Jurnal Ilmu Komputer & Agri-informatika

Tersedia secara *online* di:
http://journal.ipb.ac.id/index.php/jika

Volume 10 Nomor 2 halaman 121 – 132
e ISSN: 2654-9735, p ISSN: 2089-6026

Pengelompokan Publikasi Ilmiah Berdasarkan Bidang Kepakaran Menggunakan Latent Dirichlet Allocation dan Normalized PSO-K-means

Clustering of Scientific Publications Based on Field of Expertise Using Latent Dirichlet Allocation and Normalized PSO-K-means

FINA CHARISMA HAYATINA[1]*, SONY HARTONO WIJAYA[1], MEDRIA KUSUMA DEWI HARDHIENATA[1]

Abstrak

Salah satu cara untuk memvalidasi klaim kepakaran dosen adalah dengan meninjau dokumen publikasi ilmiah yang tersedia. Namun, menentukan kelompok kepakaran dari sejumlah dokumen memerlukan pengetahuan yang memadai dan waktu yang relatif lama, sehingga menjadi sulit dilakukan. Penelitian ini bertujuan untuk membangun suatu model yang dapat mengelompokkan dokumen berdasarkan bidang kepakaran. Penelitian ini menggunakan algoritma klasterisasi K-means untuk mengelompokkan dokumen berdasarkan bidang kepakaran dosen. *Latent dirichlet allocation* digunakan untuk mereduksi dimensi data, dan *particle swarm optimization* digunakan untuk menentukan *centroid* awal pada algoritma K-means. Hasil penelitian ini berhasil mengelompokkan dokumen publikasi ilmiah dengan nilai koefisien *silhouette* sebesar 0.42. Selain itu, penggunaan PSO sebagai penentu *centroid* optimal pada algoritma K-means dapat meningkatkan nilai koefisien *silhouette* sebesar 5.56%. Model yang dibangun dievaluasi dengan mencocokkan klaster yang dihasilkan dengan klaim yang diberikan. Hasilnya menunjukkan bahwa sebanyak 75% hasil pencocokan sesuai dan 25% tidak sesuai.

Kata Kunci**:** kepakaran, k-means, *latent dirichlet allocation*, *particle swarm optimization*, publikasi ilmiah.

Abstract

Validating a lecturer's expertise claims often involves scrutinizing their scholarly publications. However, this process can be quite demanding, requiring significant knowledge and time due to the need to assess numerous documents. To address this challenge, this study endeavors to create a model that can categorize documents based on their areas of expertise. The study employs the K-means clustering algorithm to group documents according to the lecturers' fields of expertise. In order to enhance the efficiency of this process, Latent Dirichlet Allocation is utilized to reduce data dimensions. Additionally, Particle Swarm Optimization is used to determine the optimal initial cluster centers for the K-means algorithm. The research yielded promising results, successfully categorizing scholarly publications with a silhouette coefficient of 0.42. Furthermore, by using PSO to identify the optimal cluster centers, the silhouette coefficient was improved by 5.56%. The model's performance was evaluated by comparing the resulting clusters with the provided claims, showing a 75% matching rate and a 25% non-matching rate.

*Keywords***:** *expertise, K-means, latent dirichlet allocation, particle swarm optimization, scientific publications.*

[1] Departemen Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, Bogor 16680;
* Penulis Korespondensi: Tel/Faks: 0821-10932567; Surel: finacharisma@apps.ipb.ac.id

PENDAHULUAN

Pakar atau ahli (*expert*) merujuk pada individu yang memiliki keahlian khusus yang tidak umum dimiliki oleh kebanyakan orang (Rosnelly 2012). Di lingkungan perguruan tinggi, keberadaan seorang pakar diperlukan dalam rangka pelaksanaan tridharma, kolaborasi, dan juga penentuan promotor untuk jenjang S3. Setiap pakar memiliki keahlian yang unik, yang dapat dikenali melalui dokumen yang paling mewakili profilnya (Campos *et al.* 2020).

Saat ini, platform-platform seperti Google Scholar dan *Science and Technology Index* (SINTA) menyediakan informasi terkait kepakaran di bidang akademik berdasarkan klaim dari individu yang bersangkutan. Validitas klaim ini dapat diuji dengan mengamati dokumen publikasi ilmiah yang terkait. Dalam menentukan kelompok kepakaran dalam kumpulan dokumen publikasi ilmiah, diperlukan pengetahuan yang cukup dan waktu yang relatif lama. Oleh karena itu, diperlukan pengembangan model yang dapat secara otomatis mengelompokkan dokumen publikasi ilmiah ke dalam berbagai bidang kepakaran.

Dokumen publikasi ilmiah merujuk pada kumpulan teks yang tidak terstruktur. Teknik *text mining* dapat diterapkan untuk mendapatkan informasi dari data teks tersebut. *Text mining* merupakan proses untuk mengekstraksi informasi dengan mengidentifikasi pola dan hubungan menarik dari kumpulan teks yang besar (Feldman dan Sanger 2007). Salah satu area penerapan *text mining* adalah dalam klasterisasi (Usai *et al.* 2018). Klasterisasi merupakan proses membagi sekelompok objek ke dalam klaster, di mana objek dalam satu klaster memiliki tingkat kemiripan yang tinggi, sedangkan objek antar klaster memiliki tingkat kemiripan yang rendah (Han *et al.* 2012).

K-means merupakan salah satu algoritma yang populer dalam pengelompokan dokumen (Chouhan dan Purohit 2018). Algoritma ini memiliki kelebihan, antara lain, sifatnya yang sederhana, waktu komputasi yang relatif cepat, dan kemudahan dalam implementasinya (Yuan dan Yang 2019). Dalam proses pengelompokan dokumen menggunakan K-means, dokumen diubah menjadi bentuk vektor dan dikelompokkan berdasarkan jarak terdekat. Namun, proses ini menghasilkan matriks berdimensi tinggi yang dapat mengakibatkan masalah yang dikenal sebagai "kutukan dimensi" (Bui *et al.* 2017). *Latent Dirichlet Allocation* (LDA) dapat digunakan untuk mengatasi permasalahan ini dengan mengubah dokumen ke dalam bentuk distribusi peluang topik. Setiap peluang topik pada dokumen memiliki nilai dalam rentang 0 hingga 1, dan total nilai dari semua peluang topik dalam dokumen harus sama dengan 1.

Beberapa penelitian telah dilakukan untuk mengatasi masalah "kutukan dimensi" pada K-means dengan LDA, seperti penelitian yang dilakukan oleh Saini *et al.* (2020) dan Bui *et al.* (2017). Namun, pada penelitian sebelumnya, pemilihan *centroid* awal pada K-means masih dilakukan secara acak, yang berpotensi menghasilkan lokal optimum. Chouhan dan Purohit (2018) menjalankan penelitian untuk mengatasi permasalahan lokal optimum pada K-means. Penelitian ini mengusulkan penggunaan *Particle Swarm Optimization* (PSO) tradisional untuk menemukan *centroid* awal yang optimal. Namun, PSO tradisional tidak dapat diterapkan untuk mencari *centroid* optimum dalam kasus sekumpulan dokumen yang berbentuk distribusi peluang topik karena adanya batasan dalam ruang solusi.

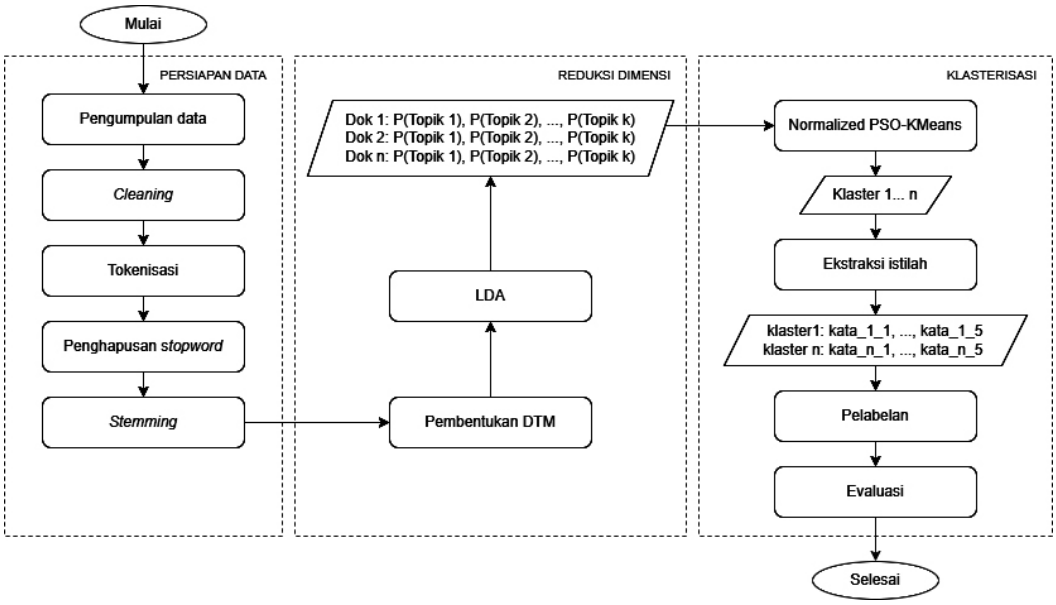
Dengan latar belakang yang telah diuraikan di atas, penelitian ini bertujuan untuk mengelompokkan dokumen publikasi ilmiah menggunakan algoritma K-means sebagai metode klasterisasi, LDA untuk mengurangi dimensi data, dan *Normalized* PSO sebagai alat untuk menentukan *centroid* awal yang optimal dalam algoritma K-means. Selain itu, klaster-klaster dokumen yang dihasilkan dalam penelitian ini diberi label berdasarkan bidang kepakaran yang sesuai dengan indeks *Universal Decimal Classification* (UDC). Penelitian ini diharapkan akan menghasilkan klaster-klaster dokumen berdasarkan bidang kepakaran, yang dapat digunakan sebagai alternatif untuk memvalidasi klaim seseorang terkait kepakaran dalam bidang akademik.

METODE

Pada penelitian ini terdapat tiga kelompok tahapan yang dilakukan yaitu persiapan data, reduksi dimensi data, dan klasterisasi. Detail tahapan pada tiap kelompok dapat dilihat pada Gambar 1.

Data Penelitian

Data penelitian yang digunakan merupakan kumpulan dokumen publikasi ilmiah dosen Fakultas Matematika dan Ilmu Pengetahuan Alam (FMIPA) Institut Pertanian Bogor (IPB), yang terbit antara tahun 2004 sampai 2022, yang tersedia pada situs sipakaril.ipb.ac.id. Data diunduh pada tanggal 2 Mei 2022 dengan total 1732 dokumen berbahasa Inggris, dengan jenis karya ilmiah jurnal internasional. Penelitian ini menggunakan seluruh isi dokumen yang dikonversi secara otomatis menggunakan *Optical Character Recognition* (OCR).



Gambar 1 Tahapan penelitian

Persiapan Data

Terdapat 5 tahapan pada kelompok persiapan data. Tahap pertama adalah mengumpulkan dokumen publikasi ilmiah dosen berbentuk *Portable Document Format* (PDF) dan mengubahnya ke dalam bentuk *plaintext*. Tahap kedua adalah *cleaning* yaitu proses menghilangkan karakter atau kata yang tidak diperlukan dengan menghilangkan tanda baca, angka, karakter *non-alphanumeric*, spasi ganda, *Uniform Resource Locator* (URL), dan email. Tahap ketiga adalah tokenisasi yaitu proses memecah sekumpulan teks menjadi token-token terpisah. Penelitian ini menggunakan token berupa kata karena LDA bekerja lebih baik pada token berbentuk kata (Suadaa dan Purwarianti 2016; Yau *et al.* 2014). Tahap keempat adalah penghapusan *stopword* yaitu sekumpulan kata yang tidak berhubungan dengan subjek utama dengan menggunakan daftar *stopword* pada *library* NLTK. Tahap kelima adalah *stemming* yaitu proses mengubah berbagai varian morfologi kata ke bentuk dasarnya menggunakan algoritma *porter stemmer.*

Reduksi Dimensi

Terdapat dua tahapan pada kelompok reduksi dimensi. Tahap pertama adalah membentuk *Document Term Matrix* (DTM). DTM merupakan sebuah matriks di mana setiap barisnya mewakili dokumen terkait dan setiap kolomnya mewakili istilah yang ada. Pembobotan istilah yang digunakan pada penelitian ini adalah *term frequency* yaitu jumlah kemunculan istilah pada dokumen.

Tahap kedua adalah reduksi dimensi data menggunakan *Latent Dirichlet Allocation*

(LDA). Pada LDA, setiap topik terdiri dari probabilitas distribusi kata yang menunjukkan seberapa penting kata tersebut dalam suatu topik (Sun 2014). Kata-kata yang membentuk struktur topik dihasilkan dalam dua tahap. Pertama, distribusi topik dipilih secara acak untuk setiap dokumen dalam korpus. Kedua, sebuah topik dipilih secara acak dari distribusi topik dan sebuah kata dipilih secara acak dari distribusi yang terpilih untuk setiap kata dalam dokumen (Li *et al.* 2018). Nilai probabilitas gabungan untuk tiap kata pada topik dihitung dengan Persamaan 1 (Khairani 2022).

$$p(\theta,z,w\,|\,\alpha,\beta)=\prod_{(j=1)}^{M}p\left(\theta_j\,|\,\alpha,\beta\right)\prod_{i=1}^{K}p(\varphi_i,\beta)\left(\prod_{t=1}^{N}p\left(z_{j,t}\,|\,\theta_j\right)p\left(w_{j,t}\,|\,\varphi_{i,}z_{j,t}\right)\right), \quad (1)$$

Pada Persamaan 1, α adalah distribusi topik dalam dokumen, β adalah distribusi kata dalam topik, *M* adalah jumlah dokumen, *N* adalah jumlah kata dalam dokumen, *K* adalah jumlah topik, θ_j adalah distribusi topik untuk dokumen ke-*j*, φ_i adalah distribusi kata untuk topik *i*, $z_{j,t}$ adalah penetapan topik pada kata ke-*j* pada dokumen *t*, dan $w_{j,t}$ adalah kata ke-*j* pada dokumen *t*.

Klasterisasi

Terdapat empat tahapan pada kelompok klasterisasi. Tahap pertama adalah klasterisasi dokumen menggunakan algoritma *normalized* PSO-K-means yang alurnya dapat dilihat pada Gambar 2. Normalisasi partikel pada penelitian ini mengadopsi perhitungan yang diusulkan oleh Guo *et al.* (2016) yang disesuaikan agar dapat digunakan untuk mencari *centroid* optimum pada sekelompok dokumen. Metrik pengukuran jarak yang digunakan adalah *bhattacharyya distance* yang formulanya dapat dilihat pada Persamaan 2, di mana *a* dan *b* adalah dua vektor dengan *k* dimensi.

$$d_{Bhat}=-ln\sum_{i=1}^{k}\sqrt{a_i b_i} \quad (2)$$

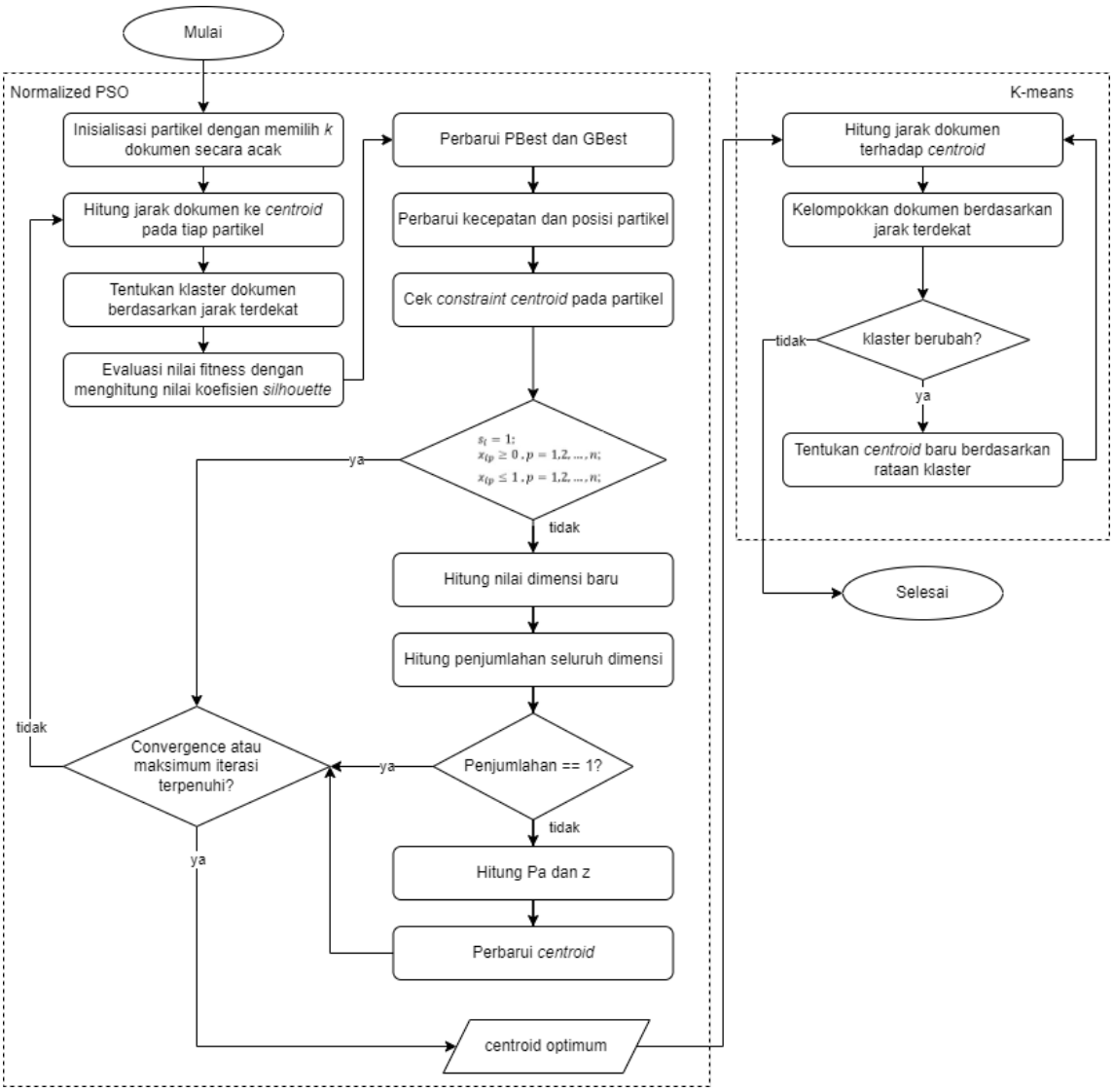
1. *Normalized* PSO
   - Langkah 1: Inisialisasi partikel dengan memilih sejumlah *k* dokumen secara acak sebagai *centroid*.
   - Langkah 2: Hitung jarak tiap dokumen ke *centroid* pada partikel.
   - Langkah 3: Tentukan klaster dokumen berdasarkan jarak terdekat.
   - Langkah 4: Evaluasi nilai *fitness* dengan memaksimumkan nilai koefisien *silhouette* pada Persamaan 3 dimana *a(i)* adalah jarak rata-rata data ke-*i* terhadap data lainnya dalam satu klaster dan *b(i)* adalah jarak rata-rata minimum data ke-*i* terhadap data lainnya dalam klaster yang berbeda.

$$s(i)=\frac{b(i)-a(i)}{max\{a(i),b(i)\}} \quad (3)$$

   - Langkah 5: Perbarui *personal best* (pbest) dan *global best* (gbest).
   - Langkah 6: Perbarui kecepatan dengan Persamaan 4 dan posisi partikel dengan Persamaan 5.

$$v_{id}=w_{id}+c_1 rand_1\left(p_{id}-x_{id}\right)+c_2 rand_2\left(p_{gd}-x_{id}\right) \quad (4)$$

$$x_{id}=x_{id}+v_{id} \quad (5)$$



Gambar 2 *Flowchart* algoritma Normalized PSO-KMeans

- Langkah 7: Untuk tiap partikel, Cek *constraint* tiap *centroid*. Untuk tiap *centroid*, hitung penjumlahan seluruh dimensi pada *centroid* dengan Persamaan 6.

$$S_i = \sum_{p=1}^{n} x_{ip}, i = 1,2,\dots,n.$$
$$\text{Jika } \left\{ \begin{array}{l} s_i = 1; \\ x_{ip} \geq 0, p = 1,2,\dots,n; \\ x_{ip} \leq 1, p = 1,2,\dots,n; \end{array} \right\} \text{ Terpenuhi, lanjut ke langkah 8. Selainnya,} \quad (6)$$

  a. Perbarui posisi *centroid* dan nilai s_i. Jika $x_{ip} < 0$, hitung nilai dimensi dengan Persamaan 7 dimana *n* adalah jumlah dimensi. Jika $x_{ip} > 1$, hitung nilai dimensi dengan Persamaan 8.

$$x_{ip} = \frac{1}{n} r_{ip}, \; dimana, \; r_{ip} \in (0,1) \quad (7)$$

$$x_{ip} = \frac{x_{ip}}{\left(x_{ip} + \frac{1}{n} \right)} \quad (8)$$

b. Jika s_i tidak memenuhi syarat, tarik *centroid* yang melewati batas kembali ke ruang pencarian dengan menghitung nilai P_a sebagai pusat *zoom* ruang PSO. Jika *centroid* berada pada partikel merupakan optimum global hitung P_a dengan Persamaan 9. Selainnya, hitung dengan Persamaan 10. Kemudian hitung *z* sebagai *attractor* dengan Persamaan 11 dimana $x_i^j(t+1)$ adalah posisi *centroid-i* pada dimensi ke-*j* pada waktu ke-*t* dan Pa_i^j adalah nilai *centroid-i* pada dimensi ke-*j*. Terakhir perbarui posisi *centroid* dengan Persamaan 12.

$$Pa_i = G_b * r_1 + r_2, \text{ dimana } r_1 \in (0.5,1.5),\ r_1 \in (-0.5,1.5) \tag{9}$$

$$Pa_i = \left(P_i + (G_b - P_i) * r_3\right) * \frac{1}{2} \text{dimana } r_3 \in (0,1) \tag{10}$$

$$z = \frac{\sum_{j=1}^{n} x_i^j(t+1) - \sum_{j=1}^{n} Pa_i^j}{1 - \sum_{j=1}^{n} Pa_i^j} \tag{11}$$

$$x_i^{'(t+1)} = \frac{x_i(t+1) - Pa}{z} + Pa \tag{12}$$

- Langkah 8: Ulangi *cycle*. Kembali ke langkah 2 hingga tidak terdapat partikel yang memiliki solusi lebih baik dalam 30 iterasi secara berturut-turut atau iterasi mencapai 1000.

2. Modul K-means
   - Langkah 1: Hitung jarak tiap dokumen ke tiap *centroid*.
   - Langkah 2: Kelompokkan dokumen ke *centroid* dengan jarak terdekat.
   - Langkah 3: Jika klaster berubah, perbarui *centroid* dengan menghitung rataan klaster.
   - Langkah 4: Ulangi *cycle*. Kembali ke langkah 1 hingga tidak terjadi perubahan pada klaster.

Setelah klaster dokumen diperoleh tahap selanjutnya adalah ekstraksi istilah dengan menghitung bobot *Term Frequency / Inverse Cluster Frequency* (TFxICF) dan mengambil 5 kata dengan bobot tertinggi. Persamaan 13 dan 14 merupakan formula dari TFxICF (Suadaa dan Purwarianti 2016) dimana $TF_{t,i}$ adalah jumlah *term t* pada klaster *i*, CF_t adalah jumlah klaster yang mengandung *term t* dan *N* adalah jumlah klaster.

$$ICF_t = log\left(\frac{N}{CF_t}\right) \tag{13}$$

$$TFxICF_{t,i} = TF_{t,i}.ICF_t \tag{14}$$

Selanjutnya, pelabelan klaster oleh pakar. Setiap klaster dilabeli berdasarkan 5 kata yang diperoleh pada tahap sebelumnya. Pakar yang berkontribusi dalam pelabelan klaster adalah Ir. Julio Adisantoso M.Kom. sebagai pakar *text mining* dibantu oleh pustakawan IPB yaitu Widiyati Kania S.Sos., M.P., Azizah S.Sos., dan Lindawati S.I.Pust. Pelabelan dilakukan dengan mempertimbangkan definisi kata pada tiap klaster serta index klasifikasi pada *Universal Decimal Classification* (UDC). UDC merupakan suatu sistem pengelompokan internasional yang sering digunakan untuk mengklasifikasikan buku atau dokumen berdasarkan subjek informasi yang terkandung di dalamnya (Andriaty dan Suryantini 2007).

Terakhir, evaluasi hasil klaster dengan mencocokkan hasil klasterisasi dokumen publikasi dengan klaim dosen. Evaluasi dilakukan untuk melihat apakah klaster-klaster dokumen yang diperoleh dapat digunakan untuk memvalidasi klaim kepakaran dosen. Klaim dosen dikumpulkan melalui situs SINTA dan Google scholar. Klaster dokumen dianggap sesuai jika klaster yang bersangkutan mencakup bidang kepakaran yang diklaim. Pencocokan klaster dilakukan secara manual dengan bantuan tim Lembaga Manajemen Informasi dan Transformasi Digital (LMITD) IPB berdasarkan deskripsi dari bidang yang bersangkutan.

HASIL DAN PEMBAHASAN

Persiapan Data

Langkah pertama pada tahap ini adalah mengunduh data berupa *file* dengan ekstensi *.csv* dari basis data IPB. Data tersebut merupakan data karya ilmiah dosen FMIPA IPB yang termasuk ke dalam jenis karya ilmiah jurnal internasional. Terdapat total 1734 data dengan sembilan atribut yaitu karya ilmiah id, dokumen karya ilmiah id, judul, jenis karya id, jenis karya, tahun terbit, nama *file*, URL, dan penulis.

Selanjutnya dokumen diunduh dan diekstrak secara otomatis menggunakan *library Tesseract* OCR pada python. Proses ekstraksi dokumen meliputi konversi dokumen menjadi gambar, pengenalan teks dalam gambar dengan OCR, dan penyimpanan hasil ekstraksi ke dalam *file* dengan ekstensi *.txt*. Setelah seluruh dokumen dikonversi, dilakukan praproses data teks yang meliputi *cleaning*, tokenisasi, penghapusan *stopword*, dan *stemming*.

Terdapat 28,858,020 karakter yang dihilangkan pada proses *cleaning*. Gambar 3 merupakan persentase karakter yang dihilangkan. Karakter yang paling banyak dihilangkan adalah spasi ganda sebanyak 62% diikuti oleh titik sebanyak 8%, koma sebanyak 5%, garis baru sebanyak 5%, dan karakter lainnya sebanyak 20%.

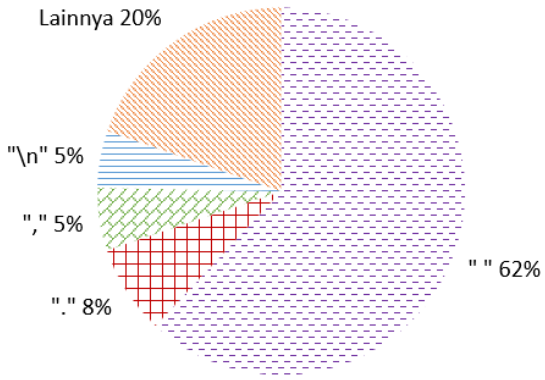
Selanjutnya sekumpulan teks dipecah menjadi token-token berbentuk kata pada proses tokenisasi. Terdapat 8,568,458 kata yang diperoleh setelah proses tokenisasi. Gambar 4 merupakan visualisasi lima kata dengan frekuensi terbanyak. Kata yang paling banyak muncul dalam korpus adalah "*the*" sebanyak 431,943, diikuti oleh "*of*" sebanyak 332,161, "*and*" sebanyak 266,101, "*in*" sebanyak 173,531, dan "*a*" sebanyak 147,089.

Kemudian kata yang dianggap tidak relevan dihilangkan dalam proses penghapusan *stopword*. Terdapat 1,442,576 *stopword* yang dihilangkan pada proses ini. Gambar 5 merupakan visualisasi lima *stopword* dengan frekuensi terbanyak. *Stopword* yang paling banyak muncul dalam korpus adalah "*the*" sebanyak 431,943, diikuti oleh "*and*" sebanyak 266,101, "*for*" sebanyak 71,781, "*with*" sebanyak 60,078, dan "*was*" sebanyak 54,819.

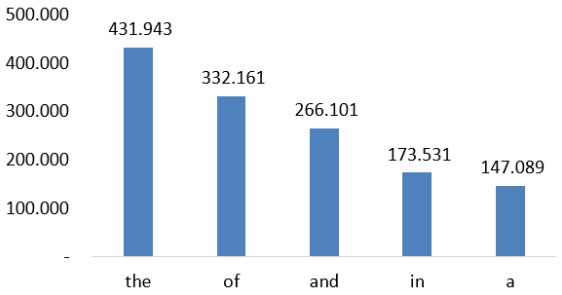
Terakhir setiap kata diubah ke dalam bentuk dasarnya dalam proses *stemming*. Terdapat total 172,619 kata unik yang diperoleh setelah proses *stemming*. Gambar 6 merupakan visualisasi kata setelah praproses dengan *wordcloud.* Pada Gambar 6, "*figure"* memiliki ukuran yang paling besar karena memiliki frekuensi tertinggi dalam korpus. Selain *figure,* kata yang paling banyak muncul adalah *one*, *effect*, dan *result*.

Reduksi Dimensi

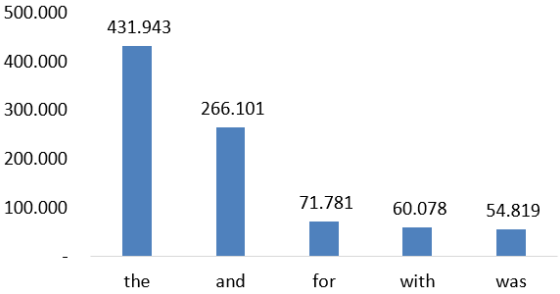
Langkah pertama pada tahap ini adalah menentukan banyaknya topik yang akan dibangun dalam pemodelan LDA. Jumlah topik ditentukan dengan memodelkan lima sampai lima puluh topik dan mengukur nilai koherensi topiknya berdasarkan pengukuran *c_v coherence*. Nilai *c_v coherence* berkisar antara 0 hingga 1 dengan nilai yang lebih tinggi menunjukkan tingkat koherensi yang lebih baik (Roder *et al.* 2015). Pada Gambar 7 terlihat bahwa nilai koherensi topik tertinggi ada pada jumlah topik 24 dengan nilai koherensi topik sebesar 0,55. Oleh karena itu 24 topik dipilih sebagai jumlah topik terbaik.



Gambar 3 Persentase karakter yang dihilangkan



Gambar 4 Top 5 frekuensi kata



Gambar 5 Top 5 *stopword*

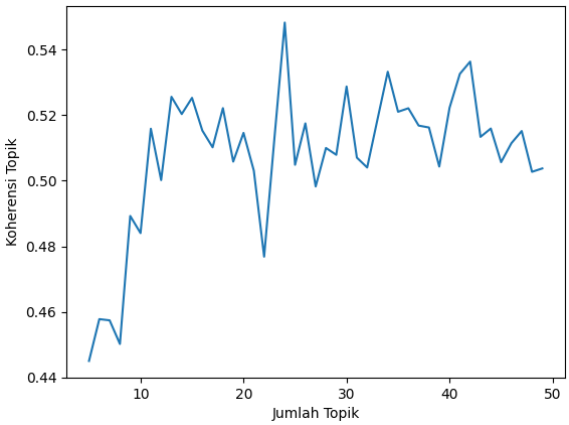
Gambar 6 *Wordcloud* setelah praproses

Setelah menentukan banyaknya topik yang akan dibangun dalam pemodelan LDA, selanjutnya dilakukan pemodelan untuk jumlah topik terpilih. Gambar 8 merupakan contoh hasil keluaran proses LDA dokumen pertama. Pada Gambar 8, peluang dokumen ke-1 masuk ke dalam topik pertama adalah 0,000053, topik ke-2 adalah 0,002750, topik ke-3 adalah 0,000053, dan seterusnya. Distribusi pada topik 1-24 merupakan distribusi dirichlet yang mana jika dilakukan penjumlahan nilai peluang topik 1-24 akan menghasilkan nilai satu.

Klasterisasi

Pada tahap ini algoritma Normalized PSO-K-means dan K-means diujikan dengan jumlah klaster yang berbeda-beda dengan menghitung nilai koefisien *silhouette*. Penelitian ini menggunakan koefisien *silhouette* karena tidak ada informasi eksternal terkait kelas aktual tiap dokumen. Input yang digunakan adalah daftar dokumen yang direpresentasikan dengan 24 peluang topik yang diperoleh pada tahap reduksi dimensi. Gambar 9 merupakan perbandingan nilai koefisien *silhouette* kedua algoritma. Pada Gambar 9 terlihat bahwa secara umum algoritma Normalized PSO-K-means memiliki hasil yang lebih baik dibandingkan dengan algoritma KMeans. Hasil klasterisasi terbaik ada pada algoritma Normalized PSO-K-means dengan nilai koefisien *silhouette* sebesar 0.42, dengan jumlah klaster 15. Gambar 10 menunjukkan visualisasi hasil klasterisasi Normalized PSO-K-means dengan jumlah klaster 15.

Tabel 1 merupakan perbandingan nilai koefisien *silhouette* kedua algoritma. Nilai rataan koefisien *silhouette* algoritma *normalized* PSO-K-means lebih tinggi dibandingkan K-means yaitu 0.38 banding 0.36. Berdasarkan nilai tersebut dapat dilihat bahwa penggunaan PSO sebagai penentu *centroid* optimum pada algoritma K-means dapat meningkatkan nilai koefisien *silhouette* sebesar 5.56%. Peningkatan ini dapat terjadi karena PSO merupakan algoritma berbasis populasi yang memiliki kemampuan untuk mengeksplorasi seluruh ruang pencarian. Setiap partikel dalam PSO bergerak melalui ruang pencarian berdasarkan pengetahuan posisi terbaik dirinya sendiri dan kelompoknya sehingga mampu mengeksplorasi berbagai solusi yang potensial. Artinya PSO dapat mencari *centroid* yang lebih baik secara global, tidak hanya berfokus pada solusi lokal.



Gambar 7 Nilai koherensi topik

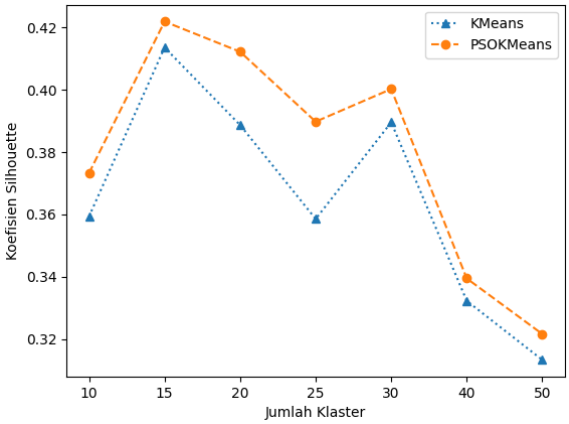
Topik1	Topik2	Topik3	Topik4	Topik5	Topik6
5,33E-05	0,00275	5,33E-05	5,33E-05	0,064614	5,33E-05

Topik7	Topik8	Topik9	Topik10	Topik11	Topik12
0,014389	5,33E-05	5,33E-05	5,33E-05	5,33E-05	0,00541

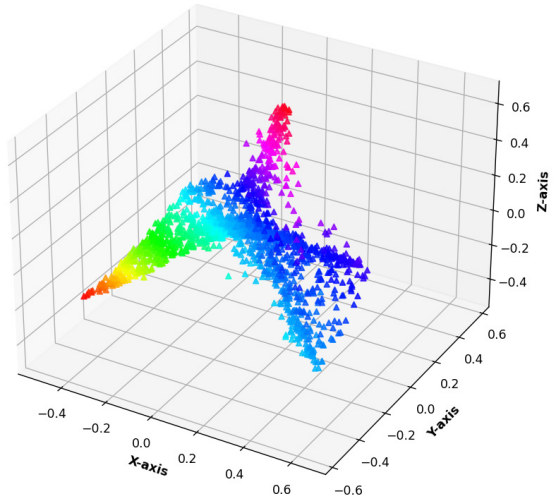
Topik13	Topik14	Topik15	Topik16	Topik17	Topik18
5,33E-05	0,012003	5,33E-05	0,002681	0,897247	5,33E-05

Topik19	Topik20	Topik21	Topik22	Topik23	Topik24
5,33E-05	5,33E-05	5,33E-05	5,33E-05	5,33E-05	5,33E-05

Gambar 8 Contoh keluaran proses LDA dokumen pertama



Gambar 9 Perbandingan nilai koefisien silhouette algoritma normalized PSO-K-means dan K-means



Gambar 10 Visualisasi hasil klasterisasi normalized PSO-K-means

Setelah klaster dokumen diperoleh, istilah-istilah penting dalam klaster diekstrak dengan menghitung nilai TF/ICF. Semakin tinggi nilai TF/ICF, semakin penting istilah tersebut dalam klaster. Tabel 2 merupakan hasil pelabelan klaster oleh pakar berdasarkan lima kata dengan nilai TF/ICF tertinggi. Sebagai contoh, Klaster 1 terdiri dari kata *chitinase, actinomycetes, endophyte, streptomyces* dan *phage*. Berdasarkan kata-kata tersebut serta index klasifikasi UDC tiap kata, dokumen-dokumen dalam klaster 1 dapat dikelompokkan ke dalam bidang kepakaran *microbiology*.

Setelah label klaster didapatkan. Hasil klaster dievaluasi dengan mencocokkan klaster dokumen dengan klaim kepakaran dosen. Terdapat total 499 dosen yang dievaluasi dari 1732 dokumen. Evaluasi dilakukan dengan bantuan tim LMITD IPB dengan melakukan pencocokan satu per satu secara manual berdasarkan deskripsi bidang yang bersangkutan.

Tabel 3 merupakan contoh evaluasi klaster yang sesuai dengan klaim dosen. Pada Tabel 3, dokumen yang ada sesuai dengan klaim yang diberikan serta klaster yang dihasilkan. Dokumen-dokumen tersebut berkaitan dengan perlindungan data, privasi, dan keamanan. *Information security* merupakan bidang yang berhubungan dengan perlindungan informasi yang meliputi perangkat keras, media penyimpanan, dan transmisi informasi (Whitmand dan Mattord 2022). *Cryptography* merupakan bidang yang berhubungan dengan seni penulisan atau pemecahan kode (enkripsi/dekripsi) untuk melindungi informasi dari akses yang tidak sah (Oxford 2020). *Steganography* merupakan bidang yang berhubungan dengan penyembunyian informasi dalam konten lain secara tidak terlihat (Irfan 2013). W*atermarking* merupakan

Tabel 1 Nilai koefisien *silhouette* kedua algoritma

Jumlah Klaster	PSO-K-means	K-means
10	0.37	0.35
15	0.42	0.41
20	0.41	0.38
25	0.38	0.35
30	0.40	0.38
40	0.33	0.33
50	0.32	0.31
Rataan	0.38	0.36

Tabel 2 Hasil pelabelan klaster oleh pakar

Klaster	Istilah	Label
1	Chitinase, Actinomycetes, Endophyte, Streptomyces, Phage	Microbiology
2	Loyalty, Dishonest, Satisfaction, MCS, Waze	Management, Earth Science and Technology
3	Allophane, Imogolite, Chitosan, OPEFB, Adsorption	Chemistry
4	Orangutan, Tarsier, Macaque, TASR, Haplotype	Primatology and Genetics
5	Jamu, Watermark, DNN, DEMNAS, CNN	Information Technology
6	Pollinate, Stingless, Desmos, Glabrous, Pollen	Botany and Entomology
7	Soliton, SHG, Handedness, SBHM, Breakwater	Physics
8	ARIMA, Poverty, LASSO, EBLUP, GWR	Statistics
9	ECF, Lipase, Neuron, Dock, BrdU	Molecular Biology and Medical Science
10	BST, LiTaO, DOPE, Ferroelectric, Film	Material Science and Engineering
11	Peatland, Mangrove, Emission, PeatCLSM, Peat	Environmental Science
12	Antioxidant, DPPH, Flavonoid, Phenol, Glucosidase	Biochemistry
13	Homegarden, Plantation, Forest, Canopy, Rainforest	Ecology
14	Biofertilizer, Transgene, Bunar, Hawara, BCF	Agricultural Biology
15	VNS, Skyline, UID, Goos, SNP	Mathematics and Computer Science

Tabel 3 Contoh hasil klaster yang sesuai dengan klaim dosen

Nama	Shelvie Nidya Neyman
Klaim	Information Security, Cryptography, Steganography, Watermarking
Klaster	Information Technology
Dokumen	1. Perlindungan Hak Cipta Baru untuk Peta Vektor dengan Menggunakan *Watermarking* Berbasis FFT 2. Perlindungan Kepemilikan pada Model Elevasi Digital (DEM) Menggunakan *Watermark* Berbasis Transformasi 3. Protokol Komunikasi Aman untuk IoT Berbasis Arduino Menggunakan *Lightweight Cryptography*

Tabel 4 Contoh hasil klaster yang tidak sesuai dengan klaim dosen

Nama	Donny Citra Lesmana
Klaim	Financial Mathematics, Numerical Methods for Partial Differential Equation
Klaster	Physics
Dokumen	1. Metode numerik untuk menentukan harga opsi dengan model volatilitas *Risk Adjusted Pricing Methodology* (RAPM) 2. Skema numerik untuk menentukan harga opsi Amerika dengan biaya transaksi dalam proses *jump diffusion*

bidang yang berhubungan dengan penyisipan *watermark* ke dalam citra digital yang bertujuan untuk melindungi hak cipta tanpa merusak citra asli (Munir 2006). Keempat bidang tersebut berhubungan erat dengan pengolahan dan keamanan informasi yang merupakan fokus utama dalam bidang *information technology*.

Tabel 4 merupakan contoh evaluasi klaster yang tidak sesuai dengan klaim dosen. Pada Tabel 4, dokumen-dokumen yang ada sesuai dengan klaim yang diberikan namun tidak sesuai dengan klaster yang dihasilkan. Sebagai contoh dokumen pertama relevan dengan klaim

yang diberikan karena pada penelitian ini, metode numerik digunakan untuk menyelesaikan persamaan matematika yang kompleks yang muncul dalam analisis harga opsi. Penelitian ini tidak sesuai dengan klaster yang dihasilkan karena secara umum bidang fisika berfokus pada studi sifat dan perilaku materi, energi, dan fenomena alam lainnya. Ketidaksesuaian ini dapat terjadi karena label klaster belum mencakup seluruh bidang ilmu pada dokumen yang tersedia karena label ditentukan berdasarkan lima kata dengan nilai TF/ICF tertinggi.

Berdasarkan hasil pencocokan klaster dokumen dengan klaim dari 499 dosen, diperoleh 372 hasil klasterisasi dokumen sesuai dengan klaim yang diberikan dan 127 tidak sesuai. Berdasarkan sampel yang diambil, diperoleh 75% klaster dokumen sesuai dengan klaim yang diberikan dan 25% tidak sesuai. Ketidaksesuaian klaster dengan klaim dapat terjadi karena label klaster belum mencakup seluruh bidang ilmu pada dokumen, atau dokumen yang tersedia mencakup bidang yang diklaim namun tidak secara khusus membahas bidang tersebut.

SIMPULAN

Pengelompokan dokumen publikasi ilmiah berdasarkan bidang kepakaran telah berhasil dilakukan dengan nilai koefisien *silhouette* sebesar 0.42. Penggunaan PSO untuk menentukan *centroid* optimum pada algoritma K-means dapat meningkatkan nilai koefisien *silhouette* sebesar 5.56%. Selanjutnya, berdasarkan hasil pencocokan klaster dengan klaim dosen, 75% klaster dokumen sesuai dengan klaim yang diberikan. Penelitian ini menunjukkan bahwa klaster dokumen yang dibangun dalam penelitian ini dapat digunakan untuk mendukung klaim kepakaran seorang dosen. Saran penelitian selanjutnya, diharapkan menggunakan lingkup dokumen yang lebih luas, tidak terbatas pada bidang ilmu dalam lingkup matematika dan ilmu pengetahuan alam serta mempertimbangkan metode pelabelan lain seperti penggunaan pendekatan *fusion* untuk menggabungkan beberapa pelabel klaster (Roitman *et al.* 2014), sehingga label yang diperoleh dapat mencakup seluruh bidang ilmu dalam klaster.

UCAPAN TERIMA KASIH

Terima kasih disampaikan kepada Ir. Julio Adisantoso M.Kom., Widiyati Kania S.Sos., M.P., Azizah S.Sos., dan Lindawati S.I.Pust., yang telah membantu dalam proses pelabelan klaster serta Lembaga Manajemen Informasi dan Transformasi Digital IPB yang telah menyediakan data penelitian serta membantu dalam proses evaluasi hasil klaster.

DAFTAR PUSTAKA

Bui QV, Sayadi K, Amor SB, Bui M. 2017. Combining Latent Dirichlet Allocation and K-means for documents clustering: effect of probabilistic based distance measures. *2017 Intelligent Information and Database Systems (ACIIDS).* 248-257. https://doi.org/10.1007/978-3-319-54472-4_24.

Campos LM, Fernandez-luna JM, Huete JF, Exposito LR. 2020. Automatic construction of multi-faceted user profiles using text clustering and its application to expert recommendation and filtering problems. *Knowledge Based System*. 190:1-18. https://doi.org/10.1016/j.knosys.2019.105337.

Chouhan R, Purohit A. 2018. An approach for document clustering using PSO and K-means algorithm. *2018 2nd International Conference on Inventive Systems and Control (ICISC).* 1380-1384. https://doi.org/10.1109/ICISC.2018.8399034.

Feldman R, Sanger J. 2007. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. New York (NY). Cambridge University Press.

Guo J, Gou J, Wang C, Luo W. 2016. The Normalized-PSO and Its Application in Attribute Weighted Optimal Problem. *2016 3rd International Conference on Trustworthy Systems and their Applications (TSA).* 48-53. https://doi.org/10.1109/TSA.2016.18.

Han J, Kamber M, Pei J. 2012. *Data Mining: Concepts and Techniques.* 3rd ed. Massachusetts (MA). Morgan Kaufmann Publishers is an imprint of Elsevier.

Irfan. 2013. Penyembunyian Informasi (steganography) Gambar Menggunakan Metode LSB (Least Significant Bit). *Rekayasa Teknologi*. 5(1):1-6.

Khairani F. 2022. Topic Modelling dan Prediksi Fenomena Ekonomi Indonesia Berdasarkan Berita Online. Thesis. Bogor (ID): Institut Pertanian Bogor.

Li C, Lu Y, Wu J, Zhang Y, Xia Z, Wang T, Yu D, Chen X, Liu P, Guo J. 2018. LDA Meets Word2Vec: A Novel Model for Academic Abstract Clustering. *Companion Proceedings of The Web Conference 2018*. 1699-1706. https://doi.org/10.1145/3184558.3191629.

Munir R. 2006. Kriptografi. Bandung (ID). Informatika.

Oxford, University of. 2020. "English Oxford Learner's Dictionary".

Roder M, Both A, Hinneburg A. 2015. Exploring the space of topic coherence measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. 399-408. https://doi.org/10.1145/2684822.2685324.

Roitman H, Hummel S, Shmueli-Scheuer M. 2014. A fusion approach to cluster labeling. *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14)*. 883-886. https://doi.org/10.1145/2600428.2609465.

Rosnelly R. 2012. *Sistem Pakar: Konsep dan Teori*. Yogyakarta (ID): Cv Andi Offset.

Saini Y, Bachchas V, Kumar Y, Kumar S. 2020. Abusive Text Examination Using Latent Dirichlet Allocation, Self Organizing Maps and K Means Clustering. *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS).* 1233-1238. https://doi.org/10.1109/ICICCS48265.2020.9121090.

Suadaa LH, Purwarianti A. 2016. Combination of Latent Dirichlet Allocation (LDA) and Term Frequency-Inverse Cluster Frequency (TFxICF) in Indonesian text clustering with labeling. *2016 4th International Conference on Information and Communication Technology (ICoICT)*. 1-6. https://doi.org/10.1109/ICoICT.2016.7571885.

Sun X. 2014. Textual document clustering using topic models. *2014 Tenth International Conference on Semantics, Knowledge and Grids (SKG)*. 1-4. https://doi.org/10.1109/SKG.2014.27.

Usai A, Pironti M, Mital M, Mejri CA. 2018. Knowledge discovery out of text data: a systematic review via text mining. *J Knowl Manag*. 22:1471-1488. https://doi.org/10.1108/JKM-11-2017-0517.

Whitmand ME, Mattord HJ. 2022. *Principle of Information Security.* Ed ke-7. Boston (US). Cengage Learning Inc.

Yau CK, Porter A, Newman N, Suominen A. 2014. Newman, N. et al. Clustering scientific documents with topic modeling. *Scientometrics 100*. 767-786. https://doi.org/10.1007/s11192-014-1321-8.

Yuan C, Yang H. 2019. Research on K-Value Selection Method of K-Means Clustering Algorithm. *J*. 2(2):226-235. https://doi.org/10.3390/j2020016.