PERANCANGAN DAN ANALISIS CLUSTERING DATA MENGGUNAKAN METODE K-MEDOIDS UNTUK BERITA BERBAHASA INGGRIS
DESIGN AND ANALYSIS OF DATA CLUSTERING USING K-MEDOIDS METHOD FOR ENGLISH NEWS

Harival Zayuka[1], Surya Michrandi Nasution, ST.,MT[2], Yudha Purwanto, ST.,MT[3].

[1,2,3] Prodi S1 Sistem Komputer, Fakultas Teknik Elektro, Universitas Telkom
harivalzayuka@students.telkomuniversity.ac.id[1], mirchandi@telkomuniversity.ac.id[2], omyudha@telkomuniversity.ac.id [3].

Abstrak

Saat ini perkembangan dunia teknologi dan informasi sangat berkembang dengan pesat. Tidak heran hal ini ini terjadi juga pada jumlah dokumen berita khususnya berita digital yang ada pada media online. Hal ini menyebabkan semakin sulitnya untuk melakukan pencarian terhadap suatu topik berita. Clustering merupakan salah satu metode data mining yang bersifat unsupervised learning untuk mengelompokkan dokumen berdasarkan kemiripannya. Untuk melakukan pengelompokan tersebut, digunakan salah satu algoritma clustering yaitu Partitioning Around Medoid (PAM). Partitioning Around Medoid merupakan suatu algoritma clustering yang berusaha melakukan pengelompokan suatu dataset dengan mencari terlebih dahulu sejumlah titik yang merepresentasikan suatu cluster (medoid). Setelah mendapatkan k-medoid dokumen pada dataset dikelompokan kedalam cluster yang memiliki jarak ke medoid terdekat. Adapun metode pendekatan yang digunakan untuk menghitung jarak antar dokumen adalah euclidean distance method. Nilai rangking yang dibangun menggunakan metode TF*IDF pada penelitian ini dapat dijalankan sehingga dapat diketahui hasil summary dari berita pertama yang mempunyai nilai rangking 2.4082399653118496 dan berita kedua yang mempunyai nilai rangking 3.4614262661931448 dan sesuai dengan penentuan kalimat utama dalam website tersebut.

Kata kunci : Partitioning Around Medoid, K-Medoids, Euclidean Distance Method.

Abstract

Currently the development of the information technology world and is growing rapidly. No wonder this is the case also in the number of digital news, especially news documents that exist in the online media. This causes more difficult it is to conduct a search of a news topics. Clustering is one method of data mining that is unsupervised learning to classify documents based on similarity. To do that grouping, use one of the clustering algorithm is Partitioning Around Medoid (PAM). Partitioning Around Medoid a clustering algorithm that seeks grouping a dataset by finding beforehand a number of points which represents a cluster (medoid). After obtaining the documents in the dataset k medoid grouped into clusters that have the distance to the nearest medoid. As for the method used to calculate the distance between the document is euclidean distance method. The ranking value built using TF*IDF methods tn this study can run so that can be known summary result from the first news that has a value rank is 2.4082399653118496 and the second news that has a rank value is 3.4614262661931448, and in accordance with main idea in the website.

Keywords: Partitioning Around Medoid, K-Medoids, Euclidean Distance Method.

I. PENDAHULUAN

1.1. Latar Belakang

Berdasarkan data dari Kementerian Komunikasi dan Informasi Indonesia yang diperoleh dari Lembaga Riset Pasar E-Marketer, populasi pengguna internet tanah air pada tahun 2017 mencapai 112 juta orang. Hal ini menunjukkan besarnya kebutuhan internet bagi penduduk Indonesia sebagai media untuk mendapatkan informasi.

Seiring dengan itu, berbagai bentuk dan jenis informasi tersedia melimpah di internet. Namun, setiap pengguna internet memiliki kebutuhan informasi tersendiri. Sebagian besar para pengguna hanya ingin memperoleh informasi dalam bidang tertentu saja. Membaca koleksi dokumen yang berjumlah banyak tentu akan membutuhkan waktu yang sangat lama.

Disamping itu, latar belakang penulis membuat Tugas Akhir ini adalah perkembangan website yang meningkat drastis sehingga mengakibatkan semakin pesat pula perkembangan artikel berita. Banyaknya informasi tersebut menyebabkan pengguna internet dapat mengalami kesulitan untuk menemukan informasi atau berita yang diinginkan, sehingga informasi tersebut tidak dapat dimanfaatkan secara optimal. Oleh karena itu dibutuhkan pengelompokan informasi

agar lebih terstruktur sehingga informasi tersebut dapat dimanfaatkan secara maksimal.

Tahapan pertama yang dilakukan dalam mengolah informasi adalah preprocessing. Preprocessing adalah tahapan untuk mentransformasi data ke suatu format yang lebih mudah dan efektif untuk diproses oleh pemakai. Langkah-langkah dalam preprocessing terdiri dari sentence, tokenization, stopworld removal, stemming, TF (Term Frequency), IDF (Inverse Document Frequency), TF*IDF.

Langkah selanjutnya adalah pengelompokan dokumen informasi dalam data mining dengan metode clustering. Clustering merupakan proses pengelompokkan data atau objek ke dalam bentuk cluster (kelompok) sehingga setiap data dalam kelompok memiliki kemiripan variabel-variabel yang diteliti, dan didapatkan kemiripan objek dalam kelompok yang sama dibandingkan sesama objek dari kelompok yang berbeda. Tidak melakukan klasifikasi, estimasi, atau memprediksi nilai variabel target. Dan teknik ini hanya melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan. Terdapat banyak metode clustering yang dapat digunakan untuk melakukan pengelompokan dokumen, diantaranya partitional clustering dan hierarchical clustering.

Masing-masing metode ini memiliki kelebihan dan kekurangan. Partitional clustering merupakan metode clustering yang prosesnya membagi objek menjadi beberapa partisi dan setiap objeknya hanya menjadi bagian dari sebuah cluster, sehingga tidak terjadi overlaping. Sementara itu, hierarchical clustering merupakan metode clustering yang direpresentasikan dengan sebuah pohon biner yang disebut dendrogram pada metode hierarchical clustering objek cluster-nya dapat terjadi overlaping. Metode hierarchical clustering ini mengelompokkan data dengan cara membuat suatu hirarki berupa dendogram dimana data yang mirip akan ditempatkan pada hirarki yang berdekatan dan yang tidak mirip di tempatkan pada hirarki yang berjauhan.

Penelitian ini menggunakan metode K-Medoids. Metode K-Medoids merupakan bagian dari partitioning clustering. Metode K-Medoids cukup efisien untuk dataset yang kecil. Langkah awal K-Medoids adalah mencari titik yang paling representatif (medoids) dalam sebuah dataset dengan menghitung jarak dalam kelompok dari semua kemungkinan kombinasi dari medoids sehingga jarak antar titik dalam suatu cluster kecil sedangkan jarak titik antar cluster besar.

Penelitian ini penulis melakukan penelitian terhadap data berita berbahasa inggris yang diambil dari beberapa situs online, kemudian disalin ke notepad dalam format .txt. Data berita tersebut berupa dokumen teks yang memiliki isi tulisan atau kumpulan kata berbahasa inggris. Metode pembobotan yang digunakan dalam penelitian ini ada metode TF-IDF. Metode tersebut digunakan untuk mendapatkan nilai rangking pada setiap kalimat. Sehingga dengan nilai rangking tersebut didapatkan kalimat utama dalam dokumen tersebut yang selanjutnya diproses untuk menghasilkan kesimpulan (*summary*) dari artikel/berita.

1.2. Rumusan Masalah

Berdasarkan permasalahan yang terdapat pada latar belakang, maka rumusan masalah yang terkain dalam penelitian ini adalah :

1. Bagaimana menentukan kalimat utama dalam suatu dokumen berita menggunakan pembobotan TF-IDF.
2. Bagaimana melakukan pengelompokkan dokumen berita menggunakan metode K-Medoids.

1.3. Tujuan

Tujuan yang dari penelitian ini adalah sebagai berikut:

1. Menentukan kalimat utama dalam sebuah dokumen berita menggunakan pembobotan TF-IDF .
2. Mengelompokkan dokumen berita dengan metode K-Medoids.

1.4. Batasan Masalah

Pada penelitian ini terdapat batasan masalah sebagain berikut :

1. Menggunakan bahasa pemrograman Java dan IDE Netbeans sebagai compiler.
2. Dataset yang digunakan adalah dataset berita online yang sudah disalin ke dalam file notepad dengan format .txt.
3. Data teks menggunakan berita berbahasa inggris.
4. Tidak menangani kesalahan pada penulisan kata dalam dokumen.
5. Kualitas cluster tidak dibandingkan dengan algoritma lainnya.

II. TEORI DASAR

2.1. Preprocessing

Sebelum di proses data mining sering kali diperlukan preprocessing. Data preprocessing menerangkan tipe-tipe proses yang melaksanakan data mentah untuk mempersiapkan prosedur yang lainnya. Tujuan preprocessing dalam data mining adalah mentransformasi data ke suatu format yang prosesnya lebih mudah dan efektif untuk kebutuhan pemakai. Tahapan preprocessing diantaranya sentence, tokenization, stopword removal, stemming, TF, IDF, TF*IDF.

2.2. Clustering

Menurut Jiawei Han (2006: 401-430) secara umum metode pada *clustering* dapat digolongkan ke dalam beberapa metode berikut diantaranya, metode partisi, metode hierarki, metode berbasis kerapatan (*density based method*), metode berbasis *grid*, metode berbasis model.

Langkah kerja metode partisi adalah apabila terdapat basis data sejumlah n objek, selanjutnya data dipartisi menjadi k partisi dari data, dimana setiap partisi mewakili sebuah cluster k ≤ n.

Syarat yang harus terpenuhi suntuk metode partisi sebagai berikut:

1. Setiap kelompok harus berisi setidaknya satu objek.
2. Setiap objek harus memiliki tepat satu kelompok.

2.3. K-Medoids

Penelitian ini menggunakan metode K-Medoids. Metode K-Medoids merupakan bagian dari partitioning clustering. Metode K-Medoids cukup efisien untuk dataset yang kecil. Langkah awal K-Medoids adalah mencari titik yang paling representatif (medoids) dalam sebuah dataset dengan menghitung jarak dalam kelompok dari semua kemungkinan kombinasi dari medoids sehingga jarak antar titik dalam suatu cluster kecil sedangkan jarak titik antar cluster besar.

Langkah-langkah K-Medoids adalah:

1. Pilih poin k sebagai inisial centroid / nilai tengah (medoids) sebanyak k cluster.
2. Cari semua poin yang paling dekat dengan medoids, dengan cara menghitung jarak vektor antar dokumen dengan menggunakan Euclidian Distance.

   Rumusnya adalah sebagai berikut

   $$d(x,y) = \sqrt{\sum_{i=1}^{n} |xi - yi|^2} \qquad (2.3)$$

   dimana
   d(x,y) = jarak antara data ke-i dan data ke-j
   xi1 = nilai atribut ke satu dari data ke-i
   yj1 = nilai atribut ke satu dari data ke-j
   n = jumlah atribut yang digunakan

3. Secara acak, pilih poin yang bukan medoids.
4. Hitung total jarak antar medoid.
5. Jika TD baru < TD awal, tukar posisi medoids dengan medoids baru, jadilah medoids yang baru.
6. Ulangi langkah 2-5 sampai medoids tidak berubah.

2.4 Rank

Rank value (bobot) dari kalimat yang akan digunakan untuk mendapatkan kalimat utama diperoleh berdasarkan hasil penjumlahan atau summing dari setiap nilai TF*IDF masing-masing kata yang menyusun kalimat tersebut. TF merupakan nilai yang menunjukkan frekuensi setiap kata (term) yang terdapat dalam dokumen. Nilai TF sebanding dengan frekuensi kemunculan kata. Semakin tinggi nilai TF maka semakin banyak kemunculan suatu kata (term) dalam suatu dokumen berita. [8]

Sementara itu nilai IDF diperoleh dari persamaan berikut

$$\text{IDF} = \log(\frac{N}{TF}) \qquad (2.4)$$

Dimana :
IDF = Inverse Document Frequency
TF = Term Frequency
N = Jumlah kalimat dalam dokumen.

Berdasarkan persamaan tersebut diperoleh nilai IDF berbanding terbalik dengan nilai TF. Artinya, semakin besar nilai TF maka nilai IDF yang dihasilkan semakin kecil untuk total jumlah kata (N) yang tetap. Akibatnya semakin kecil bobot suatu kalimat yang diperoleh dari nilai

TF*IDF menunjukkan semakin penting pula kalimat tersebut. Sehingga diperoleh kalimat utama merupakan kalimat dengan nilai bobot terkecil.

III. PERANCANGAN SISTEM

3.1. Gambaran Umum Sistem

Pada penelitian ini telah dilakukan proses pada sebuah sistem peringkasan dan *clustering* terhadap dokumen berbahasa inggris. Dokumen diambil secara manual dengan cara *copy-paste* dan di simpan ke dalam sebuah *file* dengan format .txt.

Tahapan sistem yang dibangun terdiri atas :

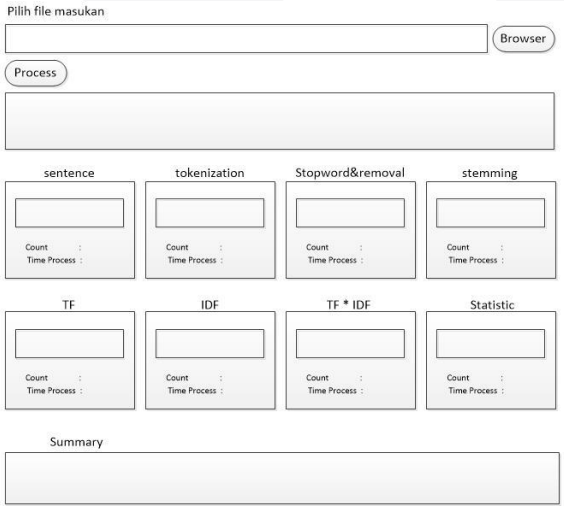
1. Sistem menerima inputan *dataset* berupa *file* .txt yang berisi berita berbahasa inggris. Kemudian dilakukan proses *preprocessing* yang terdiri atas *sentence, tokenization, stopwords removal, case folding, stemming*, TF-IDF.
2. Untuk menghitung jarak antar dokumen menggunakan metode *euclidean distance.*
3. Proses *clustering* yang digunakan adalah metode *k-medoids*, dimana mencari titik yang paling representatif *(medoids)* dalam sebuah *dataset* dengan menghitung jarak dalam kelompok dari semua kemungkinan kombinasi dari *medoids* sehingga jarak antar titik dalam suatu *cluster* kecil sedangkan jarak titik antar *cluster* besar.

.

3.2. Perancangan Sistem

Perancangan *interface* merupakan antarmuka yang digunakan pada proses *summarize* dokumen teks. Berikut adalah fitur-fitur yang terdapat pada perancangan *interface* :

1. *Browse* merupakan fitur yang berfungsi untuk mencari serta mengunggah *file* berupa dokumen teks dengan format *.txt*.
2. *Process* adalah fitur yang digunakan untuk memproses dokumen teks yang sudah kita unggah untuk ditampilkan isinya dan memprosesnya kembali agar menghasilkan *summary*.
3. *Sentence* adalah kolom untuk menampilkan hasil *sentence* dokumen.
4. *Tokenization* adalah kolom untuk menampilkan hasil *tokenisasi* dokumen.
5. *Stopword removal* adalah kolom untuk menampilkan hasil *stopword removal* dokumen.
6. *Stemming* adalah kolom untuk menampilkan hasil *stemming* dokumen.
7. *TF (Term Frequency)* adalah kolom untuk menampilkan hasil TF dokumen.
8. *IDF (Inverse Document Frequency)* adalah kolom untuk menampilkan hasil IDF dokumen.
9. *TF*IDF* adalah kolom untuk menampilkan hasil bobot TF*IDF dokumen.
10. *Statistic* adalah kolom untuk menampilkan hasil statistik dokumen.
11. *Summary* adalah kolom untuk menampilkan hasil ringkasan dokumen.



Gambar 3.2 Perancangan Sistem

IV. IMPLEMENTASI DAN PENGUJIAN

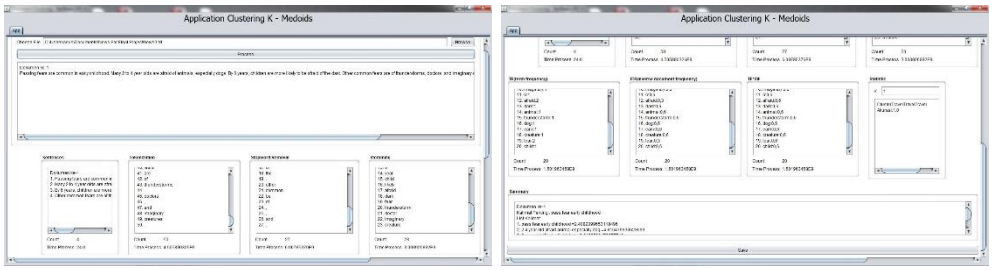
4.1 Implementasi Sistem

Tahap implementasi merupakan tahap dimana sistem yang telah dirancang, menjelaskan mengenai pembuatan sistem yang sesuai dengan analisis dan perancangan sebelomnya. Setelah tahap implementasi dilakukan maka dibutuhkan sebuah pengujian sistem untuk membuktikan bahwa apliasi dapat berjalan sesuai dengan yang diharapkan.

Implementasi sistem dapat dilihat pada contoh dokumen teks yang telah di akses dari website highered.mheducation.com berbunyi

Passing fears are common in early childhood. Many 2 to 4 year olds are afraid of animals, especially dogs. By 6 years, children are more likely to be afraid of the dark. Other common fears are of thunderstorms, doctors, and imaginary creatures.

Berdasarkan informasi dari website ini kalimat yang menjadi *main idea* dalam dokumen teks tersebut adalah *passing fears are common in early childhood*. Dan jika diproses dalam sistem ini akan menghasilkan data sebagai berikut:



Gambar 4.1 Implementasi Berita Pada Sistem

Dari gambar sebelumnya dapat dilihat pada pengujian pertama berita tersebut maka didapatkan data sebagai berikut:

Table 4.1.1 ***Hasil Dokumen Berita***

Berita Ke-2	**Counts**	**Time Process (s)**
Sentence	4	24
Tokenization	50	4.5
Stopwords Removal	27	6
Stemming	23	3
TF	20	1.5
IDF	20	1.5
TF*IDF	20	1.5

Sentence merupakan sebuah proses yang memisahkan antara kalimat dengan kalimat lainnya. Dalam dokumen teks ini didapatkan waktu proses selama 24 detik dengan kalimat sejumlah 4, hal ini sangat terbukti dengan dokumen teks ini berjumlah 4 kalimat sebagai rincian sebagai berikut:

1. *Passing fears are common in early childhood.*
2. *Many 2 to 4 year olds are afraid of animals, especially dogs.*
3. *By 6 years, children are more likely to be afraid of the dark.*
4. *Other common fears are of thunderstorms, doctors, and imaginary creatures.*

Tokenization merupakan sebuah proses yang memisahkan satu kata dengan kata lain. Dalam sistem ini, tanda baca titik, koma, dan lain-lainnya di hitung sebagai satu kata. Pada dokumen teks ini terdapat 50 kata yang diproses selama 4.5 detik, dan sangat sesuai dengan nilai counts yang terdapat dalam sistem.

Stopwords Removal adalah proses penghapusan kata yang tidak memiliki makna yang memiliki kamus tersendiri. Dalam sistem ini terdapat sejumlah 27 kata yang telah dihapus dalam waktu 6 detik.

Stemming adalah proses lanjutan dari stopwords removal. Dalam proses ini akan membentuk suatu kata menjadi kata dasar. Data yang tidak terhapus dalam stemming adalah 50 (tokenization) - 27 (stopwords removal) = 23 buah kata. Proses ini memakan waktu selama 3 detik.

TF dan IDF adalah sebuah proses pengukuran pembobotan kata untuk memperhitungkan seberapa penting sebuah token dalam kumpulan teks dalam dokumen.

Term Frequency (TF) pada kalimat *pass fear early childhood* yang telah mengalami proses sentence, tokenization, stopwords removal dan stemming dari kalimat utuh *passing fears are common in early childhood* adalah sebagai berikut:

Table 4.1.2 ***TF Hasil Dokumen Berita***

Term	Frequence (System)	Frequence (Manual)	Errors
Pass	1	1	0
Fear	2	2	0
Early	1	1	0
Childhood	1	1	0

Berdasarkan tabel diatas dapat dilihat bahwa errors antara hasil perhitungan TF dari program dengan perhitungan TF secara manual adalah 0. Artinya program telah berjalan dengan benar.

IDF (Inverse Document Frequency) kalimat pass fear early childhood yang telah mengalami proses sentence, tokenization, stopwords removal dan stemming dari kalimat utuh passing fears are common in early childhood adalah sebagai berikut:

Table 4.1.3 ***DF Hasil Dokumen Berita***

Term	TF	N	IDF (System)	IDF (Manual)	Errors
Pass	1	4	0.6	0.6	0
Fear	2	4	0.3	0.3	0
Early	1	4	0.6	0.6	0
Childhood	1	4	0.6	0.6	0

Berdasarkan tabel diatas dapat dilihat bahwa errors antara hasil perhitungan IDF dari program dengan perhitungan IDF secara manual adalah 0. Artinya program telah berjalan dengan benar.

Nilai IDF dapat diperoleh dengan persamaan berikut:

$$\text{IDF} = \log(\frac{N}{TF})$$

Sedangkan untuk penentuan bobot kalimat dapat dilihat hasil penjumlahan TF*IDF dari setiap kata yang ada pada kalimat.

Table 4.1.4 ***TF*IDF Hasil Dokumen Berita***

Term	TF	IDF	TF*IDF
Pass	1	0.6	0.6
Fear	2	0.3	0.6
Early	1	0.6	0.6
Childhood	1	0.6	0.6
TOTAL			2.4

Berdasarkan pada tabel dapat diperoleh bahwa hasil penjumlahan TF*IDF untuk kalimat pertama adalah 2.4 dan sesuai dengan rank value pada sistem yang berjumlah 2.4082399653118496.

Jika dilakukan 30 kali proses pengujian maka dapat data untuk time process sebagai berikut:

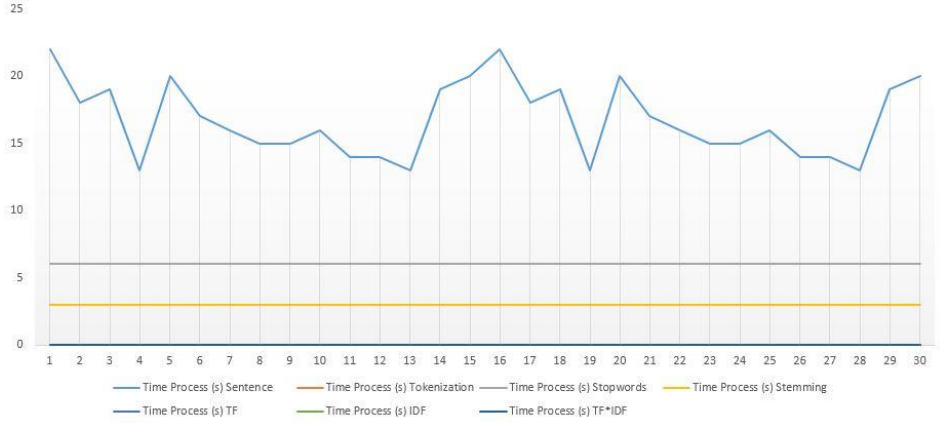
Table 4.1.4 ***Hasil Dokumen Berita dengan 30 Kali Pengujian***

	Time Process (s)						
NO	**Sentence**	**Tokenization**	**Stopwords**	**Stemming**	**TF**	**IDF**	**TF*IDF**
1	22	4.5	6	3	1.5	1.5	1.5
2	18	4.5	6	3	1.5	1.5	1.5
3	19	4.5	6	3	1.5	1.5	1.5
4	13	4.5	6	3	1.5	1.5	1.5
5	20	4.5	6	3	1.5	1.5	1.5
6	17	4.5	6	3	1.5	1.5	1.5
7	16	4.5	6	3	1.5	1.5	1.5
8	15	4.5	6	3	1.5	1.5	1.5
9	15	4.5	6	3	1.5	1.5	1.5
10	16	4.5	6	3	1.5	1.5	1.5
11	14	4.5	6	3	1.5	1.5	1.5
12	14	4.5	6	3	1.5	1.5	1.5
13	13	4.5	6	3	1.5	1.5	1.5
14	19	4.5	6	3	1.5	1.5	1.5
15	20	4.5	6	3	1.5	1.5	1.5
16	22	4.5	6	3	1.5	1.5	1.5
17	18	4.5	6	3	1.5	1.5	1.5
18	19	4.5	6	3	1.5	1.5	1.5
19	13	4.5	6	3	1.5	1.5	1.5
20	20	4.5	6	3	1.5	1.5	1.5
21	17	4.5	6	3	1.5	1.5	1.5
22	16	4.5	6	3	1.5	1.5	1.5
23	15	4.5	6	3	1.5	1.5	1.5
24	15	4.5	6	3	1.5	1.5	1.5
25	16	4.5	6	3	1.5	1.5	1.5
26	14	4.5	6	3	1.5	1.5	1.5
27	14	4.5	6	3	1.5	1.5	1.5
28	13	4.5	6	3	1.5	1.5	1.5
29	19	4.5	6	3	1.5	1.5	1.5

30	20	4.5	6	3	1.5	1.5	1.5

Maka dari itu dapat dilihat nilai rata-rata time process pada tabel sebagai berikut :
Nilai rata-rata time process sentence adalah 16.73 s
Nilai rata-rata time process tokenization adalah 4.5 s
Nilai rata-rata time process stopwords adalah 6 s
Nilai rata-rata time process stemming adalah 3 s
Nilai rata-rata time process TF adalah 1.5 s
Nilai rata-rata time process IDF adalah 1.5 s
Nilai rata-rata time process TF*IDF adalah 1.5 s
Dan data-data diatas dapat dilihat juga pada grafik dibawah ini

Dan data-data diatas dapat dilihat juga pada grafik dibawah ini:



Grafik 4.1 Grafik *Time Process* Berita

4.2 Pengujian Sistem

4.2.1 Pengujian Black Box

Untuk melakukan tes aplikasi sistem clustering dokumen menggunakan metode k-medoids, maka dilakukan menggunakan metode pengujian black box. Pengujian black box digunakan untuk melakukan pengujian fungsional terhadap sistem yang telah dibangun, apakah aplikasinya sudah sesuai dengan harapan atau belum sesuai. Berikut adalah hasil dari pengujian dengan menggunakan metode black box.

No	*Test Case*	Hasil Harapan	Hasil Keluaran	Kesimpulan
1	*Browse/* Pilih file dokumen teks yang hasil prosesnya ingin ditampilkan.	File dokumen teks yang dipilih akan terunggah pada field “choose file” untuk kemudian diproses hasil perhitungan *summarize.*	File dokumen teks yang dipilih berhasil terunggah pada field “choose file” untuk kemudian diproses hasil perhitungan summarize.	Tercapai
2	*Process* / proses file yang sudah di upload, dan choose.	Semua file dokumen teks akan terproses dan akan menghasilkan summarize. Selain itu juga akan muncul hasil dari proses perhitungan dan rnagkuman pada field “sentence, tokenization, stopword removals, stemming, tf, idf, tf*idf” dan summary sesuai dengan dokumen yang ingin diproses.	Semua file dokumen teks akan terproses dan akan menghasilkan summarize. Selain itu juga akan muncul hasil dari proses perhitungan dan rnagkuman pada field “sentence, tokenization, stopword removals, stemming, tf, idf, tf*idf” dan summary sesuai dengan dokumen yang ingin diproses.	Tercapai

V. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan dari hasil penelitian ini dapat disimpulkan bahwa

1. Sebagai validasi dari metode TF*IDF sudah dapat dijalankan pada percobaan ke-1 yang didapatkan nilai *rangking* sejumlah 2.4082399653118496 untuk kalimat ke-1 dan percobaan ke-2 yang didapatkan nilai *rangking* sejumlah 3.4614262661931448 untuk kalimat ke-1. Dan hasil sistem sama dengan hasil dari referensi.

2. Pada pengujian dengan berita yang telah di upload di www.cnn.com pada 27 Juni 2016 yang lalu di dapatkan kalimat utama dengan metode TF*IDF pada kalimat ke-17 yang mempunyai kata penting "*Handful Resorts Include Misool Eco Resort*" dengan rank value yang dihitung secara manual sejumlah 7.68 dan sesuai dengan *rank value* pada sistem yang berjumlah 7.672521605716973. Artinya perhitungan yang dilakukan oleh sistem sudah sama dengan perhitungan yang dilakukan secara manual oleh penulis.

3. Pada berita yang yang di uji pada sistem yang di plotkan sebagai 2 cluster yang terdiri dari cluster travel (1) dan cluster sport (0), maka berita yang berjudul dengan "*Scuba in Indonesia: Raja Ampat's coral reefs astound divers"* masuk ke cluster 1 sebagai cluster travel yang menunjukan kalau berita ini adalah berita berbahasa inggris yang membahas tentang traveling. Hal ini dapat dibuktikan dengan nilai akurasi sebesar 100%.

5.2 Saran

Untuk pengembangan sistem ini lebih lanjut penulis ingin menyampaikan beberapa saran, antara lain:

1. Pada penelitian selanjutnya penulis di harapkan dapat menggunakan algoritma selain TF*IDF. Hal ini disebabkan setelah melakukan penelitan ini ternyata penentuan bobot kalimat dari data berupa string (kata) dengan menggunakan teknik TF*IDF memberikan hasil yang tidak selalu akurat. Hal ini terjadi karena teknik ini menentukan bobot hanya berdasarkan presentase kemunculan suatu string dalam kalimat. Sehingga untuk pengembangan penelitan ini selanjutnya agar diperoleh hasil yang lebih akurat, bobot suatu kalimat tidak hanya ditentukan oleh presentase kemunculan string.
2. Pada penelitian selanjutnya penulis di harapkan lebih memperlihatkan nilai-nilai cluster hasil dari k-medoids sehingga didapatkan nilai hasil akurasi, recall, dan precision setelah proses clustering.

DAFTAR PUSTAKA

[1] Berry, M.W. & Kogan, J. 2010. **Text Mining Aplication and theory**. WILEY : United Kingdom.
[2] Feldman, R & Sanger, J. 2007. **The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data**. Cambridge University Press : New York.
[3] Prasetyo, Eko. 2014. **Data Mining : Mengolah Data Menjadi Informasi Menggunakan Matlab**. Yogyakarta : Penerbit Andi.
[4] Noor, M. Helmi, dan Moch. Hariadi. 2009. **Image Cluster Berdasarkan Warna Untuk Identifikasi Kematangan Buah Tomat Dengan Metode Valley Tracing**. Surabaya : ITS.
[5] Kusuma, P. A., Rini Nur Hasanah, dan Harry Soekotjo Dachlan. 2014. **DSS untuk Menganalisis pH Kesuburan Tanah Menggunakan Metode Single Linkage**. Jurnal EECCIS Vol. 8, No. 1.
[6] Handoyo, Rendy,. R. Rumani, dan M, Surya Michrandi Nasution. 2014. **Perbandingan Metode Clustering Menggunakan Metode Single Linkage Dan K - Means Pada Pengelompokan Dokumen**. Bandung : Universitas Telkom.
[7] Nugraha Adhiatma, Fachri. 2016. **Perancangan dan Analisis Clustering Data Menggunakan Metode Single Linkage untuk Berita Berbahasa Inggris**. Bandung : Universitas Telkom.
[8] Kurniawati. 2016**. Term Weighting Berbasis Indeks Kelas Menggunakan Metode TF.IDF.ICSF Untuk Perangkingan Dokumen Alqur'an**. Malang: Universitas Islam Negeri Maulana Malik Ibrahim.