ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 5, Nomor 4, Tahun 2016, Halaman 763-770
Online di: http://ejournal-s1.undip.ac.id/index.php/gaussian

ANALISIS KECENDERUNGAN INFORMASI DENGAN MENGGUNAKAN METODE *TEXT MINING* (Studi Kasus: Akun *twitter* @detikcom)

Syaifudin Karyadi[1], Hasbi Yasin[2], Moch. Abdul Mukid [3]
[1]Mahasiswa Departemen Statistika FSM Universitas Diponegoro
[2,3]Staff Pengajar Departemen Statistika FSM Universitas Diponegoro
e-mail syaifudinkaryadi@gmail.com

ABSTRACT

The internet is an extraordinary phenomenon. Starting from a military experiment in the United States, the internet has evolved into a 'need' for more than tens of millions of people worldwide. The number of internet users is large and growing, has been creating internet culture. One of the fast growing social media twitter. Twitter is a microblogging service that stores text database called tweets. To make it easier to obtain information that is dominant discussed, then sought the topic of twitter tweet using clustering. In this research, grouping 500 tweets from twitter account @detikcom using k-means clustering. The results of this study indicate that the maximum index Dunn, the best grouping K-means clustering to obtain the dominant topic as many as three clusters, namely the government, Jakarta, and politics.

Keywords: text mining, clustering,, k-means , dunn index, and twitter.

1. PENDAHULUAN

Menurut Francis dan Flynn (2010), *text mining* adalah teknologi baru yang digunakan untuk data perusahaan yang selalu bertambah sehingga data teks yang tidak terstruktur tersebut dapat dianalisis. Salah satu inovasi *software* yang dapat meringankan biaya bagi penambang teks adalah *software* yang bersifat *open source*. Dua jenis *sofware open source* yang sangat populer dan diunggulkan adalah R dan Perl. R adalah bahasa pemrograman yang mendukung hal-hal yang berkaitan dengan statistik dan digunakan pada hal-hal yang berhubungan dengan ilmu pasti, matematis.

Beberapa informasi penting yang dapat diperoleh dari twitter antara lain seperti melihat sejarah perkembangan manusia, sejarah obama terpilih menjadi presiden, dll. Tersedia dalam *tweet-tweet* yang bisa dirunut di *twitter*. Penelitian ini dilakukan pengelompokkan 500 *tweet* dari akun *twitter* @detikcom menggunakan metode *k-means clustering* yang bertujuan untuk untuk mengetahui kecenderungan topik pemberitaan dan mengetahui topik yang paling sering muncul. Hasil analisis pada akun *twitter* berita tersebut akan memberikan gambaran pemberitaan akhir-akhir ini. Penelitian ini menjadi penting mengingat akun @detikcom merupakan akun berita *online* dengan *followers* terbanyak, sehingga berita yang disampaikan juga akan mempengaruhi pengetahuan dan presepsi publik terhadap suatu masalah.

2. TINJAUAN PUSTAKA

2.1. *Twitter*

Twitter diluncurkan sebagai situs *micro-blogging* pada Maret 2006 yang memungkinkan pengguna untuk mengirim *update status* hingga 140 karakter, yang dikenal sebagai *tweets*. Sejak diluncurkan, *twitter* telah mengumpulkan basis pengguna yang besar dan sekarang memiliki lebih dari 300 juta pengguna per Juni 2011 (Goyal dan Diwakar, 2011).

2.1 Data Mining dan *Text Mining*

Menurut Susanto dan Suryadi (2010), *data mining* adalah disiplin ilmu yang tujuan utamanya adalah untuk menambang pengetahuan dari data atau informasi yang dimiliki. *Text mining* adalah salah satu solusi yang dapat membantu permasalahan diatas. Menurut Gupta dan Lehal (2009), *text mining* mirip dengan *data mining*, kecuali pada teknik *data mining* yang didesain untuk pengerjaan data yang terstruktur pada sebuah database, tapi *text mining* dapat bekerja pada data yang tidak terstruktur atau semi testruktur seperti *email*, sebuah dokumen *text* lengkap, *html* dan lain-lain. Sehingga *text mining* merupakan sebuah penemuan baru dari informasi yang belum diketahui dengan mengekstrak informasi dari sumber tertulis.

Menurut Kurniawan, *et al.* (2012), langkah-langkah yang dilakukan dalam *text mining* adalah sebagai berikut :

1. *Text Preprocessing*

   Tindakan yang dilakukan pada tahap ini adalah:
   - *To lower case*, yaitu mengubah semua karakter huruf menjadi huruf kecil.
   - *Tokenizing*, yaitu proses penguraian deskripsi yang semula berupa kalimat – kalimat menjadi kata-kata.
   - *Remove number*, yaitu menghilangkan karakter angka yang ada pada kata tersebut.
   - *Remove url*, yaitu menghilangkan *link* internet.
   - *Remove punctuation*, yaitu menghilangkan delimiter-delimiter seperti tanda titik(.), koma(,) dan spasi.
2. *Feature Selection*

   Pada tahap ini tindakan yang dilakukan adalah:
   - *stopword* (*stopword removal)* adalah kosakata yang bukan merupakan ciri (kata unik) dari suatu dokumen. *Stopword* untuk bahasa Indonesia diperoleh dari: http://www.ranks.nl/stopwords/indonesian (Doyle, 2010).
   - s*temming* adalah proses pemetaan dan penguraian berbagai bentuk (*variants*) dari suatu kata menjadi bentuk kata dasarnya *(stem).*

2.2. *Term-document Matrix*

Menurut Zhao (2012), sebuah *term-document matrix* menunjukkan hubungan antara *term* dan dokumen, dimana setiap baris berisi *term* dan setiap kolom untuk dokumen.

2.3. Pembobotan

Pada penelitian ini, *term* yang telah terbentuk dihitung bobot kemunculannya dengan menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF). TF-IDF tersebut dilakukan untuk melihat bobot keterkaitan suatu *term* dengan dokumen. *Term Frequency* (TF) merupakan banyaknya *term* yang muncul pada dokumen. Sedangkan *Inverse Document Frequency* (IDF) bertujuan untuk mengetahui apakah *term* yang dicari cocok dengan kata kunci yang diinginkan *term* yang sering muncul akan memberikan pengaruh yang kecil dalam menentukan keterkaitan kata kunci dengan dokumen. *Term* yang jarang muncul akan memberikan keterkaitan yang lebih besar jika dibandingkan dengan *term* yang sering muncul (Zhang & Tang, 2008).

TF-IDF dihitung dengan menggunakan persamaan seperti berikut (Salton and Buckley, 1988):

$$W_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} . \log_2 \frac{D}{d_i}$$

2.4. Ukuran Kedekatan Kontinu

Jarak merupakan konsep penting dalam pengembangan metode pengelompokkan. Untuk mengukur jarak antara dua titik A dan B (d(A,B)), dapat menggunakan beberapa konsep jarak. Ukuran jarak harus memenuhi syarat-syarat sebagai berikut (Santoso, 2007):

1. $d(A,B) \geq 0$ (non-negatif)
2. $d(A,B) = 0$ jika dan hanya jika A = B

Jarak antara suatu objek atau titik objek dengan objek atau titik itu sendiri adalah nol

3. $d(A,B) = d(B,A)$ (simetris)

Jarak dari A ke B adalah sama dengan jarak dari B ke A

4. $d(A,C) \leq d(A,B) + d(B,C)$ (ketidaksamaan segitiga)

Formula jarak *Euclidean* merupakan formula jarak yang paling sering digunakan dalam analisis pengelompokkan. Karena, perhitungan jarak *Euclidean* mencari jarak terpendek dari dua titik dengan prinsip orthogonal (tegak lurus). Formula jarak *Euclidean* dinyatakan sebagai berikut (Prasetyo, 2012):

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

2.5. *Clustering*

Dalam Larose (2005), disebutkan bahwa algoritma *k-means* pertama kali digagas oleh MacQueen. Langkah-langkah pada algoritma *k-means* adalah sebagai berikut (Susanto dan Suryadi, 2010):

1. Tanyakan pada pemakai algoritma *k-means*, catatan-catatan yang ada akan dibuat menjadi berapa kelompok. Sebutlah sebanyak *k* kelompok.
2. Secara sembarang, pilihlah *k* buah catatan(dari sekian catatan yang ada) sebagai pusat-pusat kelompok awal.
3. Untuk setiap catatan, tentukan pusat kelompok terdekatnya dan tetapkan catatan tersebut sebagai anggota dari kelompok yang terdekat puat kelompoknya. Hitung rasio antara besaran *Between Cluster variation* dengan *Within Cluster Variation* , lalu bandingkan rasio tersebut dengan rasio sebelumnya (bila sudah ada). Lanjutkan ke langkah berikutnya, jika rasio membesar. Hentikan prosesnya, jika rasio tidak membesar.
4. Perbarui pusat-pusat kelompok(berdasarkan kelompok yang didapat dari langkah ketiga) dan kembalilah ke langkah ketiga.

2.6. Validasi *Cluster*

Validasi *cluster* merupakan teknik yang penting dilakukan untuk memberikan nilai validitas dari *cluster* yang didapat. Menurut Prasetyo (2014), pertanyaan penting kaitannya dengan evaluasi *cluster* sebagai berikut:

1. Menentukan tendensi *cluster* set data, misalnya membedakan apakah ada struktur *non-random* yang sebenarnya ada dalam data.
2. Menentukan jumlah *cluster* yang tepat.
3. Mengevaluasi seberapa baik analisis *cluster* menyelesaikan data tanpa informasi eksternal.
4. Perbandingan hasil analisis *cluser* terhadap hasil eksternal yang diketahui, seperti label kelas yang sebenarnya juga diketahui.
5. Perbandingan dua set *cluster* untuk menentukan manakah yang lebih baik.

Nilai DI yang semakin besar menandakan hasil *clustering* yang semakin baik. *Dunn Index* (DI) didapatkan dari persamaan berikut (Prasetyo, 2014):

$$DI = \min\{\min\{\frac{\delta_{(i,j)}}{\max\{\Delta_i\}}\}\}$$

3. METODOLOGI PENELITIAN

3.1. Data

Penelitian ini menggunakan 500 *tweets* terakhir. *Tweets* tersebut berasal dari *timeline* akun *twitter* @detikcom.

3.2. Metode Pengumpulan Data

Metode pengumpulan data yang berasal dari pesan teks atau *tweet* dari *timeline* akun *twitter* @detikcom diperoleh dari API (*Application Programming Interface*) pada hari Jum'at, 3 Juni 2016 jam 18.30 WIB.

3.3. Metode Analisis

Analisis data menggunakan metode *text mining* dengan bantuan *software* R. *Package* yang digunakan adalah *twitteR*, httr, base64enc, *tm, SnowballC, Rweka, rJava, Rwekajars, ggplot2, wordcloud, fpc.* Adapun metode analisis yang digunakan untuk mencapai tujuan penelitian dalam penulisan Tugas Akhir ini diuraikan sebagai berikut:

1. Membuat akun pada API, untuk memperoleh *consumer key, consumer secret, access token*, dan *access token secret* yang akan digunakan untuk mengambil data *text* twitter dengan *software* R.
2. *Text Pre-Process*, dimana data teks yang telah diambil dari twitter diolah melalui beberapa tahap, yaitu:
   - *To lower case*
   - *Tokenizing*
   - *Remove number*
   - *Remove url*
   - *Remove punctuation*
3. *Feature selection*, dimana data *text* yang telah melalui tahap *text Pre-Process* dilakukan proses selanjutnya, yaitu:
   - *stopword* (*stopword removal).*
   - *Stemming.*
4. Data *text* yang telah disusun ulang, kemudian dibuat *term-document matrix*. Baris dari setiap *matrix* tersebut berisi term dan setiap kolomnya untuk dokumen. *Matrix* yang terbentuk merupakan matrik yang telah diberi pembobotan TF dan TF-IDF.
5. Membuat barplot dan *wordcloud* dari *term-document matrix* dengan pembobotan TF.
6. Menentukan jumlah *cluster* optimum dengan memperhatikan nilai *Dunn Index*.
7. Interpretasi.

4. HASIL DAN PEMBAHASAN

4.1. Profil Akun @detikcom

Akun *twitter* @detikcom pertama kali diluncurkan pada bulan Agustus 2009 dengan jumlah post sebanyak 972 ribu *tweet* per Juni 2016, dengan jumlah *followers*/pengikut sebanyak 12,7 juta menjadikan akun *twitter* @detikcom sebagai akun berita dalam negeri dengan jumlah pengikut terbanyak di Indonesia.

4.2. *Application Programming Interface* (API)

Penelitian yang bertujuan untuk memperkenalkan beberapa fungsi analisis dasar yang dapat diimplementasikan dengan menggunakan *twitter* API . Penelitian yang menekankan pada teknik analisis untuk data yang diambil dalam jumlah besar dari akun *twitter* @detikcom. API berfungsi untuk mendapatkan *consumer key, consumer secret, access token,* dan *access token secret* yang akan digunakan untuk mengambil data *text twitter* dengan *software* R.

4.3. *Term-document Matrix* dari 500 *tweet* @detikcom

Term-document Matrix dengan pembobotan TF digunakan untuk mellihat kecenderungan *term* yang sering muncul pada 500 *tweet* @detikcom. Sedangkan *term-document Matrix* dengan pembobotan TF-IDF digunakan untuk melakukan pengukuran jarak pada pengelompokkan k-means.

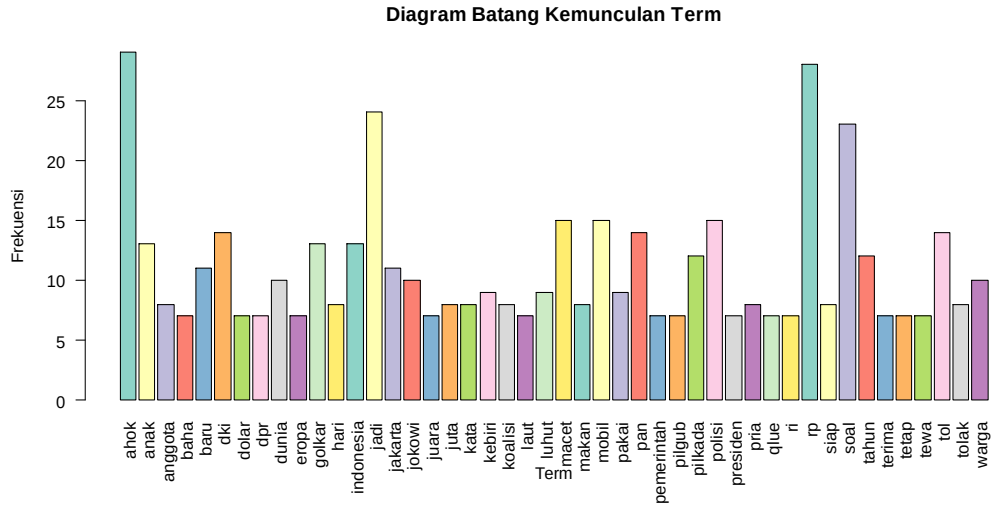
Tabel 1. *Term-document Matrix* dengan pembobotan tf untuk 500 *tweet*

Term	doc1	doc2	...	doc498	doc499	doc500
absen	0	0		0	0	0
abu	0	0		0	0	0
acar	0	0		0	0	0
acara	0	0		0	0	0
acuhkan	0	1		0	0	0
anak	0	0		0	0	0
arah	1	0		0	0	0
⋮			...			
Tersambar	0	0		1	0	0
zidan	0	0		0	0	0
zukerberg	0	0		0	0	0
zulkifli	0	0		0	0	0

Tabel 2. *Term-document Matrix* dengan pembobotan TF-IDF untuk 500 *tweet*

term	doc1	doc2	...	doc498	doc 499	doc 500
absen	0	0		0	0	0
abu	0	0		0	0	0
acar	0	0		0	0	0
acara	0	0		0	0	0
acuhkan	0	8,9658		0	0	0
anak	0	0		0	0	0
arah	6,9658	0		0	0	0
⋮			...			
tersambar	0	0		8,9658	0	0
zidan	0	0		0	0	0
zukerberg	0	0		0	0	0
zulkifli	0	0		0	0	0

4.4. *Frequent Terms* dari 500 *tweet* @detikcom



Gambar 1. Diagram Batang Kemunculan *Term* dari 500 *tweet* @detikcom (Frekuensi >= 6)

Berdasarkan diagram batang pada Gambar 1 dapat dilihat bahwa *term* yang paling sering muncul dalam 500 *tweet* teratas @detikcom adalah ahok dengan kemunculan sebanyak lebih dari 25 kali. Artinya pemberitaan @detikcom dilihat dari 500 *tweet* teratasnya cenderung memberitakan tentang Gubernur DKI Jakarta Ir. Basuki Tjahaja Purnama. M.M. atau biasa disebut ahok. Selain mengenai ahok kecenderungan pemberitaan lainnya yakni mengenai Rp(rupiah). DKI. mobil. macet. Luhut. RI. DPR. Jokowi. Eropa dll.

4.5. *Wordcloud* dari 500 *tweet* @detikcom

Gambar 2. Wordcloud 500 t*weet* dari akun @detikcom

Dengan melihat *wordcloud* pada Gambar 12 dapat disimpulkan bahwa *term* ahok merupakan yang paling besar serta berada di tengah *wordcloud* jika dibandingkan dengan

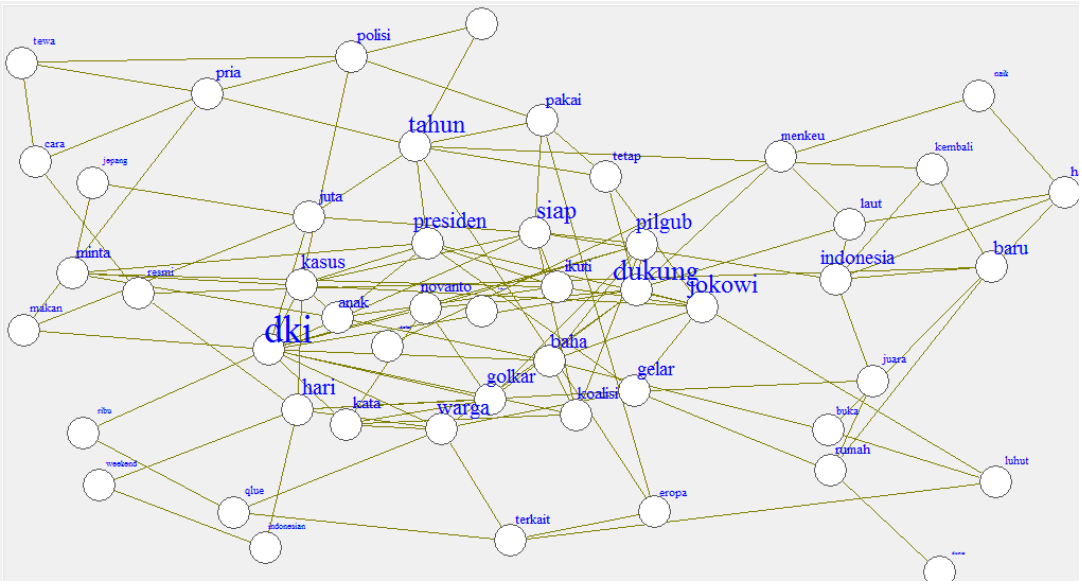
term yang lain. Diikuti dengan *term-term* yang lain yakni rp(rupiah), pan, pilkada, mobil, macet, jokowi, luhut, golkar, indonesia. dll.

4.6. *K-Means Clustering*

Berikut merupakan hasil pengelompokkan untuk *tweet* media sosial *twitter* menggunakan *K-means Clustering*:

Cluster	Anggota	Label
1	anak, bahas, baru, buka, cara, dki, dolar, dukung, dunia, eropa, gelar, golkar, harga, hari, ikuti, indonesia, indonesian, jepang, jokowi, juara, juta, kasus, kata, kebiri, kembali, koalisi, laut, luhut, makan, menkeu, minta, naik, novanto, pakai, pilgub, polisi, presiden, pria, qlue, resmi, ribu, rumah, siap, tahun, terima, terkait, tetap, tewas, warga, dan *weekend*.	Pemerintahan
2	ahok, jadi, jakarta, macet, mobil, soal, dan tol.	Jakarta
3	anggota, dpr, kota, maju, mundur, pan, pemerintah, pilkada, tolak dan wajib.	Politik

Analisis lanjut untuk menginterpretasikan *cluster* 1 pada penelitian ini dibantu dengan *network graph* seperti pada Gambar 3. *Network Graph* ini digunakan untuk melihat *term-term* yang penting dalam *cluster* 1. Berikut *Network Graph* hasil pengelompokkan pada *cluster* 1:



Gambar 3. *Network Graph K-means Clustering* untuk *Cluster* 1

4.7. Validasi *Cluster* dari 500 *tweet* @detikcom

Berdasarkan nilai *Dunn index* terbesar untuk *K-means* adalah 0,7977 yaitu pada jumlah *cluster* sebanyak tiga. Maka untuk 500 *tweet* media sosial *twitter cluster* terbaik menurut *K-means* yang dapat dibentuk adalah sebanyak tiga *cluster*.

5. PENUTUP

5.1 Kesimpulan

Berdasarkan hasil dan pembahasan pada bab IV, dapat diperoleh beberapa kesimpulan sebagai berikut:

1. Kecenderungan topik informasi yang disampaikan dari 500 *tweets* teratas akun *twitter* @detikcom dapat dilihat dari diagram batang dan *wordcloud* diperoleh hasil bahwa *term* yang paling sering muncul adalah ahok dengan kemunculan terbanyak yaitu lebih dari 25 kali. Artinya pemberitaan @detikcom cenderung memberitakan tentang Gubernur DKI Jakarta Ir. Basuki Tjahaja Purnama, M.M. atau biasa disebut ahok. Selain mengenai ahok kecenderungan pemberitaan lainnya yakni mengenai Rp (rupiah), DKI, mobil, macet, Luhut, RI, DPR, Jokowi, Eropa dan lain-lain.
2. *Cluster tweet* yang terbentuk dari 500 *tweets* teratas akun *twitter* @detikcom dengan *K-means clustering* yakni sebanyak tiga *cluster*, hal ini berdasarkan hasil dari nilai *dunn index*. Dimana, nilai *dunn index* terbesar untuk *k-means* adalah 0,7977 yaitu pada jumlah *cluster* sebanyak tiga. Dengan label untuk *cluster* 1 tentang pemerintahan, *cluster* 2 tentang Jakarta, dan *cluster* 3 tentang Politik.

5.2 Saran

Berdasarkan kesimpulan yang telah disampaikan, dapat dikemukakan beberapa saran yang dapat dilakukan oleh penelitian selanjutnya, antara lain :

1. Melakukan penelitian untuk menentukan nilai n (banyaknya data) yang tepat untuk diolah dengan *text mining* baik dari salah satu akun *twitter* ataupun dari *twitter* secara keseluruhan.
2. Melakukan penelitian menggunakan *text mining* untuk menganalisa pesan teks *tweet* pada media sosial yang didapat dari *twitter* dengan kata kunci sesuai keinginan.

DAFTAR PUSTAKA

Doyle, D. 2010. *Indonesian Stopwords.* http://www.ranks.nl/stopwords/indonesian. Diakses : 3 Juni 2016.

Francis, L., and Flynn, M.2010. Text Mining Handbook . *Casualty Actuarial Society E-Forum*. Spring.

Gupta, V., and Lehal, G, S. 2009. A Survey of Text Mining Techniques and Applications. *Emerging Technologies In Web Intelegence* Vol. 1, No. 1: Hal. 60-76.

Kurniawan, B., Effendi, S., and Sitompul, O,S. 2012. Klasifikasi Konten Berita Dengan Metode Text Mining. *Dunia Teknologi Informasi* Vol. 1, No. 1: Hal. 14-19.

Larose, D.T. 2005. *Discovering Knowledge in Data: An Introduction to Data Mining.* Hoboken: Interscience, John wiley & Sons, Inc.

Prasetyo, E. 2012. *Data Mining : Konsep dan Aplikasi Menggunakan MATLAB*. Yogyakarta : ANDI.

_____. 2014. *Data Mining : Mengolah Data Menjadi Informasi Menggunakan MATLAB*. Yogyakarta : ANDI.

Santoso, B. 2007. *Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Yogyakarta : Graha Ilmu.

Susanto, S., and Suryadi, D. 2010. *Pengantar Data Mining*. Yogyakarta : C.V Andi Offset.

Zhang, W., and Tang, X. 2008. *TF IDF, LSI and Multi-word in Information Retrieval and Text Categorization.*International Conference on Systems, Man and Cybernetics.

Zhao, Y. 2012. *R and Data Mining: Examples and Case Studies.* Elsevier .