**RABIT (Jurnal Teknologi dan Sistem Informasi Univrab)** ISSN CETAK : 2477-2062
Volume 3 No. 1 | Januari 2018 : Hal : 49-59 ISSN ONLINE: 2502-891X

# PENGEMBANGAN INTELLIGENT DATA COLLECTOR UNTUK ANALISIS BIG DATA ARTIKEL BERITA ONLINE

**Zul Indra[1)], Liza Trisnawati[2)]**

[1),2)] Teknik Informatika Universitas Abdurrab
Jl. Riau Ujung No. 73 Pekanbaru, Riau
Telp (0761) 38762 - 0811 771 4800
Alamat Email zul.indra@univrab.ac.id, liza.trisnawati@univrab.ac.id

**ABSTRAK**

Big data telah menjadi salah satu topik yg paling menarik dalam dunia teknologi informasi sekarang ini. Salah satu sumber big data yang tersedia dan bebas diakses adalah artikel berita online. Dalam sehari, sebuah situs berita populer bisa menghasilkan lebih dari 100 artikel berita baru. Bayangkan berapa banyak jumlah halaman berita yang tersedia untuk kita baca sekarang ini. Sementara itu, tahap awal untuk melakukan analisis big data terhadap artikel berita online adalah data storing dan preprocessing. Berdasarkan pemikiran tersebut maka perlu dikembangkan suatu aplikasi yang bisa mengumpulkan artikel berita online secara otomatis untuk kemudian di analisis lebih lanjut. Penelitian ini bermaksud mengembangkan suatu aplikasi yang diberi nama dengan intelligent data collector (IDC) yang memudahkan kita untuk mengumpulkan artikel berita online. Aplikasi IDC ini mengumpulkan artikel berita online kemudian melakukan preprocessing terhadap artikel-artikel tersebut dan menyimpannya dalam database lokal. Database ini kemudian bisa digunakan lebih lanjut untuk berrbagai macam data mining proses seperti opinion mining (sentiment analysis), topic classification, text summarization dan lain sebagainya.

**Kata Kunci:** big data, artikel berita online, preprocessing, text mining, IDC

**ABSTRACT**

*Big data has become one of the most popular research areas in information technology. One of the big data sources which is available and free to access is online news article. In fact, a popular news website can produce more than 100 new articles in one day. We can imagine that how many news webpage are available to read now. To perform a big data analysis for this online news article, the initial stage which is used to be done is data storing and preprocessing step. Therefore, it is a necessity to develop an application that can ease us to collect online news article. Based on this idea, this research is intended to develop an automatic data collector application that can collect online news article automatically from online news website.This application is called as intelligent data collector (IDC). The IDC application will collect online news article from news website, perform preprocessing step and then save output in local database. This local database can then be used for several data mining process such as opinion mining (sentiment analysis), topic classification, text summarization and etc.*

**Keywords:** *big data, online news article, preprocessing, text mining, IDC*

## I. PENDAHULUAN

### 1. Latar Belakang Masalah

Saat sekarang ini, dapat disimpulkan bahwa internet telah menjadi sumber utama informasi bagi keseharian kita. Bisa dikatakan bahwa mayoritas masyarakat lebih sering memperoleh dan membaca berita secara online dibadingkan memperolehnya dari media cetak seperti koran atau majalah. Informasi yang dipublikasikan melalui media online cenderung lebih dipilih karena memang memiliki kecepatan akses yang lebih tinggi

dan jumlah informasi yang bisa diakses lebih banyak apabila dibandingkan dengan media konvensional.

Sejak diperkenalkan pertama kali ke publik, jumlah informasi yang ada di internet sudah meningkat dengan sangat cepat. Pada tahun 1994, World Wide Worm (salah satu dari mesin pencari pertama) menyatakan bahwa mereka sudah mengindeks (menyimpan) 110.000 halaman dokumen web (Brin & Page, 2012). Memasuki tahun 1997, web crawler yang dikenal sebagai mesin pencari terbaik saat itu telah mengindeks 2 juta hingga 100 juta dokumen (Brin & Page, 2012). Lebih dari pada itu di bulan maret 2004, mesin pencari Google mengklaim bahwa mereka sudah 4,28 miliar dokumen web (Zareh Bidoki & Yazdani, 2008). Jumlah dokumen web ini akan terus bertambah dengan sangat cepatnya dimana sekitar 7 juta dokumen web baru yang akan bertambah setiap harinya Chong (2010).

Meledaknya jumlah dokumen online yang bisa diakses tersebut membuat masyarakat kaya akan sumber informasi. Media digital yang ada di internet telah menjadi salah satu sumber big data yang bebas kita akses setiap harinya. Berdasarkan pemikiran tersebut maka perlu dikembangkan suatu aplikasi yang bisa mengumpulkan artikel berita online secara otomatis untuk kemudian di analisis lebih lanjut. Penelitian ini bermaksud mengembangkan suatu aplikasi yang diberi nama dengan intelligent data collector (IDC) yang memudahkan kita untuk mengumpulkan artikel berita online.

Aplikasi IDC ini bertugas untuk mengumpulkan artikel berita online kemudian melakukan preprocessing terhadap artikel-artikel tersebut dan menyimpannya dalam database lokal. Database ini kemudian bisa digunakan lebih lanjut untuk berbagai macam proses data mining seperti *opinion mining (sentiment analysis)*, *topic classification*, *text summarization* dan lain sebagainya.

## II. STUDI PUSTAKA

### 1. Big Data

Saat sekarang ini, big data telah menjadi topik populer atau *research trend* dalam dunia ilmu komputer. Secara definisi, big data bisa diartikan sebagai sebuah datasets yang memiliki ukuran yang melebihi kemampuan dari software database untuk menangkapnya, menyimpannya dan mengolahnya (Zicari, R. V., 2014).

Big Data bukanlah sebuah teknologi, teknik, maupun inisiatif yang berdiri sendiri. Big Data adalah suatu trend yang mencakup area yang luas dalam dunia bisnis dan teknologi. Big Data menunjuk pada teknologi dan inisiatif yang melibatkan data yang begitu beragam, cepat berubah, atau berukuran super besar sehingga terlalu sulit bagi teknologi, keahlian, maupun infrastruktur konvensional untuk dapat menanganinya secara efektif.

Ada 3 ciri utama dalam Big Data (Chen, H., Chiang, R. H., & Storey, V. C., 2012) ini yang biasa dikenal dengan istilah 3V yakni *volume* (ukuran), *velocity*(kecepatan), dan *variety* (ragam). Dalam hal 3V ini, Big Data memiliki ukuran (volume), kecepatan (velocity), atau

ragam (variety) yang terlalu ekstrim untuk dikelola dengan teknik konvensional.

a. *Volume* (Ukuran).

Pada tahun 2000 lalu, PC biasa pada umumnya memiliki kapasitas penyimpanan sekitar 10 gigabytes. Saat ini, Facebook menyedot sekitar 500 terabytes data baru setiap harinya; sebuah pesawat Boeing 737 menghasilkan sekitar 240 terabytes data penerbangan dalam satu penerbangan melintasi Amerika; makin menjamurnya penggunaan ponsel pintar (*smartphone*), bertambahnya sensor-sensor yang disertakan pada perangkat harian, akan terus mengalirkan jutaan data-data baru, yang terus ter-*update*, yang mencakup data-data yang berhubungan dengan lingkungan, lokasi, cuaca, video bahkan data tentang suasana hati si pengguna ponsel pintar.

b. *Velocity* (kecepatan).

Clickstreams maupun ad impressions mencatat perilaku pengguna Internet dalam jutaan event per detik; algoritma jual-beli saham dalam frekwensi tinggi dapat mencerminkan perubahan pasar dalam hitungan microseconds; proses-proses yang melibatkan hubungan antara suatu mesin dengan mesin lainnya telah melibatkan pertukaran data antar jutaan perangkat; peralatan sensor dan perangkat-perangkat pada infrastruktur menghasilkan log data secara real time; sistem game online dapat melayani jutaan pengguna secara bersamaan, yang masing-masing memberikan sejumlah input per detiknya.

c. *Variety* (ragam).

Big Data tidak hanya menyangkut data yang berupa angka-angka, data tanggal, dan rangkaian teks. Big Data juga meliputi data-data ruang / geospatial, data 3D, audio dan video, dan data-data teks tak berstruktur termasuk file-file log dan media sosial. Sistem database tradisional didesain untuk menangani data-data berstruktur, yang tak terlalu sering mengalami update atau updatenya dapat diprediksi, serta memiliki struktur data yang konsisten yang volumenya tak pernah sebesar Big Data.

Selain itu, sistem database tradisional juga didesain untuk digunakan dalam satu server yang berdiri sendiri, yang berakibat pada keterbatasan dan mahalnya biaya untuk peningkatan kapasitas, sedangkan aplikasi sudah dituntut untuk mampu melayani pengguna dalam jumlah yang jauh lebih besar dari yang pernah ada sebelumnya. Dalam hal ini, database Big Data seperti halnya MongoDB maupun HBase, dapat memberikan solusi yang feasible yang memungkinkan peningkatan profit perusahaan secara signifikan.

Singkatnya, Big Data menggambarkan kumpulan data yang begitu besar dan kompleks yang tak memungkinkan lagi untuk dikelola dengan tools software tradisional.

## 2. Text Mining

Text mining bisa didefenisikan sebagai cabang ilmu yang berhubungan dengan metode dari mengolah, mendeteksi dan mengekstrak informasi dari data yang berbentuk teks (Deshpande, SP, & Thakare, Dr VM., 2010). Text mining memiliki kesamaan konsep analisis dan teknik dengan data mining konvensional. Akan tetapi text mining memiliki perbedaan dalam tujuannya untuk menemukan pola

dari data yang tidak terstruktur jika dibandingkan dengan data mining dimana data mining memerlukan data yang telah terstruktur.

Tujuan utama dari teknik text mining ini adalah untuk memenuhi permintaan pengguna untuk mengekstrak informasi yang detail dari sebuah sumber yang sangat besar (Korde & Mahender, 2012). Sebagai contoh, pada umumnya sebuah dokumen teks hanya memiliki sedikit bagian yang terstruktur seperti judul, penulis, tanggal publikasi dan kategori. Adapun konten dari dokumen teks tersebut kemungkinan besar merupakan data yang tidak terstruktur yang memiliki lebih banyak informasi. Merupakan pekerjaan yang rumit jika kita menggunakan data mining konvensional untuk mendapatkan informasi dari konten yang tidak terstruktur tadi.

Dengan perkembangan jumlah dokumen digital yang sangat cepat maka teknik text mining ini menjadi suatu kajian yang semakin penting. Text mining mempunyai banyak manfaat untuk menemukan data teks yang relevan dan data yang diinginkan dari sumber data yang tidak terstruktur. Secara garis besar, metode text mining ini digunakan untuk berbagai operasi seperti mengambil kesimpulan (meringkas) dari suatu teks, mengklasifkan dokumen, sentiment analysis dan lain sebagainya.

## 3. Preprocessing

Preprocessing merupakan tahapan awal dari proses *text mining* dimana proses ini bertujuan untuk mempersiapkan data mentah menjadi data yang akan diolah lebih lanjut. Proses preprocessing ini bertujuan agar data teks tersebut menjadi lebih terstruktur. Pada umumnya preprocessing meliputi 4 tahapan yakni (1) case folding, (2) tokenizing, (3) filtering, dan (4) stemming (Salton, 1989).

a. Case Folding

Tidak semua dokumen teks konsisten dalam penggunaan huruf kapital. Oleh karena itu, peran Case Folding dibutuhkan dalam mengkonversi keseluruhan teks dalam dokumen menjadi suatu bentuk standar (biasanya huruf kecil atau *lowercase*). Sebagai contoh, *user* yang ingin mendapatkan informasi "KOMPUTER" dan mengetik "KOMPOTER", "KomPUter", atau "komputer", tetap diberikan hasil retrieval yang sama yakni "komputer". Case folding adalah mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf 'a' sampai dengan 'z' yang diterima. Karakter selain huruf dihilangkan dan dianggap delimiter.

b. Tokenizing

Tahap Tokenizing adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya. Tokenisasi secara garis besar memecah sekumpulan karakter dalam suatu teks ke dalam satuan kata, bagaimana membedakan karakter-karakter tertentu yang dapat diperlakukan sebagai pemisah kata atau bukan. Sebagai contoh karakter whitespace, seperti enter, tabulasi, spasi dianggap sebagai pemisah kata. Namun untuk karakter petik tunggal ('), titik (.), semikolon (;), titk dua (:) atau lainnya, dapat memiliki peran yang cukup banyak sebagai pemisah kata.

c. Filtering

Tahap Filtering adalah tahap mengambil kata-kata penting dari hasil token. Tahapan ini menggunakan algoritma *stoplist* (membuang kata kurang penting) atau *wordlist* (menyimpan kata penting). Oleh karena itu tahapan ini juga biasa dikenal dengan proses *stopwords removal*. Stoplist/stopword adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan bag-of-words. Contoh stopwords adalah "yang", "dan", "di", "dari" dan seterusnya. Kata-kata seperti "dari", "yang", "di", dan "ke" adalah beberapa contoh kata-kata yang berfrekuensi tinggi dan dapat ditemukan hampir dalam setiap dokumen (disebut sebagai stopword). Penghilangan stopword ini dapat mengurangi ukuran indeks dan waktu pemrosesan. Selain itu, juga dapat mengurangi level *noise*.

d. Stemming

Suatu dokumen tidak dapat dikenali langsung oleh komputer. Oleh karena itu, dokumen tersebut terlebih dahulu perlu dipetakan ke dalam suatu representasi dengan menggunakan teks yang berada di dalamnya. Proses pemetaan ini dinamakan dengan proses pembuatan indeks. Untuk membuat indeks dari suatu dokumen, diperlukan langkah terakhir dari preprocessing yakni teknik stemming (Uysal & Gunal, 2014).

Teknik Stemming diperlukan selain untuk memperkecil jumlah indeks yang berbeda dari suatu dokumen, juga untuk melakukan pengelompokan kata-kata lain yang memiliki kata dasar dan arti yang serupa namun memiliki bentuk atau form yang berbeda karena mendapatkan imbuhan yang berbeda.

## III. METODE

Aplikasi *intelligent data collector* (IDC) bertujuan untuk memudahkan proses pengumpulan artikel berita online. Dengan menggunakan aplikasi IDC ini, pengguna cukup mengetikkan alamat *Uniform Resource Locator* (URL) dari artikel online yang diinginkan datanya kemudian aplikasi ini akan menyimpannya dalam database lokal. Setelah dokumen berhasil disimpan ke database lokal, maka dokumen tersebut akan mengalami proses preprocessing untuk kemudian bisa menjadi data yang siap diolah melalui teknik *text mining* yang diinginkan.

Secara garis besar, aplikasi IDC ini terdiri dari dari 2 konsep utama yakni proses pengumpulan data dan preprocessing. Detail untuk tiap konsep akan didiskusikan pada sub bab berikut ini.

### 1. Pengumpulan Data

Konsep utama dari tahapan pengumpulan ini hampir sama dengan konsep dari *web crawler* yang digunakan pada mesin pencari (*search engine*) untuk mendapatkan informasi dari website yang ingin mereka indeks. Pada umumnya, *web crawler* adalah sebuah *agent* atau virtual robot (*automatic program*) yang merayap ke dalam website untuk mendapatkan informasi dari webiste tersebut dan menyimpannya ke dalam database mesin pencari (Baeza-Yates & Ribeiro-Neto, 1999).

Menganut konsep dari mesin pencari ini, proses pengumpulan data pada aplikasi IDC ini akan merayap ke dalam website berita online yang URL-nya diketikkan oleh

pengguna. Data yang diperoleh dari URL artikel berita tersebut akan kemudian disimpan ke dalam database lokal aplikasi IDC. Pseudocode algoritma dari proses pengumpulan data ini adalah sebagai berikut.

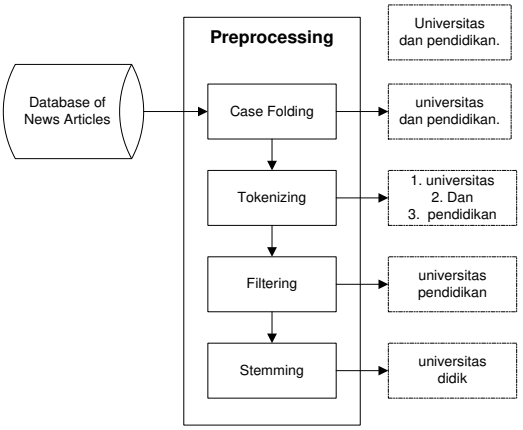
**Algoritma** Pseudocode Pengumpulan Data

```
Set URL (URL = '$url')
Open connection to $url = '$html'

Get information from URL
{
      $html = file_get_html($url);

      Parse $html
      {

          select title from $html
          $html->find('title') as $title
          store $title in database

          select content from $html
          $html->find('content')      as
      $content
          store $content in database

      }
}
```

Seperti terlihat pada pseudocode algoritma, aplikasi IDC hanya akan menyimpan judul dan isi dari artikel online. Aplikasi IDC akan menghapus data yang tidak penting selain judul dan isi dari artikel online tersebut.

e. Preprocessing

Seperti yang sudah dijelaskan pada bab tinjauan pustaka, proses preprocessing ini bertujuan untuk mempersiapkan data mentah menjadi data yang akan diolah lebih lanjut dan membuat artikel online tersebut menjadi lebih terstruktur. Proses dari setiap tahap preprocessing diilustrasikan pada gambar berikut.




Proses preprocessing diawali dengan tahapan *Case Folding* dimana semua karakter pada dokumen online akan dirubah ke dalam *lowercase*. Dokumen yang sudah dirubah menjadi lower case kemudian akan dipisah menjadi kata perkata pada tahapan *tokenizing*. Berikutnya kata dan karakter yang tidak penting (seperti tanda baca) akan dihilangkan pada proses filtering. Tahapan terakhir adalah *Stemming* untuk mengubah setiap kata ke dalam bentuk dasarnya sehingga dihasilkan indeks dokumen yang sudah siap diolah lebih lanjut untuk beragam teknik *text mining*. Proses stemming pada penelitian ini mengadopsi konsep *The Enhanced Confix-Stripping stemmer* yang dikembangkan oleh Arifin dkk., 2008.

## IV. HASIL DAN PEMBAHASAN

Bagian ini merupakan pembahasan dari hasil penelitian yang tediri dari dua sub bagian yaitu (1) persiapan percobaan yang berisikan perangkat yang digunakan untuk mengembangkan aplikasi IDC; (2) pembahasan hasil pengujian dari sistem informasi aplikasi IDC.

### 1. Persiapan Percobaan dan Datasets

Aplikasi IDC merupakan aplikasi berbasis web yang dikembangkan dengan menggunakan bahasa pemrograman PHP, AJAX dan Jquery. Sedangkan untuk basisdata, MySQL versi 5.6.37 digunakan sebagai media penyimpanan data.

Dalam penelitian ini, artikel berita online untuk pengujian aplikasi diperoleh dari portal berita ANTARA. Portal berita ANTARA dipilih sebagai sumber data karena merupakan portal berita resmi milik pemerintah Indonesia dan merupakan portal berita yang sangat update dari sisi content berita. Penelitian ini menggunakan berita dari portal berita ANTARA untuk pengujian.

Tabel 1. Dataset Pengujian Aplikasi

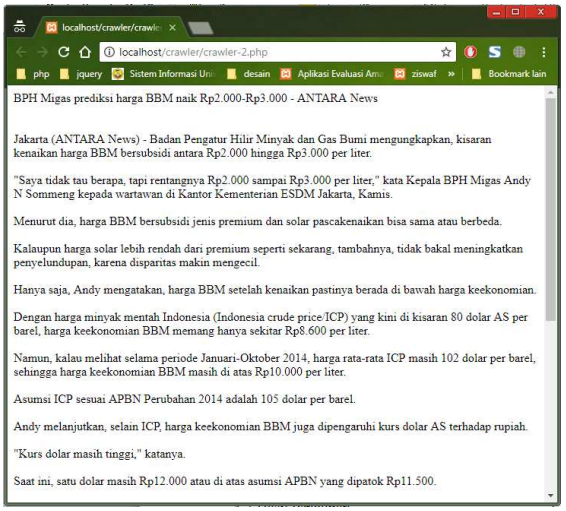
| No. | Kategori |
|---|---|
| Ekonomi | 25 Artikel |
| Olahraga | 25 Artikel |
| Hiburan | 25 Artikel |
| Teknologi | 25 Artikel |

Seperti yang terlihat pada tabel, total 100 artikel berita online dengan 4 kategori berita merupakan sumber data pengujian dalam penilitian ini.

### 2. Hasil Pengujian

Pengujian pada penelitian ini dilakukan dengan mengetikkan 100 URL artikel berita online yang berasal dari portal berita ANTARA. Keberhasilan dari aplikasi IDC ditentukan dari kemampuannya untuk menyimpan artikel berita dari URL yang diketikkan ke dalam database aplikasi IDC (database lokal) dan melakukan preprocessing terhadap artikel berita tersebut.

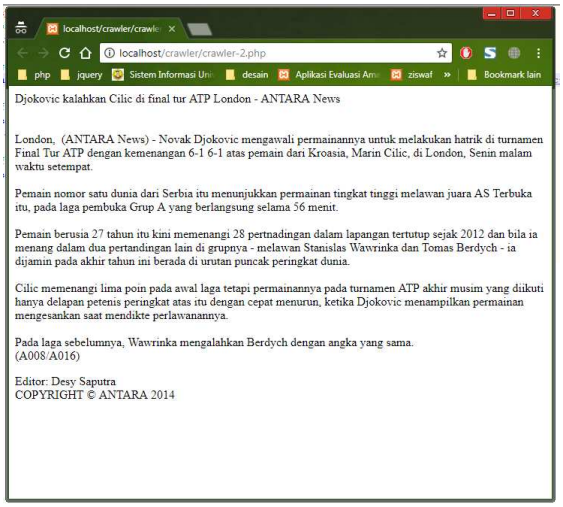
Pengujian pertama dilakukan dengan mencoba menyimpan data artikel berita dari *http://www.antaranews.com/berita/464178/bph-migas-prediksi-harga-bbm-naik-rp2000-rp3000*. Hasil dari pengujian tersebut telihat pada gambar berikut ini.



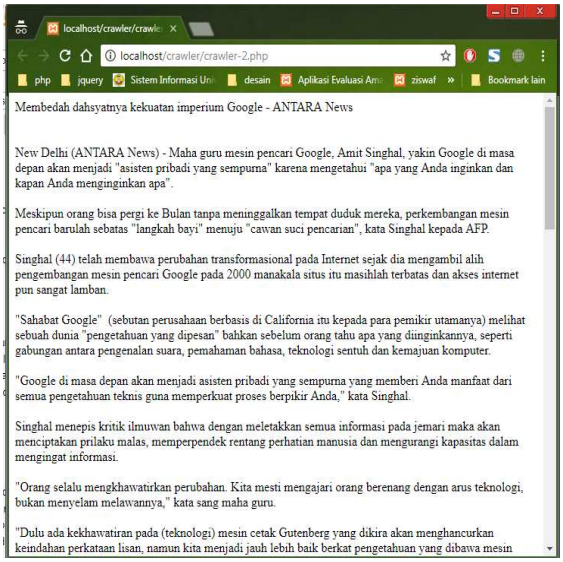

Gambar 1. Pengujian untuk URL *http://www.antaranews.com/berita/464178/bph-migas-prediksi-harga-bbm-naik-rp2000-rp3000*

Pengujian berikutnya dilanjutkan hingga semua datasets (100 artikel berita dari portal ANTARA) selesai diujicoba untuk

membuktikan apakah aplikasi IDC bisa menampilkan judul dan isi dari URL datasets.

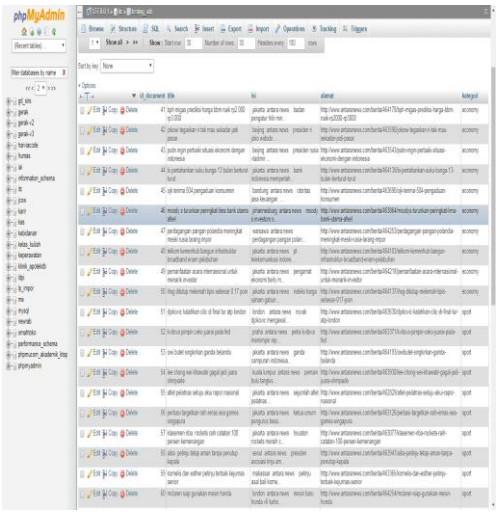



Gambar 2. Pengujian untuk URL *http://www.antaranews.com/berita/463630/djokovic-kalahkan-cilic-di-final-tur-atp-london*



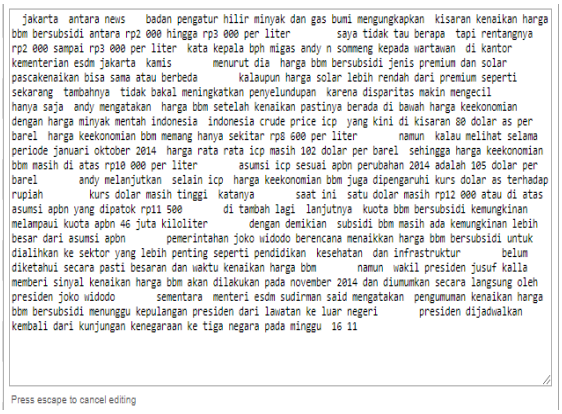

Gambar 3. Pengujian untuk URL *http://www.antaranews.com/berita/462629/membedah-dahsyatnya-kekuatan-imperium-google*

Seperti terlihat pada gambar dan hasil pengujian, dapat disimpulkan bahwa aplikasi IDC berhasil menampilkan judul dan isi berita untuk setiap URL yang diketikkan. Data isi dan judul berita dari setiap URL tersbut kemudian disimpan dalam database lokal untuk diproses lebih lanjut dalam tahap preprocessing.




Gambar 4. Penyimpanan judul dan isi artikel ke dalam database lokal

Tahapan pertama dari preprocessing adalah *case folding* untuk mengubah karakter pada dokumen berita ke dalam format *lower case*. Hasil dari proses *case folding* ini terlihat pada gambar berikut.




Gambar 5. Hasil Case Folding

Tahapan berikutnya adalah *filtering* Pada tahap ini dokumen yang telah drubah ke dalam format *lower case* kemudian dipisahkan menjadi kata-perkata untuk kemudian dibersihkan dari kata yang kurang penting (*stoplist*) dan tanda baca sehingga hanya tersisa kata-kata yang penting (*wordlist*) saja. Seperti pada gambar 5, kata sambung "dan, dengan kalaupun" yang merupakan kata *stoplist* akan dihapsukan dari artikel berita tersebut.

Tahapan terkahir dari preprocessing adalah *stemming* yakni mengubah setiap kata (*wordlist*) hasil dari proses *filtering* ke dalam bentuk dasarnya dengan membuang imbuhan pada kata tersebut. Sebagai contoh kata "pengatur" yang terdapat pada gambar 5 dirubah ke dalam bentuk dasarnya dengan membuang imbuhan (awalan) sehingga menjadi kata dasar "atur". Begitu juga dengan kata "mengungkapkan" dirubah dengan membuang awalan "me" dan akhri "an" sehingga menjadi kata dasar "ungkap". Sedangkan kata "liter" yang sudah merupakan kata dasar tidak mengalami proses stemming. Informasi dokumen yang telah mengalami proses stemming ini kemudian disimpan dalam database yang dinamakan dengan indeks dokumen.

+ Options

| | id_term | term | id_document | kategori |
|---|---|---|---|---|
| Edit Copy Delete | 1 | potensi | 1 | economy |
| Edit Copy Delete | 2 | investasi | 1 | economy |
| Edit Copy Delete | 3 | malu | 1 | economy |
| Edit Copy Delete | 4 | promosi | 1 | economy |
| Edit Copy Delete | 5 | potensi | 1 | economy |
| Edit Copy Delete | 6 | sumber | 1 | economy |
| Edit Copy Delete | 7 | daya | 1 | economy |
| Edit Copy Delete | 8 | alam | 1 | economy |
| Edit Copy Delete | 9 | malu | 1 | economy |
| Edit Copy Delete | 10 | limpah | 1 | economy |
| Edit Copy Delete | 12 | sayang | 1 | economy |
| Edit Copy Delete | 14 | potensi | 1 | economy |
| Edit Copy Delete | 15 | manfaat | 1 | economy |
| Edit Copy Delete | 17 | untung | 1 | economy |
| Edit Copy Delete | 18 | daerah | 1 | economy |
| Edit Copy Delete | 19 | direktur | 1 | economy |

Gambar 6. Tabel indeks artikel online kategori ekonomi

+ Options

| | id_term | term | id_document | kategori |
|---|---|---|---|---|
| Edit Copy Delete | 14707 | sembilan | 76 | technology |
| Edit Copy Delete | 14709 | wasiat | 76 | technology |
| Edit Copy Delete | 14711 | steve | 76 | technology |
| Edit Copy Delete | 14712 | jobs | 76 | technology |
| Edit Copy Delete | 14713 | langgar | 76 | technology |
| Edit Copy Delete | 14714 | apple | 76 | technology |
| Edit Copy Delete | 14715 | senin | 76 | technology |
| Edit Copy Delete | 14717 | 6 | 76 | technology |
| Edit Copy Delete | 14718 | 10 | 76 | technology |
| Edit Copy Delete | 14719 | 2014 | 76 | technology |
| Edit Copy Delete | 14721 | kemarin | 76 | technology |
| Edit Copy Delete | 14722 | nanda | 76 | technology |
| Edit Copy Delete | 14723 | tiga | 76 | technology |

Gambar 7. Tabel indeks artikel online kategori teknologi

+ Options

| | id_term | term | id_document | kategori |
|---|---|---|---|---|
| Edit Copy Delete | 5613 | 165 | 26 | sport |
| Edit Copy Delete | 5614 | sepeda | 26 | sport |
| Edit Copy Delete | 5615 | jalur | 26 | sport |
| Edit Copy Delete | 5616 | 181 | 26 | sport |
| Edit Copy Delete | 5617 | km | 26 | sport |
| Edit Copy Delete | 5618 | etape | 26 | sport |
| Edit Copy Delete | 5619 | itdbi | 26 | sport |
| Edit Copy Delete | 5620 | 165 | 26 | sport |
| Edit Copy Delete | 5621 | serta | 26 | sport |
| Edit Copy Delete | 5622 | international | 26 | sport |
| Edit Copy Delete | 5623 | tour | 26 | sport |
| Edit Copy Delete | 5624 | de | 26 | sport |
| Edit Copy Delete | 5625 | banyuwangi | 26 | sport |
| Edit Copy Delete | 5626 | ijen | 26 | sport |
| Edit Copy Delete | 5627 | awal | 26 | sport |

Gambar 8. Tabel indeks artikel online kategori olahraga

Options

| | id_term | term | id_document | kategori |
|---|---|---|---|---|
| Edit Copy Delete | 9925 | hadiah | 51 | entertainment |
| Edit Copy Delete | 9926 | lamborghini | 51 | entertainment |
| Edit Copy Delete | 9928 | raffi | 51 | entertainment |
| Edit Copy Delete | 9929 | ahmad | 51 | entertainment |
| Edit Copy Delete | 9930 | larang | 51 | entertainment |
| Edit Copy Delete | 9931 | lirik | 51 | entertainment |
| Edit Copy Delete | 9932 | perempuan | 51 | entertainment |
| Edit Copy Delete | 9933 | raffi | 51 | entertainment |
| Edit Copy Delete | 9934 | ahmad | 51 | entertainment |
| Edit Copy Delete | 9935 | kado | 51 | entertainment |
| Edit Copy Delete | 9936 | spesial | 51 | entertainment |
| Edit Copy Delete | 9937 | mobil | 51 | entertainment |
| Edit Copy Delete | 9938 | mewah | 51 | entertainment |
| Edit Copy Delete | 9939 | lamborghini | 51 | entertainment |
| Edit Copy Delete | 9940 | acara | 51 | entertainment |
| Edit Copy Delete | 9941 | hotman | 51 | entertainment |
| Edit Copy Delete | 9942 | paris | 51 | entertainment |
| Edit Copy Delete | 9943 | ceo | 51 | entertainment |
| Edit Copy Delete | 9944 | lamborghini | 51 | entertainment |
| Edit Copy Delete | 9945 | indonesia | 51 | entertainment |
| Edit Copy Delete | 9946 | johnson | 51 | entertainment |
| Edit Copy Delete | 9947 | yaptonaga | 51 | entertainment |
| Edit Copy Delete | 9948 | lamborghini | 51 | entertainment |
| Edit Copy Delete | 9949 | kado | 51 | entertainment |
| Edit Copy Delete | 9950 | nikah | 51 | entertainment |
| Edit Copy Delete | 9951 | raffi | 51 | entertainment |
| Edit Copy Delete | 9952 | ahmad | 51 | entertainment |
| Edit Copy Delete | 9953 | nagita | 51 | entertainment |
| Edit Copy Delete | 9954 | slavina | 51 | entertainment |
| Edit Copy Delete | 9956 | mobil | 51 | entertainment |

Gambar 9. Tabel indeks artikel online kategori hiburan

Indeks dokumen ini adalah hasil akhir dari tahap preprocessing yang menyatakan bahwa aplikasi IDC telah berhasil mengumpulkan informasi dari URL artikel online yang diketikkan oleh *user*. Selain itu indeks dokumen ini sudah siap untuk digunakan pada beragam aplikasi *text mining* lainnya seperti *opinion mining (sentiment analysis)* untuk mendapatkan respond pengguna terhadap sebuah topik, pengkategorian dokumen secara otomatis (*topic classification*), meringkas dokumen berita (*text summarization*) dan lain sebagainya.

## V. KESIMPULAN

Berdasarkan penelitian dan pengujian sistem yang telah dilakukan maka dapat ditarik kesimpulan bahwa aplikasi IDC telah berjalan dengan baik. Aplikasi ini berhasil menyimpan artikel berita online dengan hanya mengetikkan URL dari artikel online tersebut. Artikel berita yang telah disimpan dalam database kemudian berhasil diolah melalui proses preprocessing menjadi indeks dokumen yang berisi kata-kata dasar dari artikel berita. Aplikasi juga hanya menyimpan kata-kata penting dengan membuang *stopword* sehingga indeks dokumen yang dibangun menjadi lebih kecil ukurannya dan kinerjanya menjadi lebih baik.

Walaupun penelitian ini sudah berjalan dengan baik, ada beberapa masukan yang menjadi saran pengembangan lebih lanjut untuk aplikasi ini yaitu :

a. Pengumpulan berita dari *social media*

   Dikarenakan social media sudah menjadi salah satu sumber berita utama, maka pengumpulan berita dari sumber *social media* seperti facebook dan twitter menarik untuk dilakukan.

b. Pengumpulan dengan tema spesifik
   Pengumpulan informasi pada aplikasi IDC ini belum bisa secara spesifik menentukan tema apa yang ingin dicari. Belum terdapat fitur pencarian informasi sesuai topik yang diinginkan oleh *user*. Oleh karena itu penyimpan informasi dengan tema spesifik seesuai keinginan dari *user* bisa menjadi

suatu penelitian lanjutan yang layak untuk dikerjakan.

## DAFTAR PUSTAKA

[1] Arifin, Agus Zainal, Mahendra, I Putu, & Ciptaningtyas, Henning T. ( 2008). Enhanced Confix Stripping Stemmer and Ants Algorithm For Classifying News Documents in Indonesian Language. Paper presented at the The 5th International Conference on Information & Communication Technology & Systems., Surabaya.

[2] Baeza-Yates, Ricardo, & Ribeiro-Neto, Berthier. (1999). Modern information retrieval (Vol. 463): ACM press New York.

[3] Brin, Sergey, & Page, Lawrence. (2012). Reprint of: The anatomy of a large-scale hypertextual web search engine. Computer Networks, 56(18), 3825-3833.

[4] Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. MIS quarterly, 36(4).

[5] Chong, TIAN. (2010). A kind of algorithm for page ranking based on classified tree in search engine. Paper presented at the Computer Application and System Modeling (ICCASM), 2010 International Conference on.

[6] Deshpande, SP, & Thakare, Dr VM. (2010). Data mining system and applications: A review. International Journal of Distributed and Parallel systems (IJDPS), 1(1), 32-44.

[7] Korde, Vandana, & Mahender, C Namrata. (2012). Text classification and classifiers: A survey. International Journal of Artificial Intelligence & Applications (IJAIA), 3(2), 85-99.

[8] Salton, Gerard. (1989). Automatic Text Processing: The Transformation, Analysis, and Retrieval of: Addison-Wesley.

[9] Uysal, Alper Kursat, & Gunal, Serkan. (2014). The impact of preprocessing on text classification. Information processing & management, 50(1), 104-112.

[10] Zicari, R. V. (2014). Big data: Challenges and opportunities. Big data computing, 564.