

PEMBENTUKAN *DATASET* DAN KLASTERISASI DATA KEPAKARAN BERBASIS PUBLIKASI DAN SITASI GOOGLE SCHOLAR

NAUFAL FAIQ RAMADHANI



**PROGRAM SARJANA ILMU KOMPUTER
SEKOLAH SAINS DATA MATEMATIKA DAN ILMU KOMPUTER
INSTITUT PERTANIAN BOGOR
BOGOR
2026**



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



PERNYATAAN MENGENAI SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa skripsi dengan judul “Pembentukan *Dataset* dan Klasterisasi Data Kepakaran Berbasis Publikasi dan Sitasi Google Scholar” adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir skripsi ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Juli 2026

Naufal Faiq Ramadhani
G6401211023

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumunkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



ABSTRAK

NAUFAL FAIQ RAMADHANI. Pembentukan *Dataset* dan Klasterisasi Data Kepakaran Berbasis Publikasi dan Sitasi Google Scholar. Dibimbing oleh AHMAD RIDHA.

Pemetaan kepakaran peneliti penting untuk memahami perkembangan bidang penelitian di perguruan tinggi. Penelitian ini membangun dataset dan memetakan kelompok kepakaran peneliti di Sekolah Sains Data, Matematika, dan Informatika (SSMI) IPB University berdasarkan publikasi Google Scholar. Data hasil *web scraping* diproses melalui deduplikasi dan prapemrosesan teks, direpresentasikan dengan TF-IDF, direduksi dimensinya dengan LSA, lalu dikelompokkan menggunakan K-Means. Hasil penelitian menghasilkan lima kelompok kepakaran yang merepresentasikan bidang keilmuan di SSMI. Jumlah kluster ini dipilih berdasarkan kesesuaiannya dengan struktur program studi, meskipun konfigurasi enam kluster memiliki nilai pemisahan lebih tinggi. *Dataset* dan pemetaan yang dihasilkan dapat mendukung pengambilan keputusan akademik dan pengembangan kolaborasi riset.

Kata Kunci: *dataset*, Google Scholar, pemetaan kepakaran, TF-IDF, LSA, *K-Means Clustering*, *web scraping* bertingkat, *text mining*.

ABSTRACT

NAUFAL FAIQ RAMADHANI. Dataset Formation and Expertise Data Clustering Based on Google Scholar Publications and Citations. Supervised by AHMAD RIDHA.

Mapping researcher expertise is essential for understanding the evolution of research fields within higher education institutions. This study constructs a dataset and maps researcher expertise groups at the School of Data Science, Mathematics, and Informatics (SSMI) of IPB University based on Google Scholar publications. Web-scraped data underwent deduplication and text preprocessing, was represented using TF-IDF, underwent dimensionality reduction via LSA, and was subsequently clustered using K-Means. The study identified five expertise groups representing the scientific disciplines within SSMI. This number of clusters was selected based on its alignment with the academic program structure, even though a six-cluster configuration yielded a higher separation score. The resulting dataset and mapping can support academic decision-making and the development of research collaborations.

Keywords: *dataset*, Google Scholar, expertise mapping, TF-IDF, LSA, *K-Means Clustering*, multilevel *web scraping*, *text mining*.



- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

© Hak Cipta milik IPB, tahun 2026
Hak Cipta dilindungi Undang-Undang

Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan atau menyebutkan sumbernya. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik, atau tinjauan suatu masalah, dan pengutipan tersebut tidak merugikan kepentingan IPB.

Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apa pun tanpa izin IPB.

PEMBENTUKAN *DATASET* DAN KLASTERISASI DATA KEPAKARAN BERBASIS PUBLIKASI DAN SITASI GOOGLE SCHOLAR

NAUFAL FAIQ RAMADHANI

Skripsi
sebagai salah satu syarat untuk memperoleh gelar
Sarjana pada
Program Studi Ilmu Komputer

**PROGRAM SARJANA ILMU KOMPUTER
SEKOLAH SAINS DATA MATEMATIKA DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2026**



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Tim Penguji pada Ujian Skripsi:

- 1 Dr. Karlisa Priandana, S.T., M.Eng.
- 2 Dr. Ir. Sri Wahjuni, , M.T.



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



Judul Skripsi : Pembentukan *Dataset* dan Klasterisasi Data Kepakaran Berbasis
Publikasi dan Sitasi Google Scholar

Nama : Naufal Faiq Ramadhani

NIM : G6401211023

Disetujui oleh

Pembimbing:

Ahmad Ridha, S.Kom, M.S.

NIP. 198005072005011001

Diketahui oleh

Ketua Program Sarjana Ilmu Komputer:

Dr. Sony Hartono Wijaya, S. Kom., M. Kom.

NIP. 1981080920081210202

Tanggal Ujian:
3 Agustus 2026

Tanggal Lulus:

PRAKATA

Puji dan syukur penulis panjatkan kepada Allah *subhanaahu wa ta'ala* atas segala karunia-Nya sehingga karya ilmiah ini berhasil diselesaikan. Tema yang dipilih dalam penelitian yang dilaksanakan sejak bulan Januari 2025 sampai bulan Juni 2026 ini dengan judul “Pembentukan *Dataset* dan Klasterisasi Data Kepakaran Berbasis Publikasi dan Sitasi Google Scholar”.

Di samping itu, penulis mengucapkan terima kasih kepada seluruh teman dan sahabat yang telah menemani perjalanan selama masa perkuliahan, khususnya rekan-rekan di Perdesi Academy dan Pelari Kalcer, komunitas AgriUX, serta teman-teman Program Studi Ilmu Komputer IPB University. Kebersamaan, dukungan, dan berbagai pengalaman yang dibagikan selama ini menjadi penyemangat bagi penulis dalam menyelesaikan penelitian ini. Ungkapan terima kasih yang tulus penulis sampaikan kepada Ayah dan Ibu tercinta, kakak penulis Viki Fathurohman, kakak ipar Afifah Mufidati, serta Adel Maulina Zulfa atas doa, kasih sayang, kesabaran, dan dukungan yang tidak pernah putus selama proses penyusunan skripsi hingga akhirnya dapat diselesaikan.

Semoga karya ilmiah ini bermanfaat bagi pihak yang membutuhkan dan bagi kemajuan ilmu pengetahuan.

Bogor, Juli 2026

Naufal Faiq Ramadhani



DAFTAR ISI

DAFTAR TABEL	X
DAFTAR GAMBAR	X
DAFTAR LAMPIRAN	XI
I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Tujuan	3
1.4 Manfaat	3
1.5 Ruang Lingkup	4
II TINJAUAN PUSTAKA	5
2.1 Pemetaan Kepakaran Akademik	5
2.2 Google Scholar	5
2.3 <i>Web Scraping</i>	6
2.4 <i>Natural Language Processing</i>	6
2.5 <i>Term Frequency-Inverse Document Frequency (TF-IDF)</i>	7
2.6 <i>Latent Semantic Analysis (LSA)</i>	9
2.7 <i>Algoritma K-Means Clustering</i>	10
2.8 <i>Elbow Method & Silhouette Score</i>	11
III METODE	14
3.1 Tahapan Penelitian	14
3.2 Seleksi Data	14
3.3 Prapemrosesan Data	16
3.4 Transformasi Data	19
3.5 Penambangan Data	20
3.6 Evaluasi & Interpretasi Data	21
3.7 Peralatan Penelitian	22
IV HASIL DAN PEMBAHASAN	23
4.1 Seleksi Data	23
4.2 Prapemrosesan Data	23
4.3 Transformasi Data	29
4.4 Penambangan Data	32
4.5 Evaluasi & Interpretasi Data	32
V SIMPULAN DAN SARAN	40
5.1 Simpulan	40
5.2 Saran	40
DAFTAR PUSTAKA	41
LAMPIRAN	45

DAFTAR TABEL

1	Perbedaan sebelum dan sesudah penghapusan tanda petik	24
2	Contoh pembuatan ID dengan MD5	24
3	Contoh data duplikat dengan ID artikel yang sama	24
4	Perbandingan karakteristik bahasa pada korpus teks	26
5	Perbedaan sebelum dan sesudah proses translasi otomatis	26
6	Perbedaan sebelum dan sesudah <i>case folding & punctuation removal</i>	27
7	Contoh penentuan domain-specific <i>stopwords</i>	28
8	Hasil penentuan nilai <i>n_components</i>	31
9	Hasil perhitungan selisih <i>inertia</i>	33
10	Distribusi jumlah penulis per klaster per level kepakaran (K = 5)	36
11	Distribusi jumlah penulis per klaster per level kepakaran (K = 6)	36
12	Top 15 kata kunci dominan pada setiap klaster (K = 5)	37
13	Top 15 kata kunci dominan pada setiap klaster (K = 6)	38
14	Perbandingan interpretasi klaster pada setiap LLM (K = 5)	38
15	Perbandingan interpretasi klaster pada setiap LLM (K = 6)	39

DAFTAR GAMBAR

1	Metodologi <i>Knowledge Discovery in Databases</i>	14
2	Alur seleksi data bertingkat (Level 1 dan 2) dari Google Scholar	15
3	Alur tahapan prapemrosesan data	16
4	Ilustrasi reduksi dimensi <i>matrix</i>	20
5	Ilustrasi perubahan struktur <i>dataset</i>	29
6	Ilustrasi <i>sparse matrix</i>	30
7	Grafik pengujian <i>n_components</i> dengan rentang 5-300	31
8	<i>Elbow Plot</i> untuk rentang K = 2 hingga K = 25	33
9	Grafik <i>Silhouette Score</i> untuk rentang K = 2 hingga K = 25	34
10	Grafik distribusi penulis per klaster (K = 5)	35
11	Grafik distribusi penulis per klaster (K = 6)	36

DAFTAR LAMPIRAN

1	Daftar nama dosen SSMI	45
2	Mekanisme deduplikasi berbasis prioritas	49

3	ENGLISH_STOP_WORDS dari Scikit-Learn	52
4	Daftar custom <i>stopwords</i> untuk masing-masing judul & nama publikasi	54
5	Prompt yang digunakan untuk interpretasi klaster pada LLM	56
6	Luaran lengkap hasil interpretasi klaster dari setiap LLM	57

