BAB IV IMPLEMENTASI

Pada bab ini akan dijelaskan implementasi dari perancangan yang telah dibuat. Penjelasan implementasi berupa implementasi persiapan data (subbab 4.1), penentuan fitur (subbab 4.2), pembuatan *term-document matrix* sebagai input bagi teknik pengelompokan dokumen (subbab 4.3), penerapan teknik pengelompokan dokumen (subbab 4.4) serta evaluasi kinerja teknik pengelompokan dokumen (subbab 4.5). Implementasi persiapan data dilakukan dengan menggunakan Java. Implementasi penentuan fitur dan pembuatan *term-document matrix* dilakukan dengan menggunakan PERL. Sedangkan implementasi teknik pengelompokan dokumen dan evaluasi kinerja dilakukan dengan menggunakan Matlab.

4.1. Persiapan Data

Persiapan data yang dilakukan berupa pengambilan data yang ada pada sumber yang telah ditentukan. Setelah data diambil maka dilakukan pengecekan terhadap data yang diambil secara otomatis, apakah sudah layak digunakan dalam percobaan atau tidak.

Data yang digunakan dalam percobaan adalah artikel media massa berbahasa Indonesia yang diambil dari *website* Kompas dan Antara. Artikel dari media massa ini dibaca dan diambil secara otomatis oleh *crawler*. *Crawler* bekerja pada *page source* pada halaman *web* yang bersangkutan. Implementasi dari *crawler* ditunjukkan oleh *pseudocode* pada Gambar 4.1.

Pengambilan data artikel secara otomatis dilakukan dengan menggunakan *RSS* yang ada pada *website* Kompas dan *Index* berita dengan sistem *pages* yang ada pada *website* Antara. Karena pengambilan data artikel pada Kompas menggunakan *RSS* yang hanya menampilkan artikel dalam satu halaman *web* dengan jumlah artikel yang sangat terbatas maka pengambilan data dilakukan setiap hari sampai jumlah data artikel yang diinginkan tercapai. Oleh karena itu, perlu ada penamaan berkas artikel yang bisa mengidentifikasikan jika ada artikel yang sama dibaca pada hari yang berbeda. Artikel-artikel seperti ini hanya akan disimpan satu kali. Penamaan yang dipakai adalah kombinasi dari sumber artikel, tanggal artikel dipublikasikan,

serta judul artikel. Contoh nama berkas adalah kompas-2009-02-10-0733271-beragam-respons-wall-street-nantikan-kebijakan-washington.txt.

Setelah dibaca, artikel disimpan pada berkas teks pada direktori yang sesuai dengan kategori dari artikel yang bersangkutan. Pengelompokan artikel ke dalam kategori sudah dilakukan oleh Kompas dan Antara. Jadi yang perlu dilakukan hanyalah membaca artikel per kategori dan menyimpannya dalam kategori yang tepat. Kategori artikel yang diambil dari Kompas terdiri dari 8 kategori yaitu bisnis keuangan, olahraga, kesehatan, perempuan, sains, travel, properti, dan politik hukum. Untuk artikel dari Antara, artikel yang dibaca hanyalah artikel dari kategori olahraga. Namun, untuk keperluan eksperimen, artikel olahraga Kompas dikelompokkan lagi ke dalam kategori yang lebih spesifik. Kategori spesifik yang digunakan adalah balap, bola, bulutangkis, tenis, dan tinju. Pengelompokan artikel ke dalam kategori-kategori spesifik ini dilakukan secara manual.

```
Function main
    foreach category
        setPath(storing_directory);
        foreach page
            getAllArticleLink(page_link);
        end
    end

Function getAllArticleLink(link)
    setConnection(link);
    if connect
        article_links ← match(get_links_regex);
        foreach link in article_links
            getArticleContent(link);
        end
    else give notification
    end
```

```
Function getArticleContent(link)
        setConnection(link);
        if connect
                article_date ← match(get_date_regex);
                article_title ← match(get_title_regex);
                article_content ← match (get_title_regex);
                print file ← article_date, article_title, article_content;
        else give notification
        end
```

Gambar 4.1. *Pseudocode* Pengambilan Data Artikel dari *Website*

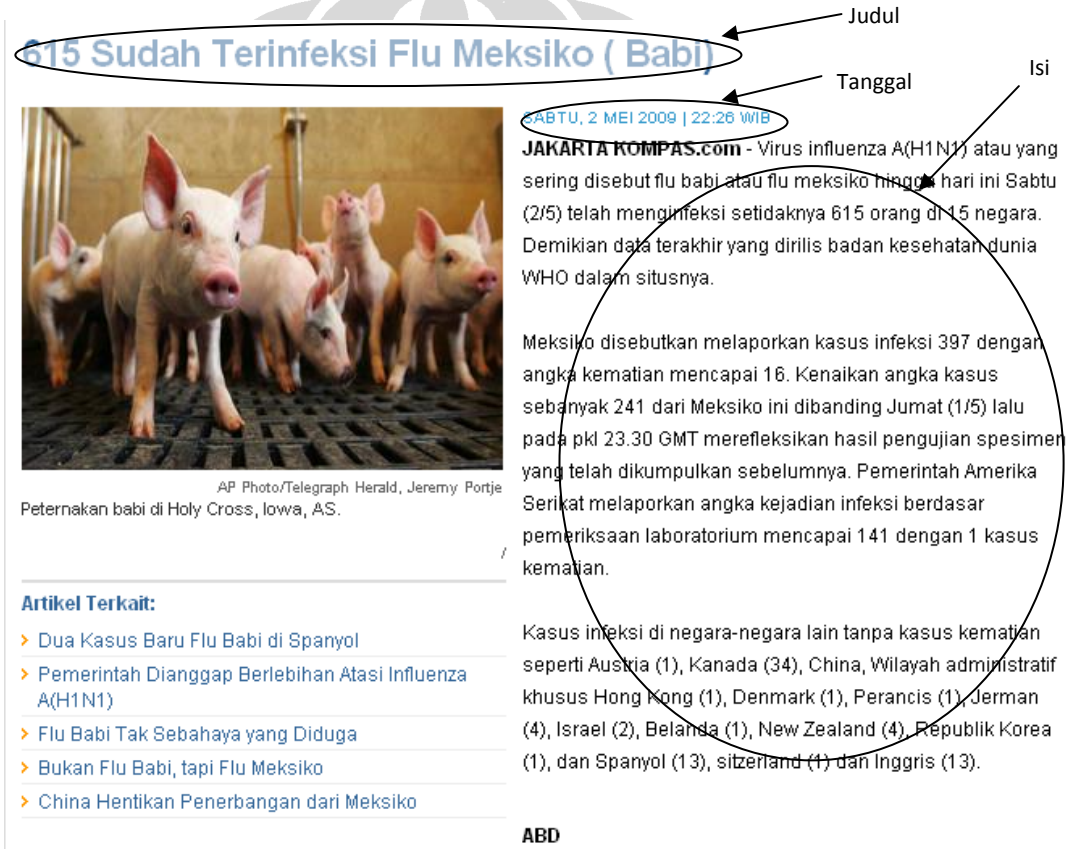
Hal yang pertama kali dilakukan oleh *crawler* adalah menentukan direktori dimana artikel yang dibaca akan disimpan. Direktori ini menunjukkan kategori dari artikel yang bersangkutan. Kemudian, semua *link* artikel untuk setiap halaman yang memuat artikel untuk kategori tertentu akan diambil. Dalam hal ini, *link* untuk setiap halaman tersebut diberikan pada *crawler*. Contoh *link* tersebut adalah http://www.kompas.com/getrss/kesehatan. *Link* ini kemudian akan merujuk pada halaman yang berisi kumpulan *link* untuk artikel yang termasuk dalam kategori kesehatan seperti yang ditunjukkan Gambar 4.2.

Gambar 4.2. Tampilan *Website* Kompas Kategori Kesehatan

Dalam pembacaan *link* untuk semua artikel pada halaman seperti yang ditunjukkan Gambar 4.2 digunakan *regular expression*. *Regular expression* yang digunakan adalah

<guid isPermaLink="false">http://www.kompas.com/read/xml/(.*)</guid>

Crawler kemudian akan menuju ke halaman yang dirujuk oleh setiap link yang didapat pada tahap sebelumnya. Gambar 4.3 menunjukkan contoh halaman yang dirujuk oleh salah satu *link* pada halaman ditunjukkan Gambar 4.2.



Gambar 4.3. Tampilan Salah Satu Artikel Kompas

Pada halaman tersebut, informasi tanggal publikasi, judul, dan isi artikel diambil dengan menggunakan *regular expression*. *Regular expression* yang digunakan untuk mengambil informasi tanggal publikasi, judul, dan isi artikel masing-masing adalah sebagai berikut:

Judul : (.*?)

Isi : (.*?)
Tanggal: (.*?)<!--end artikel -->

Kemudian informasi ini akan ditulis ke dalam satu berkas teks dengan aturan penamaan yang telah ditentukan dan disimpan pada direktori yang telah ditentukan diawal. Setelah semua artikel dibaca dan disimpan pada direktori, hal selanjutnya yang perlu dilakukan adalah mengecek apakah isi artikel sudah benar atau belum. Pengecekan dilakukan dengan melihat apakah isi berkas tidak kosong dan tidak hanya terdiri dari tanggal atau judul saja. Hal ini bisa terjadi karena pembacaan data dilakukan dengan *regular expression* dan ada kemungkinan dimana beberapa artikel memiliki pola yang berbeda dalam menandai bagian tanggal, judul, atau isi artikelnya. Apabila berkas artikel seperti ini dipakai, maka dapat mempengaruhi hasil percobaan.

Rincian jumlah data artikel yang didapat dari Kompas dan Antara yang akan dipakai sebagai data percobaan adalah sebagai berikut:

- Artikel Kompas
    1. Jumlah data artikel kategori bisnis keuangan = 478
    2. Jumlah data artikel kategori olahraga = 296 terdiri dari :
        a. Artikel balap = 47
        b. Artikel bola = 86
        c. Artikel bulutangkis = 68
        d. Artikel tenis = 61
        e. Artikel tinju = 34
    3. Jumlah data artikel kategori kesehatan = 341
    4. Jumlah data artikel kategori perempuan = 312
    5. Jumlah data artikel kategori sains = 139
    6. Jumlah data artikel kategori travel = 144
    7. Jumlah data artikel kategori properti = 148
    8. Jumlah data artikel kategori politik hukum = 113
- Artikel Antara
    1. Jumlah data artikel kategori olahraga = 302

4.2. Implementasi Penentuan Fitur

Fitur yang digunakan pada percobaan adalah fitur *unigram* yang merupakan fitur yang terdiri dari satu *token*. Pada percobaan, variasi dari fitur yang digunakan berupa persentase fitur yang digunakan dari semua fitur yang dikumpulkan dari koleksi dokumen. Hal ini dilakukan karena penggunaan persentase lebih bersifat universal yaitu tidak bergantung pada jumlah dokumen yang digunakan dalam percobaan. Persentase fitur ini bisa digunakan untuk jumlah data yang besar maupun kecil.

Hal yang perlu dilakukan pertama kali adalah mengambil semua fitur yang ada pada koleksi dokumen yaitu setiap *token* unik yang muncul dengan total frekuensi lebih besar atau sama dengan 6 atau minimal 6 pada semua dokumen. Pemilihan nilai ini didasarkan atas informasi yang diperoleh dari percobaan informal sebelumnya. Dari percobaan tersebut, terlihat bahwa pemakaian nilai dibawah 6 menghasilkan fitur-fitur yang hanya muncul di sebagian kecil dokumen dalam koleksi (bisa disebut sebagai fitur yang tidak penting) sehingga penggunaannya tidak akan memberikan efek yang signifikan terhadap penentuan kluster dari dokumen yang ada. Dari kumpulan fitur, kemudian diambil fitur dengan nilai frekuensi tertinggi sejumlah persentase yang ditentukan dalam percobaan. Misalkan variasi fitur yang digunakan adalah persentase 60% dan total fitur yang dikumpulkan dari koleksi dokumen adalah 6000, maka fitur yang akan digunakan adalah 3600 fitur dengan nilai frekuensi tertinggi. *Pseudocode* penentuan fitur ditunjukkan Gambar 4.4.

```
Function featureSelection(all_documents, percentage_to_use) return feature_list
    totalFeature=0;
    foreach document in all_documents
        foreach token in document
            token_hash{token};
        end
    end
    foreach token in token_hash
```

```
            if token_hash{token} < minimum_frequency
                delete token from token_hash;
            else
                totalFeature++;
            end
        end
        sortDescendingByValue(token_hash);
        n ← percentage_to_use * totalFeature;
        feature_list ← getTopN(token_hash,n);
        return feature_list;

Function getTopP(hash,n) return feature_list
        for i ← 1 to n
            feature_list[i] ← nextKey(hash);
        end
return feature_list;
```

Gambar 4.4. *Pseudocode* Penentuan Fitur

Penentuan fitur dilakukan untuk semua dokumen dari semua kategori yang digunakan pada percobaan. Penentuan fitur dilakukan dengan mengidentifikasi setiap token atau kata yang muncul dalam setiap dokumen dalam koleksi. Setiap kata yang muncul, dihitung frekuensi kemunculannya pada semua dokumen. Struktur data yang digunakan untuk menyimpan setiap kata yang muncul adalah *hash table* dengan kata sebagai *key* dan frekuensi kemunculan kata sebagai *value*. Pada percobaan, untuk menghindari pemrosesan atau penggunaan kata yang hanya sedikit sekali muncul dalam dokumen dan tidak memiliki arti yang penting maka perlu dilakukan pembatasan terhadap jumlah minimum frekuensi kemunculan kata. Dengan demikian, kata-kata yang memiliki frekuensi kemunculan lebih kecil dari minimum frekuensi yang ditentukan tidak akan dipakai dalam percobaan dan dihapus dari *hash table.*

Untuk variasi fitur berupa persentase fitur yang akan digunakan dalam percobaan, *feature_list* akan berisi fitur yang memiliki frekuensi tertinggi sebanyak persentase yang telah ditentukan. Oleh karena itu, dalam *hash*, token disimpan secara berurutan dari yang memiliki frekuensi kemunculan tertinggi sampai dengan yang memiliki frekuensi kemunculan terendah.

Pada percobaan juga dilakukan eksperimen untuk melihat pengaruh penggunaan kata-kata umum yang sering muncul pada dokumen seperti kata penghubung, kata keterangan waktu, tempat, dan sebagainya atau yang lebih dikenal dengan nama *stopword*. Daftar dari *stopword* yang digunakan pada percobaan dapat dilihat pada Lampiran A. Penghapusan kata-kata umum ini dilakukan setelah pengambilan semua fitur pada koleksi dokumen dengan frekuensi lebih besar atau sama dengan minimal frekuensi dan sebelum pengambilan fitur sebanyak persentase yang ditentukan. Dengan demikian, pada saat pengambilan fitur sebanyak persentase, fitur yang didapat sudah tidak mengandung kata-kata umum lagi. *Pseudocode* penghapusan *stopword* ditunjukkan Gambar 4.5.

```
Function deleteStopword(stopword_list, feature_list) return feature_list
    foreach feature in feature_list
        foreach token in stopword_list
            if feature == token
                delete feature from feature_list;
                break;
            end
        end
    end
    return feature_list;
```

Gambar 4.5. *Pseudocode* Penghapusan *Stopwords*

Untuk menghapus *stopword* dari *feature_list* perlu informasi mengenai *stopword* yang akan dihapus. Informasi ini disimpan pada *stopword_list*. Proses penghapusan dilakukan dengan membandingkan setiap fitur pada *feature_list* dengan *stopword_list*. Jika fitur pada *feature_list* ada pada *stopword_list*, maka fitur tersebut

akan dihapus dari *feature_list* sehingga pada akhirnya *feature_list* akan berisi semua fitur yang muncul pada koleksi dokumen tetapi yang bukan merupakan *stopword*.

4.3. Pembuatan *Term-Document Matrix*

Setelah melakukan penentuan fitur, maka langkah selanjutnya yang dilakukan adalah membuat *term-document matrix* yang merupakan masukan bagi teknik pengelompokan dokumen. *Term-document matrix* merupakan representasi dari setiap dokumen dalam koleksi yang digunakan pada percobaan. Untuk setiap variasi percobaan yang akan dilakukan dibentuk satu *term-document matrix*. Matriks ini merupakan matriks 2 dimensi yang terdiri dari dimensi *term* atau kata yang berasal dari *feature_list* yang sudah dibuat sebelumnya pada subbab 4.2 dan dimensi dokumen sesuai dengan penjelasan pada subbab 3.5. Dalam percobaan ini, ada 4 variasi informasi yang disimpan pada *term-document matrix* yaitu:

1. *Presence*, merupakan informasi berupa muncul tidaknya suatu *term* dalam dokumen.
2. *Frequency*, merupakan informasi berupa jumlah kemunculan *term* pada dokumen.
3. *Frequency normalized term frequency*, merupakan informasi berupa jumlah kemunculan *term* pada dokumen dibagi dengan total seluruh fitur pada dokumen tersebut.
4. *Frequency normalized term frequency-inverse document frequency*, merupakan informasi berupa *term frequency* dikalikan dengan jumlah kemunculan *term* pada keseluruhan koleksi dokumen.

Pseudocode pembuatan *term-document matrix* ditunjukkan Gambar 4.6.

```
Function createTermDocumentMatrix(documents, feature_list, information) return
    term_document_matrix

    feature_hash ← readFeature(feature_list);
    feature_size ← size(feature_hash);
```

```
total_document = count(documents);
term_document_matrix ← createMatrix(total_document,feature_size);
category_label ← createVector(total_document);
foreach documents_per_category in documents
    i ← 1;
    category_label ← category_name;
    foreach document in documents_per_category
        j ← 1;
        foreach token in feature_hash
            if information == presence and exist(token, document)
                term_document_matrix[i][j] ← 1;
            else if information == frequency
                term_document_matrix[i][j] ←
                countFrequency(token,document);
            if information == TF
                for j ← 1 to feature_size
                total_freq_in_doc ←
                sumColumn(term_document_matrix[i]);
                term_document_matrix[i][j] ←
                term_document_matrix[i][j] / total_freq_in_doc;
                end
            if information == TF-IDF
                for j ← 1 to feature_size
                total_freq_in_doc ←
                sumColumn(term_document_matrix[i]);
                term_document_matrix[i][j] ←
                (term_document_matrix[i][j] / total_freq_in_doc) *
                (log(total_document /
                token_appereance_in_documents));
            end
```

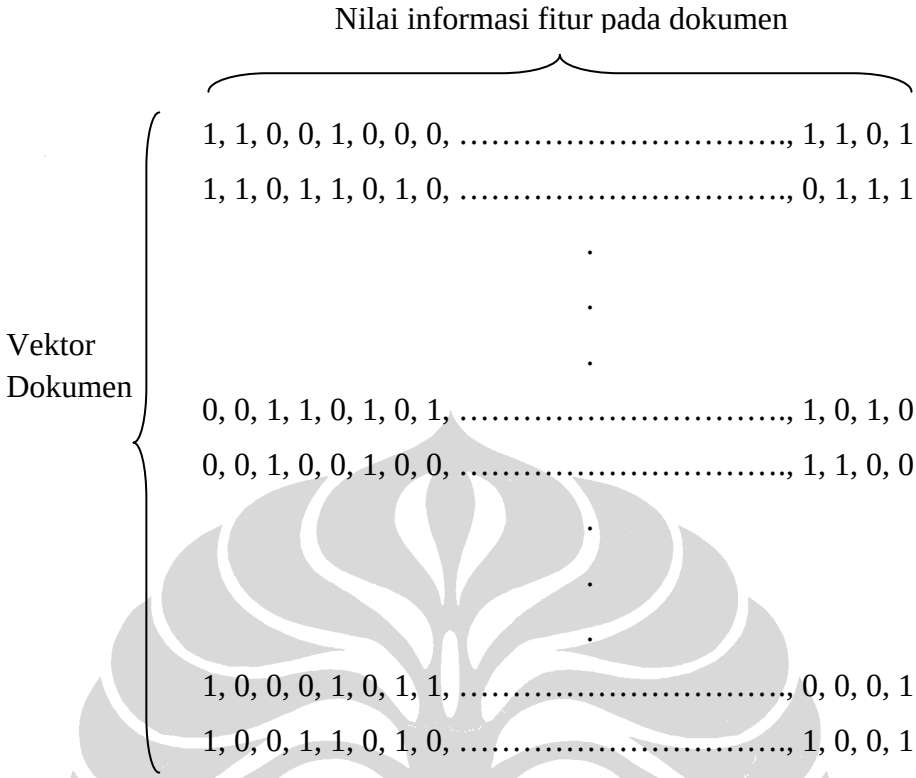
```
                end
            end
            j++;
        end
        i++;
    end

    return term_document_matrix;
```

Gambar 4.6. *Pseudocode* Pembuatan *Term-document Matrix*

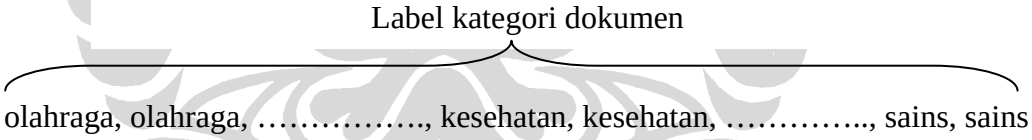
Pembuatan *term-document matrix* dimulai dengan membuat matriks kosong 2 dimensi dengan dimensi pertama adalah total dokumen yang dipakai dan dimensi yang kedua adalah jumlah fitur yang digunakan. Oleh karena itu, sebelum matriks ini dibentuk, perlu dihitung terlebih dahulu jumlah fitur dan total dokumen. Informasi yang disimpan pada matrik ini bergantung pada *information* yang diberikan yaitu salah satu dari *presence*, *frequency*, *frequency normalized term frequency*, atau *frequency normalized term frequency-inverse document frequency*. Nilai dari informasi fitur dihitung untuk tiap *token* dan tiap dokumen dalam koleksi.

Untuk keperluan evaluasi, label kategori dari tiap dokumen disimpan dalam suatu vektor. Informasi ini akan digunakan dalam pencocokan dengan hasil pengelompokan dokumen dengan teknik reduksi dimensi yang digunakan. Dari proses pencocokan ini akan dihasilkan akurasi dari pengelompokan.

Untuk setiap pembentukan *term_document_matrix* akan dihasilkan satu *term_document_matrix* yang berisi nilai dari informasi fitur dan satu vektor yang berisi informasi label kategori dari tiap dokumen. Matriks dan vektor ini disimpan dalam format *csv* (*comma separated value*). Untuk *term_document_matrix*, tiap baris merepresentasikan dokumen yang digunakan dan tiap kolom merepresentasikan fitur yang digunakan. Contoh *term_document_matrix* dan vektor berisi label kategori dokumen dengan informasi fitur berupa *presence* ditunjukkan pada Gambar 4.7 dan Gambar 4.8.



Gambar 4.7. *Term-document Matrix* dengan Informasi Fitur *Presence*



Gambar 4.8. Vektor Label Kategori Dokumen

4.4. Implementasi Teknik Pengelompokan Dokumen

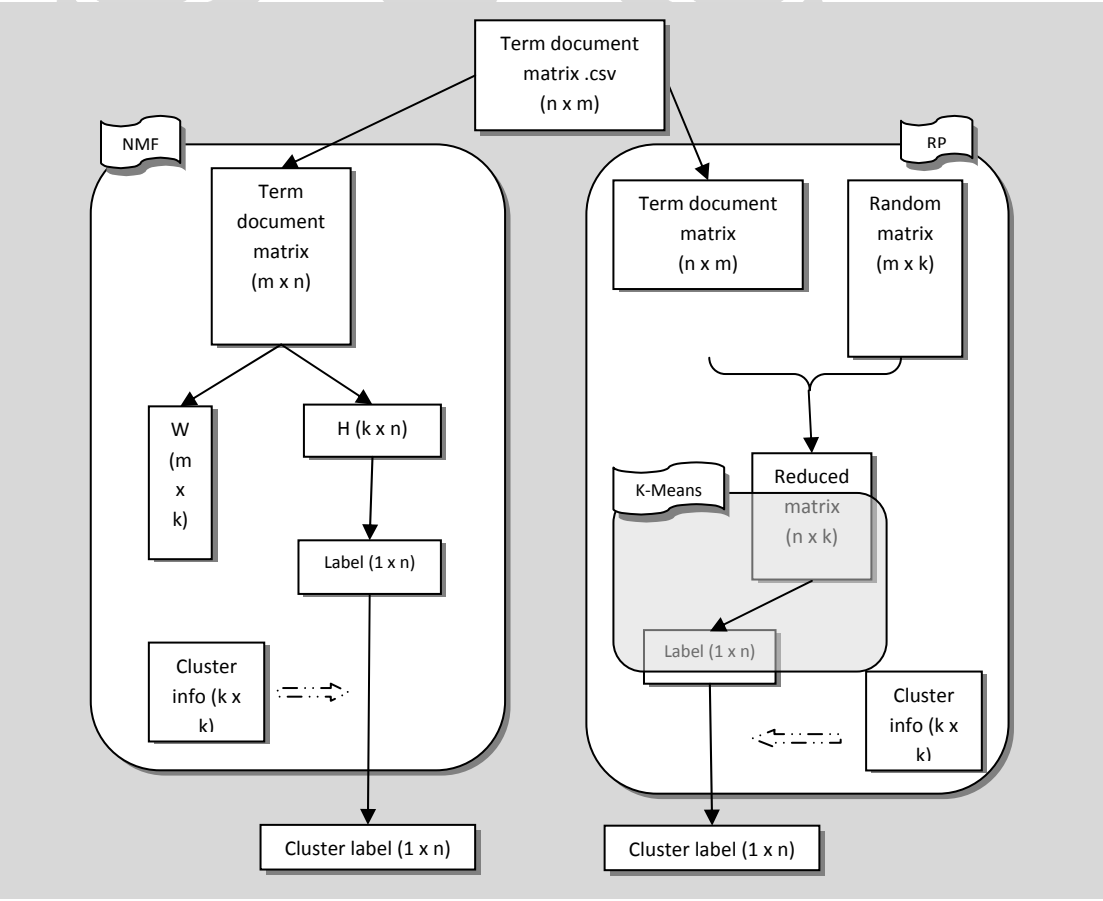
Term-document matrix yang dihasilkan pada tahap sebelumnya seperti yang telah dijelaskan pada subbab 4.3 merupakan input bagi teknik pengelompokan dokumen baik Non-Negative Matrix Factorization maupun Random Projection. Matriks dengan format *csv* ini akan dibaca dan disimpan dalam matriks dimensi pada Matlab. Matriks ini yang kemudian akan secara langsung diproses untuk menghasilkan label kluster dengan menggunakan Matlab.

Selain itu, vektor yang berisi label kategori dari tiap dokumen juga digunakan sebagai informasi untuk menerjemahkan hasil pengelompokan dengan kedua teknik

dan untuk proses evaluasi kinerja. Untuk teknik Non-Negative Matrix Factorization, label kluster dapat secara langsung ditentukan setelah proses faktorisasi selesai dilakukan. Namun, untuk Random Projection, label kluster ditentukan dengan menggunakan proses *clustering* tambahan yaitu K-Means. Gambar 4.9 menunjukkan tahapan pengelompokan dokumen dengan teknik Nonnegative Matrix Factorization dan Random Projection. Tahapan-tahapan tersebut lebih lanjut dijelaskan pada 2 subbab berikut.

4.4.1. Implementasi Teknik Non-Negative Matrix Factorization

Implementasi teknik Non-Negative Matrix Factorization dilakukan dengan menggunakan Matlab. Input yang diproses berupa *term-document matrix* seperti yang ditunjukkan Gambar 4.7. Selain itu, informasi yang disimpan pada vektor label kategori seperti yang ditunjukkan Gambar 4.8 juga digunakan. *Pseudocode* implementasi teknik Non-Negative Matrix Factorization untuk mengelompokkan dokumen ditunjukkan pada Gambar 4.11.



Gambar 4.9. Tahapan Pengelompokan Dokumen

Proses Non-Negative Matrix Factorization dimulai dengan membaca *term_document_matrix* yang telah dihasilkan pada tahap sebelumnya. Namun, untuk diproses selanjutnya, matriks ini perlu di-*transpose* terlebih dahulu. Pada awal sebelum di-*transpose*, baris dari matriks ini merepresentasikan dokumen, maka setelah di-*transpose* dokumen direpresentasikan sebagai vektor kolom.

Function GDCLS (term_document_matrix, numberOf_cluster, original_cluster_label) return cluster_label

term_document_matrix ← *transpose(term_document_matrix);*

[m,n] ← *size(term_document_matrix);*

W ← *randomMatrix(m,numberOf_cluster);*

H ← *randomMatrix(numberOf_cluster,n);*

for k ← *1 to max_iteration*

$$W_{ij} \leftarrow W_{ij} \frac{V_{ij} H_{ij}^T}{W_{ij} H_{ij} H_{ij}^T + \varepsilon}, \varepsilon = 10^{-9}$$

$$H_j = \frac{W_{ij}^T V_j}{W_{ij}^T W_{ij} + \lambda I_{ij}}, I = matriksIdentitas_{lxl}$$

end

foreach column-j in H

label[column-j] ← *max(H[column-j]);*

end

for i ← *1 to numberOf_cluster*

for j ← *1 to numberOf_cluster*

cluster_info[i][j] ← *number of documents in label[i] that appear in original_cluster_label[j];*

end

end

foreach row-i in cluster_info

cluster_info[row-i] ← *max(cluster_info[row-i]);*

end

foreach column-j in label

```
        cluster_label[column-j] ← cluster_info[label[column-j]];
    end

    return cluster_label;
```

Gambar 4.10. *Pseudocode* Implementasi Nonnegative Matrix Factorization

Proses faktorisasi dilakukan dengan membentuk matriks *W* dan *H* secara iteratif menggunakan *update rule* yang dijelaskan pada subbab 3.6.1. Proses pengelompokan dilakukan dengan menggunakan informasi pada matriks *H*. Setiap kolom matriks *H* merepresentasikan dokumen. Label hasil pengelompokan yang ditunjukkan dengan *label* pada *pseudocode* didapat dengan cara mencari nilai maksimum untuk setiap kolom. Namun, label ini masih merupakan label numerik sedangkan label yang diinginkan adalah label yang berisi nama kategori seperti bisnis keuangan, kesehatan, olahraga, dan sebagainya. Oleh karena itu, perlu dilakukan konversi dari label numerik ke label nama kategori. Untuk itu, perlu dilakukan penerjemahan terlebih dahulu untuk mengetahui korespondensi label numerik dengan label nama kategori misalnya label numerik 1 adalah sama dengan label nama kategori kesehatan, dan seterusnya. Penerjemahan dilakukan dengan menggunakan persamaan *similarity* seperti yang dijelaskan pada subbab 3.6.1 dan menggunakan informasi vektor label kategori, pada *pseudocode* ditunjukkan dengan *original_cluster_label*. Hasil penerapan persamaan *similarity* berupa matriks simetris *cluster_info*. Hasil penerjemahan didapat dengan mencari nilai maksimum untuk setiap baris pada matriks *cluster_info*.

Proses konversi selanjutnya dilakukan dengan menggunakan informasi yang didapat pada proses penerjemahan yang ada pada *cluster_info*. Hasil dari proses konversi, *cluster_label*, berisi label kategori atau kluster dari masing-masing dokumen yang merupakan hasil penerapan teknik Non-Negative Matrix Factorization. Hasil ini yang nantinya digunakan pada evaluasi kinerja teknik pengelompokan dokumen.

4.4.2. Implementasi Teknik Random Projection dengan K-Means

Teknik Random Projection untuk mereduksi dimensi dan dilanjutkan dengan K-Means untuk penentuan label kluster diimplementasikan dengan menggunakan Matlab. Input untuk teknik Random Projection adalah *term-document matrix* yang telah dibentuk seperti yang ditunjukkan Gambar 4.7 dan vektor label kategori seperti yang ditunjukkan Gambar 4.8. *Pseudocode* implementasi teknik Random Projection dan K-Means ditunjukkan pada Gambar 4.11.

```
Function RandomProjection(term_document_matrix, numberOf_cluster,
numberOfDimension_toReduce, original_cluster_label) return cluster_label
    [n,m] ← size(term_document_matrix);
    randomMatrix_distributionType = type 1 or type 2;
    if randomMatrix_distributionType = type 1
        random_matrix ← create randomMatrix(m,m-
        numberOfDimension_toReduce) from seeds = {-√3, 0, √3} with each
        probability = {1/6, 2/3, 1/6};
    else if randomMatrix_distributionType = type 2
        random_matrix ← create randomMatrix(m,m-
        numberOfDimension_toReduce) from seeds = {-1, 1} with each probability =
        {1/2, 1/2};
    end

    reduced_matrix ← term_document_matrix * random_matrix;
    label ← kmeans(reduced_matrix, numberOf_cluster);
    label ← transpose(label);

    for i ← 1 to numberOf_cluster
        for j ← 1 to numberOf_cluster
            cluster_info[i][j] ← number of documents in label[i] that appear in
            original_cluster_label[j];
        end
```

```
end
foreach row-i in cluster_info
    cluster_info[row-i] ← max(cluster_info[row-i]);
end
foreach column-j in label
    cluster_label[column-j] ← cluster_info[label[column-j]];
end

return cluster_label;
```

Gambar 4.11. *Pseudocode* Implementasi Random Projection dengan K-Means

Proses Random Projection dimulai dengan membuat matrik acak sesuai dengan tipe distribusi yang telah ditentukan. Terdapat 2 macam tipe distribusi yang digunakan sesuai dengan penjelasan pada subbab 3.6.2. Tiap baris pada matriks acak merepresentasikan vektor fitur atau *term* yang digunakan. Masing-masing vektor tersebut memiliki dimensi sesuai dengan besaran reduksi dimensi yang ditentukan, dalam *pseudocode* ditunjukkan dengan *m-numberOfDimension_toReduce*. *Term-document matrix* dengan dimensi yang telah direduksi didapat dengan cara mengalikan *term-document matrix* awal dengan matriks acak yang telah dibentuk. Proses reduksi dimensi Random Projection selesai pada tahap ini.

Untuk pengelompokan dokumen, teknik K-Means diterapkan pada *term-document matrix* yang dimensinya telah direduksi dengan teknik Random Projection. Teknik K-Means diimplementasikan dengan metode yang sudah ada pada Matlab. Seperti pada teknik Non-Negative Matrix Factorization, label hasil pengelompokan yang dihasilkan oleh K-Means masih merupakan label numerik dan bukan label nama kategori. Oleh karena itu, sama seperti yang dilakukan pada proses Non-Negative Matrix Factorization yang dijelaskan pada subbab 4.4.1, perlu dilakukan proses penerjemahan label numerik ke label nama kategori dan proses konversi dari label numerik menjadi label nama kategori. Hasil konversi yang disimpan pada *cluster_label* merupakan hasil pengelompokan dokumen dengan teknik reduksi dimensi Random Projection dan teknik pengelompokan K-Means. Hasil ini yang nantinya digunakan pada evaluasi kinerja teknik pengelompokan dokumen.

4.5. Evaluasi Kinerja

Evaluasi kinerja teknik Non-Negative Matrix Factorization dan Random Projection dengan K-Means dalam mengelompokkan dokumen dilakukan dengan cara membandingkan label kluster yang dihasilkan masing-masing teknik dengan label kategori yang asli. Hasil perbandingan dibagi dengan total dokumen merupakan nilai akurasi dari masing-masing teknik. *Pseudocode* implementasi evaluasi kinerja teknik pengelompokan dokumen ditunjukkan Gambar 4.12.

```
Function computeAccuracy(cluster_label, original_cluster_label) return accuracy

    total_match = 0;
    n ← size(cluster_label);
    foreach column-j in cluster_ label
        if cluster_label[colum-j] == original_cluster_label[column-j];
            total_match++;
        end
    end
    accuracy ← (total_match / n) * 100%;

    return accuracy;
```

Gambar 4.12. *Pseudocode* Evaluasi Kinerja

BAB V HASIL DAN PEMBAHASAN

Pada bab ini akan dijelaskan percobaan yang dilakukan mengenai pengelompokan dokumen dengan menggunakan teknik reduksi dimensi Nonnegative Matix Factorization dan Random Projection. Penjelasan meliputi variabel percobaan yang digunakan dan teknis masing-masing percobaan tersebut dilakukan. Kemudian hasil dari masing-masing percobaan yang dilakukan akan ditampilkan dan diikuti dengan pembahasannya.

5.1. Percobaan Pengelompokan Dokumen

Pada percobaan penerapan teknik reduksi dimensi Nonnegative Matrix Factorization dan Random Projection dalam aplikasi pengelompokan dokumen, ada sebelas variabel percobaan yang akan digunakan. Sebelas variabel tersebut adalah

1. Penggunaan *stopwords*

   Seperti penjelasan pada subbab 3.1, *stopwords* merupakan kata-kata umum yang sering muncul pada dokumen sehingga kata-kata ini dianggap tidak membawa informasi penting yang bisa digunakan dalam membedakan dokumen. Percobaan pengelompokan dokumen akan dilakukan dengan menyertakan *stopwords* dan menghapus *stopwords* untuk melihat pengaruh penggunaan *stopwords* pada kinerja atau akurasi pengelompokan dokumen. Penghapusan *stopwords* dilakukan pada tahap *preprocessing* yaitu tahap penentuan fitur sebelum *term-document matrix* dibentuk dan teknik pengelompokan dokumen diterapkan pada matriks tersebut.

2. Informasi fitur

   Variasi informasi fitur yang disimpan pada *term-document matrix* yang digunakan adalah *presence, frequency, frequency normalized term frequency* (TF), dan *frequency normalized term frequency-inverse document frequency* (TF-IDF). Penjelasan mengenai masing-masing informasi fitur disajikan pada subbab 3.5. Informasi fitur yang berbeda memberikan informasi yang berbeda mengenai tingkat seberapa penting (*degree of importance*) suatu *term* terhadap suatu dokumen. Percobaan dilakukan

dengan menggunakan variasi informasi fitur ini untuk menentukan informasi fitur yang paling baik digunakan dalam aplikasi pengelompokan dokumen.

3. Jumlah fitur

   Banyaknya fitur yang digunakan menunjukkan banyaknya informasi yang digunakan dalam mengelompokkan dokumen. Dalam percobaan ini, banyaknya fitur yang digunakan dinyatakan dengan persentase dari total semua fitur yang diekstrak dari koleksi yang telah diurutkan berdasarkan frekuensi kemunculan. Fitur diambil sebanyak persentase yang digunakan mulai dari fitur dengan frekuensi terbanyak. Penggunaan persentase ini dipilih karena lebih bersifat universal dibandingkan dengan jumlah fitur. Variasi persentase fitur yang digunakan adalah 90%, 70%, 50%, 30%, dan 10%.

4. Parameter khusus masing-masing teknik

   Teknik Nonnegative Matrix Factorization memiliki dua parameter khusus yaitu nilai lambda dan jumlah iterasi (lihat subbab 2.5). Variasi nilai lambda yang digunakan adalah 0.1, 0.01, dan 0.001 sesuai dengan percobaan yang dilakukan oleh Berry & Shahnaz (2004). Jumlah iterasi yang digunakan adalah beberapa variasi jumlah iterasi sebelum hasil konvergen dan beberapa variasi jumlah iterasi setelah hasil konvergen. Contohnya, Nonnegative Matrix Factorization menghasilkan faktor *W* dan *H* yang telah konvergen pada iterasi ke 10, maka variasi jumlah iterasi yang digunakan adalah 2, 4, 6, 8, 10, 15, 17, dan 30.

   Teknik Random Projection memiliki dua parameter khusus yaitu jumlah pengurangan dimensi dan jenis distribusi yang digunakan dalam matriks acak (lihat subbab 2.6). Total dimensi awal, berupa total fitur, dikurangi dengan sejumlah dimensi (dinyatakan dengan persentase dari total awal) merupakan dimensi yang direduksi. Dimensi baru ini merupakan hasil reduksi dimensi dan akan digunakan untuk pengelompokan dokumen. Variasi persentase pengurangan dimensi yang digunakan adalah 0%, 20%,

40%, 60%, dan 80%. Jenis distribusi yang digunakan adalah sesuai dengan yang diajukan oleh Aclioptas (2001) (lihat persamaan 18 dan 19 pada subbab 2.6).

5. Ukuran kluster

   Ukuran kluster adalah jumlah dokumen yang digunakan untuk tiap kategori. Variasi ukuran kluster yang digunakan adalah 30, 40, 50, 60, 70, 90, 110, dan 296.

6. Jumlah kluster

   Jumlah kluster menunjukkan seberapa banyak kluster yang akan dibentuk dari koleksi dokumen yang ada. Koleksi dokumen dibentuk dari dokumen dari kategori sejumlah kluster yang akan dibentuk dengan jumlah dokumen yang diambil dari setiap kategori sesuai dengan variasi yang ditetapkan pada variabel ukuran kluster. Variasi jumlah kluster yang digunakan adalah 2, 3, 4, 5, 6, dan 8.

7. Keseragaman ukuran kluster

   Keseragaman ukuran kluster memiliki arti bahwa jumlah dokumen dari setiap kategori yang akan dikelompokkan sama. Percobaan ini juga akan melihat pengaruh keseragaman ukuran kluster ini pada akurasi pengelompokan dokumen dengan melakukan eksperimen yang menggunakan ukuran kluster yang tidak seragam. Ukuran kluster yang tidak seragam dinyatakan dengan perbandingan. Variasi yang digunakan adalah untuk jumlah kluster 2 yaitu 5.5 : 4.5, 7 : 3, dan 9 : 1.

8. Teknik pengelompokan dokumen

   Teknik pengelompokan dokumen yang digunakan adalah Nonnegative Matrix Factorization, Random Projection dengan K-Means, dan K-Means.

9. Aspek kemiripan kluster

   Secara intuitif pengelompokan dokumen dalam domain dimana kluster yang satu memiliki tingkat kemiripan yang rendah dengan kluster yang lain seperti contohnya bisnis keuangan dengan olahraga atau kesehatan harusnya lebih mudah dibandingkan dengan pengelompokan dokumen dalam domain dimana tingkat kemiripan antar kluster cukup tinggi seperti bulutangkis, tenis yang merupakan dokumen olahraga. Oleh karena itu, percobaan ini akan menguji hipotesa tersebut dengan mengelompokkan dokumen olahraga ke dalam kluster yang lebih spesifik seperti balap, bola, bulutangkis, tenis dan tinju dan membandingkannya dengan hasil pengelompokan dokumen dalam domain dimana tingkat kemiripan antar kluster rendah.

10. Pengelompokan dokumen dari sumber yang berbeda

    Pengelompokan dokumen dari sumber yang berbeda dilakukan dengan menggabungkan koleksi dokumen dari 2 sumber yang telah dikumpulkan yaitu artikel Kompas dan artikel Antara sebelum diterapkan teknik pengelompokan dokumen.

11. Pengelompokan dokumen ke dalam kluster dengan jumlah melebihi jumlah kategori dokumen yang dipakai

    Percobaan dilakukan dengan mengelompokkan dokumen dari koleksi 2 kategori ke dalam 2, 3, 4, dan 5 kluster untuk melihat dan mengamati hasil pengelompokan yang terjadi. Salah satu kategori yang dipakai adalah olahraga karena kategori ini memiliki 5 kategori spesifik dimana setiap dokumennya telah dikelompokkan secara manual ke dalam kategori spesifik tersebut untuk keperluan evaluasi.

Secara ringkas, sebelas variabel percobaan yang digunakan ditunjukkan pada Tabel 5.1.

Tabel 5.1. Variabel Percobaan

No.	Variabel Percobaan	Nilai
1.	Penggunaan *stopwords*	Menyertakan atau menghapus *stopwords*

<table>
<tr><td>2.</td><td>Informasi fitur</td><td colspan="3">- Presence
- Frequency
- TF
- TF-IDF</td></tr>
<tr><td rowspan="2">3.</td><td rowspan="2">Jumlah fitur</td><td colspan="3">- 90% - 30%
- 70% - 10%
- 50%</td></tr>
<tr><td colspan="3">* persentase dari total semua fitur</td></tr>
<tr><td rowspan="4">4.</td><td rowspan="4">Parameter khusus teknik</td><td rowspan="2">NMF</td><td>Lambda</td><td>- 0.1
- 0.01
- 0.001</td></tr>
<tr><td>Jumlah iterasi</td><td>- Sebelum konvergen
- Setelah konvergen</td></tr>
<tr><td rowspan="2">RP</td><td>Jumlah pengurangan dimensi</td><td>- 0% - 60%
- 20% - 80%
- 40%
* persentase pengurangan dimensi dari total fitur</td></tr>
<tr><td>Tipe distribusi matriks acak</td><td>- Tipe 1
- Tipe 2</td></tr>
<tr><td>5.</td><td>Ukuran kluster</td><td colspan="3">30, 40, 50, 60, 70, 90, 110, 296</td></tr>
<tr><td>6.</td><td>Jumlah kluster</td><td colspan="3">2, 3, 4, 5, 6, 8</td></tr>
<tr><td>7.</td><td>Keseragaman ukuran kluster</td><td colspan="3">- 5.5 : 4.5
- 7 : 3
- 9 : 1
* untuk jumlah kluster 2</td></tr>
<tr><td>8.</td><td>Teknik pengelompokan dokumen</td><td colspan="3">- NMF
- RP dengan K-Means
- K-Means</td></tr>
<tr><td>9.</td><td>Aspek kemiripan kluster</td><td colspan="3">- Tingkat kemiripan kluster rendah (mis :</td></tr>
</table>

		bisnis keuangan, olahraga, kesehatan, dll) - Tingkat kemiripan kluster tinggi (mis : bulutangkis, tenis, dll)
10.	Pengelompokan dokumen dari sumber yang berbeda	Artikel Kompas dan Antara digabung, kemudian diterapkan teknik pengelompokan dokumen
11.	Pengelompokan dokumen ke kluster yang jumlahnya melebihi jumlah kategori yang dipakai	- 2 - 4 - 3 - 5 * untuk jumlah kategori 2

Secara garis besar, percobaan dibagi ke dalam tujuh kelompok percobaan. Masing-masing percobaan dilakukan dengan menggunakan teknik reduksi dimensi Nonnegative Matrix Factorization dan Random Projection. Kelompok pertama (dibahas pada subbab 5.2) adalah percobaan yang berhubungan dengan fitur yaitu percobaan terhadap variabel 1 sampai dengan 3 untuk semua kemungkinan kombinasi nilai dari masing-masing variabel. Hasil yang didapat pada kelompok percobaan pertama akan digunakan untuk kelompok percobaan selanjutnya. Hasil yang diambil adalah yang memberikan akurasi pengelompokan yang terbaik dari masing-masing variabel dengan masing-masing teknik.

Kelompok percobaan kedua (dibahas pada subbab 5.3) adalah percobaan yang berhubungan dengan parameter-parameter khusus dari masing-masing teknik, Nonnegative Matrix Factorization dan Random Projection, yaitu percobaan terhadap variabel 4. Kelompok percobaan ini menggunakan konfigurasi fitur yang memberikan hasil terbaik berdasarkan pada kelompok percobaan pertama yang telah dilakukan. Percobaan ini juga dilakukan untuk semua kemungkinan kombinasi nilai dari masing-masing variabel untuk tiap teknik.

Kelompok percobaan ketiga (dibahas pada subbab 5.4) adalah percobaan yang berhubungan dengan dokumen yaitu percobaan terhadap variabel 5 sampai dengan 7. Kelompok percobaan ini juga dilakukan untuk semua kemungkinan kombinasi nilai dari masing-masing variabel. Kelompok percobaan ini menggunakan

konfigurasi fitur dan parameter khusus teknik yang memberikan hasil terbaik berdasarkan dua kelompok percobaan sebelumnya.

Sejauh ini, konfigurasi terbaik dari masing-masing teknik, Nonnegative Matrix Factorization dan Random Projection, sudah diperoleh. Dengan demikian, perbandingan kinerja masing-masing teknik bisa dilakukan. Percobaan ini dilakukan pada kelompok percobaan keempat (dibahas pada subbab 5.5) yaitu percobaan terhadap variabel 8. Percobaan juga membandingkan kedua teknik dengan K-Means yang telah banyak digunakan pada aplikasi pengelompokan dokumen.

Kelompok percobaan kelima (dibahas pada subbab 5.6) adalah percobaan terhadap variabel 9. Hasil percobaan dalam domain dimana tingkat kemiripan kluster tinggi dibandingkan hasil yang diperoleh pada kelompok percobaan ketiga dimana tingkat kemiripan kluster rendah. Kelompok percobaan keenam (dibahas pada subbab 5.7) adalah percobaan terhadap variabel 10. Hasil percobaan ini juga dibandingkan dengan hasil kelompok percobaan ketiga. Kelompok percobaan ketujuh (dibahas pada subbab 5.8) adalah percobaan terhadap variabel 11. Untuk menjaga validitas hasil yang didapat, maka setiap percobaan akan dilakukan sebanyak 3 kali. Hasilnya kemudian akan dirata-ratakan.

5.2. Percobaan dari Aspek Fitur

Percobaan ini merupakan kelompok percobaan pertama (dijelaskan pada subbab 5.1) yang bertujuan untuk menganalisis kinerja dari teknik Nonnegative Matrix Factorization dan Random Projection berdasarkan aspek fitur yang digunakan. Dengan demikian dapat diketahui konfigurasi fitur yang memberikan hasil akurasi terbaik untuk masing-masing teknik yang dapat digunakan untuk kelompok percobaan selanjutnya. Aspek fitur yang digunakan dalam percobaan ini adalah jenis informasi fitur, jumlah fitur yang digunakan serta penggunaan *stopwords*. Percobaan ini mengelompokkan dokumen menjadi dua kluster dengan menggunakan koleksi artikel Kompas dari tiga kategori yaitu bisnis keuangan, olahraga, dan kesehatan. Dari tiga kategori ini dibentuk dua jenis koleksi yaitu koleksi pertama terdiri dari kategori bisnis keuangan dan olahraga serta koleksi kedua terdiri dari kategori kesehatan dan olahraga. Akurasi pengelompokan dokumen merupakan rata-rata dari akurasi dua jenis koleksi tersebut.

Jumlah dokumen yang digunakan adalah 296 per kategori. Untuk teknik Nonnegative Matrix Factorization, nilai parameter *lambda* yang digunakan adalah 0.01 dengan jumlah iterasi 100. Sedangkan untuk teknik Random Projection, persentase pengurangan dimensi adalah sebesar 50% dan tipe distribusi yang digunakan pada matriks acak adalah tipe 2 (lihat subbab 3.6.2). Hasil percobaan dari aspek fitur untuk teknik Nonnegative Matrix Factorization ditunjukkan pada Tabel 5.2 dan untuk teknik Random Projection ditunjukkan pada Tabel 5.3.

Pada tabel 5.2 dan 5.3, kolom kedua menyatakan penggunaan *stopwords*, kolom ketiga menyatakan jenis informasi fitur yang digunakan yaitu p yang artinya *presence*, f artinya *frequency*, tf artinya *term frequency*, dan tfidf artinya *term frequency-inverse document frequency*. Sedangkan baris pertama menunjukkan persentase fitur yang digunakan dari total fitur yang ada yaitu 90%, 70%, 50%, 30%, dan 10%. Nilai pada setiap kotak merupakan rata-rata akurasi dari percobaan yang dilakukan dengan dua jenis koleksi seperti yang telah dijelaskan diatas.

Tabel 5.2. Akurasi Hasil Percobaan dari Aspek Fitur: Jenis Informasi Fitur, Persentase Fitur yang Digunakan, dan Penggunaan *Stopwords* dengan Teknik Nonnegative Matrix Factorization

			90%	70%	50%	30%	10%
NMF	dengan *stopwords*	p	96.62	96.38	96.16	95.61	91.72
		f	86.73	86.35	86.09	84.29	82.35
		tf	73.56	70.37	58.99	72.64	70.01
		tfidf	60.14	59.61	70.65	55.91	55.74
	tanpa *stopwords*	p	97.30	97.00	96.28	96.00	94.28
		f	97.00	96.46	95.73	95.00	93.73
		tf	92.40	92.73	93.58	90.68	89.31
		tfidf	83.10	77.73	73.09	74.01	70.50

Tabel 5.3. Akurasi Hasil Percobaan dari Aspek Fitur: Jenis Informasi Fitur, Persentase Fitur yang Digunakan, dan Penggunaan *Stopwords* dengan Teknik Random Projection

			90%	70%	50%	30%	10%
RP	dengan *stopwords*	p	96.11	95.6	95.28	94.42	74.15
		f	53.88	52.37	51.68	51.00	50.84
		tf	96.45	95.20	90.37	95.95	92.39
		tfidf	51.17	50.77	50.17	50.15	50.00
	tanpa *stopwords*	p	97.13	97.00	96.60	95.28	89.28
		f	84.25	82.30	82.00	81.00	80.60
		tf	97.00	96.20	95.95	96.20	94.50
		tfidf	82.28	81.73	81.28	81.00	80.95

5.2.1. Analisa Efek *Stopwords*

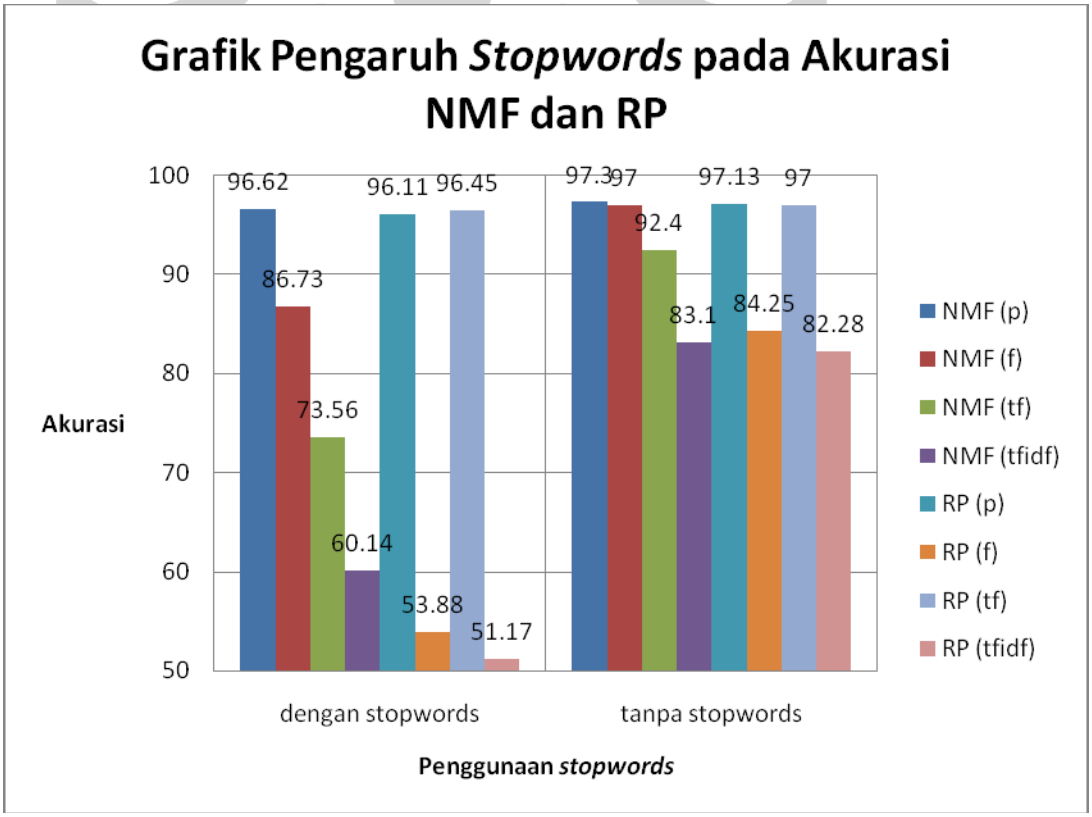
Stopwords merupakan kata-kata umum yang sering muncul pada dokumen sehingga tidak menyertakan *stopwords* dalam pengelompokan dokumen yaitu menghapus *stopwords* pada *term-document matrix* sebelum teknik pengelompokan dokumen diterapkan dapat meningkatkan kinerja dari teknik pengelompokan dokumen.

Tabel 5.4 menunjukkan perbandingan kinerja pengelompokan dokumen dari masing-masing teknik, Nonnegative Matrix Factorization dan Random Projection, pada koleksi yang menyertakan *stopwords* dan tidak menyertakan *stopwords* dengan variasi jenis informasi fitur dan persentase fitur yang digunakan 90%. Data ini merupakan hasil percobaan kelompok pertama yang ditampilkan pada Tabel 5.2.

Tabel 5.4. Efek *Stopwords* pada Akurasi NMF dan RP

	NMF				RP			
	p	f	tf	tfidf	p	f	tf	tfidf
dengan *stopwords*	96.62	86.73	73.56	60.14	96.11	53.88	96.45	51.17
tanpa *stopwords*	97.30	97.00	92.40	83.10	97.13	84.25	97.00	82.28

Dari tabel 5.4 dapat dilihat bahwa baik teknik Nonnegative Matrix Factorization maupun Random Projection menghasilkan akurasi pengelompokan dokumen yang lebih baik ketika *stopwords* dihapus dan tidak digunakan pada pengelompokan dokumen dibandingkan ketika *stopwords* digunakan. Hal ini berlaku untuk semua jenis informasi fitur yang digunakan yaitu *presence*, *frequency*, *term frequency*, dan *term frequency-inverse document frequency*. Hal ini dikarenakan dengan menghapus *stopwords*, fitur yang digunakan untuk membedakan vektor dokumen yang satu dengan yang lain hanyalah fitur-fitur atau kata-kata khusus yang tidak muncul di semua dokumen. Dengan demikian, vektor dokumen dari kategori satu lebih mudah dibedakan dengan vektor dokumen dari kategori yang lain sehingga pengelompokan pun bisa dilakukan dengan lebih akurat dan menghasilkan akurasi yang lebih tinggi. Dari percobaan ini dapat disimpulkan bahwa pengelompokan dokumen tanpa menyertakan *stopwords* memberikan akurasi yang lebih baik dibandingkan dengan menyertakannya. Grafik dari pengaruh *stopwords* pada akurasi pengelompokan dokumen ditunjukkan pada Gambar 5.1.



Gambar 5.1. Grafik Pengaruh *Stopwords* pada Akurasi NMF dan RP

5.2.2. Analisa Pengaruh Jumlah Fitur yang Digunakan

Jumlah fitur yang dinyatakan dengan persentase dari total fitur menunjukkan seberapa banyak fitur yang digunakan dalam merepresentasikan sebuah dokumen. Semakin besar persentase fitur berarti bahwa semakin banyak informasi yang digunakan dalam merepresentasikan dokumen. Dengan demikian, semakin banyak pula pengetahuan dalam membedakan satu dokumen dengan dokumen lain sehingga pengelompokan dokumen ke dalam kluster dapat dilakukan dengan lebih akurat dan menghasilkan akurasi pengelompokan yang lebih tinggi.

Tabel 5.5 menunjukkan hasil percobaan untuk melihat pengaruh jumlah fitur yang digunakan pada akurasi pengelompokan dokumen dengan teknik Nonnegative Matrix Factorization dan Random Projection dan variasi jenis informasi fitur tanpa menyertakan *stopwords*. Data ini merupakan hasil percobaan kelompok pertama yang ditampilkan pada Tabel 5.2.

Tabel 5.5. Pengaruh Jumlah Fitur dan Jenis Informasi Fitur yang Digunakan pada Akurasi NMF dan RP

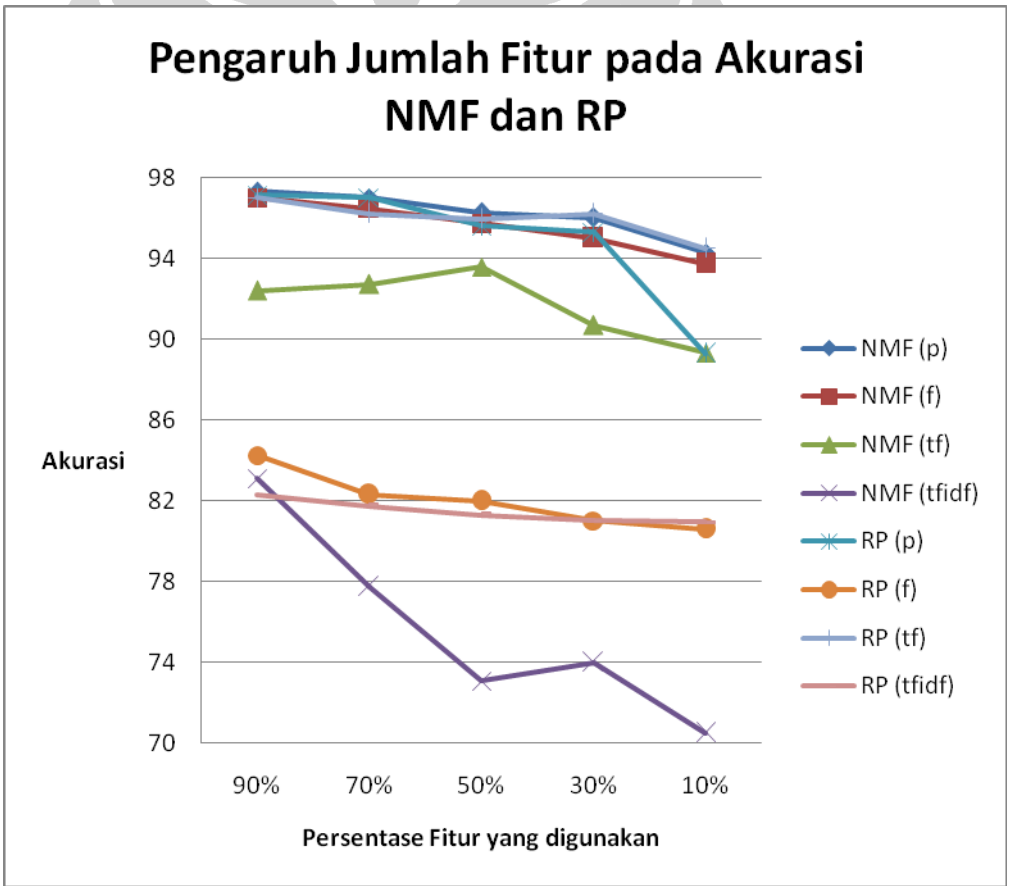
	NMF				RP			
	P	f	tf	tfidf	p	f	tf	tfidf
90%	97.30	97.00	92.40	83.10	97.13	84.25	97.00	82.28
70%	97.00	96.46	92.73	77.73	97.00	82.30	96.20	81.73
50%	96.28	95.73	93.58	73.09	96.60	82.00	95.95	81.28
30%	96.00	95.00	90.68	74.01	95.28	81.00	96.20	81.00
10%	94.28	93.73	89.31	70.50	89.28	80.60	94.50	80.95

Dari tabel 5.5 dapat dilihat bahwa untuk informasi fitur berupa *presence* dan *frequency*, persentase fitur yang digunakan yang semakin besar memberikan akurasi pengelompokan yang semakin tinggi baik dengan teknik Nonnegative Matrix Factorization maupun Random Projection. Namun, pola seperti ini tidak terjadi jika informasi fitur yang digunakan adalah *term frequency* dan *term frequency-inverse document frequency* baik dengan menggunakan teknik Nonnegative Matrix Factorization maupun Random Projection. Pada kasus ini, persentase fitur yang semakin besar tidak berakibat pada kenaikan akurasi pengelompokan dokumen.

Dari percobaan ini, dapat disimpulkan bahwa semakin banyak fitur yang digunakan, akurasi pengelompokan dokumen semakin tinggi jika informasi fitur yang digunakan adalah *presence* dan *frequency*. Grafik pengaruh jumlah fitur yang digunakan pada akurasi pengelompokan dokumen ditunjukkan pada Gambar 5.2.

5.2.3. Analisa Pengaruh Jenis Informasi Fitur

Informasi fitur yang berbeda memberikan informasi yang berbeda mengenai tingkat seberapa penting (*degree of importance*) suatu *term* terhadap suatu dokumen. Percobaan dilakukan dengan menggunakan variasi informasi fitur untuk melihat pengaruh jenis informasi fitur pada akurasi pengelompokan. Dengan demikian, dapat ditentukan informasi fitur yang paling baik dalam aplikasi pengelompokan dokumen untuk masing-masing teknik.



Gambar 5.2. Grafik Pengaruh Jumlah Fitur dan Jenis Informasi Fitur pada Akurasi NMF dan RP

Hasil percobaan untuk melihat pengaruh informasi fitur yang digunakan pada akurasi pengelompokan dokumen dengan teknik Nonnegative Matrix Factorization

dan Random Projection dan variasi jumlah fitur tanpa menyertakan *stopwords* ditunjukkan pada Tabel 5.5. Grafik pengaruh jenis informasi fitur yang digunakan pada akurasi pengelompokan dokumen ditunjukkan pada Gambar 5.2.

Pada Gambar 5.2 dan Tabel 5.5 dapat dilihat bahwa untuk teknik Nonnegative Matrix Factorization, jenis informasi fitur *presence*, *frequency*, *TF*, dan *TF-IDF* secara berturut-turut memberikan nilai akurasi yang terbaik hingga terburuk untuk variasi jumlah fitur yang digunakan. Dapat dilihat bahwa, untuk setiap variasi jumlah fitur yang digunakan, jenis informasi fitur *presence* selalu memberikan nilai akurasi yang terbaik dibandingkan dengan *frequency*, *TF*, maupun *TF-IDF*. Kemudian nilai akurasi ketika menggunakan *frequency* lebih baik daripada *TF* dan *TF-IDF*, dst. Untuk teknik Random Projection, nilai akurasi terbaik secara keseluruhan masih diperoleh ketika menggunakan *presence* dibandingkan dengan yang lain walaupun hal ini tidak berlaku untuk setiap variasi jumlah fitur yang digunakan. Namun, nilai akurasi terburuk tetap didapat ketika menggunakan jenis informasi fitur *TF-IDF*.

Dari percobaan ini, dapat disimpulkan bahwa untuk teknik Nonnegative Matrix Factorization, penggunaan informasi *presence* memberikan akurasi terbaik dibandingkan dengan *frequency*, *TF*, dan *TF-IDF*. Sedangkan untuk teknik Random Projection, penggunaan informasi *presence* secara umum (tidak berlaku untuk setiap kondisi) memberikan akurasi terbaik dibandingkan dengan *frequency, TF*, dan *TF-IDF*. Penggunaan informasi *presence* memberikan akurasi terbaik dikarenakan dengan hanya menggunakan 1 atau 0, representasi vektor dari dokumen-dokumen yang memiliki kemiripan isi akan memiliki kesamaan yang lebih tinggi dibandingkan jika informasi fitur yang lain yang digunakan. Representasi ini tidak dipengaruhi oleh berapa kali fitur tertentu muncul pada dokumen, bandingkan dengan informasi fitur yang lain dimana walaupun beberapa dokumen memiliki fitur yang sama, kemiripan masih harus dilihat dari berapa kali fitur tersebut muncul.

5.3. Percobaan dari Aspek Parameter Khusus Teknik

Percobaan ini merupakan kelompok percobaan kedua (dijelaskan pada subbab 5.1) yang bertujuan untuk menganalisis kinerja dari teknik Nonnegative

Matrix Factorization dan Random Projection dari aspek parameter-parameter khusus dari masing-masing teknik. Paramater khusus yang digunakan adalah

1. Untuk Nonnegative Matrix Factorization, parameter : nilai *lambda,* jumlah iterasi.
2. Untuk Random Projection, parameter : jumlah pengurangan dimensi, tipe distribusi matriks acak.

Percobaan ini mengelompokkan dokumen menjadi dua kluster dengan menggunakan koleksi artikel Kompas dari tiga kategori yaitu bisnis keuangan, olahraga, dan kesehatan. Dari tiga kategori ini dibentuk dua jenis koleksi yaitu koleksi pertama terdiri dari kategori bisnis keuangan dan olahraga serta koleksi kedua terdiri dari kategori kesehatan dan olahraga. Akurasi pengelompokan dokumen merupakan rata-rata dari akurasi dua jenis koleksi tersebut.

Jumlah dokumen yang digunakan adalah 296 per kategori. Percobaan ini menggunakan konfigurasi fitur yang memberikan hasil terbaik sesuai dengan percobaan yang telah dilakukan sebelumnya pada kelompok percobaan pertama mengenai fitur (subbab 5.2). Berdasarkan hasil percobaan kelompok pertama mengenai fitur, maka *stopwords* tidak diikutsertakan pada proses penentuan kluster, jumlah fitur yang digunakan adalah 90% dari total semua fitur, dan jenis informasi fitur yang digunakan adalah *presence*.

Hasil percobaan dari aspek parameter-parameter khusus untuk teknik Nonnegative Matrix Factorization ditunjukkan pada Tabel 5.6 dan untuk teknik Random Projection ditunjukkan pada Tabel 5.7.

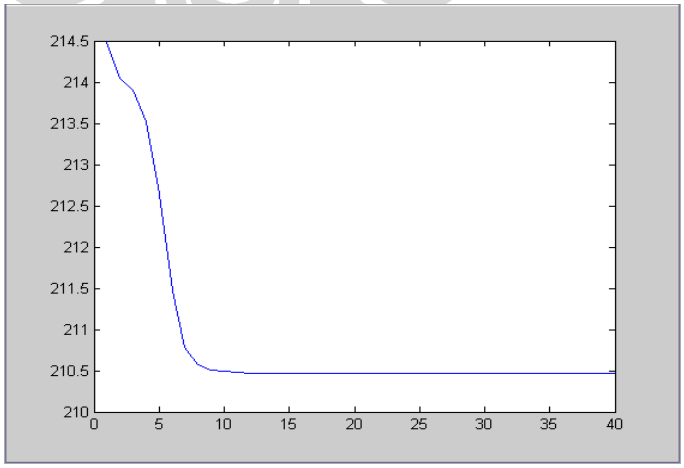
Tabel 5.6. Akurasi Hasil Percobaan dari Aspek Nilai *Lambda* dan Jumlah Iterasi Teknik Nonnegative Matrix Factorization

NMF	2	4	6	8	10	30	35	40
0.1	55.97	87.5	89.78	93.58	96.71	96.54	96.45	97.06
0.01	67.28	69.42	86.65	95.44	96.34	96.45	96.79	96.79
0.001	53.77	67.79	82.43	96.04	96.37	96.23	96.11	96.79

Tabel 5.7. Akurasi Hasil Percobaan dari Aspek Jumlah Pengurangan Dimensi dan Tipe Distribusi Matriks Acak teknik Random Projection

RP	0%	20%	40%	60%	80%
Tipe 1	97.55	96.95	95.61	93.24	73.90
Tipe 2	96.79	96.54	96.14	95.30	93.59

Pada Tabel 5.6 setiap baris menunjukkan variasi nilai *lambda* yang digunakan yaitu 0.1, 0.01, dan 0.001 sedangkan setiap kolom menunjukkan variasi jumlah iterasi yang digunakan yaitu 2, 4, 6, 8, 10, 30, 35, dan 40. Variasi jumlah iterasi ini ditentukan dengan melihat hasil percobaan yang dilakukan terhadap masing-masing koleksi dokumen dari segi konvergensi yaitu berapa iterasi yang diperlukan agar hasil faktor yang didapatkan sudah konvergen. Salah satu contoh grafik hasil percobaan yang digunakan dalam menentukan variasi jumlah iterasi yang digunakan ditunjukkan pada Gambar 5.3. Percobaan ini menggunakan nilai *lambda* 0.01. Pada grafik, sumbu-*x* menunjukkan jumlah iterasi dan sumbu-*y* menunjukkan nilai dari $\|V - WH\|$. Dari grafik pada Gambar 5.3 tersebut, dapat dilihat bahwa konvergensi dicapai ketika jumlah iterasi mencapai 10. Oleh karena itu, untuk melihat pengaruh jumlah iterasi maka variasi jumlah iterasi yang dipakai adalah beberapa iterasi sebelum konvergensi dicapai (yaitu 2, 4, 6, dan 8) dan beberapa iterasi setelah konvergensi dicapai (yaitu 10, 30, 35, dan 40). Nilai pada setiap kotak pada Tabel 5.6 merupakan rata-rata akurasi dari percobaan yang dilakukan dengan dua jenis koleksi seperti yang telah dijelaskan diatas.

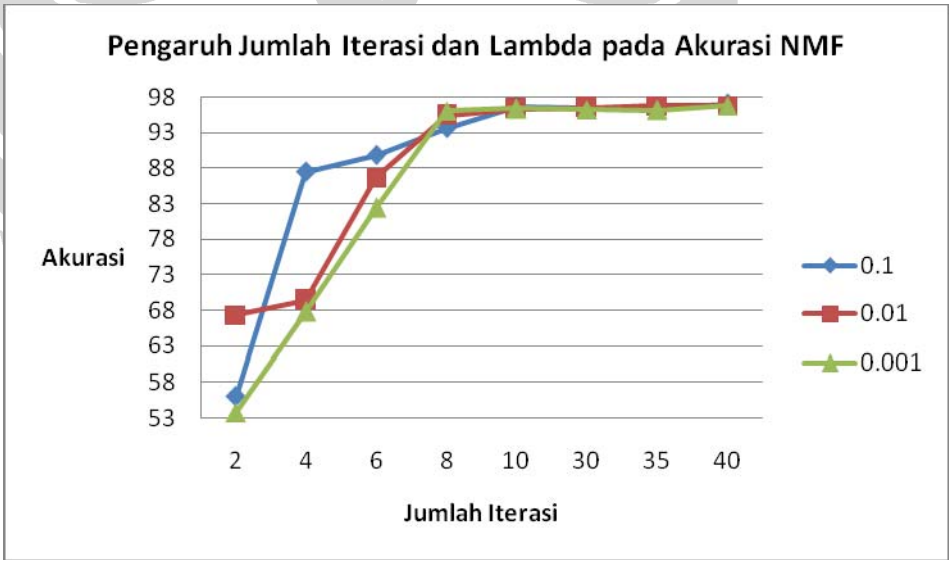


Gambar 5.3. Grafik Percobaan NMF dengan Informasi Jumlah Iterasi dan Konvergensi

Pada Tabel 5.7, setiap baris menunjukkan tipe distribusi matriks acak yang digunakan (lihat subbab 3.6.2) dan setiap kolom menunjukkan jumlah pengurangan dimensi yang dinyatakan dengan persentase dari total semua fitur yaitu 0%, 20%, 40%, 60%, dan 80%. Nilai pada setiap kotak merupakan rata-rata akurasi dari percobaan yang dilakukan dengan dua jenis koleksi seperti yang telah dijelaskan diatas.

5.3.1. Analisa Parameter Teknik Nonnegative Matrix Factorization

Dari Tabel 5.6, dapat dianalisis kinerja teknik Nonnegative Matrix Factorization dari aspek parameter-parameter khususnya yaitu nilai *lambda* dan jumlah iterasi. Dengan demikian, dapat diketahui pengaruh masing-masing parameter pada kinerja teknik sehingga dapat digunakan untuk kelompok percobaan selanjutnya. Gambar 5.4 menunjukkan grafik pengaruh nilai *lambda* dan jumlah iterasi pada akurasi teknik Nonnegative Matrix Factorization.



Gambar 5.4. Grafik Pengaruh Nilai *Lambda* dan Jumlah Iterasi pada Akurasi NMF

Informasi pada Gambar 5.3 bahwa konvergensi dicapai pada saat jumlah iterasi mencapai 10 terrefleksikan pada Gambar 5.4. Dapat dilihat pada Gambar 5.4 bahwa akurasi pengelompokan juga mencapai kestabilan setelah iterasi ke 10 yaitu pada kisaran nilai akurasi $\pm$ 96%. Pada iterasi ke 2 dimana konvergensi belum dicapai, akurasi hanya mencapai nilai 53% dan terus bergerak naik seiring dengan jumlah iterasi yang bertambah dan hingga mencapai konvergensi, nilai akurasi

bergerak stabil pada nilai $\pm$ 96%. Kondisi ini berlaku untuk semua variasi nilai *lambda* yaitu 0.1, 0.01, dan 0.001.

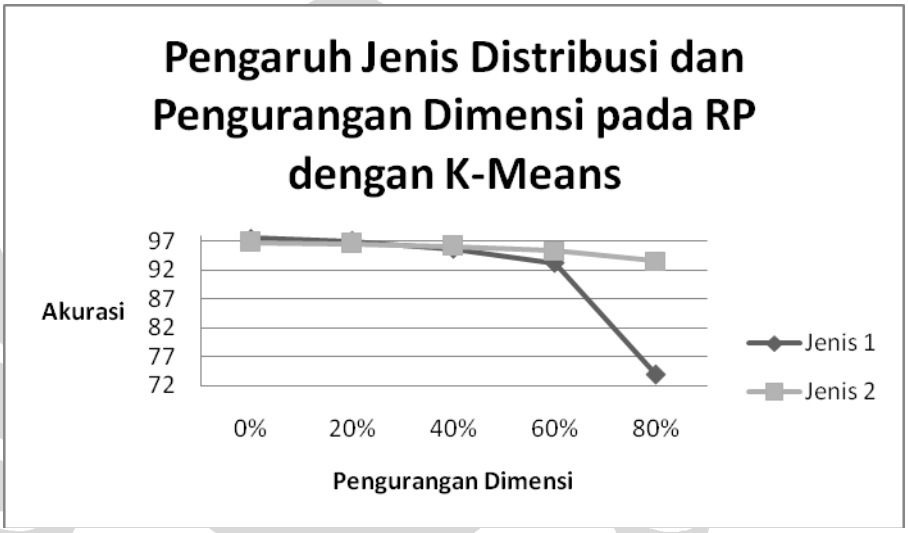
Dapat dilihat juga pada Gambar 5.4 bahwa nilai *lambda* yang lebih besar cenderung mencapai nilai akurasi yang lebih tinggi lebih dulu dibandingkan dengan nilai *lambda* yang lebih kecil. Misalnya pada iterasi ke 4 nilai *lambda* 0.1 telah memberikan akurasi 88% sedangkan nilai *lambda* 0.001 hanya memberikan akurasi 68%. Hal ini menunjukkan bahwa untuk mencapai nilai akurasi yang cukup tinggi dengan jumlah iterasi yang lebih sedikit dapat dilakukan dengan memperbesar nilai *lambda*.

Dari percobaan ini dapat disimpulkan bahwa untuk mencapai nilai akurasi yang cukup tinggi dengan jumlah iterasi yang lebih sedikit dapat dilakukan dengan memperbesar nilai *lambda*, tetapi jumlah iterasi yang jauh dari jumlah iterasi yang diperlukan untuk mencapai konvergensi akan menghasilkan akurasi yang jauh lebih rendah. Jumlah iterasi yang memungkinkan konvergensi bergantung pada ukuran data yang digunakan yaitu jumlah dokumen (yang kemudian menentukan jumlah fitur). Oleh karena itu, untuk menentukan jumlah iterasi yang memungkinkan konvergensi, perlu dilakukan percobaan terlebih dahulu yang hasilnya seperti yang ditunjukkan pada Gambar 5.3. Cara lain yang bisa dilakukan adalah menggunakan *stopping criteria* berupa selisih antara besaran matriks dengan besaran matriks iterasi sebelumnya yang telah ditentukan untuk suatu nilai tertentu. Iterasi pengubahan (*update*) matriks akan berhenti jika selisih besaran matriks lebih kecil dari nilai tersebut (matriks telah konvergen).

5.3.2. Analisa Parameter Teknik Random Projection

Dari Tabel 5.7 dapat dilakukan analisis terhadap kinerja teknik Random Projection dari aspek parameter-parameter khususnya yaitu jumlah pengurangan dimensi dan tipe distribusi matriks acak. Dengan demikian, dapat diketahui pengaruh masing-masing parameter pada akurasi teknik sehingga dapat digunakan untuk kelompok percobaan selanjutnya. Gambar 5.5 menunjukkan grafik pengaruh jumlah pengurangan dimensi dan tipe distribusi matriks acak pada akurasi teknik Random Projection.

Pada Gambar 5.5 dapat dilihat bahwa akurasi menurun seiring dengan berkurangnya dimensi. Misalnya dengan distribusi tipe 1, akurasi ketika dimensi dikurangi 20% adalah $\pm$ 97%, namun akurasi turun hingga $\pm$ 74% ketika dimensi dikurangi 80%. Namun, untuk tipe distribusi 2, akurasi yang dihasilkan cenderung stabil (tidak berkurang banyak) ketika dimensi dikurangi dengan jumlah yang signifikan banyak. Hal ini menunjukkan bahwa, Random Projection sangat efektif karena dengan mengurangi dimensi hingga 80% (yang berarti mengurangi biaya komputasi), akurasi pengelompokkan yang dihasilkan masih cukup tinggi (diatas 90%).



Gambar 5.5. Pengaruh Tipe Distribusi Matriks Acak dan Jumlah Pengurangan Dimensi pada Akurasi RP

Secara umum, tipe distribusi 2 memberikan hasil yang lebih baik. Selain itu, tipe distribusi ini juga masih memberikan akurasi yang baik walaupun dimensi dikurangi dengan jumlah yang cukup besar dimana pada saat ini tipe distribusi 1 memberikan akurasi yang tidak begitu bagus.

Dari percobaan ini dapat disimpulkan bahwa pengurangan dimensi menyebabkan penurunan akurasi pengelompokan, tetapi penurunan akurasi tidak terlalu signifikan jika tipe distribusi yang digunakan adalah tipe distribusi 2.

5.4. Percobaan dari Aspek Dokumen

Percobaan ini merupakan kelompok percobaan ketiga (lihat subbab 5.1) yang bertujuan untuk melihat kinerja teknik Nonnegative Matrix Factorization dan

Random Projection dari aspek dokumen yang digunakan. Aspek dokumen yang dilihat adalah jumlah kluster, ukuran kluster, dan keseragaman ukuran kluster. Percobaan dibagi 2 yaitu percobaan mengenai jumlah kluster dan ukuran kluster dan percobaan mengenai keseragaman ukuran kluster. Masing-masing percobaan dijelaskan secara detil pada subbab selanjutnya. Untuk percobaan 1, pengelompokan dilakukan pada dokumen artikel Kompas yang diambil dari beberapa kategori (lebih dari jumlah kluster). Misal, percobaan mengenai jumlah kluster 3, maka artikel yang digunakan diambil dari lebih dari 3 kategori. Hal ini dilakukan untuk membentuk 2 jenis koleksi yang hasilnya akan dirata-ratakan dan menjadi akurasi untuk percobaan tersebut. Penggunaan 2 jenis koleksi ini hanya dilakukan untuk jumlah kluster 2, 3, 4, 5, dan 6. Untuk jumlah kluster 8, hanya 1 jenis koleksi yang digunakan karena keterbatasan data (hanya ada 8 kategori). Untuk percobaan yang menggunakan 2 jenis koleksi, akurasi pengelompokan dokumen merupakan rata-rata dari akurasi dua jenis koleksi tersebut.

Berikut adalah kategori yang digunakan untuk masing-masing jumlah kluster yang diujicoba:

1. Jumlah kluster = 2, koleksi 1 terdiri dari bisnis keuangan dan olahraga; koleksi 2 terdiri dari kesehatan dan olahraga.
2. Jumlah kluster = 3, koleksi 1 terdiri dari bisnis keuangan, olahraga, kesehatan; koleksi 2 terdiri dari kesehatan, perempuan, dan sains.
3. Jumlah kluster = 4, koleksi 1 terdiri dari bisnis keuangan, kesehatan, olahraga, perempuan; koleksi 2 terdiri dari sains, travel, olahraga, kesehatan.
4. Jumlah kluster = 5, koleksi 1 terdiri dari bisnis keuangan, olahraga, kesehatan, perempuan, dan sains; koleksi 2 terdiri dari perempuan, sains, travel, properti, dan politik hukum.
5. Jumlah kluster = 6, koleksi 1 terdiri dari bisnis keuangan, olahraga, kesehatan, perempuan, sains, dan travel; koleksi 2 terdiri dari kesehatan, perempuan, sains, travel, properti, dan politikhukum.
6. Jumlah kluster = 8, hanya ada 1 jenis koleksi yaitu bisnis keuangan, olahraga, kesehatan, perempuan, sains, travel, properti, dan politikhukum.

Untuk percobaan ke 2 mengenai keseragaman kluster, dokumen dikelompokkan menjadi 2 kluster dengan menggunakan koleksi artikel Kompas dari

3 kategori yaitu bisnis keuangan, olahraga, dan kesehatan. Dari tiga kategori ini dibentuk dua jenis koleksi, masing-masing terdiri dari dua kategori yaitu bisnis keuangan, olahraga dan kesehatan, olahraga. Akurasi pengelompokan dokumen merupakan rata-rata dari akurasi dua jenis koleksi tersebut.

Konfigurasi fitur yang digunakan adalah yang memberikan akurasi terbaik sesuai dengan kelompok percobaan pertama (lihat subbab 5.2) mengenai fitur. Berdasarkan hasil percobaan kelompok pertama mengenai fitur, maka *stopwords* tidak disertakan pada proses penentuan kluster, jumlah fitur yang digunakan adalah 90% dari total semua fitur, dan jenis informasi fitur yang digunakan adalah *presence*.

Konfigurasi parameter khusus masing-masing teknik ditentukan berdasarkan hasil percobaan kelompok kedua mengenai parameter khusus teknik. Berdasarkan hasil percobaan kelompok kedua, untuk teknik Nonnegative Matrix Factorization, jumlah iterasi yang digunakan adalah jumlah iterasi yang memungkinkan konvergensi dicapai. Nilai eksak dari jumlah iterasi tergantung dari data yang digunakan. Untuk kelompok percobaan ini, jumlah iterasi yang digunakan adalah 50. Sedangkan untuk jumlah iterasi yang memungkinkan konvergensi, karena semua variasi nilai *lambda* memberikan akurasi yang hampir sama (lihat Gambar 5.5), maka untuk percobaan ini, nilai *lambda* manapun bisa dipilih, dan nilai 0.01 yang dipilih untuk digunakan.

Berdasarkan hasil percobaan kelompok kedua, untuk teknik Random Projection, tipe distribusi yang digunakan untuk matriks acak adalah tipe 2. Jumlah pengurangan dimensi yang digunakan adalah 60% karena tingkat pengurangan cukup besar (berhubungan dengan biaya komputasi) namun tetap memberikan akurasi yang cukup baik (lihat Gambar 5.5).

5.4.1. Analisa Pengaruh Jumlah dan Ukuran Kluster

Percobaan ini bertujuan untuk melihat pengaruh jumlah kluster dan ukuran masing-masing kluster terhadap akurasi teknik Nonnegative Matrix Factorization maupun Random Projection. Dengan bertambahnya jumlah kluster berarti bahwa kompleksitas masalah pengelompokan bertambah karena diperlukan usaha yang lebih banyak untuk membedakan setiap dokumen sehingga setiap dokumen bisa

dikelompokkan ke dalam kluster yang tepat. Ukuran kluster berkontribusi pada banyaknya informasi yang bisa digunakan dalam proses pengelompokan. Dengan semakin besarnya ukuran kluster yang berarti semakin banyak dokumen yang digunakan untuk tiap kategori menyebabkan informasi tentang suatu kluster semakin banyak. Hasil percobaan mengenai pengaruh jumlah kluster dan ukuran kluster terhadap akurasi teknik Nonnegative Matrix Factorization ditunjukkan pada Tabel 5.8 dan akurasi teknik Random Projection pada Tabel 5.9.

Tabel 5.8. Pengaruh Jumlah dan Ukuran Kluster pada Akurasi NMF

NMF	2	3	4	5	6	8
30	89.74	81.37	81.15	58.62	50.33	46.61
40	90.16	84.23	81.90	60.17	56.62	55.44
50	93.13	84.96	82.00	63.17	57.83	55.67
60	93.33	87.96	82.13	69.32	58.15	55.80
70	95.00	90.32	82.15	71.15	58.93	56.15
90	96.11	90.62	83.00	74.00	61.11	58.00
110	96.50	91.15	85.90	77.35	65.00	60.40
296	96.92					

Tabel 5.9. Pengaruh Jumlah dan Ukuran Kluster pada Akurasi RP

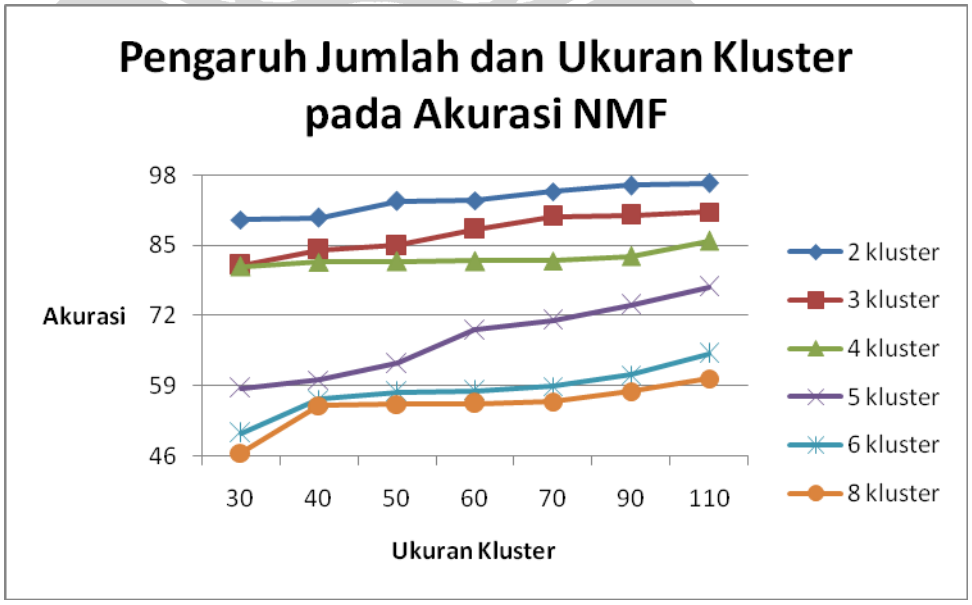
RP	2	3	4	5	6	8
30	79.78	65.00	58.54	49.50	46.50	40.78
40	80.68	67.81	60.66	50.00	47.31	42.92
50	81.00	70.81	67.00	53.17	48.67	44.70
60	82.00	72.45	70.00	53.96	51.15	44.93
70	83.93	77.15	71.07	55.37	51.43	46.50
90	85.00	82.38	74.16	59.72	52.32	48.00
110	93.63	86.12	76.59	62.62	55.91	49.66
296	94.91					

Pada Tabel 5.8 dan Tabel 5.9, baris menunjukkan variasi ukuran kluster yang digunakan yaitu 30, 40, 50, 60, 70, 90, 110, dan 296. Sedangkan kolom menunjukkan jumlah kluster yang digunakan yaitu 2, 3, 4, 5, 6, dan 8. Nilai pada setiap kotak

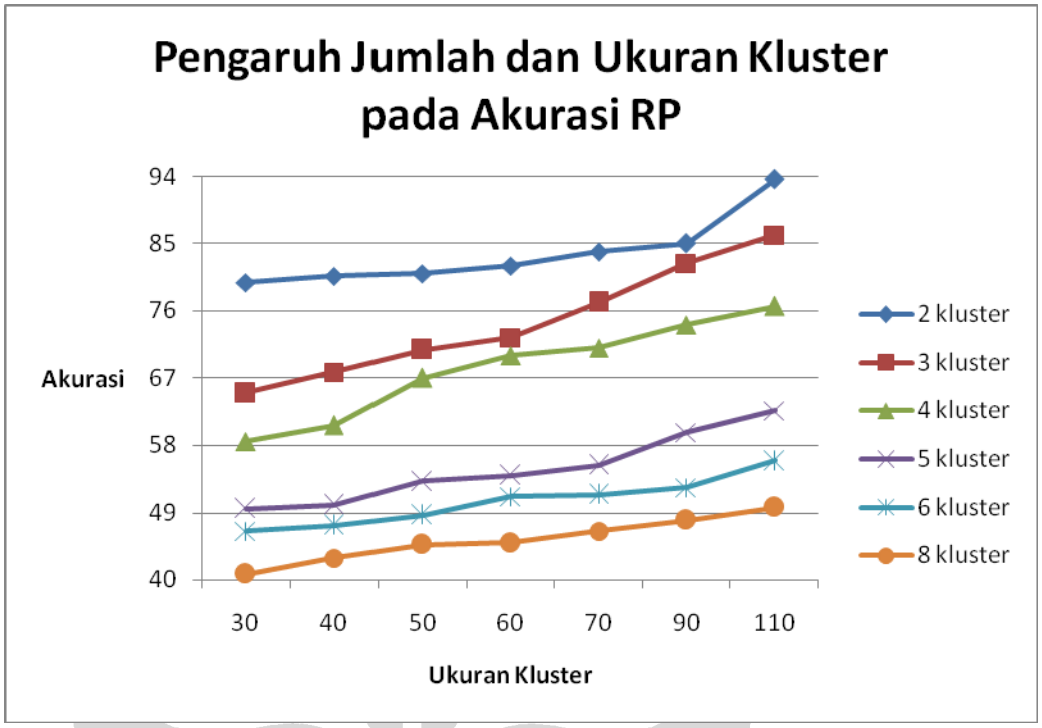
merupakan rata-rata akurasi dari percobaan yang dilakukan dengan dua jenis koleksi seperti yang telah dijelaskan pada subbab 5.4. Grafik pengaruh jumlah dan ukuran kluster pada akurasi teknik Nonnegative Matrix Factorization dan Random Projection masing-masing ditunjukkan pada Gambar 5.6 dan Gambar 5.7.

Dari Gambar 5.6 dan Tabel 5.8 dapat dilihat bahwa pengelompokan yang melibatkan jumlah kluster yang lebih sedikit dengan teknik Nonnegative Matrix Factorization memiliki akurasi yang lebih baik dibandingkan dengan jumlah kluster yang banyak. Pada saat jumlah kluster 2, akurasi yang dihasilkan sekitar 90%. Namun, pada saat jumlah kluster 8, akurasi pengelompokan turun drastis hingga mencapai sekitar 50%. Kondisi ini juga terjadi pada teknik Random Projection (lihat Gambar 5.7 dan Tabel 5.9). Pada saat jumlah kluster 2, akurasi yang dihasilkan sekitar 85%, tetapi turun drastis hingga mencapai sekitar 45% ketika jumlah kluster 8.

Ukuran kluster yang semakin besar berpengaruh pada kenaikan akurasi pengelompokan baik pada teknik Nonnegative Matrix Factorization maupun Random Projection. Tren ini dapat dilihat pada Gambar 5.6 dan Gambar 5.7. Seperti contohnya pada Gambar 5.6, pada saat ukuran kluster 30 dan jumlah kluster 2, yang artinya jumlah dokumen tiap kategori yang digunakan adalah 30, akurasi yang dihasilkan hanya sekitar 89%, tetapi akurasi ini terus mengalami kenaikan seiring dengan kenaikan ukuran kluster dan mencapai nilai 96% ketika ukuran kluster 110.



Gambar 5.6. Grafik Pengaruh Jumlah dan Ukuran Kluster pada Akurasi NMF



Gambar 5.7. Grafik Pengaruh Jumlah dan Ukuran Kluster pada Akurasi RP

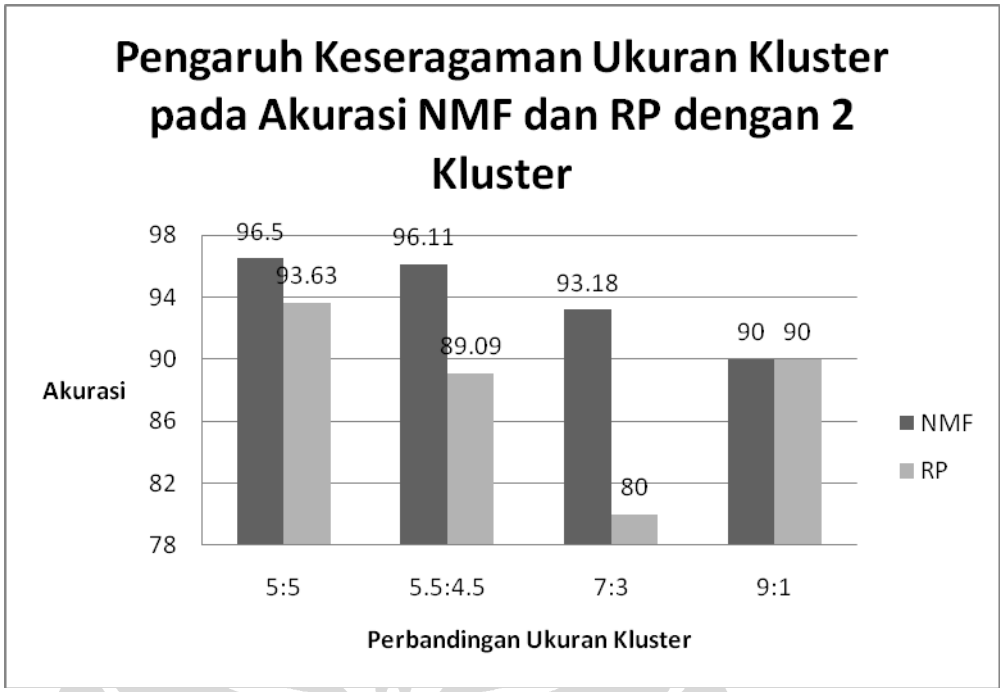
Dari percobaan ini dapat disimpulkan bahwa jumlah kluster yang semakin banyak menyebabkan penurunan pada akurasi pengelompokan baik dengan teknik Nonnegative Matrix Factorization dan Random Projection. Akan tetapi, ukuran kluster yang semakin besar menyebabkan kenaikan pada akurasi pengelompokan, baik untuk teknik Nonnegative Matrix Factorization maupun Random Projection.

5.4.2. Analisa Pengaruh Keseragaman Ukuran Kluster

Percobaan ini bertujuan untuk melihat pengaruh keseragaman ukuran kluster pada akurasi teknik Nonnegative Matrix Factorization dan Random Projection. Hasil percobaan ditunjukkan pada Tabel 5.10 dan Gambar 5.8.

Tabel 5.10. Pengaruh Keseragaman Ukuran Kluster pada Akurasi NMF dan RP

	5 : 5	5.5 : 4.5	7 : 3	9 : 1
NMF	96.50	96.11	93.18	90.00
RP	93.63	89.09	80.00	90.00



Gambar 5.8. Grafik Pengaruh Keseragaman Kluster pada Akurasi NMF dan RP

Pada Tabel 5.10, baris menunjukkan teknik yang digunakan yaitu Nonnegative Matrix Factorization dan Random Projection, sedangkan kolom menunjukkan perbandingan ukuran kluster yang digunakan. Variasi perbandingan ukuran kluster adalah 5 : 5, 5.5 :4.5, 7 : 3, dan 9 : 1 (dari total dokumen yang dipakai yaitu 220). Nilai pada setiap kotak merupakan rata-rata akurasi dari percobaan yang dilakukan dengan dua jenis koleksi seperti yang telah dijelaskan pada subbab 5.4.

Dari Tabel 5.10 dan Gambar 5.8 dapat dilihat bahwa akurasi menurun seiring dengan semakin besarnya perbedaan ukuran 2 kluster yang digunakan (rasio perbandingan semakin besar). Namun, ketika rasio perbandingan mencapai 9 (9 : 1), akurasi menjadi suatu nilai yang konstan yaitu 90% yang berarti bahwa setiap dokumen dikelompokkan ke dalam kluster yang memiliki jumlah dokumen 90% dari total dokumen yang digunakan. Dengan kata lain, 10% dokumen dari kategori lain tidak cukup untuk membangun informasi kluster sehingga kluster yang terbentuk hanya 1 (yang jumlah dokumennya 90% dari total dokumen). Kondisi ini berlaku baik untuk teknik Nonnegative Matrix Factorization maupun Random Projection.

Dari percobaan ini dapat disimpulkan bahwa ketidakseragaman ukuran kluster berpengaruh pada penurunan akurasi pengelompokan baik dengan teknik

Nonnegative Matrix Factorization maupun teknik Random Projection. Semakin besar rasio ketidakseragaman, maka penurunan akurasi semakin besar. Namun, akurasi akan menjadi suatu nilai konstan bila rasio ketidakseragaman terlalu besar.

5.5. Percobaan dari Aspek Teknik Pengelompokan

Percobaan ini merupakan kelompok percobaan keempat (lihat subbab 5.1) yang bertujuan untuk membandingkan kinerja teknik yang digunakan untuk pengelompokan dokumen. Dari 3 kelompok percobaan sebelumnya, semua variabel yang penting dari teknik pengelompokan sudah dianalisa sehingga pada kelompok percobaan keempat ini, perbandingan kinerja teknik pengelompokan dokumen dapat dilakukan. Teknik yang dibandingkan adalah teknik reduksi dimensi sekaligus pengelompokan dokumen Nonnegative Matrix Factorization, teknik reduksi dimensi Random Projection yang dilanjutkan teknik pengelompokan K-Means, dan teknik pengelompokan K-Means. Percobaan ini mengelompokkan dokumen menjadi dua kluster dengan menggunakan koleksi artikel Kompas dari tiga kategori yaitu bisnis keuangan, olahraga, dan kesehatan. Dari tiga kategori ini dibentuk dua jenis koleksi, yaitu koleksi pertama terdiri dari kategori bisnis keuangan dan olahraga serta koleksi kedua terdiri dari kategori kesehatan dan olahraga. Akurasi pengelompokan dokumen merupakan rata-rata dari akurasi dua jenis koleksi tersebut.

Berdasarkan hasil percobaan kelompok pertama mengenai fitur, maka *stopwords* tidak disertakan dalam proses, jumlah fitur yang digunakan adalah 90%, dan informasi fitur yang digunakan adalah *presence*.

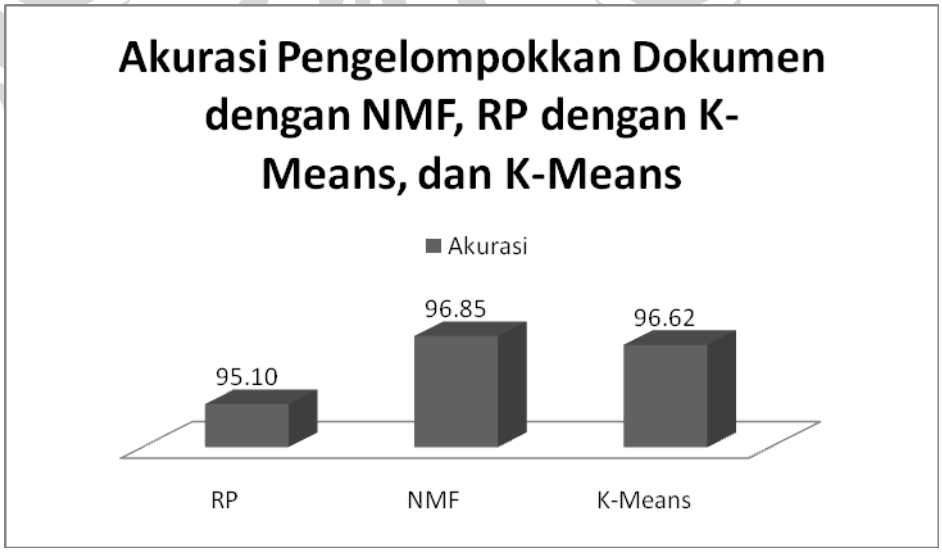
Berdasarkan hasil percobaan kelompok kedua, untuk teknik Nonnegative Matrix Factorization, jumlah iterasi yang digunakan adalah jumlah iterasi yang memungkinkan konvergensi dicapai. Nilai eksak dari jumlah iterasi tergantung dari data yang digunakan. Untuk kelompok percobaan ini, digunakan jumlah iterasi 30. Sedangkan nilai *lambda* yang digunakan adalah 0.01. Sedangkan untuk teknik Random Projection, jumlah pengurangan dimensi yang digunakan adalah 60% dan tipe distribusi matriks acak adalah tipe 2. Jumlah dokumen per kategori yang digunakan adalah 296.

Hasil percobaan mengenai perbandingan kinerja antara teknik Nonnegative Matrix Factorization, teknik Random Projection dengan K-Means, dan teknik K-Means ditunjukkan pada Tabel 5.11. Grafik perbandingan 3 teknik tersebut ditunjukkan pada Gambar 5.9.

Tabel 5.11. Perbandingan Akurasi Pengelompokan dengan NMF, RP dengan K-Means, dan K-Means

	RP	NMF	K-Means
Akurasi	95.10	96.85	96.62

Pada Tabel 5.11 dan Gambar 5.9 dapat dilihat bahwa pengelompokan dokumen dengan menggunakan teknik Nonnegative Matrix Factorization menghasilkan akurasi yang sedikit lebih baik dibandingkan dengan teknik Random Projection (yang dilanjutkan dengan K-Means) maupun teknik K-Means. Hal ini menunjukkan bahwa teknik Nonnegative Matrix Factorization merupakan teknik yang cukup efektif dalam pengelompokan dokumen karena walaupun mereduksi dimensi data, teknik ini tetap menghasilkan akurasi yang baik.



Gambar 5.9. Grafik Perbandingan Akurasi Pengelompokan dengan NMF, RP dengan K-Means, dan K-Means

5.6. Percobaan dari Aspek Kemiripan Kluster

Percobaan ini merupakan percobaan kelompok kelima (lihat subbab 5.1) yang bertujuan melihat pengaruh kemiripan kluster terhadap kinerja teknik Nonnegative

Matrix Factorization dan Random Projection. Percobaan ini menggunakan dokumen dari kategori yang memiliki tingkat kemiripan tinggi.

Hasil percobaan akan dibandingkan dengan hasil percobaan terhadap dokumen dari kategori yang tingkat kemiripannya rendah (pada percobaan kelompok ketiga subbab 5.4). Dokumen yang digunakan adalah artikel Kompas dari kategori bola, balap, bulutangkis, tenis, dan tinju. Variasi jumlah dokumen yang digunakan per kategori (ukuran kluster) adalah 30, 40, 50, dan 60. Sedangkan variasi jumlah kluster yang digunakan adalah 2, 3, 4, dan 5. Namun, karena keterbatasan data, maka tidak semua variasi jumlah dan ukuran kluster bisa diujicoba. Untuk percobaan dengan jumlah kluster 2, digunakan 2 jenis koleksi. Akurasi merupakan nilai rata-rata dari hasil percobaan 2 jenis koleksi. Sedangkan percobaan yang lain hanya menggunakan 1 jenis koleksi karena keterbatasan data (dokumen) yang ada.

Berikut adalah kategori dengan tingkat kemiripan tinggi yang digunakan untuk masing-masing jumlah kluster yang diujicoba:

1. Jumlah kluster = 2, koleksi 1 terdiri dari bulutangkis dan tenis, koleksi 2 terdiri dari bulutangkis dan bola.
2. Jumlah kluster = 3, kategori yang digunakan adalah bulutangkis, tenis, dan bola.
3. Jumlah kluster = 4, kategori yang digunakan adalah bulutangkis, tenis, bola, dan balap.
4. Jumlah kluster = 5, kategori yang digunakan adalah bulutangkis, tenis, bola, balap, dan tinju.

Konfigurasi fitur dan parameter masing-masing teknik sama dengan percobaan kelompok keempat (lihat subbab 5.5). Hasil percobaan mengenai pengaruh kemiripan kluster pada akurasi teknik Nonnegative Matrix Factorization dan Random Projection masing-masing ditunjukkan pada Tabel 5.12 dan Tabel 5.13. Grafik dari masing-masing Tabel 5.12 dan Tabel 5.13 ditunjukkan pada Gambar 5.10 dan Gambar 5.11.

Pada Tabel 5.12 dan Tabel 5.13, label pada kolom pertama menunjukkan ukuran kluster yaitu 30, 40, 50, dan 60. Sedangkan label pada baris kedua

menunjukkan jumlah kluster yaitu 2, 3, 4, dan 5. Label *T* dan *R* masing-masing menunjukkan dokumen yang digunakan, *T* = dokumen berasal dari kategori dengan tingkat kemiripan tinggi, dan *R* = dokumen berasal dari kategori dengan tingkat kemiripan rendah. Data untuk *R* diambil dari Tabel 5.8 dan Tabel 5.9.

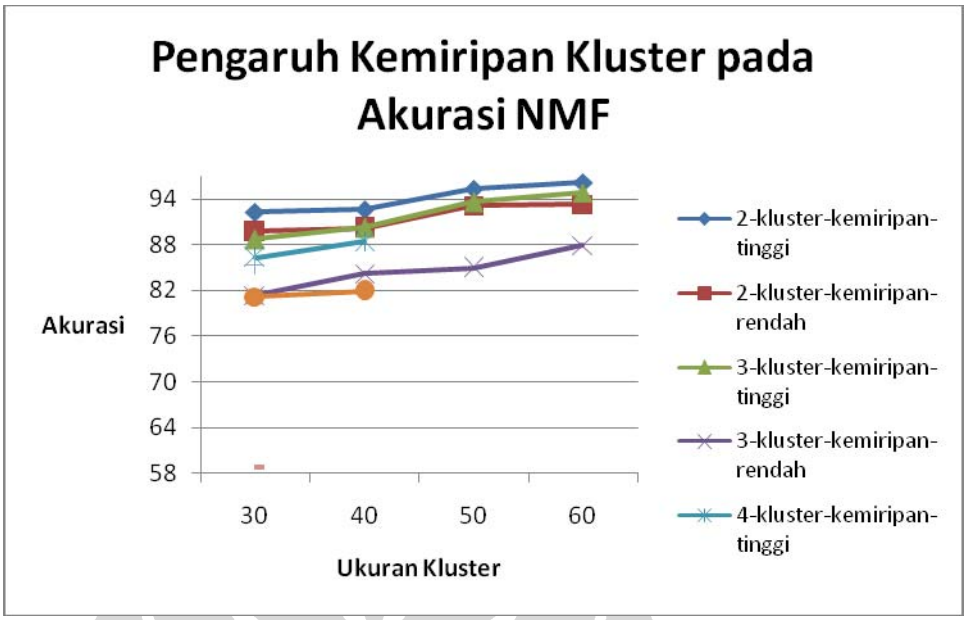
Tabel 5.12. Pengaruh Kemiripan Kluster pada Akurasi NMF

	NMF							
	2		3		4		5	
	T	R	T	R	T	R	T	R
30	92.33	89.74	88.73	81.37	86.33	81.15	85.33	58.62
40	92.67	90.16	90.25	84.23	88.39	81.90		
50	95.38	93.13	93.62	84.96				
60	96.22	93.33	94.82	87.96				

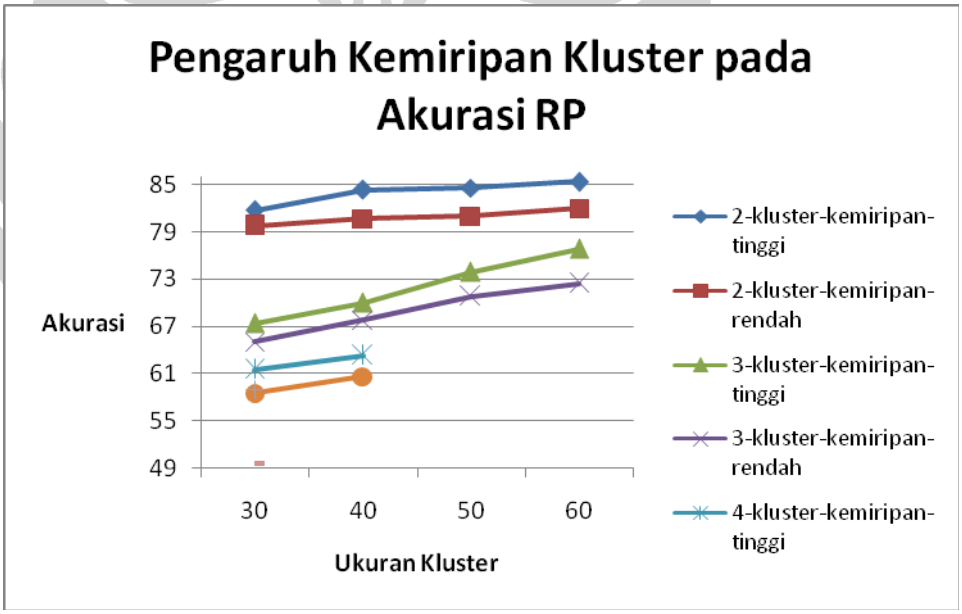
Tabel 5.13. Pengaruh Kemiripan Kluster pada Akurasi RP

	RP							
	2		3		4		5	
	T	R	T	R	T	R	T	R
30	81.69	79.78	67.41	65.00	61.50	58.54	59.00	49.50
40	84.33	80.68	70.00	67.81	63.33	60.66		
50	84.58	81.00	73.90	70.81				
60	85.34	82.00	76.85	72.45				

Dari Tabel 5.12 dan Tabel 5.13 dapat dilihat bahwa baik untuk teknik Nonnegative Matrix Factorization maupun Random Projection, pengelompokan dalam domain dimana tingkat kemiripan kluster tinggi justru menghasilkan akurasi yang lebih baik dibandingkan dengan pengelompokan dalam domain dimana tingkat kemiripan kluster rendah. Hasil percobaan ini sesuai dengan percobaan yang dilakukan Berry & Shahnaz (2004).



Gambar 5.10. Grafik Pengaruh Kemiripan Kluster pada Akurasi NMF



Gambar 5.11. Grafik Pengaruh Kemiripan Kluster pada Akurasi RP

Alasan pengelompokan dalam domain dimana tingkat kemiripan kluster tinggi justru menghasilkan akurasi yang lebih baik adalah karena dokumen dalam kluster ini memiliki kesamaan pemakaian kata-kata khusus yang lebih tinggi. Contohnya pada kluster bola, kata-kata khusus yang berhubungan dengan kategori bola banyak muncul di keseluruhan dokumen bola. Bandingkan dengan apa yang terjadi pada kluster olahraga dimana kata-kata khususnya lebih bersifat luas, mencakup bola, bulutangkis, tenis, dll. Jika dilihat dari segi *vector space model* yang

digunakan, kata-kata khusus yang lebih spesifik akan semakin baik digunakan dalam merepresentasikan vektor-vektor dokumen. Dan pada akhirnya, akurasi pengelompokan bisa lebih baik.

5.7. Percobaan dari Aspek Sumber Dokumen

Percobaan ini merupakan percobaan kelompok keenam (lihat subbab 5.1) yang bertujuan untuk melihat kinerja teknik Nonnegative Matrix Factorization dan Random Projection dalam mengelompokkan dokumen dari sumber yang berbeda. Artikel yang digunakan adalah artikel dari Kompas dan Antara. Percobaan ini mengelompokkan dokumen menjadi dua kluster dengan menggunakan koleksi artikel tiga kategori yaitu bisnis keuangan, olahraga, dan kesehatan. Dari tiga kategori ini dibentuk dua jenis koleksi yaitu koleksi pertama terdiri dari kategori bisnis keuangan dan olahraga serta koleksi kedua terdiri dari kategori kesehatan dan olahraga. Untuk kategori olahraga, dokumen yang digunakan berasal dari Kompas dan Antara dengan perbandingan jumlah 1:1. Akurasi pengelompokan dokumen merupakan rata-rata dari akurasi dua jenis koleksi tersebut.

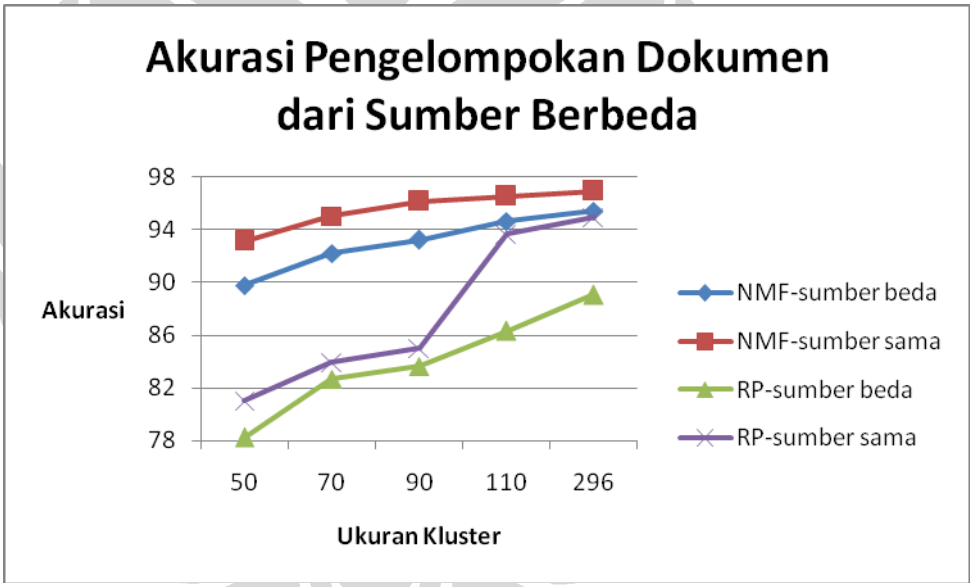
Hasil percobaan akan dibandingkan dengan hasil yang diperoleh pada percobaan kelompok ketiga (lihat subbab 5.4) dimana semua dokumen yang digunakan hanya berasal dari satu sumber yaitu Kompas. Konfigurasi fitur dan parameter masing-masing teknik sama dengan percobaan kelompok keempat (lihat subbab 5.5). Hasil percobaan mengenai pengelompokan dokumen dari sumber berbeda pada akurasi teknik Nonnegative Matrix Factorization dan Random Projection ditunjukkan pada Tabel 5.14. Grafik dari Tabel 5.14 ditunjukkan pada Gambar 5.12. Pada Tabel 5.14, label pada kolom pertama menunjukkan ukuran kluster, sedangkan label pada baris kedua yaitu *B* dan *S* berarti bahwa dokumen berasal dari sumber berbeda (*B*) atau sumber sama (*S*).

Dari Tabel 5.14 dan Gambar 5.12 dapat dilihat bahwa baik dengan teknik Nonnegative Matrix Factorization maupun Random Projection, pengelompokan dokumen dari sumber yang berbeda menghasilkan akurasi yang lebih rendah dibandingkan dengan pengelompokan dokumen dari sumber yang sama. Namun, perbedaan akurasi yang terjadi tidak terlalu besar yaitu dengan rata-rata 3%. Hal ini menunjukkan bahwa teknik Nonnegative Matrix Factorization dan Random

Projection masih cukup efektif untuk digunakan mengelompokkan dokumen walaupun dari sumber yang berbeda.

Tabel 5.14. Akurasi Pengelompokan Dokumen dari Sumber Berbeda

	NMF		RP	
	B	S	B	S
50	89.75	93.13	78.25	81.00
70	92.17	95.00	82.67	83.93
90	93.19	96.11	83.65	85.00
110	94.63	96.50	86.30	93.63
296	95.38	96.92	89.05	94.91



Gambar 5.12. Akurasi Pengelompokan Dokumen dari Sumber Berbeda

5.8. Percobaan Pengelompokan Dokumen ke Kluster yang Jumlahnya Melebihi Jumlah Kategori yang Dipakai.

Percobaan ini merupakan percobaan kelompok ketujuh (lihat subbab 5.1) yang bertujuan untuk melihat pola yang terjadi ketika mengelompokkan dokumen ke kluster yang jumlahnya melebihi jumlah kategori yang digunakan. Dokumen yang digunakan berasal dari 2 kategori yang terdiri dari bisnis keuangan dan olahraga. Jumlah kluster yang digunakan adalah 2, 3, 4, dan 5. Dengan jumlah kluster ini, percobaan dilakukan untuk melihat bagaimana kluster yang terbentuk. Untuk

keperluan evaluasi, kategori olahraga telah dibagi kedalam 5 kategori yang lebih spesifik yaitu balap, bola, bulutangkis, tenis, dan tinju. Untuk masing-masing variasi jumlah kluster, dilakukan 2 kali percobaan.

Konfigurasi fitur dan parameter masing-masing teknik sama dengan yang digunakan pada kelompok percobaan keempat (lihat subbab 5.5). Jumlah dokumen yang digunakan per kategori adalah 150 dimana untuk kategori olahraga, jumlah dokumen untuk masing-masing kategori spesifiknya adalah 30. Hasil percobaan dengan menggunakan teknik Nonnegative Matrix Factorization dan Random Projection masing-masing ditunjukkan pada Tabel 5.15 dan Tabel 5.16.

Pada Tabel 5.15 dan Tabel 5.16, label pada kolom pertama menunjukkan kategori dokumen yaitu BK (Bisnis Keuangan), BLP (Balap), BO (Bola), BT (BuluTangkis), TNS (Tenis), dan TNJ (Tinju). Label pada baris pertama menunjukkan jumlah kluster yaitu 2, 3, 4, dan 5. Sedangkan label pada baris kedua menunjukkan percobaan yang dilakukan yaitu percobaan 1 dan 2. Masing-masing kotak menunjukkan label kluster hasil pengelompokan. Label kluster didapat dari mayoritas label dokumen pada kategori yang bersangkutan. Jadi misalkan 140 dokumen bisnis keuangan memiliki label 1 dan 10 dokumen sisanya memiliki label 2 maka untuk dokumen bisnis keuangan, label kluster hasil pengelompokannya adalah 1. Hal ini sesuai dengan rumus *similarity* yang telah dijelaskan pada subbab 3.6.1.

Tabel 5.15. Hasil Percobaan Pengelompokan Dokumen ke Kluster yang Jumlahnya Melebihi Jumlah Kategori yang Dipakai dengan Teknik NMF

NMF	2		3		4		5	
	1	2	1	2	1	2	1	2
BK	1	1	3	2	4	3	4	4
BLP	2	2	1	1	3	4	2	1
BO	2	2	1	1	2	4	5	2
BT	2	2	1	1	3	2	2	2
TNS	2	2	1	1	3	2	2	2
TNJ	2	2	2	3	1	1	1	2
Akurasi	96.67	96.33	95.67	95.67	89.67	90.33	87.00	78.33

Tabel 5.16. Hasil Percobaan Pengelompokan Dokumen ke Kluster yang Jumlahnya Melebihi Jumlah Kategori yang Dipakai dengan Teknik RP

RP	2		3		4		5	
	1	2	1	2	1	2	1	2
BK	2	1	2	3	3	3	1	4
BLP	1	2	3	3	2	4	4	2
BO	1	2	3	1	4	4	2	5
BT	1	2	3	2	4	4	2	2
TNS	1	2	3	2	4	4	2	2
TNJ	1	2	1	2	4	2	5	5
Akurasi	96.00	96.00	95.67	94.67	93.33	92.67	82.67	86.33

Dari Tabel 5.15 dan Tabel 5.16, dapat dilihat bahwa pada saat jumlah kluster 2, dokumen bisnis keuangan selalu memiliki label kluster yang berbeda dengan dokumen olahraga. Hal ini menunjukkan bahwa dokumen bisnis keuangan dan olahraga dikelompokkan ke dalam kluster yang berbeda dengan rata-rata akurasi 96%. Untuk jumlah kluster 3, dokumen bisnis keuangan masih dikelompokkan ke kluster yang berbeda dengan dokumen olahraga, sedangkan pada dokumen olahraga, pola umum yang terlihat adalah dokumen bulutangkis dan tenis selalu berada dalam 1 kluster. Rata-rata akurasi untuk jumlah kluster ini masih cukup tinggi yaitu 95%. Hal ini menunjukkan bahwa pola pengelompokan hampir sama dengan pengelompokan manual dimana dokumen bulutangkis dan tenis memang akan dikelompokkan ke dalam 1 kluster.

Untuk jumlah kluster 4, pola yang sama masih terlihat yaitu dokumen bisnis keuangan masih membentuk kluster sendiri dan dokumen bulutangkis dan tenis juga masih membentuk kluster sendiri. Dokumen balap, bola, dan tinju yang pada awalnya masih dalam berada satu kelompok dengan dokumen bulutangkis dan tenis, mulai terpisah dari kelompok tersebut.

Dengan pola yang muncul pada jumlah kluster 2 sampai dengan 4, pada saat jumlah kluster 5 seharusnya kluster yang terbentuk adalah kluster bisnis keuangan, balap, bola, bulutangkis dan tenis, serta tinju. Namun, ternyata percobaan

memperlihatkan bahwa ada beberapa kluster yang kosong yang artinya pembagian kelompok tidak seperti yang diperkirakan sebelumnya. Ada dokumen balap yang masuk dalam kelompok bulutangkis dan tenis, atau dokumen bola dan tinju menjadi satu kelompok, dsb. Ini menunjukkan bahwa fitur yang digunakan masih belum cukup untuk membedakan dokumen-dokumen tersebut.

Dari percobaan ini dengan akurasi yang cukup tinggi menunjukkan bahwa pemakaian fitur dalam membedakan satu dokumen dengan dokumen masih cukup efektif. Dengan memakai kemunculan fitur sebagai dasar dalam membedakan dokumen, pengelompokan yang cukup baik masih bisa diperoleh.

5.9. Rangkuman Hasil Percobaan

Subbab ini berisi rangkuman dari hasil semua percobaan yang dilakukan. Penjelasan masing-masing hasil percobaan dijelaskan pada subbab 5.2 sampai dengan subbab 5.8. Beberapa hal yang dapat dirangkum adalah

1. Baik untuk teknik Nonnegative Matrix Factorization maupun Random Projection, penghapusan *stopwords* dari fitur yang digunakan menyebabkan kenaikan akurasi pengelompokan dokumen (lihat subbab 5.2.1). Disamping itu, penggunaan fitur yang semakin banyak juga berpengaruh pada kenaikan akurasi pengelompokan dokumen (lihat subbab 5.2.2). Informasi fitur yang paling efektif digunakan dalam pengelompokan dokumen adalah *presence* (lihat subbab 5.2.3).
2. Untuk teknik Nonnegative Matrix Factorization, dengan jumlah iterasi yang memungkinkan konvergensi dicapai, akurasi akan memiliki nilai yang lebih tinggi dibandingkan dengan jumlah iterasi yang belum cukup membuat konvergensi dicapai. Di sisi lain, percobaan terhadap nilai *lambda* tidak memberikan tanda bahwa semakin besar nilai *lambda*, akurasi yang dihasilkan semakin tinggi atau sebaliknya. Namun, satu hal yang bisa disimpulkan adalah untuk mencapai nilai akurasi yang cukup tinggi dengan jumlah iterasi yang lebih sedikit dapat dilakukan dengan memperbesar nilai *lambda* (lihat subbab 5.3.1).
3. Untuk teknik Random Projection, penggunaan tipe distribusi 2 memberikan hasil akurasi yang cenderung lebih baik. Disamping itu, pengurangan dimensi

menyebabkan penurunan nilai akurasi. Namun, untuk jumlah tertentu (yang cukup signifikan besar), akurasi yang dihasilkan masih cukup baik (lihat subbab 5.3.2).

4. Semakin banyak dokumen yang digunakan, akurasi semakin tinggi. Namun, jumlah kluster yang semakin banyak menyebabkan penurunan akurasi (lihat subbab 5.4.1). Disamping itu, semakin besar rasio ketidakseragaman ukuran kluster, maka penurunan akurasi semakin besar. Namun, akurasi akan menjadi suatu nilai konstan bila rasio ketidakseragaman terlalu besar (lihat subbab 5.4.2).
5. Teknik Nonnegative Matrix Factorization memberikan akurasi yang lebih baik dibandingkan Random Projection maupun K-Means (lihat subbab 5.5).
6. Pengelompokan dokumen dalam domain dengan tingkat kemiripan kluster tinggi memberikan akurasi yang lebih tinggi dibandingkan dengan pengelompokan dokumen dalam domain dengan tingkat kemiripan kluster rendah (lihat subbab 5.6).
7. Pengelompokan dokumen dimana dokumen berasal dari sumber yang berbeda memberikan akurasi yang lebih rendah dibandingkan pengelompokan dokumen dari sumber yang sama (lihat subbab 5.7).
8. Pemakaian fitur (yaitu *unigram*) dalam membedakan dokumen satu dengan dokumen lain cukup efektif. Hal ini ditunjukkan pada hasil percobaan pada subbab 5.8 dan keseluruhan percobaan yang dilakukan. Secara umum, pengelompokan dokumen dengan teknik reduksi dimensi Nonnegative Matrix Factorization dan Random Projection pada dokumen bahasa Indonesia memberikan hasil yang cukup baik (hampir sama dengan penerapannya pada dokumen bahasa Inggris).