**MAKALAH**

**INFORMATION RETRIEVAL**

**K-Means Clustering, Klasifikasi Dokumen Teks , dan**

**Klastering Twiter dan Undang-Undang**

Makalah ini disusun sebagai bukti hasil tugas akhir

****

Disusun Oleh :

**ISTIQOMAH 15.01.55.0023**

**PROGRAM STUDI SISTEM INFORMASI**

**FAKULTAS TEKNOLOGI INFORMASI**

**UNIVERSITAS STIKUBANK (UNISBANK) SEMARANG**

**DAFTAR ISI**

Kata Pengantar...................................................................................................................

Daftar Isi.............................................................................................................................

BAB I

PENDAHULUAN.................................................................................................................

1.1 Latar Belakang..............................................................................................................

1.2 Rumusan Masalah........................................................................................................

BAB II

PEMBAHASAN...................................................................................................................

2.1 Klasifikasi Dokumen Teks ..............................................................................................

2.2Klaster tweet dan undang-undang.............................................................................

2.3 Kmeans Clustering.......................................................................................................

BAB III

PENUTUP...........................................................................................................................

3.1 Kesimpulan...................................................................................................................

**Kata Pengantar**

Puji syukur selalu dipanjatkan atas kehadirat Tuhan Yang Maha Esa karena berkat rahmat dan hidayahnya makalah ini dapat dibuat. Makalah ini dibuat dengan tujuan untuk memenuhi tugas mata kuliah Information Retrieval yang diampu oleh Bapak Dr. Drs. Eri Zuliarso,M.Kom. Tidak lupa diucapkan lupa terima kasih kepada teman-teman dan keluarga yang selalu mendukung saya dalam menyelesaikan makalah.

Saya menyadari bahwa dalam proses pembuatan makalah ini dan hasil dari makalah terdapat banyak kekurangan dan kesalahan. Sehingga saya sangat membuka bagi siapa pun yang ingin memberikan kritik dan saran yang membangun bagi saya. Saya berharap dengan selesainya makalah dengan judul Peran Klasifikasi Dokumen Teks dan Berita Hoax dapat bermanfaat bagi seluruh pembaca, amin.

penulis

**BAB 1**

**PENDAHULUAN**

* 1. **Latar Belakang Masalah**

Clustering merupakan suatu metode untuk mencari dan mengelompokkan data yangmemiliki kemiripan karakteristik (similarity) antara satu data dengan data yang lain.Clustering merupakan salah satu metode data mining yang bersifat tanpa arahan(unsupervised).Yang dimaksud metode Unsupervised yaitu metode ini diterapkan tanpa adanya latihan (training) dan guru (teacher) serta tidak memerlukan target output. Dalam data mining ada dua jenis metode clustering yang digunakan dalam pengelompokan data, yaitu hierarchical clustering dan non-hierarchical clustering (Santosa, 2009).Teknik Clustering saat ini juga telah banyak digunakan untuk mengatasi permasalahan yang terkait dengan segementasi data. Implementasi clustering ini dapat diterapkan pada berbagai bidang sebagai contoh dalam hal text mining.Teknik clustering dapat digunakan sebagai metode dalam mengelompokkandokumen teksyang memiliki kesamaan konten/isi dan tema dari teks tersebut.Tujuan utama dari Metode cluster adalah pengelompokan sejumlah data atau obyek ke dalam cluster(kelompok) sehingga dalam setiap Cluster akan berisi data yang memiliki kesamaan karakteristik dari data tersebut.

Berkembangnya teknologi telah membuat banyak informasi bermunculan.Informasi-informasi tersebut tertuang dalam bentuk dokumen terutama dokumen digital. Semakin banyak informasi yang ada maka semakin banyak dokumen-dokumen yang digunakan. Untuk bisa mengorganisir informasi-informasi tersebut dengan mudah, maka dibutuhkan klasifikasi dokumen secara otomatis. Dalam penelitian yang akan dilakukan sebagai tugas akhir ini, penulis akan membuat sistem klasifikasi dokumen teks dan berita hoax secara otomatis dengan Bahasa R. Melalui program ini, dokumen dan berita yang ada akan secara otomatis terklasifikasi sehingga mempermudah pengguna dalam melakukan pencarian dokumen dan berita sesuai dengan kelas yang sudah ditentukan untuk klasifikasi artikel berita, sehingga dapat diketahui tingkat keakuratan metode yang digunakan untuk klasifikasi.

**1.2 Rumusan Masalah**

Adapun rumusan masalah dalam penelitian ini adalah

1. Bagaimana menentukan kategori sesuai topic dari sebuah dokumen.
2. Bagaimana membuat program dengan bahasa R untuk mengklaster tweet dan undang-undang
3. Bagaimana membuat klaster menggunakan algoritma k-means.

**BAB 1I**

**PEMBAHASAN**

**2.1 Klasifikasi Dokumen Teks**

Klasifikasi dokumen teks adalah masalah sederhananamun sangat penting karena manfaatnya cukup besar mengingat jumlah dokumen yang ada setiap hari semakin bertambah.Manfaat dari klasifikasi dokumen adalah untuk pengorganisasian dokumen. Dengan jumlah dokumen yang sangat besar, untuk mencari sebuah dokumen akan lebih mudah apabila kumpulan dokumen yang dimiliki terorganisir dan telah dikelompokkan sesuai kategorinya masing-masing. Contoh aplikasi penggunaan klasifikasi dokumen teks yang banyak digunakan adalah e-mail spam filtering. Pada aplikasi spam filtering sebuah e-mail diklasifikasikan apakah e-mail tersebut termasuk spam atau tidak dengan memperhatikan kata-kata yang terdapat di dalam e-mail tersebut. Aplikasi ini telah digunakan oleh banyak e-mail provider. Sebuah dokumen dapat dikelompokkan ke dalam kategori tertentu berdasarkan kata-kata dan kalimat-kalimat yang ada di dalam dokumen tersebut. Kata atau kalimat yang terdapat di dalam sebuah dokumen memiliki makna tertentu dan dapat digunakan sebagai dasar untuk menentukan kategori sesuai topic dari dokumen tersebut. Perhatikan beberapa kalimat berikut ini:

1.Pemilih yang namanya tercantum dalam daftar pemilih tetap dan daftar pemilih tambahan,apabila sampai dengan 3 (tiga) hari sebelum hari dan tanggal pemungutan suara belum menerima surat pemberitahuan untuk memberikan suara di TPS (Model C4), diberi kesempatan untuk meminta Pemanfaatan dokumen..., Bayu Distiawan Trisedya, FASILKOM UI, 20097 Universitas Indonesiakepada Ketua KPPS selambat-lambatnya 24 (dua puluh empat) jam sebelum hari dan tanggal pemungutan suara dengan menunjukkan kartu tanda penduduk atau identitas lain yang sah.[PERATURAN KOMISI PEMILIHAN UMUM NOMOR 03 TAHUN 2009]

2.Pemerintah Pusat dan bank sentral berkoordinasi dalam penetapan dan pelaksanaan kebijakan fiskal dan moneter. [UNDANG-UNDANG REPUBLIK INDONESIA NOMOR 17 TAHUN 2003]

3.Pada universitas, institut, dan sekolah tinggi dapat diangkat guru besar atau professor sesuai dengan peraturan perundang-undangan yang berlaku.[UNDANG-UNDANG REPUBLIK INDONESIA NOMOR 20 TAHUN 2003]

Pada kalimat

(1)terdapat kata pemilih, TPS, dan KPPS. Kata-kata tersebut memiliki keterkaitan erat dengan masalah pemilu, sehingga dapat disimpulkan bahwa kalimat (1) membahas masalah pemilu. Kalimat (2) memiliki kata bank, fiskal, dan moneter. Dari kata-kata tersebut akan muncul dugaan bahwa kalimat (2) sedang membahas masalah keuangan. Terakhir, pada kalimat (3) terdapat kata sekolah, guru, dan universitas yang menunjukkan bahwa kalimat tersebut membahas bidang pendidikan.Kata bank yang terdapat pada dokumen lain belum dapat dijadikan sebagai acuan bahwa dokumen lain tersebut membahas mengenai keuangan. Apabila dokumen lain tersebut memiliki kata-kata lain yang mengarahkan kepada pembahasan keuangan secara bersamaan, maka dapat disimpulkan bahwa dokumen tersebut membahas mengenai keuangan.Untuk dapat menentukan kategori dari sebuah dokumen haruslah dilihat semua kata-kata yang terkait pada dokumen tersebut.

* 1. **`Klaster tweet dan undang-undang**
* **Tweet**

|  |
| --- |
| install.packages("twitteR") |
|  |

|  |
| --- |
| install.packages("ROAuth") |
|  |

|  |
| --- |
| install.packages("tm") |
|  |

|  |
| --- |
| install.packages("ggplot2") |
|  |

|  |
| --- |
| install.packages("wordcloud") |
|  |

|  |
| --- |
| install.packages("plyr") |
|  |

|  |
| --- |
| install.packages("RTextTools") |
|  |

|  |
| --- |
| install.packages("devtools") |
|  |

|  |
| --- |
| install.packages("e1071") |
|  |

|  |
| --- |
| install.packages("fpc") |
|  |

|  |
| --- |
| install.packages("cluster") |
|  |

|  |
| --- |
| install.packages("datasets") |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| require(devtools) |
|  |

|  |
| --- |
| library(e1071) |
|  |

|  |
| --- |
| library(twitteR) |
|  |

|  |
| --- |
| library(ROAuth) |
|  |

|  |
| --- |
| library(tm) |
|  |

|  |
| --- |
| library(ggplot2) |
|  |

|  |
| --- |
| library(wordcloud) |
|  |

|  |
| --- |
| library(plyr) |
|  |

|  |
| --- |
| library(RTextTools) |
|  |

|  |
| --- |
| library(fpc) |
|  |

|  |
| --- |
| library(cluster) |
|  |

|  |
| --- |
| library(datasets) |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| setup\_twitter\_oauth("rrpBd5ibxl1Q61Dr4FijeEhyN", "Ke3yBRRddpx2T51Dwt5oksmU59n5xd1vPlAyEUntYfkZmmLOTZ", "949124526-e9fj06fzQfC7tnML8NNUBfRJyOuf7scSpA8MfDar","PAZim0auITRhgamUJkWBLCro5WPd5rT42mkCiE1IKHYCY") |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| #WORDCLOUD |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| tweets <- userTimeline("presiden", n = 250) |
|  |

|  |
| --- |
| show(tweets) |
|  |

|  |
| --- |
| n.tweet <- length(tweets) |
|  |

|  |
| --- |
| # convert tweets to a data frame |
|  |

|  |
| --- |
| tweets.df <- twListToDF(tweets) |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| myCorpus <- Corpus(VectorSource(tweets.df$text)) |
|  |

|  |
| --- |
| # convert to lower case |
|  |

|  |
| --- |
| myCorpus <- tm\_map(myCorpus, content\_transformer(tolower)) |
|  |

|  |
| --- |
| # remove URLs |
|  |

|  |
| --- |
| removeURL <- function(x) gsub("http[^[:space:]]\*", "", x) |
|  |

|  |
| --- |
| myCorpus <- tm\_map(myCorpus, content\_transformer(removeURL)) |
|  |

|  |
| --- |
| # remove anything other than English letters or space |
|  |

|  |
| --- |
| removeNumPunct <- function(x) gsub("[^[:alpha:][:space:]]\*", "", x) |
|  |

|  |
| --- |
| myCorpus <- tm\_map(myCorpus, content\_transformer(removeNumPunct)) |
|  |

|  |
| --- |
| # remove stopwords |
|  |

|  |
| --- |
| myStopwords <- c(setdiff(stopwords('english'), c("r", "big")),"use", "see", "used", "via", "amp") |
|  |

|  |
| --- |
| myCorpus <- tm\_map(myCorpus, removeWords, myStopwords) |
|  |

|  |
| --- |
| # remove extra whitespace |
|  |

|  |
| --- |
| myCorpus <- tm\_map(myCorpus, stripWhitespace) |
|  |

|  |
| --- |
| # keep a copy for stem completion later |
|  |

|  |
| --- |
| myCorpusCopy <- myCorpus |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| tdm <- TermDocumentMatrix(myCorpus) |
|  |

|  |
| --- |
| term.freq <- rowSums(as.matrix(tdm)) |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| term.freq <- subset(term.freq, term.freq >= 20) |
|  |

|  |
| --- |
| df <- data.frame(term = names(term.freq), freq = term.freq) |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| ggplot(df, aes(x=term, y=freq)) + geom\_bar(stat="identity") + |
|  |

|  |
| --- |
| xlab("Terms") + ylab("Count") + coord\_flip() + |
|  |

|  |
| --- |
| theme(axis.text=element\_text(size=7)) |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| m <- as.matrix(tdm) |
|  |

|  |
| --- |
| # calculate the frequency of words and sort it by frequency |
|  |

|  |
| --- |
| word.freq <- sort(rowSums(m), decreasing = T) |
|  |

|  |
| --- |
| # colors |
|  |

|  |
| --- |
| pal <- brewer.pal(9, "BuGn")[-(1:4)] |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| # plot word cloud |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| wordcloud(words = names(word.freq), freq = word.freq, min.freq = 3, |
|  |

|  |
| --- |
| random.order = F, colors = pal) |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| #k-means clustering |
|  |

|  |
| --- |
| d <- dist(term.freq, method="euclidian") |
|  |

|  |
| --- |
| carsCluster <- kmeans(term.freq, 3) |
|  |

|  |
| --- |
| clusplot(as.matrix(d), carsCluster$cluster, color=T, shade=T, labels=3, lines=0) **Hasil Kmeans :** |
| * **Undang-Undang**  |  | | --- | | install.packages("RTextTools") | |  |  |  | | --- | | install.packages("e1071") | |  |  |  | | --- | | install.packages("wordcloud") | |  |  |  | | --- | | install.packages("fpc") | |  |  |  | | --- | | install.packages("cluster") | |  |  |  | | --- | | library(tm) | |  |  |  | | --- | | library(ggplot2) | |  |  |  | | --- | | library(wordcloud) | |  |  |  | | --- | | library(plyr) | |  |  |  | | --- | | library(RTextTools) | |  |  |  | | --- | | library(e1071) | |  |  |  | | --- | | library(cluster) | |  |  |  | | --- | | library(fpc) | |  |  |  | | --- | |  | |  |  |  | | --- | |  | |  |  |  | | --- | |  | |  |  |  | | --- | | #referensi https://rstudio-pubs-static.s3.amazonaws.com/31867\_8236987cf0a8444e962ccd2aec46d9c3.html | |  |  |  | | --- | | cname <- file.path("/resources/data/pakeri/uu") | |  |  |  | | --- | | cname | |  |  |  | | --- | | dir(cname) | |  |  |  | | --- | | docs <- Corpus(DirSource(cname)) | |  |  |  | | --- | |  | |  |  |  | | --- | | summary(docs) | |  |  |  | | --- | | inspect(docs[2]) | |  |  |  | | --- | | docs <- tm\_map(docs, removePunctuation) | |  |  |  | | --- | | docs <- tm\_map(docs, removeNumbers) | |  |  |  | | --- | | docs <- tm\_map(docs, tolower) | |  |  |  | | --- | | docs <- tm\_map(docs, removeWords, c("dapat", "yang","adalah","untuk","dan")) | |  |  |  | | --- | | docs <- tm\_map(docs, PlainTextDocument) | |  |  |  | | --- | | dtm <- DocumentTermMatrix(docs) | |  |  |  | | --- | | dtm | |  |  |  | | --- | | tdm <- TermDocumentMatrix(docs) | |  |  |  | | --- | | tdm | |  |  |  | | --- | |  | |  |  |  | | --- | | freq <- colSums(as.matrix(dtm)) | |  |  |  | | --- | | length(freq) | |  |  |  | | --- | |  | |  |  |  | | --- | | m <- as.matrix(dtm) | |  |  |  | | --- | | dim(m) | |  |  |  | | --- | | write.csv(m, file="dtm.csv") | |  |  |  | | --- | | dtms <- removeSparseTerms(dtm, 0.1) # This makes a matrix that is 10% empty space, maximum. | |  |  |  | | --- | | inspect(dtms) | |  |  |  | | --- | | ord <- order(freq) | |  |  |  | | --- | | freq[head(ord)] | |  |  |  | | --- | | findFreqTerms(dtm, lowfreq=10) | |  |  |  | | --- | |  | |  |  |  | | --- | |  | |  |  |  | | --- | |  | |  |  |  | | --- | | ##dtmss <- removeSparseTerms(dtm, 0.00001) # This makes a matrix that is only 15% empty space, maximum. | |  |  |  | | --- | | ##inspect(dtmss) | |  |  |  | | --- | |  | |  |  |  | | --- | | d <- dist(t(dtms), method="euclidian") | |  |  |  | | --- | | fit <- hclust(d=d, method="ward.D2") | |  |  |  | | --- | | fit | |  |  |  | | --- | | plot(fit, hang=-1) | |  |  |  | | --- | |  | |  |  |  | | --- | |  | |  |  |  | | --- | | d <- dist(t(dtm), method="euclidian") | |  |  |  | | --- | | kfit <- kmeans(d, 5) | |  |  |  | | --- | | clusplot(as.matrix(d), kfit$cluster, color=T, shade=T, labels=2, lines=0) | |  |  |  | | --- | |  | |  |  |  | | --- | | pamCluster <- pam(d, 3) | |  |  |  | | --- | | clusplot(as.matrix(d), pamCluster$cluster, color=T, shade=T, labels=3, lines=0) | |  |  |  | | --- | | si <- silhouette(pamCluster) | |  |   plot(si) # silhouette plot  **Hasil Klastering Abstrak :** |
|  |

* 1. **K-Means Clustering**

Metode **K-Means Clustering** berusaha mengelompokkan data yang ada ke dalam beberapa kelompok, dimana data dalam satu kelompok mempunyai karakteristik yang sama satu sama lainnya dan mempunyai karakteristik yang berbeda dengan data yang ada di dalam kelompok yang lain.

Data clustering menggunakan metode **K-Means Clustering** ini secara umum dilakukan dengan algoritma dasar sebagai berikut:

1. Menenukan jumlah cluster
2. Alokasikan data ke dalam cluster secara random
3. Menghitungitung rata-rata dari data yang ada di masing-masing cluster
4. Alokasikan masing-masing data ke centroid/rata-rata terdekat
5. Kembali ke Step 3, apabila masih ada data yang berpindah cluster atau apabila perubahan nilai centroid, ada yang di atas nilai threshold yang ditentukan atau apabila perubahan nilai pada objective function yang digunakan di atas nilai threshold yang ditentukan

**Tabel Data**

|  |  |  |
| --- | --- | --- |
| Nodok | Pilkada | Asian games |
| 1 | 2 | 1 |
| 2 | 4 | 3 |
| 3 | 3 | 4 |
| 4 | 1 | 5 |
| 5 | 5 | 2 |

|  |  |  |
| --- | --- | --- |
| n | a | b |
| 1 | 2 | 1 |
| 2 | 4 | 3 |
| 3 | 3 | 4 |
| 4 | 1 | 5 |
| 5 | 5 | 2 |

**C1 = {4,3}**

**C2 = {1,5}**

**Persamaan Euclidian Distance**

**PENGULANGAN KE-1**

Jarak Data dengan Centroid C1 :

Jarak Data dengan Centroid C2 :

Hasil Perhitungan Pengulanan Ke-1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| n | a | b | dc1 | dc2 |
| 1 | 2 | 1 | 2.83 | 4.12 |
| 2 | 4 | 3 | 0 | 3.61 |
| 3 | 3 | 4 | 1.41 | 2.24 |
| 4 | 1 | 5 | 3.61 | 0 |
| 5 | 5 | 2 | 1.41 | 5 |

Pengelompokkan Data pada Pengulangan Ke-1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| n | a | b | dc1 | dc2 | c1 | c2 |
| 1 | 2 | 1 | 2.83 | 4.12 | Ok |  |
| 2 | 4 | 3 | 0 | 3.61 | Ok |  |
| 3 | 3 | 4 | 1.41 | 2.24 | Ok |  |
| 4 | 1 | 5 | 3.61 | 0 |  | Ok |
| 5 | 5 | 2 | 1.41 | 5 | Ok |  |

Nilai Rata-Rata Centroid pada Pengulanan Ke-1

|  |  |  |
| --- | --- | --- |
|  | a | b |
| c1 | 3.5 | 2.5 |
| c2 | 1 | 5 |

**PENGULANGAN KE-2**

Jarak Data dengan Centroid C1 :

.71

Jarak Data dengan Centroid C2 :

Hasil Perhitungan Pengulanan Ke-2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| n | a | b | dc1 | dc2 |
| 1 | 2 | 1 | 2.83 | 4.12 |
| 2 | 4 | 3 | 0 | 3.61 |
| 3 | 3 | 4 | 1.41 | 2.24 |
| 4 | 1 | 5 | 3.61 | 0 |
| 5 | 5 | 2 | 1.41 | 5 |

Pengelompokkan Data pada Pengulangan Ke-2

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| n | a | b | dc1 | dc2 | c1 | c2 |
| 1 | 2 | 1 | 2.12 | 4.12 | Ok |  |
| 2 | 4 | 3 | 0.71 | 3.61 | Ok |  |
| 3 | 3 | 4 | 1.58 | 2.24 | ok |  |
| 4 | 1 | 5 | 3.54 | 0 |  | Ok |
| 5 | 5 | 2 | 1.58 | 5 | ok |  |

Nilai Rata-Rata Centroid pada Pengulanan Ke-2

|  |  |  |
| --- | --- | --- |
|  | a | b |
| c1 | 3.5 | 2.5 |
| c2 | 1 | 5 |

1. **Hierarchical Clustering**

**Tabel Data**

|  |  |  |
| --- | --- | --- |
| Nodok | Pilkada | Asian games |
| 1 | 2 | 1 |
| 2 | 4 | 3 |
| 3 | 3 | 4 |
| 4 | 1 | 5 |
| 5 | 5 | 2 |

**Persamaan Euclidian Distance**

D*man* (*Data1*, *Data1*) =

D*man* (*Data1*, *Data2*) =

D*man* (*Data1*, *Data3*) =

D*man* (*Data1*, *Data4*) =

D*man* (*Data1*, *Data5*) =

D*man* (*Data2*, *Data3*) =

D*man* (*Data2*, *Data4*) =

D*man* (*Data2*, *Data5*) =

D*man* (*Data3*, *Data4*) =

D*man* (*Data3*, *Data5*) =

D*man* (*Data4*, *Data5*) =

**Tabel Dman**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dman | 1 | 2 | 3 | 4 | 5 |
| 1 | 0 | 2.8 | 3.1 | 4.1 | 3.1 |
| 2 | 2.8 | 0 | 1.4 | 3.6 | 1.4 |
| 3 | 3.1 | 1.4 | 0 | 2.8 | 2.8 |
| 4 | 4.1 | 3.6 | 2.8 | 0 | 5 |
| 5 | 3.1 | 1.4 | 2.8 | 5 | 0 |

**Persamaan Average Linkage (Jarak Rata-Rata)**

min(D*man*) = min(d23) = 1.4

d(23)1 = average {d21 , d31} = average {2.8 , 3.1} = (2.8+3.1) / 2 = 2.9

d(23)4 = average {d24 , d34} = average {3.6 , 2.8} = (3.6+2.8) / 2 = 3.2

d(23)5 = average {d25 , d35} = average {1.4 , 2.8} = (1.4+2.8) / 2 = 2.1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dman | (23) | 1 | 4 | 5 |
| (23) | 0 | 2.9 | 3.2 | 2.1 |
| 1 | 2.9 | 0 | 4.1 | 3.1 |
| 4 | 3.2 | 4.1 | 0 | 5 |
| 5 | 2.1 | 3.1 | 5 | 0 |

min(D*man*) = min(d(23)5) = 2.1

d(235)1 = average {d21 , d31 , d51} = average {2.9 , 2.9 , 3.1} = (2.9+2.9+3.1) / 3 = 2.9

d(235)4 = average {d24 , d34 , d54} = average {3.2 , 3.2 , 5} = (3.2+3.2+5) / 3 = 3.8

|  |  |  |  |
| --- | --- | --- | --- |
| Dman | (235) | 1 | 4 |
| (235) | 0 | 2.9 | 3.8 |
| 1 | 2.9 | 0 | 4.1 |
| 4 | 3.8 | 4.1 | 0 |

min(D*man*) = min(d(235)1) = 2.9

d(2351)4 = average {d24 , d34 , d54 , d14} =

average {3.8 , 3.8 , 3.8 , 4.1} = (3.8+3.8+3.8+4.1) / 4 = 3.8

|  |  |  |
| --- | --- | --- |
| **Dman** | **(2351)** | **4** |
| **(2351)** | **0** | **3.8** |
| **4** | **3.8** | **0** |
|  |  |  |

**Dendogram**

**J**

**A**

**R  
A  
K**

4

3

2

1

4

1

5

3

2

**BAB 1II**

**PENUTUP**

* 1. **Kesimpulan**

- Analisis cluster dilakukan untuk mengelompokan objek-objek

yang memiliki kemiripan (homogen). Berdasarkan karakteristik yang dimiliki,

dengan analisis cluster sekelompok objek dapat dikelompokkan. Terdapat

beberapa algoritma cluster yang dapat digunakan untuk mengelompokkan objek-objek.

* klasifikasi topik juga membandingkan hasil klasifikasi topik dari segi keseragaman data yaitu keseragaman jumlah data training untuk setiap kategori dan aspek kemiripan data yaitu dengan menggunakan data yang memiliki tingkat kemiripan rendah dari artikel media massa dan data yang memiliki tingkat kemiripan tinggi dari abstrak tulisan ilmiah dalam bidang computer science.