditya AND dika atau pandji AND pragiwaksono**"

```

03:49 AM Sat Oct 12
~/Desktop/KULIAH/RM/tugas1/TEXT
TEXT --
./Desktop/KULIAH/RM/tugas1/TEXT
# SWISH Format: 2.5.0
# Search words: raditya dika AND NOT pandji pragiwaksono
# Removed stopwords:
# Number of hits: 0
# Search time: 0.001 seconds
# Run time: 0.001 seconds
1000 ALL_2/antara/25872018_1_3-entertainment.txt "25872018_1_3-entertainment.txt" 3485
920 ALL_2/kompas/313_9-entertainment.txt "313_9-entertainment.txt" 1455
982 ALL_2/tempo/18832018_seleb_6-entertainment.txt "18832018_seleb_6-entertainment.txt" 2061
791 ALL_2/tempo/06862018_seleb_1-entertainment.txt "06862018_seleb_1-entertainment.txt" 1568
784 ALL_2/kompas/545_8-entertainment.txt "545_8-entertainment.txt" 1356
784 ALL_2/kompas/585_5-entertainment.txt "585_5-entertainment.txt" 1135
710 ALL_2/kompas/545_8-entertainment.txt "545_8-entertainment.txt" 1541
710 ALL_2/tempo/18832018_seleb_4-entertainment.txt "18832018_seleb_4-entertainment.txt" 1685
710 ALL_2/tempo/18832018_seleb_11-entertainment.txt "18832018_seleb_11-entertainment.txt" 1882
782 ALL_2/tempo/20872018_seleb_18-entertainment.txt "20872018_seleb_18-entertainment.txt" 1212

```

Parameter `-m [n]` digunakan untuk menampilkan *top n* dokumen, misalkan `-m 10` menampilkan 10 dokumen teratas.

## 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 proyek ini dapat di unduh pada link berikut [UNDUH FILE](#)

github

Versi enak dilihat -> [link](#)

## Data

File konfigurasi dapat dilihat [di sini](#)