

Laporan Tugas 1

Teks dan Web Mining

Diajukan untuk melengkapi tugas Mata Kuliah Teks dan Web Mining

Oleh:

MUAMMAR ZIKRI AKSANA

1608107010045



JURUSAN INFORMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS SYIAH KUALA
DARUSSALAM, BANDA ACEH
MARET, 2019

Daftar Isi	Halaman
I. Penjelasan Tugas	3
II. Proses Pembagian Isi Konten	4
III. Crawling	6
A. Mendapatkan Link Dengan Mechanize	6
B. Melakukan Crawling Dengan Wget	7
IV. 5 File hasil ekstraksi	9
V. Lampiran	11

I. Penjelasan Tugas

Pada tugas 1 ini, diberikan case untuk mendownload 25000 file html dengan 2 kelas konten, pada tugas ini diambil dua kelas yaitu sepak bola dan basket. File didownload dari beberapa portal berita antara lain :

- www.bolasport.com
- indeks.kompas.com
- www.tempo.co/indeks/
- index.okezone.com
- republika.co.id/
- www.liputan6.com
- www.bola.com
- www.suaramerdeka.com
- www.viva.co.id
- www.mainbasket.com

dengan file yang berhasil didapat sebanyak 208113 :

```
0 total
mza@zx:/media/mza/Pro/DataPro/data-clean$ for D in */*; do printf "$D : ";
find $D -type f | wc -l; done
antaranews/basket : 6
bola/basket : 3514
bola/bola : 10441
bolasport/basket : 690
bolasport/bola : 126
juara/basket : 4421
kompas/bola : 9
liputan6/bola : 27824
mainbasket/basket : 3483
okezone/bola : 33541
republika/bola : 30255
sindo/basket : 187
suara/basket : 134
suaramerdeka/bola : 2898
tempo/bola : 66765
viva/basket : 2179
viva/bola : 21607
mza@zx:/media/mza/Pro/DataPro/data-clean$ ls */* | wc -l
208113
```

II. Proses Pembagian Isi Konten

Proses extract file dari hasil crawling dilakukan dengan beberapa tahapan :

1. Didapat semua path dari file yang ingin di extract dan disimpan pada sebuah array. (extractContent.pl : line 33)

```
my $listFile=`find $ARGV[0] -type f -follow`;
# setiap path file disimpan kedalam array
my @filePaths= split("\n",$listFile);
```

2. Setiap path file yang di dapat, dilakukan looping untuk setiap anggota array dan didapat *path-file* lalu dijalankan proses **extractFile(path-file)**, sampai semua anggota array selesai dibaca. (extractContent.pl : line 38)

```
foreach my $file (@filePaths){
    # print $filePaths[0];
    print "-----\n";
    print "Start at : $startAt\n";
    print "Time      : ".localtime."\n";
    print "countf : ".$countf++." : ".scalar @filePaths."\n";
    print "from : $file\n";
    # ubah kedalam format clean lalu di save
    saveFile($ARGV[1],$file,extractFile($file));
}
```

extractFile(path-file) > Tahapan untuk pembagian isi konten :

1. File dari *path-file* dibaca lalu ditampung pada sebuah variabel
2. Lalu diambil title dan url dengan regex dari variabel file yang telah ditampung sebelumnya
3. Isi variabel dibersihkan dari tag-tag html dan simbol yang tidak diperlukan dan diambil teksnya saja dan ditampung sebagai content (untuk step 2 -3 , extractContent.pl : line 57-73)

```
if( $file =~ /<title.*?>(.*?)<\title>/){
    $title = $1;
    $title = clean_str($title);
    $title="<title>$title</title>\n";
    # bagian title dihapus dari file agar tidak ikut dalam content
    $file =~ s/<title.*?>(.*?)<\title>//;
}
# get URL
if( $file =~ /<url>(.*?)<\url>/){
    $url = $1;
    $url="<url>$url</url>\n";
    # bagian url dihapus dari file agar tidak ikut dalam content
    $file =~ s/<url>(.*?)<\url>//;
}
```

```
$extractor->extract($file);
# get BODY (Content)
my $content = $extractor->as_text;
```

4. Content yang telah didapat dinormalisasi karakter pemisah (tanda bacanya) untuk penanda kalimat.(util.pl : line 113-134)

```
# Fungsi ini untuk normalisasi string
# Misal setelah titik harus ada spasi, tapi untuk url dan titel tidak dan lain sebagai
# Proses ini dilakukan agar tahap pembagian konten jumlah kata dapat diprediksikan de
# return string
sub simanticToken{
    $var=$_[0];
    my @matches = ( $_[0] =~ m/\s+[a-z]+\.\s*[A-Z]+/g );
    # @varr=split \s+[a-z]+\.\s*[A-Z]+,$var;

    foreach $data (@matches){
        # print $data;
        my $token=simanticTokenSentence($data);
        $var=~ s/$data/$token/g;
    }

    return $var;
}

# Fungsi untuk split dan menambahkan spasi pada sepenggal string
# ex: sekolah.Zikri => sekolah. Zikri
# return string
sub simanticTokenSentence{
    my @token = split "\\.",$_[0];
    # return index 0 dan 1 karena hasil split telah pasti hanya terdiri 2 anggota
    return $token[0].". ".$token[1];
}
```

5. Content yang telah dinormalisasi tanda bacanya di-split menjadi 3 bagian sesuai case dari tugas 1 , pada kasus ini dipisah dengan titik, tanda seru dan tanda tanya. (util.pl : line 83)

```
# Fungsi untum case-soal dimana content dipecah
# return string
sub fragmentContent{
    my @kalimat=split /(?!=[\.\!\?])\s+/,$_[0];
    my $splice = int(scalar @kalimat/3);
    my $html="<atas>".iterateFor(@kalimat,0,$splice*1)."</atas>\n";
    $html.="<tengah>".iterateFor(@kalimat,$splice*1+1,$splice*2)."</tengah>\n";
    $html.="<bawah>".iterateFor(@kalimat,$splice*2+1,$splice*3+@kalimat%3)."</bawah>\n";

    return $html;
}
```

6. Lalu file di-save dengan isi url - title - dan content yang telah dibagi sebelumnya.

III. Crawling

Pada proses crawling untuk mendapatkan source file yang dibutuhkan untuk menyelesaikan tugas 1, dilakukan 2 tahapan yaitu , proses untuk mendapatkan link dan proses download untuk setiap link yang telah didapat.

a. Mendapatkan Link Dengan Mechanize

Untuk mengambil link pada sebuah halaman web dengan Mechanize dibutuhkan link awal dari web tersebut, pada kasus ini web yang di crawling adalah portal berita yang memiliki indeks berita yang sesuai dengan class yang dipilih(pada tugas ini dipilih class sepak bola dan basket) portal berita dipilih karena terdapat pola pada setiap url indeksinya sehingga memungkinkan untuk di-crawling dengan skala besar meski kita tidak memiliki list link untuk di crawling dengan hanya merubah sedikit dari pola link sebelumnya.

Pada kasus ini dibuat proses :

1. Input format link dan tentukan jumlah hari yang ingin di-crawling
2. Dibuat perulangan dari 0 sampai jumlah hari lalu untuk masing perulangan didapat url berbeda dari hasil *generate-date*. (crawl.pl : line 38)

```
# looping sebanyak jumlah day yang ditentukan
for(my $i=0; $i <= $ARGV[2]; $i++){
    # proses mendapatkan url
    fly(generator($ARGV[1]), $ARGV[4]);
}
```

3. Untuk link dengan date yang telah di-generate, diambil yang ada dari url yang dihasilkan tersebut dengan method *get* dari kelas *mechanize*. (crawl.pl line : 54)

```
my $temp = eval{$mech->get($_[0])};
```

4. List url yang didapat dari hasil *get* sebelumnya di-hash untuk setiap anggota list sehingga url dapat dipastikan unik dan url tersebut memenuhi pattern url yang dibutuhkan, setelah perulangan tahap 2 selesai maka lanjut tahap 5. (crawl.pl line :70)

```
# untuk setiap link yang didapat
for my $i ($mech->links()){
    # difilter dengan filer url yang ditentukan
    if($i->url=~$_[1]){
        print "Link get : ".$go++."\n";
        $urls{$i->url}=1;
        print $hashLink $i->url."\n";
        # print $i->url."\n";
    }
}
```

- Setelah perulangan tahap 2 selesai, maka dilanjutkan ke proses crawling dari url yang terdapat pada hash.

```
# Fungsi untuk melakukan crawling
sub crawl{
    #folder-file-url

    # perintah wget melakukan crawling terhadap url yang diberikan dan hasilnya disimpan pada sebuah scalar
    my $RESULT=`wget -qO- --user-agent="Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36
    (KHTML, like Gecko) Chrome/51.0.2704.103 Safari/537.36" $_[0]/$_[1].html $_[2]`;
    # print $RESULT;
    open(my $web,'>>:encoding(UTF-8)','$_[0]/$_[1].html');

    my $uri=<url>$_[2]</url>\n";
    # concat url dengan hasil crawling sebelumnya, agar data url dari web terdata
    print $web $uri.$RESULT;
    close $web;
    # print $CMDOUT;
}
```

b. Melakukan Crawling Dengan Wget

Jika list url yang akan di-crawling telah selesai didapatkan maka dilanjutkan ke proses crawling, dengan tahapan :

- Key dari hash diambil dan disimpan pada sebuah array, lalu dilakukan perulangan untuk mendapatkan url dari anggota array sebelumnya. untuk setiap anggota array di panggil fungsi **crawling**(path-save,url).

```
# Procedure go untuk melakukan proses each dari url
sub go{
    # static-name - destination
    foreach my $url (keys %urls) {
        print "-----\n";
        print "Crawl      : $_[0] : $loop : $go\n";
        print "From       : $url\n";
        print "To         : $_[1]\n";
        print "Start at   : $startAt\n";
        print "Time       : ".localtime."\n";
        print "-----\n";
        # proses crawling untuk setiap url dari hash
        crawl($_[1],($_[0]."-".$loop++),$url);
    }
    return 1;
}
```


crawling(path-save,url) > Tahapan mendasar untuk pembagian isi konten :

1. Dijalankan perintah wget : (crawl.pl : line 105)

```
wget -qO- --user-agent="Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/51.0.2704.103 Safari/537.36" $_[0]/$_[1].html $_[2]
```

 - index 0 adalah path-save
 - index 1 adalah nama file
 - index 2 adalah alamat dari web yang akan di-crawl
 - diperlukan user agent untuk beberapa situs berita.
2. Setiap hasil wget disimpan pada sebuah variabel (wget menggunakan flag -q untuk no download).
3. Sebelum disimpan kedalam file variabel sebelumnya di concat

```
open(my $web,'>>:encoding(UTF-8)','$_[0]/$_[1].html');  
my $uri="<url>$_[2]</url>\n";  
# concat url dengan hasil crawling sebelumnya, agar data url dari web terdata  
print $web $uri.$RESULT;  
close $web;
```

dengan url (crawl.pl : line 107)

4. Setelah hasil wget di concat dengan urlnya barulah di-save.

IV. File hasil ekstraksi

File hasil ekstraksi terdiri dari :

- <url></url>
- <title></title>
- <atas></atas>
- <tengah></tengah>
- <bawah></bawah>

berikut screenshot file beberapa hasil file hasil ekstraksi :

1. File laman dengan kategori basket dari situs viva

```
mza@zx: /media/mza/Pro/DataPro/data-clean/viva/basket
mza@zx: /media/mza/Pro/DataPro/data-clean/viva/basket 80x28
markas baru mereka? kata Irving seperti dilansir ESPN . 7</bawah>
mza@zx: /media/mza/Pro/DataPro/data-clean/viva/basket$ cat viva-basket-2001.berse
h.dat
<url>https://www.viva.co.id/sport/basket/619478-playoff-nba-bulls-diseruduk-si-k
ijang-nets-bangkit</url>
<title>Playoff NBA Bulls Diseruduk Si Kijang Nets Bangkit ndash VIVA</title>
<atas>VIVA.co.id Chicago Bulls kembali menelan kekalahan saat menghadapi Milwauk
ee Bucks dalam game 5 playoff NBA 2015. Keunggulan tiga game Sang Banteng Merah
pun mulai terkejar oleh Si Kijang. Sempat unggul 3 0 Bulls mengalami kekalahan d
alam dua game terakhir termasuk pada pertandingan Senin 27 April 2015 atau Selas
a WIB. Tampil di markas sendiri United Center Bulls tak mampu berbuat banyak men
ghadapi serangan Bucks dan harus menelan kekalahan 88 94.Keunggulan mereka pun k
embali terkikis jadi 2 3.</atas>
<tengah>Michael Carter Williams mencetak 22 poin dan sembilan assists untuk Buck
s ditambah 21 poin dari Khris Middleton. Sedangkan 25 poin untuk Bulls yang dice
tak Pau Gasol tak mampu memberikan perlawanan berarti. Kedua tim akan kembali be
rtemu pada game 6 yang akan digelar Kamis 30 April atau Jumat pagi WIB. Bulls be
rpeluang merebut tiket semifinal Wilayah Timur sedangkan Bucks ingin memaksa ser
i lanjut ke game 7.</tengah>
<bawah>Di pertandingan lain Brooklyn Nets juga berhasil bangkit dalam dua game t
erakhir untuk menyamakan kedudukan jadi 2 2. Terakhir Nets menang lewat overtime
120 115 di Barclays Center. Dari Wilayah Barat Portland Trail Blazers berhasil
memperpanjang nafas saat menghadapi Memphis Grizzlies. Tampil di depan publik se
ndiri Blazers berhasil menang dengan keunggulan tujuh poin 99 92 untuk membuat k
edudukan jadi 1 3. Hasil Lengkap Playoff NBA Senin 28 April 2015 atau Selasa WIB
Brooklyn Nets 120 Atlanta Hawks 115 Chicago Bulls 88 Milwaukee Bucks 94 Memphis
Grizzlies 92 Portland TrailBlazer 99 one 9</bawah>
mza@zx: /media/mza/Pro/DataPro/data-clean/viva/basket$
```

2. File laman dengan kategori bola dari situs bolaspport

```
mza@zx: /media/mza/Pro/DataPro/data-clean/bolasport/bola$ cat bolasport-bola-2.be
rsih.dat
<url>https://www.bolasport.com/read/311635552/eks-gelandang-persis-solo-resmi-be
rlabuh-ke-psis-semarang-untuk-2019</url>
<title>Eks Gelandang Persis Solo Resmi Berlabuh ke PSIS Semarang untuk 2019 Bola
sport.com</title>
<atas>Eks Gelandang Persis Solo Resmi Berlabuh ke PSIS Semarang untuk 2019 BOLAS
PORT.COM Eks gelandang Persis Solo Heru Setyawan resmi melabuhkan kariernya bers
ama PSIS Semarang untuk musim 2019. Setelah menjalani trial bersama dengan PSIS
Semarang pada dua leg babak 32 besar Piala Indonesia melawan Persibat Batang PSI
S Semarang melirik Heru Setyawan.Diakui oleh Heru ia bersyukur bisa mendapatkan
kesempatan untuk trial di PSIS Semarang .</atas>
<tengah>Apalagi selama babak 32 besar Piala Indonesia Heru selalu diturunkan seb
agai starting XI. Baca Juga Kasus Marko Simic dan Aturan Tarkam dari PSSI Ini Ka
ta Ratu Tisha Baca Juga Bhayangkara FC Ujian Nyata PSIS Semarang di Piala Indone
sia Saya sangat bersyukur bisa mendapatkan kesempatan untuk trial bersama PSIS u
jar Heru.</tengah>
<bawah>Kesempatan itu tentu tidak saya sia siakan ucapnnya kepada BolaSport.com R
abu 1322019. 5</bawah>
mza@zx: /media/mza/Pro/DataPro/data-clean/bolasport/bola$
```

3. File laman dengan kategori bola dari situs viva

```
mza@zx:/media/mza/Pro/DataPro/data-clean/viva/bola$ cat viva-bola-4001.bersih.dat
<url>https://www.viva.co.id/bola/liga-inggris/503560-manchester-city-juara-premier-league-2013-2014</url>
<title>Manchester City Juara Premier League 2013 2014 ndash VIVA</title>
<atas>VIVAbola Manchester City memastikan diri menjadi juara Premier League musim 2013 2014 usai menundukkan West Ham United dengan skor 2 0 Minggu 11 Mei 2014. Dalam laga yang berlangsung di Etihad Stadium dua gol ManCity dicetak Samir Nasri dan Vincent Kompany. Dengan hasil ini Manchester City juara dengan mengoleksi 86 poin dari 38 pertandingan di Premier League sedangkan Liverpool menjadi runner up Premier League dengan mengumpulkan 84 poin. Di babak pertama ManCity langsung menunjukkan dominasi mereka dengan serangan cepat ke pertahanan The Hammers. Gol Samir Nasri pada menit ke 39 membuat City memimpin 1 0 atas West Ham.</atas>
<tengah>ManCity langsung mengebrak di awal babak kedua dengan serangan cepat melalui Nasri dan David Silva. Gawang West Ham yang dikawal Adrian San Miguel berulangkali mendapat ancaman dari pasukan Manuel Pellegrini. Pada menit ke 49 akhirnya gawang The Hammers jebol untuk kedua kalinya dalam laga ini. Berawal dari sepak pojok bola jatuh ke kaki Edin Dzeko sebelum disambar oleh Vincent Kompany pada menit ke 49 yang membuat ManCity memimpin untuk sementara 2 0 atas West Ham.</tengah>
<bawah>West Ham mendapat peluang pada menit ke 55 ketika Matthew Stewart gagal menyambut umpan silang Mohamed Diame dengan sempurna. City balas menyerang kala sundulan Agüero masih melenceng pada menit ke 63. Selang empat menit Adrian menyematkan gawangnya ketika Agüero mendapat kesempatan usai memanfaatkan umpan Pablo Zabaleta. Pada menit ke 78 aksi Adrian lagi lagi membuat The Citizens gagal menambah gol lewat Nasri. 9</bawah>
mza@zx:/media/mza/Pro/DataPro/data-clean/viva/bola$
```

4. File laman dengan kategori basket dari situs bolasport

```
mza@zx:/media/mza/Pro/DataPro/data-clean/bolasport/basket$ cat bolasport-basket-2.bersih.dat
<url>https://www.bolasport.com/read/311625208/ibl-pertamax-2018-2019-hangtuah-tutup-seri-ke-6-dengan-kemenangan</url>
<title>IBL Pertamax 2018 2019 HangTuaH Tutup Seri Ke 6 dengan Kemenangan Bolasport.com</title>
<atas>Pebasket asing HangTuaH Gary Jacobs Jr berupaya melewati penjagaan pemain Pacific Caesar Surabaya pada laga seri ke 6 IBL Pertamax 2018 2019 di GOR Pacific Surabaya Jawa Timur Minggu 322019. BOLASPORT.COM Klub bola basket HangTuaH berhasil menutup seri keenam IBL Pertamax 2018 2019 dengan mengalahkan tim tuan rumah Pacific Caesar Surabaya. Tampil di GOR Pacific Surabaya Jawa Timur Minggu 322019 malam tim asuhan Andika Supriadi Saputra sukses meraih kemenangan 83 76. Kemenangan ini menjadi penebus kekalahan yang diterima HangTuaH pada 3 laga sebelumnya sekaligus menjaga asa mereka untuk lolos ke play off.</atas>
<tengah>Pemain asing HangTuaH Gary Jacobs menjadi aktor utama kemenangan timnya. Pebasket berambut gimbal itu mencatatkan double double hasil dari 36 poin dan 11 rebound. HangTuaH sudah mendominasi sejak awal pertandingan.</tengah>
<bawah>Performa ciamik Jacobs dan pemain lokal Abraham Renoldi Wenas membawa mereka unggul 12 5 atas Pacific. Lewat serangan efektif dan eksekusi yang baik HangTuaH pun mengakhiri kuartir pertama dengan skor 20 15. Skuat HangTuaH tampil semakin impresif pada kuartir kedua. Masuknya Luca Lioteza membuat permainan mereka kian hidup. 7</bawah>
```

5. File laman dengan kategori basket dari situs bolasport

```
mza@zx:/media/mza/Pro/DataPro/data-clean/bolasport/basket$ cat bolasport-basket-S200.bersih.dat
<url>https://www.bolasport.com/read/311427051/hasil-nba-2018-2019-double-double-kawhi-leonard-antar-toronto-raptors-menangi-laga-perdana</url>
<title>Hasil NBA 2018 2019 Double double Kawhi Leonard Antar Toronto Raptors Menangi Laga Perdana Bolasport.com</title>
<atas>Forward Toronto Raptors Kawhi Leonard jersey merah 2 mendribel bola seraya di jaga pemain Cleveland Cavaliers Cedi Osman saat melakoni laga pembuka NBA 2018 2019 di Scotiabank Arena Toronto Kanada Rabu 17102018. Raptors menang 116 104 atas Cavaliers. Torehan double double dari Kawhi Leonard sukses mengantarkan Toronto Raptors memenangi laga perdana NBA 2018 2019 mereka di Scotiabank Arena Toronto Kanada Rabu 17102018 malam waktu setempat atau Kamis pagi WIB. Dilansir BolaSport.com dari ESPN kontribusi 24 poin dan 12 rebound yang dibukukan Kawhi Leonard menjadi salah satu kunci kemenangan Toronto Raptors atas Cleveland Cavaliers dengan skor 116 104. Saya merasa baik.</atas>
<tengah>Saya senang kami bermain lepas dan meraih kemenangan tutur Leonard yang mendapat kepercayaan bermain selama 37 menit. Pertandingan ini berlangsung hebat. Mereka penonton memberi saya energi lebih kata Leonard lagi. Raptors memulai perjalanan musim reguler NBA kali ini dengan menjamu Cavaliers yang notabene merupakan finalis musim lalu.</tengah>
<bawah>Meski lawan yang dihadapi memiliki status mentereng Raptors tidak gentar. Sebaliknya mereka mampu mengalahkan Cavaliers. Baca juga Jadwal NBA 11 Laga Tersaji Duel Rookie Jempolan Jadi Sorotan 9</bawah>
mza@zx:/media/mza/Pro/DataPro/data-clean/bolasport/basket$
```


V. Lampiran

Script untuk menyelesaikan tugas 1 terdapat pada file MuammarZikriAksana_1608107010045_twm.tar.gz yang terdiri dari :

folder :

- link : folder ini berisi file link yang berhasil di-*crawling* dan file link yang gagal pada saat di-*crawling*.
- log : folder ini berisi file yang mencatat file yang gagal di-*extract*
- ref : folder ini berisi referensi link ,syntax dan lainnya pada saat tugas dikerjakan
- src : folder ini berisi script yang digunakan untuk menyelesaikan tugas 1 :
 - crawl.pl , script untuk melakukan crawling
 - extractContent.pl, script untuk membersihkan file hasil crawling
 - generateUrl.pl, script untuk meng-*generate* url untuk dicrawling
 - util.pl , script yang berisikan fungsi-fungsi yang diperlukan untuk proses crawling dan extract

* beberapa folder diatas dibutuhkan kan untuk run program meski dibuat sebagai empty folder