

**PENGELOMPOKAN SKILL REQUIREMENT PADA LOWONGAN
PEKERJAAN ALUMNI MENGGUNAKAN ALGORITMA K-MEANS**

SKRIPSI

Diajukan untuk Memenuhi Sebagian dari
Syarat Memperoleh Gelar Sarjana Komputer
Program Studi Ilmu Komputer



oleh
Rizki Nugraha
1506748

**PROGRAM STUDI ILMU KOMPUTER
DEPARTEMEN PENDIDIKAN ILMU KOMPUTER
FAKULTAS PENDIDIKAN MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS PENDIDIKAN INDONESIA
BANDUNG
2020**

**PENGELOMPOKAN *SKILL REQUIREMENT* PADA LOWONGAN
PEKERJAAN ALUMNI MENGGUNAKAN ALGORITMA *K-MEANS***

Oleh
Rizki Nugraha
NIM 1506748

Sebuah skripsi yang diajukan untuk memenuhi salah satu syarat memperoleh gelar
Sarjana Komputer pada Program Studi Ilmu Komputer Fakultas Pendidikan
Matematika dan Ilmu Pengetahuan Alam

© Rizki Nugraha 2020

Universitas Pendidikan Indonesia

Januari 2020

Hak Cipta dilindungi Undang-Undang

Skripsi ini tidak boleh diperbanyak seluruhnya atau sebagian, dengan dicetak
ulang, difotokopi, atau cara lainnya tanpa izin dari penulis

RIZKI NUGRAHA

1506748

**PENGELOMPOKAN *SKILL REQUIREMENT* PADA LOWONGAN
PEKERJAAN ALUMNI MENGGUNAKAN ALGORITMA *K-MEANS***

Disetujui dan disahkan oleh pembimbing:

Pembimbing I,



Dr. Rani Megasari, M.T.

NIP. 198705242014042002

Pembimbing II,

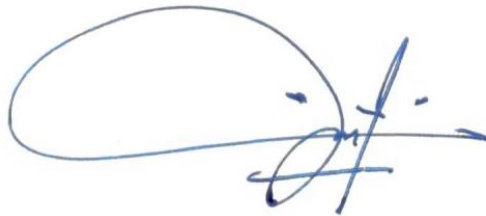


Erna Piantari, S. Kom., M.T.

NIP. 920171219890224201

Mengetahui,

Ketua Departemen Pendidikan Ilmu Komputer



Lala Septem Riza, M.T., Ph. D.

NIP. 197809262008121001

PENGELOMPOKAN *SKILL REQUIREMENT* PADA LOWONGAN PEKERJAAN ALUMNI MENGGUNAKAN ALGORITMA *K-MEANS*

Oleh

Rizki Nugraha – nugraharizki80@gmail.com

1506748

ABSTRAK

Pengangguran merupakan salah satu permasalahan besar di Indonesia. Badan Pusat Statistik menunjukkan penurunan angka pengangguran di tingkat SD, SMP, SMK, dan SMA, namun justru mengalami peningkatan pada tingkat Diploma dan Sarjana. Salah satu faktor penyebab permasalahan ini adalah kesenjangan yang terjadi antara industri dengan pendidikan tinggi. Permasalahan ini dapat diatasi dengan memberikan lulusan pendidikan tinggi pengetahuan mengenai kebutuhan industri saat ini. Salah satunya dapat diperoleh melalui lowongan pekerjaan yang tersebar dalam sebuah media sosial, salah satunya grup alumni Departemen Pendidikan Ilmu Komputer UPI. Untuk mempermudah lulusan mendapatkan pengetahuan tersebut dilakukan pengelompokan lowongan pekerjaan dengan metode k-Means. Pemrosesan lowongan pekerjaan. Sebelum dapat dikelompokkan harus mengalami praproses terlebih dahulu. Praproses yang dilakukan menggunakan kerangka kerja text mining, langkah-langkahnya terdiri atas case folding, tokenizing, filtering, dan stemming. Khusus untuk stemming menggunakan algoritma Nazief dan Adriani karena menggunakan bahasa Indonesia. Hasil praproses direpresentasikan menjadi angka menggunakan algoritma TF-IDF dan ekstraksi fitur PCA untuk mereduksi dimensi. Setelah itu dilakukan pengelompokkan menggunakan algoritma k-Means dengan jumlah cluster yang bervariasi dan evaluasi menggunakan metode Elbow dan Silhouette Coefficient. Dari hasil analisis evaluasi cluster didapatkan nilai Sum of Squared Error sebesar 2.6517 dan nilai silhouette sebesar 0.5795 pada jumlah cluster 3.

Kata Kunci: *Clustering, Text Mining, TF-IDF, lowongan pekerjaan, Silhouette coefficient, Elbow method*

**SKILL REQUIREMENT CLUSTERING IN ALUMNI JOB VACANCIES
WITH K-MEANS ALGORITHM**

Arranged by

Rizki Nugraha – nugraharizki80@gmail.com

1506748

ABSTRACT

Unemployment is one of the big problems in Indonesia. The Central Statistics Agency showed a decrease in unemployment at the elementary, junior high, vocational, and high school levels, but instead experienced an increase at the Diploma and Bachelor level. One factor causing this problem is the gap between the industry and higher education. This problem can be overcome by giving graduates of tertiary education knowledge of current industry needs. This knowledge can be obtained through job vacancies spread in a social media, one of them is the alumni group of UPI Computer Science Education Department. To make it easier for graduates to obtain this knowledge, a grouping of job vacancies is conducted using the k-Means method. Job vacancy processing Before it can be grouped, it must be preprocessed. The pre-processing is carried out using a text mining framework, the steps consist of case folding, tokenizing, filtering, and stemming. Especially for stemming using the Nazief and Adriani algorithm because it uses Indonesian. Preprocess results are represented as numbers using the TF-IDF algorithm and PCA feature extraction to reduce dimensions. After that the grouping is done using the k-Means algorithm with a variety of clusters and evaluation using the Elbow and Silhouette Coefficient methods. From the cluster evaluation analysis results obtained a Sum of Squared Error of 2.6517 and a silhouette value of 0.5795 on the number of clusters 3.

Keyword: *Clustering, Text Mining, TF-IDF, Job Vacancy, Silhouette Coefficient, Elbow Method*

DAFTAR ISI

PERNYATAAN.....	iv
ABSTRAK.....	i
<i>ABSTRACT</i>	ii
KATA PENGANTAR.....	iii
UCAPAN TERIMAKASIH.....	iv
DAFTAR ISI.....	vi
DAFTAR TABEL.....	ix
DAFTAR GAMBAR.....	x
DAFTAR LAMPIRAN.....	xii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang Masalah.....	1
1.2 Rumusan Masalah.....	4
1.3 Tujuan Penelitian.....	5
1.4 Manfaat Penelitian.....	5
1.5 Batasan Masalah Penelitian.....	5
1.6 Sistematika Penulisan.....	6
BAB II KAJIAN PUSTAKA.....	7
2.1 Preprocessing Text.....	7
2.1.1 Case Folding.....	7
2.1.2 Tokenizing.....	8
2.1.3 Filtering.....	9
2.1.4 Stemming.....	10
2.2 Pembobotan Data Menggunakan TF-IDF.....	12
2.3 PCA (<i>Principal Component Analysis</i>).....	18
2.4 Clustering.....	19
2.4.1 Algoritma k-Means.....	20
2.4.2 Euclidean Distance.....	25
2.5 Evaluasi Clustering.....	25
2.6 Penelitian Terkait.....	28
BAB III METODOLOGI PENELITIAN.....	34

3.1	Desain Penelitian	34
3.1.1	Perumusan Masalah	35
3.1.2	Pengumpulan Data	35
3.1.3	Studi Literatur	35
3.1.4	Model <i>clustering</i>	35
3.1.5	Pengembangan Perangkat Lunak <i>clustering</i> lowongan pekerjaan ..	37
3.1.6	Skenario Eksperimen dan Eksperimen	37
3.1.7	Hasil Eksperimen	37
3.1.8	Analisis dan Kesimpulan	38
3.2	Metode Penelitian	38
3.2.1	Metode Pengumpulan Data	38
3.2.2	Metode Pengembangan Perangkat Lunak	38
3.3	Perangkat dan Data Penelitian	40
3.3.1	Perangkat Penelitian	40
3.3.2	Data Penelitian	41
BAB IV HASIL PENELITIAN DAN PEMBAHASAN		42
4.1	Pengumpulan Data	42
4.2	Pengembangan Perangkat Lunak	42
4.2.1	Analisis	42
4.2.2	Desain	43
4.2.3	Implementasi	44
4.2.4	Pengujian	49
4.3	Rancangan Skenario Eksperimen	50
4.3.1	Skenario Pertama Tanpa Menentukan Batas Minimal DF	51
4.3.2	Skenario Kedua Menentukan Batas Minimal DF	51
4.4	Hasil Eksperimen	52
4.4.1	Skenario Pertama Tanpa Menentukan Batas Minimal DF	53
4.4.2	Skenario Kedua Menentukan Batas Minimal DF	60
4.5	Analisis Hasil Eksperimen	67
BAB V KESIMPULAN DAN SARAN		72
5.1	Kesimpulan	72
5.2	Saran	72

DAFTAR PUSTAKA	74
LAMPIRAN	78

DAFTAR TABEL

Tabel 2.1 Contoh data lowongan pekerjaan.....	13
Tabel 2.2 Pemetaan jumlah kata terhadap dokumen lowongan pekerjaan.....	13
Tabel 2.3 Perhitungan DF beserta IDF dari keseluruhan term.....	15
Tabel 2.4 Hasil akhir perhitungan TF-IDF	16
Tabel 2.5 Pemetaan data awal clustering	22
Tabel 2.6 Pemetaan data pada iterasi pertama	22
Tabel 2.7 Daftar anggota cluster pada iterasi pertama.....	23
Tabel 2.8 Pemetaan data pada iterasi kedua.....	23
Tabel 2.9 Daftar anggota cluster pada iterasi kedua	24
Tabel 2.10 Pemetaan data pada iterasi ketiga	24
Tabel 2.11 Daftar anggota cluster pada iterasi ketiga	24
Tabel 2.12 Hasil pengelompokan dengan jumlah cluster berbeda.....	26
Tabel 2.13 Perhitungan jarak titik A dengan titik lain dalam satu cluster	26
Tabel 2.14 Perhitungan jarak titik A dengan titik lain dalam cluster berbeda.....	26
Tabel 2.15 Nilai $a(i)$ untuk setiap titik dan cluster.....	27
Tabel 2.16 Nilai $b(i)$ untuk setiap titik dan cluster.....	27
Tabel 2.17 Perhitungan silhouette untuk keseluruhan cluster.....	27
Tabel 4.1 Perancangan fungsi beserta kegunaan.....	44
Tabel 4.2 Pengujian error handling pada program clustering	50
Tabel 4.3 Hasil iterasi pengelompokan pada skenario pertama.....	55
Tabel 4.4 Hasil iterasi pengelompokan pada skenario kedua	63
Tabel 4.5 Perbandingan jumlah iterasi pada tiap skenario.....	67
Tabel 4.6 Hasil evaluasi SSE dan <i>silhouette</i> pada skenario pertama.....	70
Tabel 4.7 Hasil evaluasi SSE dan <i>silhouette</i> pada skenario kedua	70

DAFTAR GAMBAR

Gambar 1.1 Jumlah pengangguran pada berbagai jenjang pendidikan	1
Gambar 2.1 <i>Pseudocode case folding</i> pada tahap praproses	8
Gambar 2.2 <i>Pseudocode filtering</i> pada tahap praproses.....	10
Gambar 2.3 <i>Pseudocode stemming</i> tahap praproses.....	12
Gambar 2.4 <i>Pseudocode TF-IDF</i> tahap pembobotan	18
Gambar 2.5 <i>Clustering</i> dengan Metode Partisi	20
Gambar 3.1 Desain Penelitian	34
Gambar 3.2 Model Pengembangan Perangkat Lunak <i>Waterfall</i>	39
Gambar 4.1 Contoh 10 data lowongan pekerjaan sebelum diproses	42
Gambar 4.2 Kode program untuk import data dari file ke dalam variabel.....	45
Gambar 4.3 Kode program praproses data	45
Gambar 4.4 Kode program untuk pembobotan	46
Gambar 4.5 Kode program pereduksi dimensi matriks	46
Gambar 4.6 Kode program dari fungsi <i>findCentroid</i>	47
Gambar 4.7 Kode program dari fungsi <i>findMemberCentroid</i>	48
Gambar 4.8 Kode program evaluasi <i>clustering</i>	49
Gambar 4.9 Kode program <i>labelling</i> terhadap <i>cluster</i> yang terbentuk	49
Gambar 4.10 Kode program untuk menampilkan anggota dan label <i>cluster</i>	49
Gambar 4.11 Kode program pada skenario pertama	51
Gambar 4.12 Kode program pada skenario kedua	52
Gambar 4.13 Contoh data sebelum dilakukan praproses	52
Gambar 4.14 Contoh data setelah dilakukan praproses.....	52
Gambar 4.15 Hasil pembobotan TF-IDF skenario pertama pada 5 data terakhir.	53
Gambar 4.16 Hasil PCA skenario pertama pada 5 data terakhir	54
Gambar 4.17 Contoh data akhir sebelum pengelompokan skenario pertama	54
Gambar 4.18 Sebaran data setelah dilakukan PCA	54
Gambar 4.19 Hasil pengelompokan dengan jumlah $k=2$	56
Gambar 4.20 Hasil pengelompokan dengan jumlah $k=3$	56
Gambar 4.21 Hasil pengelompokan dengan jumlah $k=4$	57
Gambar 4.22 Hasil pengelompokan dengan jumlah $k=5$	58

Gambar 4.23 Hasil pengelompokan dengan jumlah $k=6$	58
Gambar 4.24 Hasil pengelompokan dengan jumlah $k=7$	59
Gambar 4.25 Hasil pengelompokan dengan jumlah $k=8$	59
Gambar 4.26 Hasil pengelompokan dengan jumlah $k=9$	60
Gambar 4.27 Hasil pengelompokan dengan jumlah $k=10$	60
Gambar 4.28 Hasil pembobotan TF-IDF skenario kedua pada 5 data terakhir	61
Gambar 4.29 Contoh Hasil PCA skenario kedua	61
Gambar 4.30 Sebaran data setelah dilakukan PCA skenario kedua	62
Gambar 4.31 Contoh data akhir sebelum pengelompokan skenario kedua	62
Gambar 4.32 Hasil pengelompokan dengan jumlah $k=2$ kedua	63
Gambar 4.33 Hasil pengelompokan dengan jumlah $k=3$ kedua	64
Gambar 4.34 Hasil pengelompokan dengan jumlah $k=4$ kedua	64
Gambar 4.35 Hasil pengelompokan dengan jumlah $k=5$ kedua	65
Gambar 4.36 Hasil pengelompokan dengan jumlah $k=6$ kedua	65
Gambar 4.37 Hasil pengelompokan dengan jumlah $k=7$ kedua	66
Gambar 4.38 Hasil pengelompokan dengan jumlah $k=8$ kedua	66
Gambar 4.39 Hasil pengelompokan dengan jumlah $k=9$ kedua	66
Gambar 4.40 Hasil pengelompokan dengan jumlah $k=10$ kedua	67
Gambar 4.40 Nilai SSE pada skenario pertama	68
Gambar 4.41 Nilai <i>silhouette</i> pada skenario pertama	68
Gambar 4.42 Nilai SSE pada skenario kedua	69
Gambar 4.43 Nilai <i>silhouette</i> pada skenario kedua	69
Gambar 4.44 Pusat <i>cluster</i> dan jumlah anggota pada $k=3$ skenario kedua	71
Gambar 4.45 Anggota <i>cluster</i> dengan labelnya	71

DAFTAR LAMPIRAN

Lampiran 1 Data yang digunakan pada penelitian	78
Lampiran 2 Data lowongan pekerjaan setelah praproses	84
Lampiran 3 Data akhir setelah pelabelan	89

DAFTAR PUSTAKA

- Abualigah, L. M., Khader, A. T., Al-Betar, M. A., & Alomari, O. A. (2017). Text Feature Selection With a Robust Weight Scheme and Dynamic Dimension Reduction to Text Document Clustering. *Expert Systems with Applications*, 84, 24–36.
- Afzali, M., & Kumar, S. (2019). Text Document Clustering: Issues and Challenges. *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 263–268.
- Agnihotri, D., Verma, K., & Tripathi, P. (2014). Pattern and Cluster Mining on Text Data. *2014 Fourth International Conference on Communication Systems and Network Technologies*, 428–432.
- Agusta, L. (2009). Perbandingan Algoritma Stemming Porter dengan Algoritma Nazief & Adriani untuk Stemming Dokumen Teks Bahasa Indonesia. *Konferensi Nasional Sistem Dan Informatika, 2009*, 196–201.
- Beil, F., Ester, M., & Xu, X. (2002). Frequent Term-Based Text Clustering. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 436–442.
- Bholowalia, P., & Kumar, A. (2014). EBK-means: A clustering technique based on elbow method and k-means in WSN. *International Journal of Computer Applications*, 105(9).
- Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.
- Handayani, T. (2015). Relevansi lulusan perguruan tinggi di Indonesia dengan kebutuhan tenaga kerja di era global. *Jurnal Kependudukan Indonesia*, 10(1), 53–64.
- Hartono, O. R. (2016). Indonesian Stoplist. Retrieved January 12, 2019, from <https://www.kaggle.com/oswinrh/indonesian-stoplist>
- Huang, A. (2008). Similarity Measures For Text Document Clustering. *Proceedings of The Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008)*, Christchurch, New Zealand, 4, 9–56.
- Hudin, M. S., Fauzi, M. A., & Adinugroho, S. (2018). Implementasi Metode Text

- Mining dan K-Means Clustering untuk Pengelompokan Dokumen Skripsi (Studi Kasus: Universitas Brawijaya). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 2(11), 5518–5524. Retrieved from <http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/3332>
- Iswahyuni, A. D. (2018). Desain Kurikulum Perguruan Tinggi untuk Mengeliminasi Gap Persepsi Perguruan Tinggi degan Industri. *Ratih: Jurnal Rekayasa Teknologi Industri Hijau*, 2(2), 14.
- Jolliffe, I. (2011). *Principal Component Analysis*. Springer.
- Kusrini, K. (2015). Grouping of Retail Items by Using K-Means Clustering. *Procedia Computer Science*, 72, 495–502. <https://doi.org/10.1016/j.procs.2015.12.131>
- Langgeni, D. P., Baizal, Z. K. A., & AW, Y. F. (2015). Clustering Artikel Berita Berbahasa Indonesia Menggunakan Unsupervised Feature Selection. *Seminar Nasional Informatika (SEMNASIF)*, 1(4).
- Liu, X.-W., He, P.-L., & Wang, H.-Y. (2005). The Research of Text Clustering Algorithms Based on Frequent Term Sets. *2005 International Conference on Machine Learning and Cybernetics*, 4, 2352–2356.
- MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1(14), 281–297.
- Mahmood, S., & Al-Rufaye, F. M. L. (2017). Arabic Text Mining Based on Clustering and Coreference Resolution. *2017 International Conference on Current Research in Computer Science and Information Technology (ICCRIT)*, 140–144.
- Nazief, B., & Adriani, M. (1996). Confix Stripping: Approach to Stemming Algorithm for Bahasa Indonesia. *Internal Publication, Faculty of Computer Science, University of Indonesia, Depok, Jakarta*, 41.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, 12, 2825–2830.
- Pusparisa, Y. (2019). Angka Pengangguran Lulusan Universitas Meningkat. Retrieved from <https://katadata.co.id/infografik/2019/05/17/angka->

pengangguran-lulusan-perguruan-tinggi-meningkat

- Robertson, S. (2004). Understanding Inverse Document Frequency: On Theoretical Arguments for IDF. *Journal of Documentation*, 60(5), 503–520.
- Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to The Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.
- Slamet, C., Rahman, A., Ramdhani, M. A., & Dharmalaksana, W. (2016). Clustering The Verses of The Holy Qur'an Using K-Means Algorithm. *Asian Journal of Information Technology*, 15(24), 5159–5162. <https://doi.org/10.3923/ajit.2016.5159.5162>
- Soleh, A. (2017). Masalah Ketenagakerjaan dan Pengangguran di Indonesia. *Cano Ekonomos*, 6(2), 83–92.
- Sommerville, I. (2010). *Software engineering*. New York: Addison-Wesley.
- Statistik, B. P. (1999). *Berita Resmi Statistik*.
- Statistik, B. P. (2019). *Berita Resmi Statistik*.
- Sugianto, D. (2018). BI Catat Jumlah Lowongan Kerja di Kuartal II-2018 Naik 19,2%. Retrieved January 10, 2020, from <https://finance.detik.com/berita-ekonomi-bisnis/d-4111902/bi-catat-jumlah-lowongan-kerja-di-kuartal-ii-2018-naik-192>
- Sukamto, R. A., & Shalahuddin, M. (2011). Rekayasa Perangkat Lunak Menggunakan CodeIgniter dan JQuery. *Yogyakarta: Andi*.
- Thinsungnoen, T., Kaoungku, N., Durongdumronchai, P., Kerdprasop, K., & Kerdprasop, N. (2015). *The Clustering Validity with Silhouette and Sum of Squared Errors*. 44–51. <https://doi.org/10.12792/iciae2015.012>
- Thomas, R. E., & Khan, S. S. (2016). Improved Clustering Technique Using Metadata for Text Mining. *2016 International Conference on Communication and Electronics Systems (ICCES)*, 1–5.
- Tunali, V., & Bilgin, T. T. (2012). Examining the Impact of Stemming on Clustering Turkish Texts. *2012 International Symposium on Innovations in Intelligent Systems and Applications*, 1–4.
- Vijayarani, S., Ilamathi, J., Nithya, M., Ilamathi, M. J., Nithya, M., Professor, A., & Research Scholar, M. P. (2015). Preprocessing Techniques for Text Mining

- An Overview. *International Journal of Computer Science & Communication Networks*, 5(1), 7–16. <https://doi.org/10.1016/j.procs.2013.05.286>

Zhang, Y., & Jiang, M. (2010). Chinese Text Mining Based on Subspace Clustering. *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, 4, 1617–1620.