JURNAL PINTER
VOL 2. NO.1 JUNI 2018
e-ISSN: 2597- 4475
http://doi.org/10.21009/pinter.2.1.6

Kinerja Algoritma Kmeans++ Pada Pengelompokkan Dokumen Teks Pendek pada Abstrak di Jurusan Teknik Elektro Fakultas Teknik UNJ

Catur Rahma Sistiani, Widodo[2], Bambang Prasetya Adhi[3]

[1] Mahasiswa Prodi Pendidikan Teknik Informatika dan Komputer, Teknik Elektro, FT – UNJ
[2,3] Dosen Prodi Pendidikan Teknik Informatika dan Komputer, Teknik Elektro, FT – UNJ
[1] caturrahmasistiani@gmail.coml, [2] widodo03@yahoo.com, [3] bambangpadhi7@gmail.com

Abstrak

Pengelompokkan pada dokumen teks pendek masih sulit ini dikarenakan di sparsity kata. Tujuan penelitian ini adalah untuk mengetahui kinerja algoritma k-means++ pada teks pendek dan untuk mengetahui proses pengelompokkan algoritma k-means++ pada teks pendek di abstrak skripsi Jurusan Teknik Elektro Fakultas Teknik UNJ dilaksanakan pada semester genap tahun ajaran 2014-2015. Penelitian ini menggunakan metode penelitian eksperimen. Data abstrak yang digunakan sebanyak 200 abstrak. Penelitian meneliti 4 data yaitu Data pertama adalah abstrak ilmiah di jurusan Teknik Elektro, Universitas Negri Jakarta pada paragraf 1 sampai paragraf 3. Data kedua adalah paragraf 1 pada abstrak ilmiah di jurusan Teknik Elektro, Universitas Negri Jakarta. Data ketiga adalah paragraf 2 pada abstrak ilmiah di jurusan Teknik Elektro, Universitas Negri Jakarta. Data keempat adalah paragraf 3 pada abstrak ilmiah di jurusan Teknik Elektro, Universitas Negri Jakarta. Pengujian kinerja algoritma k-means++ menggunakan matrix confusion. Berdasarkan hasil penelitian, didapatkan kesimpulan bahwa keakurasian pada abstrak, paragraf 1 di abstrak, paragraf 2 di abstrak, dan paragraf 3 di abstrak mencapai lebih dari 80% . Didapatkan juga kesesuaian antar data yang diprediksi dengan hasil yang benar dari data yang sebenarnya(presisi) pada abstrak, paragraf 1 di abstrak, paragraf 2 di abstrak, dan paragraf 3 di abstrak mencapai lebih dari 50% . Didapatkan juga peluang munculnya data relevan yang diambil sesuai dengan query (recall) pada abstrak, paragraf 1 di abstrak, paragraf 2 di abstrak, dan paragraf 3 di abstrak mencapai lebih dari 80%.

Kata kunci : **Algoritma kmenas++, Teks Pendek, *Matrix Confusion***

1. Pendahuluan

Banyaknya dokumen teks yang tersimpan dalam komputer membuat pencarian informasi menjadi sulit. *Clustering* menjadi salah satu solusi untuk mengelompokkan dokumen yang berjumlah besar sehingga membantu proses pencarian informasi yang dibutuhkan.

Clustering ini adalah metode pengelompokkan data yang menemukan kelompoknya berdasarkan kemiripan atau kesamaan secara natural. Ada banyak sekali macam-macam *clustering* contohnya algoritma k-means yang paling banyak dikenal, mudah diterapkan dan juga algoritma k-means menjadi salah satu 10 top algoritma pada data mining menurut IEEE ICDM pada tahun 2006. Tetapi k-means ini mempunyai kekurangan yang sangat tergantung dengan kondisi inisialisasi awal *clustering*. Pada algoritma k-means, pemilihan kondisi inisialisasi awal di lakukan secara acak, jika inisialisasi kurang baik, maka waktu proses pengelompokkan yang dihasilkan pun menjadi kurang optimal. maka algoritma k-means ini perlu pengembangan untuk inisialisasi pada *cluster* awal.

Pada tahun 2007 David Arthur dan Sergei Vassilvitskii mengembangkan algoritma k-means++ untuk mengatasi masalah kekurangan pada proses pengelompokkan yang kurang optimal, yang terdapat pada algoritma k-means.

Avalaiable at:
http://journal.unj.ac.id/unj/index.php/pinter/article/view/9772

Tipe teks yang berjenis teks panjang mudah saat proses pengelompokannya dan sering digunakan untuk penelitian, tetapi pengelompokkan untuk teks pendek masih jarang. Pengelompokkan teks pendek masih dianggap sulit secara statistik dan teks pendek ini mempunyai masalah di *sparsity* (keterbatasan) kata-kata dan teks pendek ini mengandung tidak lebih dari 100 kata.

2. Clustering

Menurut Berkhin, Pavel diacu dalam Sri Andayani (2007), *clustering* adalah membagi data ke dalam group-group yang mempunyai obyek yang karakteristiknya sama. Menurut Garcia-Molina Et al., diacu dalam Sri Andayani (2007), *clustering* adalah mengelompokkan item data ke dalam sejumlah kecil grup sedemikian sehingga masing-Masing grup mempunyai sesuatu persamaan yang esensial. *Clustering* adalah pengelompokkan objek yang mirip satu sama lain dan objek yang berbeda tergabung dengan kelompok lain (Max Bramer,2007). Berdasarkan definisi di atas dapat disimpulkan bahwa *clustering* adalah pengelompokkan data yang menemukan kelompokkannya dengan karakteristik yang sama dan objek yang berbeda akan tergabung dengan kelompok yang lain.

3. Algoritma kmeans++

Metode *clustering* ada banyak salah satunya adalah k-means++, telah dikembangkan dari algoritma k-means di mana salah satu metode *clustering* menggunakan pendekatan yang didasari pada partisi (D.Arthur,2007). K-means menentukan nilai awal *centroid* secara acak yang kadang-kadang memerlukan waktu pemrosesan lebih lama. Oleh karena itu, k-means++ digunakan untuk mengurangi kelemahan k-means dari pemrosesan waktu yang lama. Algoritma k-means++ merupakan algoritma pengelompokan secara partisi yang merupakan pengembangan dari algoritma k-means (Wayan Surya Prianta,dkk., 2011). Berdasarkan pendapat beberapa ahli di atas, dapat di simpulkan bahwa k-means++ adalah pengembangan dari algoritma kmeans untuk mengurangi kelemahan k-means yang kadang-kadang memerlukan waktu pemrosesan *clustering* lebih lama. Berikut ini adalah algoritma k-means++

1. Menentukan satu *centroid* awal pada seluruh data dokumen secara acak
2. Tambahkan sebuah center baru c_i dari semua data yang belum terpilih sebagai *centroid*, dengan menggunakan *weighted propability distribution* dimana data yang dipilih dengan probabilitas tertinggi dengan menggunakan rumus : $\frac{D(x)^2}{\sum x \in X D(x)^2}$
3. Ulang langkah ke 2, hingga sejumlah *k centroids* telah terpilih
4. Lalu di lanjutkan dengan algoritma K-Means, yaitu
   A. Hitung jarak antar *centroid* dengan menggunakan rumus sebagai berikut:
   $d_{(x,y)} = \sqrt{(xi - \text{yi})^2 + (xi - \text{yi})^2}$
   Keterangan:
   d= titik dokumen
   x=data *record*
   y=data *centroid*
   B. Kelompokkan setiap data berdasarkan jarak terpendek antar *centroid* dengan dokumen, untuk menentukkan posisi cluster suatu dokumen. Misalnya dokumen A mempunyai jarak yang paling dekat dengan *centroid* 1 dibanding dengan yang lain maka dokumen A masuk ke kelompok 1.
   C. Hitung kembali posisi *centroid* baru untuk tiap tiap *centroid* ($C_{i...j}$) dengan cara menghitung nilai rata rata dokumen yang masuk pada *cluster* awal ($G_{i...j}$). rumus sebagai berikut:
   $$C(i) = \frac{X_1 + X_2 + X_n}{\sum x}$$
   Keterangan:
   X_1= nilai data *record* ke-1
   X_2= nilai data *record* ke-2
   X_n= nilai data *record* ke-n
   $\sum x$= jumlah data *record*
   D. Ulangi langkah a,b,c hingga posisi *centroid* tidak berubah

4. Dokumen Teks Pendek

Jenis jenis dokumen menurut sifatnya ada 2 yaitu dokumen tekstual dengan dokumen nontekstual. Dokumen tekstual adalah menyajikan informasi dalam bentuk tulisan, contohnya jurnal, majalah, buku, dan sebagainya. Dokumen teks pendek termasuk jenis dokumen tekstual. Menurut, dokumen teks pendek adalah dokumen yang berisi tidak lebih dari 100 kata, contohnya pada setiap paragraf abstrak ilmiah, dan *tweet* pada *twitter* (Mika Timonen,dkk., 2012). Menurut buku pedoman skripsi Jurusan Teknik Elektro Fakultas Teknik UNJ(2012), abstrak merupakan tulisan singkat menyeluruh dari isi skripsi/KI/komprehensif sehingga dengan membaca abstrak pembaca dapat menilai isi nya dengan cepat karena abstrak berisi pokok masalah, tujuan, metode penelitian, hasil penelitian dan kesimpulan. Panjang abstak maksimal 1 halaman berjumlah 200 kata (lebih kurang 20 kalimat).

5. Metodologi

Metode penelitian yang digunakan adalah metode eksperimen.

Langkah langkah penelitian

A. Pengumpulan Data

pengelompokkan abstrak berdasarkan kelompok ini akan dipakai 4 data yang berbeda. Data pertama adalah abstrak ilmiah di jurusan Teknik Elektro, Universitas Negri Jakarta pada paragraf 1 sampai paragraf 3. Data kedua adalah paragraf 1 pada abstrak ilmiah di jurusan Teknik Elektro, Universitas Negri Jakarta. Data ketiga adalah paragraf 2 pada abstrak ilmiah di jurusan Teknik Elektro, Universitas Negri Jakarta. Data keempat adalah paragraf 3 pada abstrak ilmiah di jurusan Teknik Elektro, Universitas Negri Jakarta. Setelah itu data diubah format nya menjadi .txt.

B. Pengelompokkan Secara Manual

Setelah data dibagi bagi maka data di kelompokkan dahulu sacara manual dengan berdasarkan 2 cluster yaitu pendidikan dan non pendidikan. Pengelompokkan berdasarkan pendidikan memiliki keyword yang dapat memudahkan untuk proses pengelompokkan secara manual. Keywordnya seperti siswa/peserta didik, pembelajaran, pengajar/guru. Kalau pengelompokkan berdasarkan non pendidikan juga memiliki keyword dan keywordnya adalah selain dari keyword pengelompokkan pendidikan. Setelah di kelompokkan maka mendapatkan hasil 66 data pendidikan dan 134 data non pendidikan.

C. Pra Proses Data

Tahapan pra proses data, meliputi:

1. *toLowerCase*

   Pada tahap *toLowerCase*, mengubah semua kata yang ada di setiap dokumen menjadi huruf kecil

2. Tokenisasi

   Pada tahap tokenisasi ini, penguraian deskripsi yang semula berupa kalimat-kalimat menjadi kata-kata dan menghilangkan delimiter-delimiter seperti tanda (.), koma(,), spasi dan karakter angka yang ada pada kata tesebut. Setiap dokumen dan query direpresentasikan dengan model *bag-of-words*

D. Feature Selection

Setelah selesai dengan tahap pra proses data maka kata kata yang tidak deskriptif dapat dihilangkan dalam pendekatan *bag-of-words* di sebut juga *stopword* . Contoh *stopword* adalah dan, atau, yang, adalah, yaitu.

E. Membuat Vector Space Model

Setelah menghilangkan stopword pada setiap dokumen, kemudian tahap berikutnya membuat *vector space model*. Pembuatan *vector space model*, yaitu mengumpulkan kata pada semua dokumen, kemudian mencari *document frequency* setiap kata pada *document*, kemudian menghitung *invers document frequency* dengan cara menggunakan rumus sebagai berikut:

$$idf = log \frac{total\ document}{document\ frequency}$$

kemudian mencari *term frequency* setiap kata per *document*, *term frequency* ini didapatkan dari banyaknya per kata per dokumen,

$$TF = Jumlah\ kata\ per\ dokumen$$

setelah itu mencari TFIDF(*Term Frequency Invers Document Frequency*) dengan cara menggunakan rumus sebagai berikut:

$$\text{TFIDF}\,(w,d) = \text{TF}(w,d) \times (1 + (\log(\frac{N}{\text{DF}(w)})))$$

Setelah melakukan tahap pembutan vector space model maka normalkan dengan cara persamaan sebagai berikut

$$Normal(i) = \frac{TFIDF(i)}{\sqrt{TFIDF(i)^2 + TFIDF(i+1)^2 + \cdots + TFIDF(i+n)^2}}$$

F. Mengurangi Dimensi

Setelah data sudah dinormalkan maka data terlebih dahulu di PCA(*Principal Componen Analysis*), untuk mendapatkan principal component pada matlab adalah dengan menggunakan fungsi *princomp* yang telah disiapkan pada statics *toolbox*. Pada fungsi *princomp* akan menghitung *eigen function* dari *covarians* dan menghasilkan tiga buah variabel, yaitu *coeffisien, score dan latent.*

$$[Coeff\ Score\ Latent] = princomp(data\ yang\ di\ input)$$

Keterangan

Coeffisien : menyimpan nilai koefisien dari *principal component*

Latent : menyimpan varians dari *principal component*

Score : *data principal componen* dari data yang di input yang telah di urutkan dari baris pertama sampai baris terakhir, dimana baris pertama mengandung informasi data (*principal component*) yang

paling penting pertama , kemudian paling penting nomor 2 berada di baris kedua dan seterusnya hingga baris terakhir yang merupakan data yang kurang berarti.

Setelah di *princomp* ambil 2 data pertama dari hasil *coeffisien.* Kemudian dikali dengan data yang yang sudah dinormalkan.

G. Algoritma Kmeans++

Setelah tahap mengurasi dimensi kemudian selanjutnya langkah berikutnya data dapat di cluster dengan menggunakan algoritma k-means++. Langkah langkah *clustering* dengan algoritma k-means++ sebagai berikut

A. Memilih *centroid* awal dengan cara memilih k_i secara acak setelah itu memilih k_{i+1} dengan cara menggunakan *weighted propability distribution,* dengan menggunakan rumus : $\frac{D(x)^2}{\sum x \in X^{D(x)^2}}$ Setelah itu cari data dengan probabilitas tertinggi, sampai sejumlah *k centroids* telah terpilih.

B. Kemudian hitung iterasi dengan menggunakan rumus sebagai berikut:
$d_{(x,y)} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$
Keterangan:
d= titik dokumen
x=data *record*
y=data *centroid*

C. Kelompokkan setiap data berdasarkan jarak terpendek antar *centroid* dengan dokumen, untuk menentukkan posisi *cluster* suatu dokumen.

D. Hitung kembali posisi *centroid* baru untuk tiap-tiap *centroid* ($C_{i...j}$) dengan cara menghitung nilai rata rata dokumen yang masuk pada *cluster* awal ($G_{i...j}$). rumus sebagai berikut:
$$C(i) = \frac{x1 + x2 + x \ldots + x_n}{\sum x}$$
Keterangan:
X_1= nilai data *record* ke-1
X_2= nilai data *record* ke-2
$\sum x$= jumlah data *record*

E. Ulangi langkah b,c,d hingga posisi *centroid* tidak berubah, yaitu titik *centroid* sama dengan titik sebelumnya.

6. Hasil dan Analisis

Tabel 6.1 Pengukuran pada abstrak

Predicted Class				Total
Actual Class		Non Pendidikan	Pendidikan	
	Non Pendidikan	133	20	153
	Pendidikan	1	46	47
TOTAL		134	66	200

Berdasarkan tabel 4.5 ,didapatkan data berupa matriks untuk mengukur nilai akurasi, presisi dan *recall* pada abstrak setiap kategori. Perhitungan untuk akurasi, presisi dan *recall* adalah

Akurasi= $\frac{133+46}{200} = 0{,}90$

Presisi Non Pendidikan = $\frac{133}{134} = 0{,}99$

Presisi Pendidikan = $\frac{46}{66} = 0{,}70$

Recall Non Pendidikan = $\frac{133}{153} = 0{,}87$

Recall Pendidikan = $\frac{46}{47} = 0{,}98$

Tabel 6.2 Pengukuran pada paragraf 1 abstak

Predicted Class				TOTAL
Actual Class		Non Pendidikan	Pendidikan	
	Non Pendidikan	127	18	145
	Pendidikan	7	48	55
TOTAL		134	66	200

Berdasarkan tabel 4.6 ,didapatkan data berupa matriks untuk mengukur nilai akurasi paragraf 1 pada abstrak , presisi dan *recall* paragraf 1 pada abstrak setiap kategori. Perhitungan untuk akurasi, presisi dan *recall* adalah

Akurasi= $\frac{127+48}{200} = 0{,}88$

Presisi Non Pendidikan = $\frac{127}{134} = 0{,}95$

Presisi Pendidikan = $\frac{48}{66} = 0{,}73$

Recall Non Pendidikan = $\frac{127}{145} = 0{,}88$

Recall Pendidikan = $\frac{48}{55} = 0{,}87$

Tabel 6.3 Pengukuran pada paragraf 2 abstrak

Predicted Class				TOTAL
Actual Class		Non Pendidikan	Pendidikan	
	Non Pendidikan	133	32	165
	Pendidikan	1	34	35
TOTAL		134	66	200

Berdasarkan tabel 4.7 ,didapatkan data berupa matriks untuk mengukur nilai akurasi, presisi dan *recall* paragraf 2 pada abstrak setiap kategori. Perhitungan untuk akurasi, presisi dan *recall* adalah

Akurasi= $\frac{133+34}{200} = 0{,}84$

Presisi Non Pendidikan = $\frac{133}{134} = 0{,}99$

Presisi Pendidikan = $\frac{34}{66} = 0{,}51$

Recall Non Pendidikan = $\frac{133}{165} = 0{,}81$

Recall Pendidikan = $\frac{34}{35} = 0{,}97$

Tabel 6.4 Pengukuran pada paragraf 3 abstrak

Predicted Class				TOTAL
Actual Class		Non Pendidikan	Pendidikan	
	Non Pendidikan	132	17	149
	Pendidikan	2	49	51
TOTAL		134	66	200

Berdasarkan tabel 4.8 ,didapatkan data berupa matriks untuk mengukur nilai akurasi, presisi dan *recall* paragraf 3 pada abstrak setiap kategori.

Perhitungan untuk akurasi, presisi dan *recall* adalah

Akurasi= $\frac{49+132}{200} = 0{,}91$

Presisi Non Pendidikan = $\frac{132}{134} = 0{,}99$

Presisi Pendidikan = $\frac{49}{66} = 0{,}74$

Recall Non Pendidikan = $\frac{132}{149} = 0{,}89$

Recall Pendidikan = $\frac{49}{51} = 0{,}96$

7. Kesimpulan dan Saran

7.1. KESIMPULAN

Berdasarkan hasil penelitian, didapatkan akurasi pada abstrak sebesar 90%, akurasi pada paragraf 1 di abstrak sebesar 88%, akurasi pada paragraf 2 di abstrak sebesar 84%, akurasi pada paragraf 3 di abstrak sebesar 91%. Jadi tingkat akurasi yang baik adalah pada paragraf 3 di abstrak skripsi Jurusan Teknik Elektro Fakultas Teknik UNJ sebesar 91%. Walaupun menurut D.manning and Hinrich schutze mengatakan kemunculan kata hanya 1 kali disebut *hapax legomena*. Pada penelitian ini terjadi di setiap dokumen kemunculan kata hanya 1 kali dan itulah kekurangan dari dokumen teks pendek akan tetapi dengan algoritma kmeans++ untuk pengelompokkan teks pendek mempunyai keakurasian diatas 80% dan presisi diatas 50% dan *recall* diatas 80%. Jadi pengelompokkan teks pendek mempunyai akurasi, presisi dan *recall* yang tinggi. Maka kesimpulannya, pengelompokkan pada dokumen teks pendek dengan menggunakan algoritma kmeans++ baik untuk 2 cluster yaitu pendidikan dan non pendidikan untuk hanya di Jurusan Teknik Elektro Fakultas Teknik UNJ.

7.2 SARAN

Berdasarkan penelitian dapat dikemukkan saran, yaitu memperhatikan cara penulisan pada abstrak skripsi di Jurusan Teknik Elektro Fakultas Teknik UNJ, untuk penelitian selanjutnya dapat menambahkan k nya dan jika k ingin ditambah maka data harus di tambah dan jika penelitian ini dilanjutkan maka harus menggunakan algoritma lain dalam proses pengelompokkannya.

Daftar Pustaka:

Adi Wibowo, d. (2013). *Implementasi Generalized Vector Space Model Menggunakan WordNet*.

Adiningsih,E.S.; Mahmud; Effendi,I. (2004). *Aplikasi Analisis Komponen Utama Dalam Pemodelan Penduga Lengas Tanah Dengan Data Satelit Multi Spektral*. Matematika dan Sains , 215-222.

Amalia Indranandita, d. (2008). *Sistem Klasifikasi dan Pencarian Jurnal dengan Menggunakan Metode Naive Bayes dan Vector Space Model. Informatika* , 9-18.

Amin, f. (2011). *Implementasi Search Engine(mesin pencari) Menggunakan Metode Vector Space Model. Dinamika Teknik* , 45-58.

Andayani, S. (2007). *Pembentukan Cluster dalam Knowledge Discovery in Database dengan Algoritma k-means*.

Arthur, D., & Vassilvitskii. (2007). *K-menas++:The advantages of Careful Seeding*.

Brammer, M. (2007). *Principlesd of Data Mining*. londong: Springer.

Christopher, D. M., & Hinrich, S. (1999). *Foundations of Statistical Natural Language Processing*. London: Cambridge University Press.

Dragut, E., Fang, F., Sistla, P., Yu, S., & Meng, w. (2009). *Stop Word and Rekated Problem in WebInterfac Integration*.

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook*. New York: cambridge University Press.

Han, J., & Kamber, M. (2006). *Data Mining Concept and Tehniques*. San Fransisco: Morgan Kauffman.

Intan,R.&Defeng,A.2006.HARD:*Subject Based Search Engine Menggunakan TFIDF dan Jaccard's Coefficient*.https://puslit.petra.ac.id/journals/pdf.php

Ismail Djakaria, d. (2010). *Visualisasi Data Iris Menggunakan Analisis Komponen Utama dan Analisis Komponen Utama Kernel*. Ilmu Dasar.

Istiany, Ari.(2012).*Buku Pedoman Skripsi Jurusan Teknik Elektro Fakultas Teknik UNJ*. Jakarta: UNJ.

Karandikar, A. (2010). *Clustering Short Status Message A Topuc Model BAsed Approach*.

Kumar, R., & Mathur, R. P. (2014). *Short Text Clustering Using Numeric Data Based on N-gram*. *IEEE* , 274-276.

Kurniawan,B; Effendi,S; Sitompul,O.S. (2012). *Klasifikasi Konten Berita dengan Metode Text Mining*. *Dunia Teknologi Informasi* , 14-19.

Manning, C., Raghavan, p., & Schutze, H. (2008). *Introduction of Information Retrieval* . New York: Cambridge University Press.

Oktafia, D., dan Pardede, D.L.C., 2010, *Perbandingan Kinerja Algoritma Decision Tree dan Naïve Bayes dalam Prediksi Kebangkrutan*, *Proceeding Seminar Ilmiah Nasional KOMMIT 2010*, Universitas Gunadarma

Poerwadarminta, W. (2007). *Kamus Umum Bahasa Indonesia.*

RENIER, G.J (1997). *History its purpose and Method.* Yogyakarta: Pustaka Pelajar.

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

Simanjuntak, P.J. 2005. *Manajemen dan Evaluasi Kerja*. Lembaga Penerbit FEUI, Jakarta.

Setiawan, E. (2012). *Kamus Besar Bahasa Indonesia Online.* KEMDIKBUD.

Shrestha, P., Jacquin, C., & Daille, B. (2012). *Clustering Short Text and Its Evaluation*. *A. Gelbukh (Ed.): CICLing* , 169-180.

Smith,L.I.A Tutorial on Principal Component Analysis.Internet: http://www.csotago.ac.nz/cosc453/student_tutorials/principal_component.pdf,[19 Juni 2015]

Swastina, L. (2013). *Penerapan Algoritma C4.5 Untuk Penentuan Jurusan Mahasiswa*. *Gema Aktualita* , 93-98.

Timonen, M., Toivanen, T., Kasari, M., Teng, Y., Cheng, C., & He, L. (2012). *Keyword Extraction from Short Documents Using Three Levels of Word Evaluation*. *knowledge Discovery, Knowledge Engineering and Knowledge Management* , 130-146.

Wayan surya priantara, d. (2011). *Implementasi Deteksi Penjiplakan dengan Algoritma Winnowing pada Dokumen Terkelompok*. *Seminar Tugas Akhir* , 1-9.

Wibowo. 2007. *Manajemen Kinerja* : Jakarta : Raja Grafindo Persada.

Weiss, S., Indurkhya, n., zhang, T., & Damerau, F. (2005). *Text Mining: Predictive Methode to Analyzing Unstructed Information.* New York: Springer