



UNDERGRADUATE THESIS

Implementation of Content-Based Filtering and HDBSCAN for Developer Recommendation in a Task Management Application (Case Study: PT Tunas Kreasi Digital)

MUHAMMAD FAIRUS RAMADHANI

NPM 21081010090

THESIS ADVISOR

Eva Yulia Puspaningrum, S.Kom, M.Kom

Fetty Tri Anggraeny, S.Kom, M.Kom

**MINISTRY OF HIGHER EDUCATION, SCIENCE, AND TECHNOLOGY
UNIVERSITAS PEMBANGUNAN NASIONAL VETERAN JAWA TIMUR
FACULTY OF COMPUTER SCIENCE
INFORMATICS STUDY PROGRAM
SURABAYA
2026**

APPROVAL SHEET

Implementation of Content-Based Filtering and HDBSCAN for Developer Recommendation in a Task Management Application (Case Study: PT Tunas Kreasi Digital)

By :
Muhammad Fairus Ramadhani
NPM. 21081010090

Has been defended before, and accepted by, the Board of Assessors of the Thesis Examination of the Informatics Study Program, Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jawa Timur, on Maret 13 2026

Approved

Eva Yulia Puspaningrum, S.Kom., M.Kom
NIP. 19890705 2021212 002



(Advisor I)

Fetty Tri Anggraeny, S.Kom. M. Kom
NIP. 19820211 2021212 005



(Advisor II)

Retno Mumpuni, S.Kom., M.Sc
NIP. 19870716 2025 21 2045



(Head Assessor)

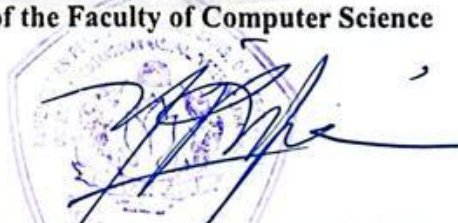
Henni Endah Wahanani, ST, M.Kom
NIP. 19780922 2021212 005



(Assessor 1)

Acknowledge by,

Dean of the Faculty of Computer Science



Prof. Dr. Ir. Novirina Hendrasarie, MT
NIP. 19681126 199403 2 001

APPROVAL SHEET

**Implementation of Content-Based Filtering and HDBSCAN
for Developer Recommendation in a Task Management Application
(Case Study: PT Tunas Kreasi Digital)**

By :
Muhammad Fairus Ramadhani
NPM. 21081010090

Approved to proceed to the Thesis Examination



Approved by,
Coordinator of Informatics Study Program
Faculty of Computer Science

A handwritten signature in black ink, appearing to read 'Fetty Tri Anggraeny', is written over the text of the coordinator's name.

Fetty Tri Anggraeny, S.Kom., M.Kom
NIP. 19820211 2021212 005

STATEMENT OF ORIGINALITY

I am the undersigned:

Student Name : Muhammad Fairus Ramadhani
NPM : 21081010090
Degree Program : Bachelor (S1)
Study Program : Informatics
Faculty : Faculty of Computer Science

Hereby declares that this undergraduate thesis contains no part of any other scientific work that has been submitted to obtain an academic degree at any higher education institution. Furthermore, it does not contain any work or opinions previously written or published by others, except for those which are explicitly cited in this thesis and listed completely in references

And I declare that this scientific document is free from elements of plagiarism. If in the future indications of plagiarism are found in this Thesis, I am willing to accept sanctions in accordance with the applicable laws and regulations.

Thus, I made this statement without any coercion from anyone and to be used as it should.

Surabaya, May 13 2026

Declarant



Muhammad Fairus Ramadhani

NPM. 21081010090

ABSTRACT

Student Name / NPM : Muhammad Fairus Ramadhani / 21081010090
Thesis Title : Implementation of Content-Based Filtering and HDBSCAN for Developer Recommendation in a Task Management Application
(Case Study: PT. Tunas Kreasi Digital)
Advisor : 1. Eva Yulia Puspaningrum, S.Kom, M.Kom
2. Fetty Tri Anggraeny, S.Kom, M.Kom.

The process of assigning developers in software development is generally still performed manually, making it prone to subjective bias and inefficiency. This study implements a developer recommendation system based on Content-Based Filtering (CBF) combined with the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) algorithm in a web-based task management application at PT. Tunas Kreasi Digital. The data used were obtained from GitHub commit logs and the company's daily logs for the 2024–2025 period. The system works by clustering historical tasks using HDBSCAN based on the similarity of task description content, then applying Content-Based Filtering to recommend developers who previously handled tasks within clusters relevant to new tasks. The optimal HDBSCAN configuration was achieved with `min_cluster_size = 2` and `min_samples = 2`, resulting in a Silhouette Score of 0.559 and the lowest noise percentage of 18.71%. Under this configuration, the system achieved a Hit Ratio of 0.88, Recall of 0.80, and Mean Reciprocal Rank (MRR) of 0.563. Testing on cosine similarity thresholds showed that using the entire historical dataset without filtering (4,505 records) produced the highest Hit Ratio and Recall values. Compared to the pure CBF method, the hybrid CBF+HDBSCAN approach outperformed almost all evaluation metrics, demonstrating that clustering developer expertise patterns can significantly improve recommendation accuracy. This system is considered feasible to implement as a decision-support tool for developer assignment in software projects.

Keywords: recommendation system, Content-Based Filtering, HDBSCAN, task management, developer, clustering

ACKNOWLEDGEMENT

All praise is devoted to Allah SWT for His blessings, guidance, and grace bestowed upon the author, so that the thesis entitled “Implementation of Content-Based Filtering and HDBSCAN for Developer Recommendation in a Task Management Application (Case Study: PT. Tunas Kreasi Digital)” could be completed successfully.

The preparation of this thesis is submitted as one of the requirements for obtaining a Bachelor’s degree in the Informatics Study Program, Faculty of Computer Science, Universitas Pembangunan Nasional “Veteran” Jawa Timur. The author would like to express sincere appreciation to all parties who have provided support and contributions throughout the research process by dedicating their time, effort, and thoughts in guiding and motivating the author in completing this thesis. The author realizes that the completion of this thesis would not have been possible without the assistance, guidance, and encouragement, both morally, spiritually, and materially, from various parties. Therefore, the author would like to express sincere gratitude to:

1. Prof. Dr. Ir. Novirina Hendrasarie, MT., as the Dean of the Faculty of Computer Science, Universitas Pembangunan Nasional “Veteran” Jawa Timur.
2. Fetty Tri Anggraeny, S.Kom., M.Kom., as the Head of the Informatics Study Program, Faculty of Computer Science, Universitas Pembangunan Nasional “Veteran” Jawa Timur and Second Supervisor, who has guided the author throughout the completion of this research.
3. Eva Yulia Puspaningrum, S.Kom., M.Kom., as the First Supervisor, who has guided the author throughout the completion of this research.
4. Retno Mumpuni, S.Kom., M.Sc., as the First Examiner, who has provided valuable insights and new perspectives for this research.
5. Henni Endah Wahanani, ST., M.Kom., as the Second Examiner, who has provided valuable insights and new perspectives for this research
6. Firza Prima Aditiawan, S.Kom., MTI., as the Academic Advisor, who has guided the author throughout the academic period and the completion of this research.

7. The lecturers and staff of the Informatics Study Program, Faculty of Computer Science, Universitas Pembangunan Nasional “Veteran” Jawa Timur, for all the assistance provided in facilitating the completion of this research.
8. The author’s father, mother, siblings, and relatives, who have continuously prayed for, encouraged, and provided material support to the author throughout the study period.
9. Friends who have provided support throughout the academic journey and the preparation of this research.

The author realizes that there are many shortcomings in the preparation of this thesis. Therefore, constructive criticism and suggestions from all parties are highly expected for the improvement and refinement of this thesis. Finally, despite the limitations possessed by the author, it is hoped that this thesis will be beneficial to all parties in general and to the author in particular.

Surabaya, May 13 2026

Author,



Muhammad Fairus Ramadhani

NPM. 21081010090

TABLE OF CONTENT

APPROVAL SHEET	i
APPROVAL SHEET	iii
STATEMENT OF ORIGINALITY	v
ABSTRACT	vii
ACKNOWLEDGEMENT	ix
TABLE OF CONTENT	xi
LIST OF TABLE	xv
LIST OF FIGURE.....	xvii
CHAPTER I INTRODUCTION	1
1.1. Background of Study	1
1.2. Problem Formulation	3
1.3. Research Objectives	3
1.4. Significance Study	3
1.5. Scope and Limitation.....	3
CHAPTER II LITERATURE REVIEW	5
2.1. Previous Review	5
2.2. Theoretical Basis	8
2.2.1. Recommendation	8
2.2.2. Task Management Application.....	8
2.2.3. Natural Language Processing (NLP).....	8
2.2.4. Machine Learning.....	8
2.2.5. Data Mining	9
2.2.6. NLTK	9
2.2.7. Pandas.....	9

2.2.8.	Numpy	10
2.2.9.	Content Based Filtering	10
2.2.10.	Euclidean Distance	11
2.2.11.	Cosine Similarity	11
2.2.12.	TF-IDF	12
2.2.13.	UMAP Dimensionality Reduction.....	13
2.2.14.	HDBSCAN	13
2.2.15.	Shilhoutte Score.....	15
2.2.16.	Recall@k	16
2.2.17.	Hit Ratio	16
2.2.18.	MRR	17
CHAPTER III SYSTEM DESIGN AND IMPLEMENTATION.....		19
3.1.	Literature Study	19
3.2.	Requirement Analysis	20
3.2.1.	System Specification	20
3.3.	Data Preparation	21
3.4.	Data Pre-processing	22
3.4.1	Translate	23
3.4.2	Case Folding.....	24
3.4.3	Remove Punctuation.....	24
3.4.4	Number Removal	25
3.4.5	Tokenization.....	25
3.4.6	Stopword Removal	26
3.4.7	Stemming.....	27
3.5.	Filtering	27
3.6.	TF-IDF	28

3.7.	UMAP.....	32
3.8.	HDBSCAN	33
3.9.	Recomendation Stage	43
3.10.	Evaluation.....	46
3.10.1	Metrics	46
3.10.2	Testing Scheme	48
CHAPTER IV RESULT AND DISCUSSION		51
4.1.	Method Implementation	51
4.1.1.	Data Preparation	51
4.1.2.	Preprocessing.....	53
4.1.3.	Filtering	57
4.1.4.	TF-IDF.....	58
4.1.5.	UMAP.....	61
4.1.6.	HDBSCAN Clustering	61
4.1.7.	Recomendation	62
4.2	Clustering Result	66
4.2.1	Variation of "min_cluster_size"	66
4.2.2	Variation of "min_sample".....	71
4.3	Recomendation Testing Result.....	77
4.4	Evaluation.....	79
4.5	System Implementation.....	87
4.5.1	Clustering in the System	87
4.5.2	Recomendation in the System.....	88
CHAPTER V CONCLUSION AND RECOMENDATION		91
5.1	Conclusion	91
5.2	Recomendation.....	91

LIST OF TABLE

Table 3.1 System Specification.....	20
Table 3. 2 Translate Result	23
Table 3. 3 Case Folding Result	24
Table 3. 4 Remove Punctuation Result	24
Table 3. 5 Number Removal Result	25
Table 3. 6 Hasil Tokenization	26
Table 3. 7 Stopword Removal Result	26
Table 3. 8 Stemming Result	27
Table 3. 9 Texts to be Vectorized Using TF-IDF	28
Table 3. 10 Term Frequency Calculation.....	28
Table 3. 11 Document Frequency Calculation.....	29
Table 3. 12 IDF Calculating.....	30
Tabel 3. 13 Perhitungan TF-IDF	30
Table 3. 14 L2 Norm Calculation for Each Document	31
Table 3. 15 <i>TF-IDF Results After L2 Normalization</i>	32
Table 3. 16 UMAP	33
Table 3. 17 Distance Between Each Point	35
Table 3. 18 Nearest neighbor	36
Table 3. 19 MRD	37
Table 3. 20 MST Order	39
Table 3. 21 Conversion of Distance to λ value	39
Table 3. 22 Condensed Tree Structure (min_cluster_size = 2).....	41
Table 3. 23 Points That Fall Out from the Condensed Tree	41
Table 3. 24 Cluster Persistence in the Condensed Tree	41
Table 3. 25 summarizes the overall stability of all clusters	42
Table 3. 26 Final Cluster Assignment for Each Document	43
Tabel 3. 27 Deskripsi Tugas.....	43
Table 3. 28 Recommendation Calculation Results	46
Table 3. 29 Cluster Evaluation Quality.....	48
Table3. 30 Recommendation Ranking Results	49

LIST OF FIGURE

Figure 2. 1 Mutual reachability distance ilustration.....	14
Figure 3. 1 Research Stage.....	19
Figure 3. 2 Developer Log Data.....	22
Figure 3. 3 Testing Data.....	22
Figure 3. 4 Data Pre-Processing Stage.....	23
Figure 4. 1 Number of Log Data per Developer	51
Figure 4. 2 Sample of Developer Activity Log Data	52
Figure 4. 3 Sample Test Data in the Form of Task Descriptions.....	52
Figure 4. 4 Distribution of Tasks per Developer in the Test Data	53
Figure 4. 5 Data Before Preprocessing	55
Figure 4. 6 Data After Preprocessing.....	55
Figure 4. 7 Filter Result	58
Figure 4. 8 TF-IDF Results on the Dataset	60
Figure 4. 9 Visualization with “min_cluster_size” = 2 and “min_sample” = 2....	67
Figure 4. 10 Visualization “min_cluster_size” = 5 and “min_samples” = 2	68
Figure 4. 11 Visualization “min_cluster_size” = 10 and “min_samples” = 2	69
Figure 4. 12 Visualization “min_cluster_size” = 15 and “min_samples” 2.....	70
Figure 4. 13 Visualization “min_cluster_size” = 15 and “min_samples” 2.....	70
Figure 4. 14 Visualization “min_cluster_size” = 2 and “min_samples” = 2	72
Figure 4. 15 Visualization “min_cluster_size” = 2 and “min_samples” = 5	73
Figure 4. 16 Visualization “min_cluster_size” = 2 and “min_samples” 10.....	74
Figure 4. 17 Visualization “min_cluster_size” = 2 and “min_samples” = 15	75
Figure 4. 18 Visualization “min_cluster_size” = 2 and “min_samples” = 20	76
Figure 4. 19 Recomendation Result of CBF	77
Figure 4. 20 Recomendation Result of CBF+HDBSCAN.....	78
Figure 4. 21 Evaluation of "min_cluster_size" Variations.....	80
Figure 4. 22 Evaluation of "min_samples"	81
Figure 4. 23 Evaluation Metrics for HDBSCAN Parameter Configurations.....	82
Figure 4. 24 CBF+HDBSCAN: Metrik vs Top-k Log (mCS=2, mS=2).....	83
Figure 4. 25 CBF: Metrik vs Top-k Log.....	84
Figure 4. 26 Comparison of Metrics: CBF+Clustering vs CBF	85

Figure 4. 27 Recommendation Comparison Based on Cosine Similarity Threshold Variations	86
Figure 4. 28 Label Log.....	88
Figure 4. 29 System Testing Input	88
Figure 4. 30 Recommendation Result.....	89