26 JUNI
2015
UNIVERSITAS CIPUTRA
PROCEEDING

SNAPTI

SEMINAR NASIONAL APLIKASI DAN PENGEMBANGAN TEKNOLOGI INFORMASI

UNIVERSITAS CIPUTRA
CREATING WORLD CLASS ENTREPRENEURS

- (253-258) Tuwanku Aria - Penerapan Sistem E-Procurement Pada Proses Pengadaan Pt Petrokimia Gresik.pdf (311.14 Kb) 2021-03-30 07:01:11 pdf
- (266-270) Muhammad Fuad - Pengenalan Gestur Semaphore Menggunakan Sensor Kinect.pdf (482.98 Kb) 2021-03-30 07:01:12 pdf
- (271-278) Yulius Hari - Kajian terhadap Technology Acceptance Model pada Sistem Mobile Learning untuk Menunjang Pembelajaran Bahasa Mandarin.pdf (296.24 Kb) 2021-03-30 07:01:11 pdf
- (279-287) Adi Chandra - Desain Arsitektur Enterprise Sistem Informasi Manajemen Kampus Menggunakan Zachman Framework (Studi Kasus Universitas Atma Jaya Makassar).pdf (326.09 Kb) 2021-03-30 07:01:12 pdf
- (288-297) Theodore Darell - Rancang Bangun Media Pembelajaran Kord Gitar Dengan Rhythm Game Berbasis Android.pdf (326.21 Kb) 2021-03-30 07:01:11 pdf
- (298-307) Rico Nova - Rancang Bangun Admin Management pada Aplikasi tetanggaBaik Sebagai Jejaring Sosial Untuk Perumahan di Indonesia Berbasis PHP.pdf (512.94 Kb) 2021-03-30 07:01:12 pdf
- (308-314) Elizabeth Irenne - Rancang Bangun Permainan Dribel Bola Basket "Basketball Jam" Untuk Remaja Menggunakan Kinect.pdf (937.62 Kb) 2021-03-30 07:01:12 pdf
- (315-323) Edwin Kurniawan - Pengenalan Tipe Tas Tangan Wanita Pada Citra Digital Menggunakan Jaringan Syaraf Tiruan Perambatan Balik.pdf (265.68 Kb) 2021-03-30 07:01:12 pdf
- (324-331) Citra Lestari - Pengembangan Model Pengklasifikasi Naäve Bayes untuk Seleksi Penelusuran URL Halaman Detail Informasi Produk Tas Wanita pada Situs X.pdf (405.72 Kb) 2021-03-30 07:01:12 pdf
- (34-44) Prima Sanjaya - Rancang Bangun Property Management System untuk Budget Hotel – Room Division (Housekeeping).pdf (1008.58 Kb) 2021-03-30 07:01:12 pdf
- (45-52) Iwa - Faktor yang Mempengaruhi Adopsi Aplikasi Edmodo sebagai Media Pembelajaran.pdf (389.64 Kb) 2021-03-30 07:01:12 pdf
- (53-59) Michael Kristianto - Perancangan Pedoman Audit Sistem Informasi pada Industri Perhotelan dengan Studi Kasus Hotel Bintang 4 Berbasis Framework.pdf (276.57 Kb) 2021-03-30 07:01:12 pdf
- (60-67) Mohammad Caesar Rahmadian - Implementasi Sistem Informasi Akuntansi dengan Software Accurate pada Perusahaan Manufaktur.pdf (277.93 Kb) 2021-03-30 07:01:12 pdf
- (68-75) Michael Sugiarto - Rancang Bangun Sistem Kunci Berbasis Android Dan Web.pdf (469.82 Kb) 2021-03-30 07:01:11 pdf
- (76-83) Ivana Thiodora - Rancang Bangun Game Edukasi "Mommy's Care" Untuk Merawat Bayi Menggunakan Teknologi Adobe Flash.pdf (596.46 Kb) 2021-03-30 07:01:11 pdf
- (84-89) Clarien Rumbayan - Perancangan Panduan Kerja Audit Sistem Informasi Pada Perusahaan Jasa Web Hosting Berbasis Framework Cobit 4.1 Studi Kasus PT XYZ.pdf (435.8 Kb) 2021-03-30 07:01:12 pdf

Pengembangan Model Pengklasifikasi Naïve Bayes untuk Seleksi Penelusuran URL Halaman Detail Informasi Produk Tas Wanita pada Situs X

Citra Lestari Teknik Informatika Universitas Ciputra, UC Town CitraRaya, Surabaya 60219

ABSTRAK

Pencarian informasi dengan mesin pencari telah umum dilakukan. Umumnya, untuk informasi tentang detail sebuah produk, pengguna mencari pada situs-situs yang menjual produk tersebut. Selain mesin pencari seperti Google yang bersifat universal, sebenarnya terdapat mesin pencari yang menelusuri web tertentu saja untuk kebutuhan yang sangat spesifik. Mesin seperti ini melakukan penelusuran secara sekaligus sehingga membutuhkan biaya, waktu dan ruang, yang besar, sehingga perlu dilakukan seleksi halaman yang hendak ditelusuri dan unduh. Karya ilmiah ini membuat sebuah model pengklasifikasi menggunakan Naïve Bayes. Model pengklasifikasi ini digunakan untuk seleksi URL halaman yang akan ditelusuri dan diunduh oleh penelusur web. Halaman yang diingikan adalah halaman detail informasi tas wanita pada situs X, salah satu situs toko online terbesar di Indonesia. Dengan pendekatan klasifikasi teks, maka sebuah URL halaman dianggap sebagai dokumen. Dokumen atau URL direpresentasikan dalam model Boolean, yang melihat muncul atau tidaknya sebuah istilah pada suatu URL. Kumpulan dokumen dilabeli sebagai tas wanita (***bw***) atau bukan tas wanita (***nbw***). Model dibangun dengan melatihkan 800 dokumen. Model ini menemukan istilah "bags" sebagai istilah dengan probabilitas posterior tertinggi (0.99), sedangkan istilah "bag" meskipun lebih kerap muncul namun memiliki kekuatan yang sama pada kedua kelas (probabilitas posterior pada kelas ***bw*** = 0.57) sehingga dapa menyebabkan bias pada hasil klasifikasi. Model pengklasifikasi ini kemudian diuji menggunakan 325 dokumen yang berbeda dengan data dokumen latih. Akurasi dari pengujian tersebut adalah sebesar 93,2%.

Kata kunci: model pengklasifikasi, Naïve Bayes, penelusuran web, tas wanita, seleksi url

1. Pendahuluan

Pencarian informasi kini telah umum dilakukan dengan memanfaatkan teknologi internet. Informasi tersebut dapat diperoleh dari situs, baik komersial maupun non-komersial. Umumnya, untuk informasi tentang detail sebuah produk, seperti merek, spesifikasi, ukuran, harga, pengguna internet mencarinya pada situs-situs yang menjual produk tersebut. Pencarian informasi secara manual biasanya dilakukan dengan bantuan mesin pencari, seperti Google. Pengguna menuliskan kata kunci dari informasi yang ingin diketahui kemudian mesin pencari akan memberikan daftar situs yang sesuai.

Mesin pencari seperti Google bersifat universal. Mesin pencari seperti ini menelusuri dan menyimpan semua situs secara berkala (Menczer, 2011). Untuk tujuan yang lebih spesifik dan khusus, dapat dibuat sebuah mesin pencari bertipe topikal yang hanya menelusuri beberapa website tertentu dan hanya menyimpan halaman tertentu, misalnya halaman detail informasi produk.

Tidak seperti penelusur Universal yang mengunjungi secara berangsur, mesin pencari bertipe topikal menelusuri halaman-halaman sekaligus pada satu waktu. Jika semua halaman ditelusuri dan diunduh, tentunya mesin pencari perlu memiliki memori yang cukup besar (Menczer, 2011). Alternatif lain adalah membuat mesin pencari yang

Citra Lestari. Tel.: +62317451699.
E-mail: caecilia.citra@ciputra.ac.id

selektif, yaitu hanya menelusuri jalur yang sesuai dan mengunduh halaman yang tepat.

Seleksi halaman yang akan ditelusuri dan diunduh dapat dilakukan dengan berbagai cara, antara lain dengan melihat URL halaman, judul halaman, isi halaman, atau struktur halaman. Untuk tiga cara terakhir, mesin perlu mengunduh halaman terlebih dahulu, sedangkan untuk cara pertama, seleksi dilakukan dengan menentukan kelayakan sebuah halaman untuk ditelusuri atau diunduh.

Seleksi URL halaman dapat dilakukan dengan pendekatan klasifikasi teks. Apabila URL halaman dianggap sebagai sebuah teks atau kalimat, maka istilah dalam URL adalah sebuah kata. Dengan asumsi tidak ada keterkaitan antar istilah dalam suatu URL, algoritma Naïve Bayes dapat menghitung probabilitas kemunculan istilah-isitilah tersebut, kemudian membangun model pengklasifikasi yang dapat memprediksi sebuah URL adalah halaman yang diinginkan, dalam hal ini adalah halaman detail informasi produk.

Karya ilmiah ini adalah bagian dari pengembangan pembuatan mesin untuk pencarian halaman detail informasi sebuah produk. Halaman detail informasi yang dicari adalah produk tas wanita Mesin ini akan mencari dari beberapa situs toko online besar di Indonesia. Pada karya ilmiah ini pencarian hanya difokuskan pada satu situs toko online X.

2. Teori Penunjang

2.1. Penelusuran Web

Penelusuran web (*web crawling*) adalah upaya atau proses untuk mengunduh secara otomatis halaman-halaman web yang tersebut di jutaan mesin server (Menczer, 2011). Program penelusur, dikenal dengan *spider* atau *robot*, mengumpulkan informasi yang kemudian dapat dianalisa dan ditambang baik secara *online* yaitu saat diunduh atau pun *offline* yaitu setelah disimpan. Karena sifat web yang sangat dinamis yang dapat berubah dan bertambah dalam hitungan mili detik, maka penelusuran harus terus dilakukan agar aplikasi dapat terus terbarui.

Berdasarkan fungsinya terdapat tiga tipe penelusur web, yaitu:

1. Penelusur Universal, yaitu penelusur yang digunakan oleh mesin pencari umum untuk melakukan kunjungan secara berangsur sekaligus memelihara dan memperbarui indeks.
2. Penelusur Terfokus melakukan penelusuran dengan menitikberatkan pada halaman-halaman dengan kategori tertentu yang diminati pengguna.
3. Penelusur Topikal memulai penelusuran pada sejumlah kecil halaman, disebut *seed*. Berbeda dengan penelusur Universal. penelusur ini melakukan penelusuran secara *real-time* dan tidak mengandalkan rangking. Dengan demikian tidak ada hasil halaman yang "basi" dan halaman baru yang belum terindeks pun akan diambil.

Beberapa pustaka penelusur web telah banyak tersedia. Salah satunya adalah Website-Specific Processors for HTML Information Extraction, disingkat WebSPHINX. Menurut (Miller, n.d.),WebSPHINX adalah sebuah pustaka *class* Java. Di dalam pustaka WebSPHINX telah disediakan *class-class* yang dapat digunakan ulang sepert Crawler, Page, Link, Classifier, DownloadParameter, dan lain sebagainya. WebSPHINX toleran pada *parsing* HTML dan menyokong standar eksklusi robot. WebSPHINX juga dapat mencocokkan pola termasuk ekspresi regula, Unix *shell wildcards*, dan ekspresi *tag* HTML.

2.2. Klasifikasi Teks

Klasifikasi teks, atau dokumen, adalah salah satu bagian dari pembelajaran mesin yang bertujuan untuk memberikan label secara otomatis pada setiap dokumen. Hal ini penting, sebab pelabelan secara manual membutuhkan biaya mahal untuk dikembangkan (Manning,Raghavan, & Scutze, 2009).

Menurut Manning, dkk (Manning, Raghavan, & Scutze, 2009) pada klasifikasi teks terdapat dokumen d yang merupakan anggota dari koleksi dokumen $\boldsymbol{X}$, $d \in \boldsymbol{X}$ dimana $\boldsymbol{X}$ adalah ruang berdimensi tinggi. Selain itu terdapat kumpulan kelas $\boldsymbol{C} = \{c_1, c_2, \ldots, c_m\}$. Kelas-kelas ini disebut juga label yang didefinisikan oleh manusia sesuai dengan kebutuhan aplikasi. Set data yang dibutuhkan oleh klasifikasi teks, $\boldsymbol{D}$, adalah kumpulan dokumen dengan label $< d,c >$.

Klasifikasi teks melakukan pendekatan pembelajaran tersupervisi terhadap sebuah set data dokumen $\boldsymbol{D}$. Dikatakan tersupervisi sebab masih dibutuhkan kebijakan manusia untuk membagi dokumen menjadi beberapa kelas atau label serta menentukan label dari beberapa dokumen awal. Beberapa dokumen awal inilah yang digunakan sebagai data pelatihan mesin. Pada prinsipnya, pelatihan ini dapat dilakukan dengan semua algoritma klasifikasi, termasuk di antaranya adalah Naïve Bayes.

2.3. Algoritma Naïve Bayes

Naïve Bayes adalah sebuah algoritma klasifikasi secara statistika, yaitu berdasarkan teorema Bayes. Algoritma Naïve Bayes mengasumsikan bahwa efek dari nilai sebuah atribut terhadap class tertentu tidak berpengaruh/ dipengaruhi pada/oleh nilai atribut lain. Hal ini dibuat untuk menyederhanakan komputasi dengan anggapan komputasi yang naif (Han & Kamber, 2006).

Menurut Han dan Kamber (Han & Kamber, 2006), algoritma Naïve Bayes bekerja sebagai berikut:

1. Dengan D adalah set tupel (*tuple*) pelatihan yang terasosiasi dengan label-label kelas, setiap tupel diwakili oleh vektor atribut berdimensi n, $\boldsymbol{X} = (x_1,x_2,\ldots,x_n)$.
2. Diasumsikan terdapat m kelas, $C_1,C_2,\ldots,C_m$. Terhadap sebuah tupel X, algoritma ini akan memprediksi kelas dari X dengan mencari kelas C_i yang memiliki probabilitas posterior tertinggi, dimana:

$$P(C_i|X) > P(C_j|X) \; untuk \; 1 \leq j \leq m, j \neq i \quad (1)$$

Adapun nilai $P(Ci|X)$ diperoleh berdasarkan teorema Bayes seperti persamaan 2 berikut

$$P(C_i|X) = P(X|C_i)P(C_i)/P(X) \quad (2)$$

3. Karena $P(X)$ adalah konstan untuk semua kelas, maka yang perlu dimaksimalkan adalah $P(X|C_i)P(C_i)$. Probabilitas prior sebuah kelas, $P(C_i)$, dapat dihitung dengan mencari jumlah kemunculan sebuah kelas C_i di set data terhadap jumlah seluruh set data, $|C_i,D|/|D|$. Apabila tidak diketahui, maka umumnya semua kelas memiliki probabilitas yang sama.
4. Dengan asumsi naïf *class conditional independence*, yaitu tidak ada relasi kebergantungan antar atribut, maka $P(X|C_i)$ dapat dihitung sebagai perkalian produk dari probabilitas atribut-atribut yang pada sebuah tupel X terhadap sebuah kelas C_i (Persamaan 3).

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \quad (3)$$

Estimasi probabilitas $P(x_k|C_i)$ dapat diperoleh dari tupel-tupel pelatihan dengan memperhatikan jenis atribut, kategorikal atau kontinyu. Jika sebuah atribut A_k adalah kategorikal, maka $P(x_k|C_i)$ adalah jumlah tupel di D dengan kelas C_i yang memiliki atribut A_k bernilai x_k. Jika atribut A_k adalah kontinyu, maka $P(x_k|C_i)$ dihitung dengan persamaan 4, dengan asumsi berdistribusi Gaussian seperti persamaan 5.

$$P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}) \quad (4)$$

$$g(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (5)$$

2.4. Representasi Dokumen dengan Model Boolean.

Dalam lacak balik informasi (*information retrieval*), sebuah dokumen dianggap sebagai se"bungkus" kata atau istilah yang urutan dan posisinya diacuhkan (Liu, 2011). Masih menurut Bing Liu (Liu, 2011), sebuah dokumen dideskripsikan oleh sejumlah istilah yang berbeda. Pada sebuah koleksi dokumen D, kumpulan istilah berbeda adalah $V = \{t_1,t_2,...,t_{|v|}\}$ yang disebut ***vocabulary*** dengan $|v|$ adalah jumlah istilah yang ada di dalamnya. Sebuah bobot $w_{ij} > 0$ diasosiasikan dengan setiap istilah t_i pada dokumen $\mathbf{d}_j \in D$. Sebuah istilah yang tidak ada dalam $\mathbf{d}_j$ memiliki bobot $w_{ij} = 0$. Setiap dokumen direpresentasikan dalan sebuah vector $\mathbf{d}_j = (w_{1j}, w_{2j}, ... , w_{|v|j})$. Dengan representasi ini, maka sebuah koleksi dokumen dapat direpresentasikan sebagai sebuah tabel relasional atau matriks.

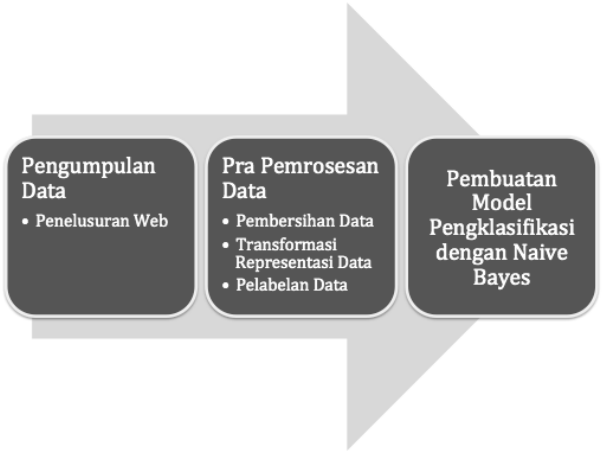
Terdapat empat model utama lacak balik informasi yaitu: model Boolean, model Ruang Vector (*vector space model*), model Bahasa (*language model*), dan model Probabilitas. Tiga model pertama adalah yang umum digunakan dan menggunakan rangka kerja seperti yang telah dijelaskan pada paragraph di atas (Liu, 2011).

Model Boolean adalah model yang paling sederhana. Pada model ini dokumen direpresentasikan sebagai kumpulan istilah yang hanya diperhitungkan ada atau tidaknya suatu istilah dalam sebuah dokumen, seperti pada persamaan 6.

$$w_{ij} = \begin{cases} 1 & \text{if } t_i \text{ appears in } d_j \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

3. Pembangunan Model Pengklasifikasi URL Halaman Detail Informasi Produk Tas Wanita

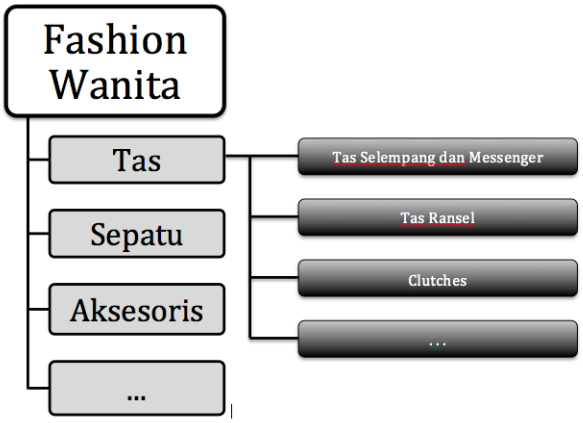
Seperti pada Gambar 2, sistem pembuatan model pengklasifikasi URL halaman detail informasi ini dimulai dengan proses pengumpulan data yang nantinya digunakan sebagai data pelatihan dan data uji. Proses ini melakukan penelusuran web dengan bantuan Websphinx. Penelusuran web dimulai dari halaman benih (*seed*) sebagai masukan (*input*). Hasil dari proses ini adalah sekumpulan URL halaman yang berekstensi HTML. Proses selanjutnya adalah pra-pemrosesan data, yaitu mengubah representasi data menjadi model Boolean dan menentukan label atau kelas dari tiap-tiap URL. Proses selanjutnya adalah pembuatan model pengklasifikasi dengan Naïve Bayes yang telah diimplementasikan oleh alat bantu Weka. Beberapa sub-bab selanjutnya memberikan penjelasan lebih lengkap mengenai tiap proses.



Gambar 2. Sistem Pembuatan Model Klasifikasi URL Halaman Detail Informasi Produk Tas Wanita

3.1. Pengumpulan Data

Set data yang dibutuhkan untuk pembuatan model klasifikasi pada karya ilmiah ini adalah sekumpulan URL halaman detail informasi produk dari situs X. Pengumpulan data dilakukan dengan penelusuran web topikal dengan bantuan Websphinx. Sebelum pengumpulan data, penulis melakukan analisa awal terhadap struktur pemetaan halaman situs X.



Gambar 1. Struktur Kategorisasi Produk Fashion Wanita pada situs X.

Pengkategorian produk pada situs X dilakukan hingga tiga level. Sebagai contoh, produk tas wanita berada di level kedua dengan "Fashion Wanita" sebagai *parent* dan beragam jenis tas wanita sebagai *children*-nya, seperti pada Gambar 1. Setiap produk pada situs X memiliki halaman detail informasi produk. URL halaman detail informasi produk merupakan halaman berekstensi HTML dengan awalan domain situs X. Domain tersebut diikuti beberapa istilah kunci dari produk. Istilah-istilah tersebut dipisahkan dengan tanda garis "-". Tabel 1 merupakan contoh dari URL halaman detail informasi produk dengan domain asli situs diganti www.x.co.id

Tabel 1. Contoh URL Halaman Detail Informasi Produk.

No	URL Halaman Detail Informasi Produk
1.	http://www.x.co.id/viyar-citrus-sling-bag-black-386488.html
2.	http://www.x.co.id/women-men-weight-lifting-gloves-fitness-gym-exercise-soft-glove-rose-1097378.html
3.	http://www.x.co.id/lotus-speculoos-crunchy-1-buah-471471.html

Karena karya ilmiah ini terfokus pada produk tas wanita, maka dilakukan pengaturan penelusuran sebagai berikut:

1. Penelusuran diawali dari halaman benih http://www.x.co.id/tas-wanita/
2. Penelusuran dilakukan dengan tingkat kedalaman sebesar 3 (tiga). Hal ini disesuaikan dengan kategorisasi produk. Apabila dimulai dari kategori "Tas Wanita", maka penelusur perlu menelusuri hingga dua level di bawah untuk mencapai masing-masing produk tas wanita, sehingga kedalaman yang dibutuhkan adalah 2 + 1.
3. Penelusur hanya mengunjungi halaman dengan URL yang memiliki kata "html", "tas", dan "bag".
4. Penelusur hanya menyimpan URL yang berekstensi "html"
5. Penelusuran web dilakukan selama 90 menit atau hingga tidak ada lagi halaman yang dapat ditelusuri.

Hasil penelusuran web ini adalah 1182 URL. Beberapa data URL yang terkumpulkan disajikan pada Tabel 2.

3.2. Pra Pemrosesan Data

Proses ini adalah proses yang cukup krusial. Pada proses ini data yaitu URL yang telah berhasil diunduh dipersiapkan untuk dapat menjadi set data yang layak diklasifikasi. Beberapa persiapaan yang dilakukan adalah: pembersihan data, transformasi data dalam model representasi Boolean, dan pelabelan data.

3.2.1. Pembersihan Data

Setelah dilakukan analisa terhadap data yang terkumpul, diketahui terdapat beberapa data yang redundan. Data redundan tersebut adalah *link* menuju pengaturan alat *mobile* dari sebuah halaman detail informasi produk. Contoh data redundan tersebut adalah URL ke-8 pada Tabel 2. URL tersebut hanya berbeda pada istilah "?setDevice=mobile" dengan URL ke-7. URL tersebut menampilkan halaman detail informasi produk yang sama, namun berbeda tata letaknya.

Tabel 2. Contoh Hasil Penelusuran

No	URL Halaman Detail Informasi Produk
1.	http://www.x.co.id/viyar-citrus-sling-bag-black-386488.html
2.	http://www.x.co.id/royal-polo-backpack-8996-06-coffee-384221.html
3.	http://www.x.co.id/sayota-sv-809-portable-vacuum-cleaner-merah-437106.html
4.	http://www.x.co.id/aosimani-1529-black-ransel-laptop-multi-fungsi-104722.html
5.	http://www.x.co.id/bgc-disney-frozen-tas-ransel-elsa-ana-3d-timbul-3-kantong-import-pink-blue-kotak-pensil-dan-alat-tulis-frozen-912374.html
6.	http://www.x.co.id/bags-hearted-pocket-clutch-brown-463152.html?mp=1
7.	http://www.x.co.id/lzd-slouchy-clutch-green-951905.html
8.	http://www.x.co.id/lzd-slouchy-clutch-green-951905.html?setDevice=mobile
9.	http://www.x.co.id/mayonette-ruena-mini-sling-hitam-416378.html
10.	http://www.x.co.id/baglis-dompet-simple-wanita-cokelat-189973.html

Dengan penemuan tersebut maka dilakukan pembersihan data. Terdapat 57 data redundan yang berhasil dibersihkan. Setelah pembersihan, besar koleksi URL menjadi 1125.

3.2.2. Transformasi Representasi Data

Proses ini mengubah representasi URL ke dalam model Booelan. Dengan menganggap sebuah URL sebagai dokumen, maka istilah-istilah yang unik dari kumpulan URL tersebut menjadi kumpulan atribut. Terdapat dua istilah yang diacuhkan, yaitu:

1. Domain dari situs X.

Hal ini disebabkan seluruh URL berasal dari domain yang sama sehingga istilah tersebut sudah pasti ada pada setiap data.

2. Ekstensi .html.
   Karena URL yang diunduh adalah yang mengandung istilah ".html", maka bisa dipastikan istilah tersebut ada di seluruh data.

Sebelum proses transformasi dilakukan, koleksi URL dibagi menjadi dua secara acak yaitu untuk 800 URL untuk set data pelatihan dan 235 URL untuk set data pengujian. Untuk setiap set data, dilakukan proses transformasi representasi data URL menjadi model Boolean yang memiliki langkah-langkah sebagai berikut:

1. Sub-proses pengumpulan atribut pada koleksi atribut A.
   Untuk setiap URL *u* pada set data *U*
   a. potong awalan "http://www.x.co.id/"
   b. potong bagian url setelah tanda "."
   c. pisahkan istilah-istilah pada *u* dengan penanda "-"
   d. masukkan istilah-istilah yang belum ada pada koleksi atribut ***A***
2. Sub-proses pembobotan atribut-atribut dokumen d.
   Untuk setiap URL u_i pada set data *U*
   a. Buat dokumen baru d_i
   b. Untuk setiap atribut a_j pada koleksi atribut A
      i. Jika atribut a_j ada pada u_i maka w_{ij} = 1, selain itu w_{ij} =0.
   c. Masukkan d_i pada koleksi dokumen D

3.2.3. Pelabelan Data

Pada karya ilmiah ini, dokumen dibagi menjadi dua kelas, yaitu Tas Wanita (dengan notasi ***bw***) dan Bukan Tas Wanita (***nbw***). Sebuah URL dilabeli ***bw*** apabila produk yang ditampilkan halaman tersebut adalah sebuah tas berjenis: 1) tas selempang wanita, 2) tas messenger wanita, 3) tas ransel wanita, 4) *clutch*, 5) tas bahu (*shoulder bag*), 6) tas *tote* wanita, 7) tas selempang badan wanita, 8) tas *satchel* wanita, 9) tas *weekender* wanita, 10) tas kerja wanita. Dompet, tas kosmetik, dan tas alat komunikasi tidak termasuk dalam kategori ***bw***, melainkan ***nbw.*** Label ***nbw*** juga diberikan pada URL yang menampilkan produk antara lain tas pria, dompet pria, tas laptop, tas sepatu, tas kamera, tas anak-anak, tas bayi, dan pembersih debu.

Proses pelabelan data dilakukan secara manual oleh penulis. Penulis membuka URL pada *browser* dan menentukan label dari URL. Penentuan label berdasarkan produk yang ditampilkan oleh halaman detail informasi bersangkutan. Penentuan tersebut didasari oleh pengetahuan awal mengenai tas wanita. Dari proses pelabelan ini diketahui pada data pelatihan terdapat 479 dokumen berlabel ***bw*** dan 321 dokumen berlabel ***nbw***.

3.3. Pembuatan Model Pengklasifikasi

Model klasifikasi dibuat dengan memasukkan data pelatihan pada algoritma Naïve Bayes. Seperti yang telah disebutkan sebelumnya, data pelatihan adalah hasil pemecahan set data yang telah dikumpulkan pada proses 3.1. Data pelatihan memiliki 1986 atribut, termasuk atribut label, dan 800 instan dengan pembagian 479 berlabel ***bw*** dan 321 berlabel ***nbw***.

Keluaran dari proses ini adalah sebuah model pengklasifikasi yang disertakan pada Lampiran A. Sebagian kecil dari model tersebut ditampilkan oleh Tabel 3.

Beberapa hal menarik dapat diperoleh dari model klasifikasi di atas, antara lain:

1. Atribut "bags" memiliki probabilitas posterior pada label ***bw*** tertinggi, *P(bw|x="bags")* = 0,99. Artinya, tanpa menghiraukan istilah-istilah lain di dalamnya, sebuah URL yang mengandung istilah "bags" diprediksi kuat sebagai halaman detail informasi produk tas wanita. Sembilan istilah lain yang juga mempunyai probabilitas posterior tinggi pada label ***bw*** tercantum pada Tabel 4.

Tabel 3. Sebagian Kecil Model Pengklasifikasi.

Atribut	Class	
	bw (0.6)	nbw (0.4)
viyar		
0	436.0	320.0
1	45.0	3.0
[total]	481.0	323.0
citrus		
0	474.0	322.0
1	7.0	1.0
[total]	481.0	323.0
bag		
0	342.0	217.0
1	139.0	106.0
[total]	481.0	323.0

Tabel 4. Sepuluh Atribut dengan Probabilitas Posterior Tertinggi pada label *bw*.

Atribut	p(x)	p(x|bw)	P(bw|x)
bags	0.12	0.19	0.99
hearted	0.08	0.14	0.98
bagtitude	0.05	0.08	0.98
bahu	0.04	0.06	0.97
yongki	0.03	0.05	0.96
komaladi	0.03	0.05	0.96
hers	0.03	0.05	0.96
brown	0.03	0.05	0.96
viyar	0.06	0.09	0.94
lzd	0.02	0.03	0.93

Tabel 5. Sepuluh Atribut dengan Probabilitas Posterior Tertinggi pada label *nbw*.

Atribut	p(x)	p(x|nbw)	P(nbw|x)
kamera	0.04	0.09	0.97
polo	0.03	0.07	0.96
cover	0.03	0.07	0.96
rain	0.03	0.07	0.95
shoes	0.03	0.07	0.95
stuff	0.05	0.13	0.95
vaccum	0.02	0.06	0.95
club	0.02	0.05	0.94
cleaner	0.02	0.05	0.94
travel	0.04	0.09	0.94

2. Sebaliknya, atribut "kamera" memiliki probabilitas posterior pada ***nbw*** tertinggi, *P(nbw|x="kamera")* = 0,97. Sehingga sebuah URL yang mengandung istilah "kamera", tanpa menghiraukan istilah-istilah lain di dalam URL tersebut, diprediksi sangat kuat sebagai halamana detail informasi produk bukan tas wanita. Sembilan istilah lain yang juga mempunyai probabilitas posterior tinggi pada label ***bw*** tercantum pada Tabel 5.
3. Seperti yang ditunjukkan oleh Tabel 6. attribut "bag" memiliki probabilitas kemunculan tertinggi (*P(x="bag")* = 0,31). Namun atribut ini muncul hampir merata di kedua kelas sehingga tidak memberikan prediksi yang kuat bagi masing-masing kelas (*P(bw|x="bag")* = 0,57 dan *P(nbw|x="bag")* = 0,43). Penulis menduga bahwa atribut-atribut ini akan mempengaruhi kesalahan klasfikasi dari model.

Tabel 6. Sepuluh Atributdengan Probabilitas Kemunculan Tinggi dan Probabilitas Posterior Rendah

Atribut	p(x)	P(bw|x)
bag	0.31	0.57
tas	0.21	0.52
hitam	0.16	0.36
backpack	0.09	0.24
biru	0.06	0.33
wallet	0.06	0.46
merah	0.06	0.53
pink	0.06	0.69
dompet	0.05	0.55
black	0.05	0.53

4. Hasil Uji Coba dan Pembahasan

Uji coba akurasi model pengklasifikasi dilakukan pada data pengujian yang telah disiapkan sebelumnya. Data tersebut terdiri dari 325 instan yang terbagi menjadi 200 instan berlabel ***bw*** dan 125 instan berlabel ***nbw***.

Pada uji coba ini model pengkasifikasi yang dibangun mempunyai akurasi 93,2%. Model tersebut berhasil mengkasifikasikan 303 instan secara benar namun masih melakukan salah klasifikasi terhadap 11 instan berlabel ***bw*** dan 11 instan berlabel ***nbw***. Tabel 7. adalah daftar instan teruji yang tidak diklasifikasikan secara benar oleh model.

Seperti yang tertera pada Tabel 8., hampir semua atribut dengan probabilitas posterior tinggi yang muncul pada set data pengujian tidak muncul pada set instan yang gagal diklasfikasi. Pengecualian terjadi pada atribut "stuff" yang muncul satu kali.

Tabel 9. menunjukkan kemunculan atribut-atribut yang terdaftar pada Tabel 6. pada instan-instan yang gagal diklasifikasi. Dari sepuluh atribut yang terdaftar, hanya atribut "biru" dan "merah" yang tidak muncul. Meskipun tidak dapat memberikan pembuktian yang valid, namun hasil ini memperkuat dugaan penulis tentang pengaruh atribut-atribut tersebut pada kesalahan klasifikasi model. Perlu ada suatu tindakan untuk menangani hal ini.

5. Simpulan dan Saran Pengembangan

Karya ilmiah ini berhasil membuat sebuah model pengklasifikasi URL halaman detail informasi produk tas wanita pada situs X dengan akurasi 93.2%. Model pengklasifikasi ini dapat digunakan sebagai saringan dalam penelusur web topikal untuk mengunjungi dan mengunduh halaman terkait.

Sebagai tambahan, dari model pengklasifikasi ini ditemukan praduga istilah-istilah yang terkait erat dengan halaman detail informasi tas wanita. Selain itu juga ditemukan praduga istilah-istilah yang membuat hasil klasifikasi model menjadi bias. Temuan tersebut masih berupa praduga dan memerlukan penelitian lebih lanjut.

Pengembangan yang juga perlu dilakukan adalah pembangunan model pengklasfikasi yang lebih general, yaitu untuk beberapa situs toko *online* lainnya. Perlu juga melakukan komparasi efisiensi waktu dan memori atas kinerja penelusur web sebelum dan sesudah penggunaan model pengklasifikasi ini.

Tabel 7. Daftar Instan yang Gagal Diklasifikasikan.

No. Instan	Istilah-istilah pada URL
16	delonghi-dl-xlr24li-violet-sco-intruskatr-violet-476913
21	mayonette-bryan-sling-coffee-192514
53	unique-tas-cross-body-elegant-runner-1128052
77	mars-collection-diapers-bag12-black-white-dbs057-487819
79	nixels-mommy-bag-longchamp-hk-large-fushia-931729
97	audyshop-shoe-tote-maroon-241055
163	bag-stuff-crocodile-tote-free-mini-pouch-hitam-386875
190	hilistork-h1666-bronze-tas-fashion-wanita-bronze-free-dompet-444418
191	nana-blanche-jam-tangan-wanita-silver-strap-stainless-steel-sw-025-982701
217	huer-temari-printed-one-zipper-wallet-green-love-1018772
220	womens-real-leather-wallet-purse-clips-clutch-phone-bag-black-1098091
224	womens-matte-long-wallets-watermelon-red-1097508
257	bloomy-rucksack-01-viola-backpack-409446
273	yadas-korea-wallet-878-40-fashion-wallet-hijau-962822
277	yadas-korea-wallet-890-16-fashion-wallet-hijau-1048307
280	whiz-iconic-3-way-easy-to-carry-korean-bag-green-tas-multifungsi-hijau-155665
283	esgotado-bag-corduro-segundo-w-tas-backpack-light-grey-427186
307	yadas-korea-wallet-889-10-fashion-wallet-rose-982809
308	yadas-korean-wallet-6802-7-fucshia-106805
309	bluetech-iconic-3-way-easy-to-carry-korean-bag-pink-821968
323	coco-pink-zoe-dompet-wanita-turquoise-873418

Tabel 8. Frekuensi Kemunculan Atribut Berprobabilitas Posterior Tinggi pada Set Data Tes dan Instan yang Gagal Diklasifikasi.

Attribut	Frekuensi Kemunculan pada	
	Set Data Tes	Gagal Klasifikasi
bags	22	0
hearted	13	0
bagtitude	19	0
bahu	12	0
yongki	13	0
komaladi	13	0
hers	9	0
brown	13	0
viyar	14	0
lzd	10	0
kamera	3	0
polo	10	0
cover	10	0
rain	9	0
shoes	6	0
stuff	13	1
vaccum	0	0
club	6	0
cleaner	9	0
travel	7	0

Tabel 9. Frekuensi Kemunculan Atribut Berfrekuensi Muncul Tinggi pada Instan yang Gagal Diklasifikasi.

Atribut	No. Instan
bag	79, 163, 220, 280, 283, 309
tas	53, 190, 283
hitam	163
backpack	257, 283
wallet	217, 220, 273, 277, 307, 308
pink	309
dompet	190, 323
black	77, 220

DAFTAR PUSTAKA

Han, J., Kamber, M. (2006). Data Mining Concepts and Techniques Second Edition. Morgan Kauffman Pub.

Liu, B. (2011) Web Data Mining: Exploring Hyperlinks, Contents. and Usage Data Second Edition. Springer.

Manning, C. D., Raghavan, P., & Scutze, H. (2009) Introduction to Information Retrieval. Cambridge: Cambridge University Press.

Menczer, F. (2011) Web Crawling. *Web Data Mining: Exploring Hyperlinks, Contents. and Usage Data Second Edition, Chapter 8*. Springer.

Miller, R.C. (n.d.). WebSPHINX: A Personal, Customizable Web Crawler. Diakses dari: *https://www.cs.cmu.edu/~rcm/websphinx/*