SKRIPSI

PERBANDINGAN FUZZY C-MEANS DAN K-MEANS UNTUK TEXT CLUSTERING MENGGUNAKAN LSI

Dini Puspita Sukma Ariyanti

NPM: 2016730084

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN
2021

UNDERGRADUATE THESIS

COMPARISON OF FUZZY C-MEANS AND K-MEANS FOR TEXT CLUSTERING USING LSI

Dini Puspita Sukma Ariyanti

NPM: 2016730084

DEPARTMENT OF INFORMATICS
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES
PARAHYANGAN CATHOLIC UNIVERSITY
2021

LEMBAR PENGESAHAN

PERBANDINGAN FUZZY C-MEANS DAN K-MEANS UNTUK TEXT CLUSTERING MENGGUNAKAN LSI

Dini Puspita Sukma Ariyanti

NPM: 2016730084

Bandung, 2 Februari 2021

Menyetujui,

Pembimbing

Dr.rer.nat. Cecilia Esti Nugraheni

Ketua Tim Penguji	Anggota Tim Penguji
Luciana Abednego, M.T.	**Mariskha Tri Adithia, P.D.Eng**

Mengetahui,

Ketua Program Studi

Mariskha Tri Adithia, P.D.Eng

PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

PERBANDINGAN FUZZY C-MEANS DAN K-MEANS UNTUK TEXT CLUSTERING MENGGUNAKAN LSI

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
Tanggal 2 Februari 2021

Dini Puspita Sukma Ariyanti
NPM: 2016730084

ABSTRAK

Salah satu cara untuk meningkatkan efektivitas dan efisiensi dalam pemrosesan data adalah dengan melakukan *clustering*. Data yang akan digunakan dalam penelitian ini berupa teks. *Text clustering* dilakukan untuk mengelompokkan dokumen berdasarkan kemiripan topik yang dibahas dalam teks-teks tersebut. Teks yang memiliki kemiripan topik akan berada dalam satu *cluster*.

Kumpulan dokumen akan dibersihkan dengan melakukan *text pre-processing*. Setelah itu kumpulan dokumen yang sudah melalui *text pre-processing* akan dimodelkan dengan menggunakan *Vector Space Model* sehingga terbentuk matriks *document-term*. Matriks *document-term* memiliki dimensi yang besar. *Latent Semantic Indexing* (LSI) akan digunakan untuk mengurangi dimensi vektor matriks dan merepresentasikan dokumen ke dalam konsep (bukan kata-kata). Hasil dari LSI nantinya akan digunakan untuk melakukan *clustering*. *Clustering* dokumen akan dilakukan dengan metode *Fuzzy C-Means*. Hasil *clustering* dokumen dengan *Fuzzy C-Means* akan dibandingkan dengan hasil *clustering* dokumen dengan *K-Means* dengan parameter jarak *intercluster*, jarak *intracluster*, dan waktu yang dibutuhkan dalam pemrosesan.

Dalam penelitian ini, akan dibuat sebuah *desktop application* dengan menggunakan bahasa Java. Perangkat lunak tersebut mengimplementasikan Algoritma LSI, Fuzzy C-Means, dan K-Means. Pengujian dibagi menjadi dua bagian, yaitu pengujian fungsionalitas untuk melihat apakah masukan dari pengguna direspon dengan baik oleh program, dan pengujian performa untuk mengukur bagaimana hasil dari algoritma yang diimplementasikan.

Hasil pengujian menunjukan secara performa LSI-FCM bekerja lebih baik dibanding LSI-KMeans. Hasil *text clustering* dengan menggunakan algoritma LSI-FCM setiap anggotanya memiliki jarak yang lebih dekat dengan titik pusat *cluster*nya dibanding dengan menggunakan algoritma LSI dan *K-Means*. Tetapi secara kecepatan LSI-FCM bekerja lebih lambat dibanding LSI-KMeans.

Kata-kata kunci: *Text Clustering*, *Latent Semantic Indexing*, LSI, *Fuzzy C-Means*, FCM, *K-Means*

ABSTRACT

One of the ways to improve the effectiveness and efficiency in data processing is do clustering. Data that will be used in this research is text file. Text clustering is used to group documents based on the similarity of topics that discussed in these texts.Text file that has similar topic will be in one cluster.

The document set will be cleaned by doing text pre-processing. After that, the document set that has been through text pre-processing will be modeled using Vector Space Model so that the document-term matrix is formed. The document-term matrix has a large dimension. Latent Semantic Indexing (LSI) will be used to reduce matrix vector dimensions and represent documents into concepts (not words). The result will be used to do clustering. Text Clustering will be done by Fuzzy C-Means method. The result of text clustering with Fuzzy C-Means will be compared with the result of text clustering with *K-Means* by parameter intercluster distance, intracluster distance, and execution time.

In this research a desktop application will be created using Java language. The software implements LSI, Fuzzy C-Means, and K-Means Algorithms. Testing is divided into two parts, functional testing to see if input from user is responded well by the program, and performance testing to measure how the results of algorithms are implemented.

The results showed that LSI-FCM performed better than LSI-KMeans. The result of text clustering with LSI-FCM algorithm showed that each data has a closer distance to the centroid than using LSI-KMeans algorithm. But LSI-FCM works slower than LSI-KMeans.

Keywords: *Text Clustering*, *Latent Semantic Indexing*, LSI, *Fuzzy C-Means*, FCM, *K-Means*

Dipersembahkan untuk Tuhan YME, keluarga, teman-teman, serta diri sendiri.

KATA PENGANTAR

Puji syukur penulis panjatkan kepada Tuhan Yang Maha Esa, karena dengan rahmat dan karunia-Nya, penulis dapat menyelesaikan penyusunan skripsi berjudul "Perbandingan *Fuzzy C-Means* dan *K-Means* untuk *Text Clustering* Menggunakan LSI" sebagai salah satu tugas akhir dalam melengkapi persyaratan dalam menyelesaikan akademik serta memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Universitas Katolik Parahyangan. Selain itu, penulisan skripsi ini bertujuan untuk memberikan pengetahuan kepada pembaca mengenai *text clustering* dengan *Fuzzy C-Means* dan *K-Means* menggunakan *Latent Semantic Indexing*. Selama penulisan skripsi ini, penulis menyadari bahwa penulisan skripsi ini dapat selesai karena bantuan dan dukungan beberapa pihak. Oleh karena itu, penulis mengungkapkan rasa terima kasih kepada:

1. Dr.rer.nat. Cecilia Esti Nugraheni selaku dosen pembimbing yang telah membimbing dan mendukung penulis selama proses penyusunan skripsi ini.
2. Ibu Luciana Abednego, M.T. dan Ibu Mariskha Tri Adithia, P.D.Eng. selaku dosen penguji yang telah memberikan kritik dan saran yang membangun selama penelitian dan juga penyusunan skripsi ini.
3. Keluarga yang selalu memberikan dorongan serta bantuan sehingga penulis dapat menyelesaikan skripsi ini dengan baik.
4. Dian, Lara, Jojo, Zaki, Chrissandi, dan Vinson sebagai rekan seperjuangan selama menempuh studi S1 dan telah memberikan saran serta bantuan dalam penyusunan skripsi.
5. Anna, Izi, Eto, dan Nurfan yang telah menemani selama penyusunan skripsi ini.
6. Seluruh pihak terkait yang telah memberikan motivasi, doa, saran dan perhatiannya, sehingga penulis dapat menyelesaikan skripsi ini dengan baik dan tepat pada waktunya.

Penulis menyadari bahwa skripsi ini masih jauh dari kata sempurna. Oleh karena itu, penulis memohon maaf jika terdapat kekurangan pada skripsi ini. Penulis juga mengharapkan kritik dan saran yang membangun untuk menyempurnakan skripsi ini. Semoga skripsi ini dapat bermanfaat bagi segenap pihak yang berkepentingan.

Bandung, Februari 2021

Penulis

DAFTAR ISI

Kata Pengantar xv

Daftar Isi xvii

Daftar Gambar xix

Daftar Tabel xxi

Daftar Kode Program xxiii

1 Pendahuluan 1
- 1.1 Latar Belakang . . . 1
- 1.2 Rumusan Masalah . . . 2
- 1.3 Tujuan . . . 2
- 1.4 Batasan Masalah . . . 3
- 1.5 Metodologi . . . 3
- 1.6 Sistematika Pembahasan . . . 3

2 Landasan Teori 5
- 2.1 Text Mining[1] . . . 5
- 2.2 Vector Space Model (VSM) [1] . . . 6
  - 2.2.1 Term Frequency dan Inverse Document Frequency (TF-IDF) . . . 6
  - 2.2.2 Cosine Similarity . . . 7
- 2.3 Latent Semantic Indexing (LSI) [2] . . . 7
  - 2.3.1 Singular Value Decomposition [3] . . . 8
  - 2.3.2 Algoritma *Singular Value Decomposition* . . . 9
- 2.4 Fuzzy C-Means [4] . . . 10
- 2.5 K-Means[4] . . . 13
- 2.6 Penilaian Performa[4] . . . 14
  - 2.6.1 Precision, Recall, dan Accuracy . . . 14
  - 2.6.2 F-Measurement . . . 15
  - 2.6.3 Jarak Intracluster dan Intercluster . . . 15
- 2.7 Library Java . . . 15
  - 2.7.1 Library Text Mining . . . 15
  - 2.7.2 Library Jama (Java Matrix) . . . 15

3 Analisis 17
- 3.1 Analisis Masalah . . . 17
  - 3.1.1 Text Pre-processing . . . 17
  - 3.1.2 Reduksi Dimensi dengan *Singular Value Decomposition* . . . 18
  - 3.1.3 Text Clustering dengan LSI dan Fuzzy C-Means . . . 20
  - 3.1.4 Text Clustering dengan LSI dan K-Means . . . 22
- 3.2 Gambaran Umum Perangkat Lunak . . . 24

3.2.1 Diagram Aktivitas . . . 24
3.3 Diagram Kelas . . . 25

4 Perancangan **29**
4.1 Perancangan Antarmuka Text Clustering . . . 29
4.2 Penjelasan Kelas . . . 30
4.2.1 Kelas MainFrame . . . 30
4.2.2 Kelas Term . . . 31
4.2.3 Kelas TermList . . . 31
4.2.4 Kelas StopWordList . . . 32
4.2.5 Kelas Document . . . 33
4.2.6 Kelas DocumentReader . . . 34
4.2.7 Kelas Tokenizer . . . 35
4.2.8 Kelas Weight . . . 36
4.2.9 Kelas Matrix . . . 37
4.2.10 Kelas SingularValueDecomposition . . . 38
4.2.11 Kelas DimensionReduction . . . 40
4.2.12 Kelas CosineDistance . . . 41
4.2.13 Kelas FuzzyCMeans . . . 41
4.2.14 Kelas KMeans . . . 42
4.3 Masukan dan Keluaran Perangkat Lunak . . . 43

5 Pengujian **45**
5.1 Implementasi Antarmuka . . . 45
5.2 Pengujian Fungsional . . . 48
5.2.1 Teknik LSI dan Fuzzy C-Means . . . 48
5.2.2 Teknik LSI dan K-Means . . . 49
5.3 Pengujian Eksperimental . . . 51
5.3.1 Pengujian Eksperimen Algoritma LSI dan Fuzzy C-Means . . . 52
5.3.2 Pengujian Eksperimen Algoritma LSI dan K-Means . . . 54
5.3.3 Analisis Perbandingan . . . 57

6 Kesimpulan dan Saran **59**
6.1 Kesimpulan . . . 59
6.2 Saran . . . 59

Daftar Referensi **61**

A Kode Program **63**

B Hasil Eksperimen **95**

DAFTAR GAMBAR

1.1 *Clustering* Data 1

2.1 Matriks *Document-Term* 6
2.2 Ilustrasi matriks A yang telah didekomposisi dan direduksi 8

3.1 Diagram Aktivitas Perangkat Lunak *Text Clustering* 25
3.2 Diagram Kelas 26

4.1 Tampilan Antarmuka *Text Clustering* 29
4.2 Tampilan Antarmuka *Distance* 30
4.3 Diagram Kelas MainFrame 30
4.4 Diagram Kelas Term 31
4.5 Diagram Kelas TermList 31
4.6 Diagram Kelas StopWordList 32
4.7 Diagram Kelas Document 33
4.8 Diagram Kelas DocumentReader 34
4.9 Diagram Kelas Tokenizer 35
4.10 Diagram Kelas Weight 36
4.11 Diagram Kelas Matrix 37
4.12 Diagram Kelas SingularValueDecomposition 38
4.13 Diagram Kelas DimensionReduction 40
4.14 Diagram Kelas CosineDistance 41
4.15 Diagram Kelas FuzzyCMeans 41
4.16 Diagram Kelas KMeans 42

5.1 Tampilan Utama Program *Text Clustering* 45
5.2 Tampilan untuk melakukan *Input Data* 46
5.3 Tampilan Utama Setelah *Input Data* 46
5.4 Tampilan Hasil *Text Clustering* dengan metode FCM dan SVD 47
5.5 Tampilan Jarak Titik Pusat *Cluster* dengan Titik Data 47
5.6 Hasil Uji Fungsional pada *Text Clustering* dengan LSI-FCM 48
5.7 Derajat Keanggotaan setiap Data pada *Text Clustering* dengan LSI-FCM 49
5.8 Hasil Uji Fungsional pada *Text Clustering* dengan LSI dan K-Means 50
5.9 Jarak *Intracluster* pada *Text Clustering* dengan LSI dan K-Means 50
5.10 Hasil 1 57
5.11 Plot Perbandingan Varian Jarak *Intracluster* FCM dan K-Means 58

DAFTAR TABEL

2.1 *Confusion Matrix* 14

3.1 Contoh Hasil *Text Clustering* LSI-FCM 22
3.2 Contoh Hasil *Text Clustering* LSI-KMeans 23

4.1 Contoh Tabel Hasil *Text Clustering* 44
4.2 Contoh Tabel *distance* 44

5.1 Contoh Hasil *Text Clustering* Manual dan Perangkat Lunak LSI-FCM 49
5.2 Contoh Hasil *Text Clustering* Manual dan Perangkat Lunak LSI-FCM 51
5.3 Tabel Dataset 51
5.4 10 Baris Pertama Tabel *distance* dengan metode LSI dan FCM pada Pengujian 1 . 52
5.5 Tabel Hasil *Text Clustering* dengan LSI dan FCM pada Pengujian 1 52
5.6 Tabel Rata-Rata Waktu Eksekusi LSI-FCM 53
5.7 Tabel Varian Jarak *Intracluster* LSI dan FCM pada Pengujian 1 53
5.8 Tabel Rata-Rata Varian Jarak *Intracluster* LSI-FCM 53
5.9 *Confusion Matrix* LSI dan FCM pada Pengujian 1 53
5.10 Tabel Rata-rata F1 dan *Accuracy* 54
5.11 Tabel *distance* dengan metode LSI dan K-Means pada Pengujian 1 55
5.12 Tabel Hasil *Text Clustering* dengan LSI dan K-Means pada Pengujian 1 55
5.13 Tabel Rata-Rata Waktu Eksekusi LSI dan K-Means 55
5.14 Tabel Varian Jarak *Intracluster* LSI-K-Means pada Pengujian 1 56
5.15 Tabel Rata-Rata Varian Jarak *Intracluster* LSI-FCM 56
5.16 *Confusion Matrix* LSI-K-Means pada Pengujian 1 56
5.17 Tabel Rata-rata F1 dan *Accuracy* 57
5.18 Tabel Perbandingan Varian Jarak *Intracluster* FCM dan K-Means 58
5.19 Tabel Perbandingan *Confusion Matrix* 58

B.1 Tabel *distance* dengan metode LSI dan FCM pada Pengujian 1 95
B.2 Tabel Hasil *Text Clustering* dengan LSI dan FCM pada Pengujian 1 96
B.3 Tabel Varian Jarak *Intracluster* LSI dan FCM pada Pengujian 1 96
B.4 *Confusion Matrix* LSI dan FCM pada Pengujian 1 96
B.5 Tabel *distance* dengan metode LSI dan FCM pada Pengujian 2 97
B.6 Tabel Hasil *Text Clustering* dengan LSI dan FCM pada Pengujian 2 97
B.7 Tabel Varian Jarak *Intracluster* LSI dan FCM pada Pengujian 2 98
B.8 *Confusion Matrix* LSI dan FCM pada Pengujian 2 98
B.9 Tabel *distance* dengan metode LSI dan FCM pada Pengujian 3 98
B.10 Tabel Hasil *Text Clustering* dengan LSI dan FCM pada Pengujian 3 99
B.11 Tabel Varian Jarak *Intracluster* LSI dan FCM pada Pengujian 3 99
B.12 *Confusion Matrix* LSI dan FCM pada Pengujian 3 99
B.13 Tabel *distance* dengan metode LSI dan FCM pada Pengujian 4 100
B.14 Tabel Hasil *Text Clustering* dengan LSI dan FCM pada Pengujian 4 100
B.15 Tabel Varian Jarak *Intracluster* LSI dan FCM pada Pengujian 4 100

B.16 *Confusion Matrix* LSI dan FCM 101
B.17 Tabel *distance* dengan metode LSI dan FCM pada Pengujian 5 101
B.18 Tabel Hasil *Text Clustering* dengan LSI dan FCM pada Pengujian 5 101
B.19 Tabel Varian Jarak *Intracluster* LSI dan FCM pada Pengujian 5 102
B.20 *Confusion Matrix* LSI dan FCM pada Pengujian 5 102
B.21 Tabel *distance* dengan metode LSI dan K-Means pada Pengujian 1 102
B.22 Tabel Hasil *Text Clustering* dengan LSI dan K-Means pada Pengujian 1 103
B.23 Tabel Varian Jarak *Intracluster* LSI-K-Means pada Pengujian 1 103
B.24 *Confusion Matrix* LSI-K-Means pada Pengujian 1 103
B.25 Tabel *distance* dengan metode LSI dan K-Means pada Pengujian 2 104
B.26 Tabel Hasil *Text Clustering* dengan LSI-K-Means pada Pengujian 2 104
B.27 Tabel Varian Jarak *Intracluster* LSI-K-Means pada Pengujian 2 104
B.28 *Confusion Matrix* LSI-K-Means pada Pengujian 2 105
B.29 Tabel *distance* dengan metode LSI dan K-Means pada Pengujian 3 105
B.30 Tabel Hasil *Text Clustering* dengan LSI-K-Means pada Pengujian 3 105
B.31 Tabel Varian Jarak *Intracluster* LSI-K-Means pada Pengujian 3 106
B.32 *Confusion Matrix* LSI-K-Means pada Pengujian 3 106
B.33 Tabel *distance* dengan metode LSI dan K-Means pada Pengujian 4 106
B.34 Tabel Hasil *Text Clustering* dengan LSI dan K-Means pada Pengujian 4 107
B.35 Tabel Varian Jarak *Intracluster* LSI dan K-Means pada Pengujian 4 107
B.36 *Confusion Matrix* LSI dan K-Means pada Pengujian 4 107
B.37 Tabel *distance* dengan metode LSI dan K-Means pada Pengujian 5 108
B.38 Tabel Hasil *Text Clustering* dengan LSI dan K-Means pada Pengujian 5 108
B.39 Tabel Varian Jarak *Intracluster* LSI dan K-Means pada Pengujian 5 108
B.40 *Confusion Matrix* LSI dan K-Means pada Pengujian 5 109

DAFTAR KODE PROGRAM

A.1 Main.java 63
A.2 Term.java 68
A.3 TermList.java 69
A.4 StopWordList.java 69
A.5 Document.java 70
A.6 DocumentReader.java 71
A.7 Tokenizer.java 72
A.8 Weight.java 73
A.9 Matrix.java 74
A.10 SingularValueDecomposition.java 84
A.11 DimensionReduction.java 89
A.12 CosineDistance.java 90
A.13 FuzzyCMeans.java 91
A.14 KMeans.java 92

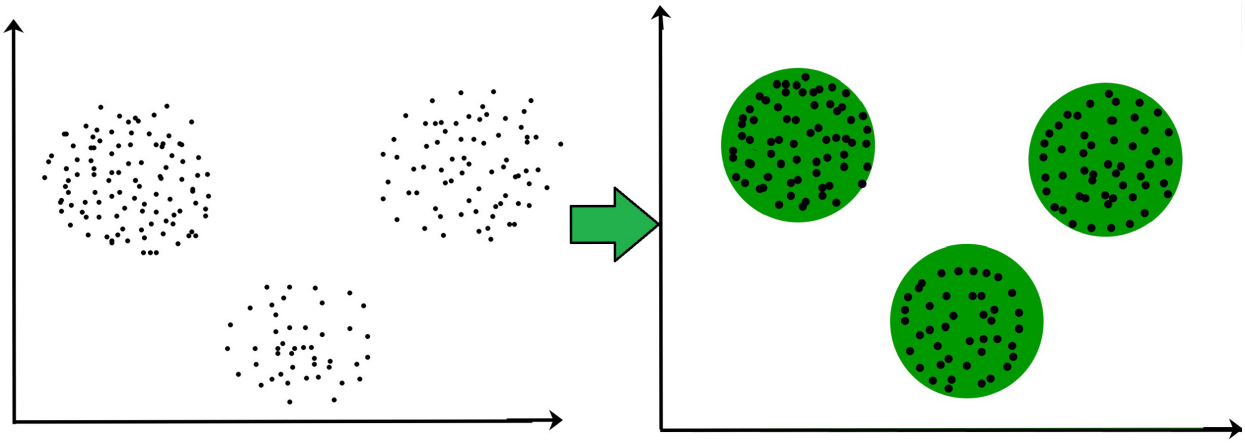
BAB 1

PENDAHULUAN

1.1 Latar Belakang

Saat ini keberagaman data semakin meningkat seiring perkembangan teknologi. Hal tersebut memberikan tantangan untuk mengatur keberagaman data secara efektif dan efisien. Salah satu cara untuk meningkatkan efektivitas dan efisiensi dalam pemrosesan data tersebut adalah dengan melakukan *clustering*.

Clustering merupakan pengelompokan sejumlah data berdasarkan kemiripan karakteristik yang dimiliki[3]. Data akan dibagi ke dalam beberapa *cluster*. Setiap *cluster* akan berisi data dengan karakteristik semirip mungkin. Gambar 1.1 menunjukkan data yang dikelompokkan menjadi *cluster* berdasarkan kemiripan karakteristik tertentu.



Gambar 1.1: *Clustering* Data

Banyak data yang dapat dikelompokkan menjadi *cluster*. Salah satunya berupa dokumen teks. Dokumen teks memiliki banyak format, beberapa contohnya adalah *Portable Document Format* (.pdf), *Microsoft Word Document* (.doc atau .docx), *Web Document* (.html), dan *Plain Text* (.txt). *Text clustering* akan dilakukan untuk mengelompokkan dokumen berdasarkan kemiripan topik yang dibahas dalam teks-teks tersebut. Teks yang memiliki kemiripan topik akan berada dalam satu *cluster*.

Sebelum diproses, dokumen teks akan dimodelkan dengan *Vector Space Model* (VSM). VSM merupakan pendekatan dengan merepresentasikan dokumen menjadi sebuah vektor. VSM memiliki beberapa kekurangan, seperti:

- Kata yang terdaftar sangat banyak sehingga menyebabkan dimensi vektor yang besar.
- Tidak memperhatikan kemiripan antara dua kata.

Latent Semantic Indexing (LSI) akan digunakan untuk mengurangi dimensi vektor dan merepresentasikan dokumen ke dalam konsep (bukan kata-kata). LSI adalah pendekatan yang digunakan untuk memproses persamaan kata atau sinonim dan memahami hubungan konsep antar kata. Hasil dari LSI nantinya akan digunakan untuk melakukan *clustering*.

Ada beberapa algoritma untuk melakukan *clustering* data, di antaranya adalah sebagai berikut:

1. ***Fuzzy C-Means*** yang merupakan metode pengelompokan data yang keberadaannya ditentukan oleh nilai derajat keanggotaan tertentu. Derajat keanggotaan sebuah data bernilai diantara 0 sampai 1.

2. ***K-Means*** yang merupakan metode pengelompokan data dengan menghitung jarak setiap data dengan titik pusat tiap *cluster* yang sudah ditentukan.

Clustering dokumen akan dilakukan dengan metode *Fuzzy C-Means*. Hasil *clustering* dokumen dengan *Fuzzy C-Means* akan dibandingkan dengan hasil *clustering* dokumen dengan *K-Means* dengan parameter jarak *intercluster*, jarak *intracluster*, dan waktu yang dibutuhkan dalam pemrosesan.

Pada skripsi ini, akan dibuat sebuah perangkat lunak untuk melakukan *text clustering* dengan *Fuzzy C-Means* dan *K-Means* menggunakan LSI. Data yang akan digunakan pada skripsi ini berupa *plain text* (.txt). Bahasa yang digunakan dalam pembuatan perangkat lunak ini adalah bahasa pemrograman *Java*. Perangkat lunak yang dibuat dapat menghasilkan *cluster* yang dibentuk, dokumen dalam setiap *cluster*, waktu eksekusi, dan jarak intercluster.

1.2 Rumusan Masalah

Berdasarkan latar belakang, rumusan masalah dari skripsi ini adalah sebagai berikut:

1. Bagaimana melakukan pendekatan *Latent Semantic Indexing* (LSI) untuk mereduksi dimensi pada VSM?

2. Bagaimana cara pengelompokan dokumen dengan menggunakan *Fuzzy C-Means* dan *K-Means* setelah matriks direduksi?

3. Bagaimana cara membandingkan metode *Fuzzy C-Means* dan *K-Means*?

4. Bagaimana perbandingan hasil pengelompokan dokumen yang diperoleh dengan menggunakan *Fuzzy C-Means* dengan *K-Means*?

1.3 Tujuan

Berdasarkan rumusan masalah, tujuan dari skripsi ini adalah sebagai berikut:

1. Mempelajari reduksi dimensi pada VSM sehingga mendapatkan representasi dokumen yang dimensinya lebih kecil dengan LSI.

2. Mempelajari pengelompokan dokumen dengan menggunakan *Fuzzy C-Means* dan *K-Means*.

3. Membangun perangkat lunak untuk melakukan *text clustering* dengan *Fuzzy C-Means* dan *K-Means* menggunakan LSI.

4. Membandingkan hasil pengelompokan dokumen yang diperoleh dengan menggunakan *Fuzzy C-Means* dengan *K-Means* dengan melihat jarak *intercluster*, jarak *intracluster*, dan waktu yang dibutuhkan dalam pemrosesan.

1.4 Batasan Masalah

Batasan-batasan masalah dari skripsi ini adalah sebagai berikut:

- Dokumen teks yang diproses merupakan dokumen .txt.
- Dokumen teks yang diproses berjumlah 25 dokumen.

1.5 Metodologi

Metodologi yang digunakan dalam penyusunan skripsi ini adalah sebagai berikut:

1. Melakukan studi literatur mengenai *Vector Space Model* (VSM).
2. Melakukan studi literatur mengenai *Latent Semantic Indexing* (LSI).
3. Melakukan studi literatur mengenai *Fuzzy C-Means*.
4. Melakukan studi literatur mengenai *K-means*.
5. Melakukan pencarian data set.
6. Melakukan analisis mengenai reduksi dimensi menggunakan *Singular Value Decomposition* untuk LSI.
7. Melakukan analisis mengenai *clustering* dokumen dengan menggunakan *Fuzzy C-Means*.
8. Melakukan analisis mengenai *clustering* dokumen dengan menggunakan *K-Means*.
9. Membuat desain perangkat lunak yang akan dibuat.
10. Mengimplementasikan perangkat lunak yang akan dibuat.
11. Melakukan perbandingan kinerja sistem yang dibuat dengan *Fuzzy C-Means* dan *K-Means* dengan parameter :
    - jarak *intercluster*,
    - jarak *intracluster*,
    - waktu yang dibutuhkan.
12. Menarik kesimpulan berdasarkan hasil perbandingan.

1.6 Sistematika Pembahasan

Skripsi ini akan tersusun dalam enam bab secara sistematis. Enam bab tersebut terdiri dari pendahuluan, landasan teori, analisis, perancangan, pengujian, serta kesimpulan. Berikut merupakan sistematika pembahasan dalam skripsi ini.

1. Bab 1 Pendahuluan
   Bab 1 berisi latar belakang, rumusan masalah, tujuan, batasan masalah, metodologi, dan sistematika pembahasan.
2. Bab 2 Landasan Teori
   Bab 2 berisi dasar teori mengenai *Text Mining*, *Vector Space Model*, *Latent Semantic Indexing*, *Fuzzy C-Means*, *K-means*, dan penilaian performa yang membahas ukuran apa saja yang digunakan untuk menilai performa program. dari hasil pencarian.

3. Bab 3 Analisis
   Bab 3 berisi analisis masalah, analisis mengenai reduksi dimensi menggunakan *Singular Value Decomposition* untuk LSI, analisis mengenai *clustering* dokumen dengan menggunakan *Fuzzy C-Means* dan *K-Means*, analisis mengenai *clustering* dokumen dengan menggabungkan algoritma LSI dengan *Fuzzy C-Means* dan algoritma LSI dengan *K-Means*.

4. Bab 4 Perancangan
   Bab 4 berisi perancangan antarmuka pogram *text clustering*, diagram kelas beserta dengan penjelasannya.

5. Bab 5 Pengujian
   Bab 5 berisi hasil pengujian untuk melihat apakah program sudah berfungsi dengan benar. Pengujian dibagi menjadi dua bagian, yaitu pengujian fungsional dan pengujian eksperimental.

6. Bab 6 Kesimpulan dan Saran
   Bab 6 berisi tentang kesimpulan yang didapatkan dari pengujian ini serta membahas saran untuk penelitian ini.