

Final Lab PI

MEMBANGUN SEARCH ENGINE

disusun untuk memenuhi
tugas praktikum Penelusuran Informasi

Oleh:

WIRDAYANI
1708107010008





























**PROGRAM STUDI INFORMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS SYIAH KUALA
DARUSSALAM, BANDA ACEH
2020**

1. Melakukan crawling

Tahap pertama yaitu melakukan crawling dalam bahasa pemrograman python untuk memperoleh beberapa data pada halaman berita web online. Data yang diperoleh kemudian dimasukkan ke dalam folder. Berikut proses melakukan crawling.

```
wirda@DESKTOP-7R09IST:/mnt/c/xampp/htdocs/finalpi/code$ python3 crawling.py
Sedang Diproses...
selesai 11 berita
selesai 21 berita
selesai 31 berita
selesai 41 berita
selesai 51 berita
selesai 61 berita
selesai 71 berita
selesai 81 berita
selesai 91 berita
selesai 100 berita
```

Url yang diambil adalah <https://antaranews.com>, dengan limit halaman 100. Setelah scraping berhasil dilakukan, hasil dari proses crawling dimasukkan ke dalam satu folder dengan format .txt.

Name	Date modified	Type	Size
 berita1.txt	1/3/2020 6:46 AM	Text Document	3 KB
 berita2.txt	1/3/2020 6:46 AM	Text Document	3 KB
 berita3.txt	1/3/2020 6:46 AM	Text Document	3 KB
 berita4.txt	1/3/2020 6:46 AM	Text Document	3 KB
 berita5.txt	1/3/2020 6:46 AM	Text Document	2 KB
 berita6.txt	1/3/2020 6:46 AM	Text Document	3 KB
 berita7.txt	1/3/2020 6:46 AM	Text Document	3 KB
 berita8.txt	1/3/2020 6:46 AM	Text Document	3 KB
 berita9.txt	1/3/2020 6:46 AM	Text Document	2 KB
 berita10.txt	1/3/2020 6:46 AM	Text Document	4 KB
 berita11.txt	1/3/2020 6:46 AM	Text Document	3 KB
 berita12.txt	1/3/2020 6:46 AM	Text Document	4 KB
 berita13.txt	1/3/2020 6:46 AM	Text Document	3 KB
 berita14.txt	1/3/2020 6:46 AM	Text Document	3 KB
 berita15.txt	1/3/2020 6:46 AM	Text Document	3 KB
 berita16.txt	1/3/2020 6:46 AM	Text Document	1 KB
 berita17.txt	1/3/2020 6:46 AM	Text Document	2 KB
 berita18.txt	1/3/2020 6:46 AM	Text Document	1 KB
 berita19.txt	1/3/2020 6:46 AM	Text Document	2 KB
 berita20.txt	1/3/2020 6:46 AM	Text Document	1 KB
 berita21.txt	1/3/2020 6:46 AM	Text Document	1 KB
 berita22.txt	1/3/2020 6:46 AM	Text Document	2 KB
 berita23.txt	1/3/2020 6:46 AM	Text Document	2 KB
 berita24.txt	1/3/2020 6:46 AM	Text Document	3 KB
 berita25.txt	1/3/2020 6:46 AM	Text Document	3 KB
 berita26.txt	1/3/2020 6:46 AM	Text Document	1 KB

2. Clean data

Selanjutnya, lakukan pembersihan data setelah proses crawling data. Pembersihan data dilakukan untuk menghilangkan simbol dan tanda baca. Pada tahap cleaning ini juga dilakukan pemisahan isi dan judul. Berikut proses melakukan cleaning.

```
wirda@DESKTOP-7R09IST:/mnt/c/xampp/htdocs/finalpi/code$ python3 Clean.py
Directory : ../data/crawling
Process...
wirda@DESKTOP-7R09IST:/mnt/c/xampp/htdocs/finalpi/code$
```

3. Score data

Setelah proses cleaning data selesai, selanjutnya beri skor untuk setiap kata yang unik, juga bobot tiap dokumen untuk kata tersebut. Pada tahap score data ini juga dilakukan proses penghapusan stopwords, stemming, dan sebagainya.

```
wirda@DESKTOP-7R09IST:/mnt/c/xampp/htdocs/finalpi/code$ python3 Score.py
Directory : ../data/clean
200it [00:02, 69.35it/s]
unique words : 4575
100%|
done
```

4. Melakukan testing pencarian

Selanjutnya, lakukan proses pencarian untuk menampilkan dokumen yang akan muncul pertama jika suatu *query* dimasukkan. Buat file python Query.py yang berfungsi untuk menampilkan json dengan tampilan nama dokumen, skor, serta url dari document tersebut.

```
wirda@DESKTOP-7R09IST:/mnt/c/xampp/htdocs/finalpi$ python3 code/Query.py 3 tewas
[{"doc": "doc84.txt", "score": 1.6161764699893073, "url": "https://www.antaranews.com/berita/1225120/a-sia-kenang-230000-korban-jiwa-akibat-tsunami"}, {"doc": "doc99.txt", "score": 1.6161764699893073, "url": "https://www.antaranews.com/berita/1223936/topan-phanfone-kacaukan-momen-natal-di-filipina"}, {"doc": "doc81.txt", "score": 0.5387254899964358, "url": "https://www.antaranews.com/berita/1225131/gempa-tsunami-gerhana-adalah-ayat-ayat-allah-swt-sebut-ulama-aceh"}]
wirda@DESKTOP-7R09IST:/mnt/c/xampp/htdocs/finalpi$
```

5. Menjalankan program pada browser dengan melakukan pencarian

Untuk melakukan pencarian di browser, buat file index.php. File ini berfungsi untuk mengeksekusi query yang dicari oleh pengguna saat web akan dijalankan.

```
← → ↻ ⓘ localhost/finalpi/index.php?t=3&s=korban
[
  - {
    doc: "doc15.txt",
    score: 5.326914150036977,
    url: "https://www.antaranews.com/berita/1233168/f-pks-instruksikan-anggotanya-bantu-korban-bencana-banjir"
  },
  - {
    doc: "doc51.txt",
    score: 5.326914150036977,
    url: "https://www.antaranews.com/berita/1228023/prcpb-yonif-126-kc-evakuasi-korban-banjir-labuhanbatu-utara"
  },
  - {
    doc: "doc52.txt",
    score: 5.326914150036977,
    url: "https://www.antaranews.com/foto/1228019/konjungsi-bulan-dan-venus"
  }
]
```

6. Baguskan tampilan web

Selanjutnya buat tampilan web untuk menampilkan kotak inputan query dan jumlah dokumen yang akan muncul. Tambah html pada file index.php yang berfungsi untuk menampilkan search box, untuk inputan query (kata yang ingin di search) berupa teks dan untuk inputan jumlah berita berupa angka(integer). Perbagus tampilan tersebut dengan menambahkan file CSS.

