Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.2023106899
Vol. 10, No. 4, Agustus 2023, hlm. 781-790 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

PENGELOMPOKAN HASIL PENCARIAN SKRIPSI BERBAHASA INDONESIA MENGGUNAKAN METODE DBSCAN DENGAN PEMBOBOTAN BM25

Rangga Adi Satria[*1], Indriati[2], Sutrisno[3]

[1,2,3]Universitas Brawijaya, Malang
Email: [1]adirangga922@gmail.com, [2]indriati.tif@ub.ac.id, [3]trisno@ub.ac.id
[*]Penulis Korespondensi

(Naskah masuk: 05 Januari 2023, diterima untuk diterbitkan: 25 Juli 2023)

Abstrak

Skripsi merupakan tugas akhir yang disusun oleh mahasiswa sebagai persyaratan untuk memperoleh gelar sarjana. Mesin pencari untuk mempermudah pencarian dokumen skripsi yang disimpan pada perpustakaan maupun penyimpanan digital umumnya menggunakan metode sederhana dengan mengembalikan dokumen yang mengandung potongan kata atau identik dengan kata kunci, sehingga dokumen yang diperoleh kurang relevan. Hasil pencarian dapat dikelompokan sehingga dokumen tersaji dengan lebih terperinci dan memudahkan pencarian lebih lanjut. Guna mengelompokan hasil pencarian skripsi berbahasa Indonesia, dengan menggunakan judul dan abstrak skripsi, digunakan pembobotan kata BM25 dan pengelompokan DBSCAN, metode pengelompokan yang mempertimbangkan kepadatan titik sampel dokumen. Pengujian dilakukan dengan mengukur hasil pengelompokan menggunakan rata-rata *silhouette coefficient* terhadap parameter epsilon dan MinPts pada metode DBSCAN, serta k1 dan b pada pembobotan BM25 dengan 4 skenario yang berbeda. Hasil pengujian menunjukan bahwa parameter k1 dan b pada pembobotan BM25 cukup mempengaruhi kualitas pengelompokan dengan metode DBSCAN. Hasil rata – rata *silhouette coefficient* terbaik untuk masing masing skenario secara berurutan adalah 0.722, 0.762, 0.945 dan 0.907 dengan parameter terbaik berupa k1=1.8, b=0.5, epsilon=0.1 dan MinPts=5 pada skenario pertama. k1=1.9, b=0.5, epsilon=0.1 dan MinPts=5 pada skenario kedua. k1=1.4, *b*=0.55, epsilon=0.1 dan MinPts=5 pada skenario ketiga dan k1=1.8, *b*=0.65, epsilon=0.1 dan MinPts=5 pada skenario keempat.

Kata kunci: *Skripsi, Judul, Abstrak, DBSCAN, BM25, Silhouette coefficient*

THE CLUSTERING OF THESIS SEARCH RESULTS IN INDONESIAN USING THE DBSCAN METHOD WITH BM25 WEIGHTING

Abstract

Thesis is a final project that must be completed by students as requirement to obtain a bachelor degree. Search engines used for searching thesis documents stored in libraries or digital storage generally use a simple method by returning documents that contain a snippet of the word or are identical to the keywords, so the obtained documents become less relevant. Search results can be clustered with the purpose of presenting the documents in more detailed way and to ease further searches. In order to cluster the search results of Indonesian language thesis, using the title and abstract of the thesis, BM25 word weighting and DBSCAN clustering were used, a clustering method that considers the document sample density point. The test performed by measuring the clustering results using the average silhouette coefficient on the epsilon and MinPts parameters in the DBSCAN method, as well as k1 and b in the BM25 weighting on 4 different scenarios. The test results show that k1 and b parameters on BM25 weighting is quite affecting the quality of the clustering results using DBSCAN method. The best average silhouette coefficient results for each scenario sequentially are 0.722, 0.762, 0.945 and 0.907 by using the best parameters in the form of k1=1.8, b=0.5, epsilon=0.1 and MinPts=5 in the first scenario. k1=1.9, b=0.5, epsilon=0.1 and MinPts=5 in the second scenario. k1=1.4, b=0.55, epsilon=0.1 and MinPts=5 in the third scenario and k1=1.8, b=0.65, epsilon=0.1 and MinPts=5 in the fourth scenario

Keywords: *Thesis, Title, Abstract, DBSCAN, BM25, Silhouette coefficient*

1. PENDAHULUAN

Skripsi adalah istilah umum di Indonesia untuk menyebut karya tulis ilmiah mahasiswa strata satu (sarjana) sebagai hasil penelitian yang membahas suatu permasalahan di bidang ilmu tertentu (Hadi, 2017). Seperti yang dijelaskan oleh Pramudita, et al.

(2021), mahasiswa bertanggung jawab sepenuhnya atas penentuan topik penelitian skripsi begitu pula dalam penyusunan skripsi.

Menurut Ramadhana, Cut, dan Husna (2019) perpustakaan menampung skripsi yang telah diselesaikan dalam bentuk buku oleh mahasiswa sehingga dapat digunakan sebagai referensi bagi mahasiswa lainnya. Menurut Sari, et al. (2021), skripsi juga dapat disimpan pada perpustakaan digital atau repository sehingga dapat mempermudah pencarian skripsi. Baik skripsi yang telah dibukukan maupun dalam bentuk digital memerlukan sebuah mesin pencari guna mempermudah pihak tertentu dalam menemukan sebuah dokumen skripsi.

Salah satu metode sederhana yang digunakan dalam mesin pencari adalah dengan mengembalikan dokumen dokumen yang mengandung potongan kata dari kata kunci. Priandono, Hakimah, dan Rozi (2020), menjelaskan bahwa mesin pencarian skripsi yang menggunakan SQL query hanya akan mengembalikan dokumen yang memiliki kata atau potongan kata sesuai dengan query dari pengguna sehingga hasil yang diperoleh kurang relevan. *Information retrieval* digunakan untuk memutuskan dokumen yang harus diambil dari koleksi untuk memenuhi kebutuhan informasi dari pengguna (Hermawan & Ismiati, 2020). Menurut B, dan Hetami (2015) proses *information retrieval* terdiri dari beberapa bagian yang diantaranya adalah pembobotan pada kata dan pemeringkatan dokumen berdasarkan kesesuaiannya dengan query.

Penelitian terkait pembobotan dokumen sebelumnya telah dilakukan oleh Tinega, Mwangi, dan Rimiru (2018) dengan membandingkan 3 model *information retrieval* yang diantaranya adalah *Boolean Model*, *Vector Space Model* (VSM) dan BM25 menunjukan bahwa BM25 memiliki hasil yang lebih unggul daripada kedua model lainnya. Dengan menggunakan 300 dokumen jurnal tahapan evaluasi menunjukan bahwa *boolean model* tidak memiliki keluaran yang relevan sehingga ditinggalkan. Hasil evaluasi dengan 3 *query* yang berbeda menunjukan bahwa BM25 lebih unggul pada nilai *precision* dan *recall* dibandingkan VSM pada keseluruhan *query.*

Pengelompokan dokumen dapat dilakukan sebelum atau sesudah proses temu kembali (Zhang, et al., 2001). Dokumen hasil pencarian dapat dikelompokan kembali dengan tujuan dokumen tersaji dengan lebih terperinci berdasarkan kesamaan topik dan memudahkan pencarian lebih lanjut. Dokumen-dokumen yang relevan dengan suatu *query* cenderung memiliki kemiripan satu sama lain dari pada dokumen yang tidak relevan (Sugiyamto, Surarso, & Sugiharto., 2014)

Penelitian yang dilakukan oleh Rachman, Goejantoro, dan Amijaya (2020) mengelompokan dokumen skripsi menggunakan metode K-Means. Dokumen skripsi yang ada diproses dengan pembobotan TF-IDF dan dikelompokan dengan metode K-Means dengan nilai k pada rentang 2 hingga 6. Proses evaluasi diukur dengan *silhouette coefficient* dan didapati bahwa dengan k = 2 memiliki nilai *silhouette coefficient* terbaik sebesar 0,12. Namun didapati bahwa dengan nilai 0,12 termasuk dalam kategori hasil pengelompokan *no structure* sesuai kategori hasil *structure* pada validitas *cluster*.

Salah satu metode pengelompokan selain K-Means adalah DBSCAN (*Density-based Spatial Clustering of Applications with Noise*). Karami dan Johansson (2014) menjelaskan bahwa *cluster* yang terbentuk dari DBSCAN berbasis pada kepadatan objek data yang diukur menggunakan dua parameter yaitu radius *cluster* $epsilon$ dan objek data minimum yang diperlukan di dalam *cluster*. Algortime ini mengumpulkan objek data dengan kepadatan cukup tinggi kedalam *cluster-cluster* dan mengumpulkan objek data dengan kepadatan yang rendah sebagai *noise* (Ayu, 2015).

Penelitian yang dilakukan oleh Isnarwaty dan Irhamah (2019) membandingkan hasil pengelompokan teks dengan metode K-Means dan DBSCAN. Penelitian ini menggunakan data berupa *tweet* yang ditujukan pada 3 layanan ekspedisi. Setiap data teks yang ada diproses dengan pembobotan TF-IDF sebelum dikelompokan dengan metode K-Means dan DBSCAN. Hasil yang diperoleh dengan evaluasi *silhoutette coefficient* menunjukan bahwa metode pengelompokan DBSCAN lebih unggul pada ketiga layanan ekspedisi dibandingkan dengan metode pengelompokan K-Means.

Dari penelitian-penelitian yang telah diuraikan, maka penelitian terkait pengelompokan hasil pencarian skripsi berbahasa Indonesia akan memanfaatkan metode pembobotan BM25 dan metode pengelompokan *Density-based Spatial Clustering of Application with Noise* (DBSCAN). Terdapat beberapa batasan masalah agar penelitian ini mendapatkan hasil yang sesuai dengan yang diharapkan, diantaranya: (1) Permasalahan yang diteliti merupakan pengaruh penggunaan metode pembobotan dan pemeringkatan BM25 terhadap pengelompokan hasil pencarian skripsi dengan metode DBSCAN. (2) Data yang digunakan merupakan data teks pada bagian judul dan abstrak dari dokumen skripsi berbahasa Indonesia. (3) Algoritme *stemming* pada tahapan *preprocessing* menggunakan algoritme Nazief dan Adriani yang telah diterapkan dalam *library* PySastrawi.

Information Retrieval

Information Retrieval merupakan bidang ilmu komputer yang mengkaji mengenai proses pengambilan informasi berdasarkan isi dan konteks dari dokumen-dokumen yang ada (Hasanah, 2017). Menurut penjelasan dari Suhardi, et al. (2021), menemu-kembalikan informasi yang relevan sesuai dengan minat pengguna yang ditargetkan adalah salah satu dari fungsi utama *information retrieval.*

***Best Match* 25 (BM25)**

BM25 adalah sebuah metode untuk melakukan pembobotan dan pemeringkatan pada sebuah kumpulan dokumen. Ghawi dan Pfeffer (2019), menjelaskan bahwa metode BM25 merupakan salah satu metode pengukuran kesamaan yang mirip seperti metode pembobotan TF-IDF untuk melakukan *information retrieval* terhadap sebuah dokumen yang terdiri dari aspek *term frequency transformation* dan *document length normalization*. Menurut Sakariana, Indriati, dan Dewi (2020), metode BM25 dapat dihitung dengan menggunakan Persamaan (1)

$$\boldsymbol{BM25} = \sum_{i=1}^{n} \boldsymbol{IDF}(\boldsymbol{q_i}).\frac{\boldsymbol{f(q_i,D).(k1+1)}}{\boldsymbol{f(q_i,D)+k1.(1-b+b.\frac{|D|}{avgdl})}} \quad (1)$$

Keterangan:
$f(q_i, D)$: Frekuensi *term* q_i pada dokumen *D*
$k1$: Konstanta dengan nilai umumnya sebesar 1,2
b : Konstanta dengan nilai umumnya sebesar 0,75
$|D|$: Jumlah kata pada dokumen *D*
$avgdl$: Rata-rata panjang dokumen dari keseluruhan dokumen
n : Jumlah *term* pada koleksi

Persamaan perhitungan *IDF (Inverse Document Frequency)* yang digunakan pada Persamaan (1) dijabarkan pada Persamaan (2)

$$\boldsymbol{IDF(q_i)} = \boldsymbol{log}\,\frac{\boldsymbol{N-n(q_i)+0.5}}{\boldsymbol{n(q_i)+0.5}} \quad (2)$$

Keterangan:
N : Jumlah seluruh dokumen yang digunakan
$n(q_i)$: Jumlah dokumen yang mengandung *term* q_i
$IDF(q_i)$: *Inverse document frequency* untuk *term* q_i

***Density Based Spatial Clustering of Application with Noise* (DBSCAN)**

DBSCAN membangun area (*cluster*) berdasarkan kepadatan yang terkoneksi dimana setiap objek dari sebuah radius area harus mengandung sejumlah minimum data (Devi, et al., 2015). Algortime DBSCAN membutuhkan dua parameter penting, yaitu parameter radius (Eps) dan jumlah titik minimum untuk membentuk kelompok ($MinPts$) (Isnarwaty & Irhamah, 2019). Menurut Birant dan Kut (2007), algortime dari metode pengelompokan DBSCAN sebagai berikut.

1. Menentukan $epsilon$ (ε) dan keanggotaan minimum ($MinPts$)
2. Memilih sebuah titik sampel p dari data
3. Mengukur jarak antara titik sampel p dengan titik sampel lain menggunakan fungsi jarak. Fungsi jarak yang digunakan pada penelitian ini adalah *euclidean distance.* Menurut Nishom (2019), Rumus perhitungan jarak *euclidean* dapat dihitung dengan menggunakan Persamaan (3)

$$\boldsymbol{d(x,y)} = \sqrt{\sum_{i=1}^{n}(\boldsymbol{x_i} - \boldsymbol{y_i})^2} \quad (3)$$

Keterangan:
$d(x, y)$: Jarak *euclidean* antara titik x dan titik y
x_i : Nilai ke-i pada pada titik data x
y_i : Nilai ke-i pada pada titik data y
n : Banyaknya data

4. Setiap titik yang jaraknya berada dalam jangkuan nilai $epsilon$ (ε) dari titik p menjadi anggota tetangga titik p.
5. Apabila banyaknya anggota tetangga titik p beserta titik p itu sendiri lebih dari sama dengan nilai $MinPts$ maka titik p ditandai sebagai titik inti *(core point)* dan sebuah *cluster* terbentuk.
6. Titik p beserta anggota tetangganya akan dimasukan dalam *cluster* yang telah terbentuk.
7. Secara iteratif melakukan perluasan anggota ketetanggaan sesuai dengan jangkauan nilai $epsilon$ (ε) dari titik inti baru yang sebelumnya merupakan anggota tetangga dari titik inti lama, sehingga setiap anggota ketetanggaan masuk pada *cluster* yang sama.
8. Proses dilanjutkan hingga semua titik sampel sudah dikunjungi.

Silhouette Coefficient

Silhouette Coefficient adalah salah satu metode untuk melakukan evaluasi dalam permasalahan pengelompokan atau *clustering*. *Silhouette coefficient* terdiri dari metode *cohesion* dan *separation*, *cohesion* bertujuan mengukur kedekatan objek dalam *cluster* sedangkan *separation* bertujuan mengukur sebarapa jauh sebuah *cluster* terpisah dengan *cluster* lain (Simanjuntak & Khaira, 2021). Menurut Struyf, Hubert, dan Rousseeuw (1997), tahapan dalam menghitung *silhouette coefficient* sebagai berikut.

Langkah pertama adalah menghitung jarak rata-rata antara titik sampel i terhadap seluruh sampel pada kelompok yang sama. Misalkan untuk setiap titik sampel i dinyatakan A yang merupakan *cluster* dari titik sampel i, maka dapat dinyatakan dengan Persamaan (4)

$$\boldsymbol{a(i)} = \frac{\boldsymbol{1}}{\boldsymbol{|A|-1}}\sum_{\boldsymbol{j\in A, j\neq i}} \boldsymbol{d(i,j)} \quad (4)$$

Keterangan:
A : *Cluster* milik sampel i
$|A|$: Banyak sampel pada *cluster A*
j : Sampel lain selain sampel i dalam *cluster*

sampel A
$d(i,j)$: Jarak antara sampel i dan sampel j
$a(i)$: Rata – rata jarak sampel i dengan seluruh sampel lain pada *cluster* yang sama

Tahap selanjutnya adalah menghitung rata-rata jarak yang paling kecil antara sampel i dengan semua sampel yang berada pada *cluster* yang berbeda. Misalkan C adalah *cluster* yang berbeda dengan A, maka dapat dinyatakan dengan menggunakan Persamaan (5)

$$\boldsymbol{b(i) = \min_{C \neq A} \{\frac{1}{|C|} \sum_{j \in C} d(i,j)\}} \quad (5)$$

Keterangan:
A : *Cluster* milik sampel i
C : *Cluster* selain milik sampel i
$|C|$: Banyak sampel pada *cluster* selain *cluster* milik sampel i
j : Anggota sampel *cluster* C
$d(i,j)$: Jarak antara sampel i dan sampel j

$b(i)$: Rata – rata jarak terkecil sampel i dengan setiap sampel pada *cluster* yang berbeda

Tahapan terakhir adalah dengan menghitung nilai *silhouette coefficient* dari sampel i dengan Persamaan (6)

$$\boldsymbol{s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}} \quad (6)$$

Keterangan:
$a(i)$: Rata – rata jarak sampel i dengan seluruh sampel lain pada *cluster* yang sama
$b(i)$: Rata – rata jarak terkecil sampel i dengan setiap sampel pada *cluster* yang berbeda
$s(i)$: Nilai *silhouette coefficient* dari sampel i

2. METODE PENELITIAN

Metode penelitian berisi tahapan penelitian secara terstruktur dalam melakukan pengelompokan dokumen dengan metode DBSCAN terhadap dokumen hasil pencarian skripsi berbahasa Indonesia yang telah melalui tahapan pembobotan serta pemeringkatan BM25.

Teknik Pengumpulan Data

Penelitian ini menggunakan data primer yang dikumpulkan dengan metode *web scrapping* pada laman *web* http://repository.ub.ac.id yang diakses pada tanggal 06 September 2021 dengan menggunakan *library* beautifulsoup4. Pasangan judul dan abstrak dari skripsi mahasiswa yang diambil adalah dari skripsi mahasiswa Strata 1 (S1), Fakultas Ilmu Komputer, Universitas Brawijaya untuk setiap program studi yaitu Pendidikan Teknologi Informasi, Sistem Informasi, Teknik Informatika, Teknik Komputer, dan Teknologi Informasi, yang selesai disusun pada tahun 2019 hingga 2021. Total keseluruhan data dokumen yang didapatkan adalah 1426 pasangan judul dan juga abstrak.

Metode *Text Preprocessing*

Menurut Işik dan Dağ (2020), *text preprocessing* adalah hal yang perlu dilakukan pertama kali setelah mendapatkan *dataset* karena data tekstual umumnya bersifat tidak terstruktur, sehingga mengandung banyak noise dan informasi yang tidak diperlukan. Tahapan pertama pada penelitian ini adalah dengan membaca daftar *stopwords* milik Tala yang akan digunakan pada tahapan *stopword removal* sebagai daftar acuan *term* atau kata yang akan dihapus dari judul maupun abstrak pada keseluruhan sampel. Tahapan berikutnya dilakukan penggabungan judul dan abstrak menjadi satu data teks sebelum seluruh hurufnya akan diubah huruf kecil pada tahapan *case folding.* Tahapan selanjutnya berupa *stemming* yang berguna untuk mengubah seluruh kata menjadi bentuk dasar. Tahapan berikutnya yaitu *stopword removal* akan melakukan penghapusan terhadap kata atau *term* yang terdapat pada daftar *stopwords* yang sebelumnya telah dibaca. Tahapan terakhir merupakan *tokenization* yang akan memisahkan kalimat menjadi bentuk terkecilnya yaitu *term* atau kata dengan menggunakan pemisah atau *delimiter* berupa karakter spasi.

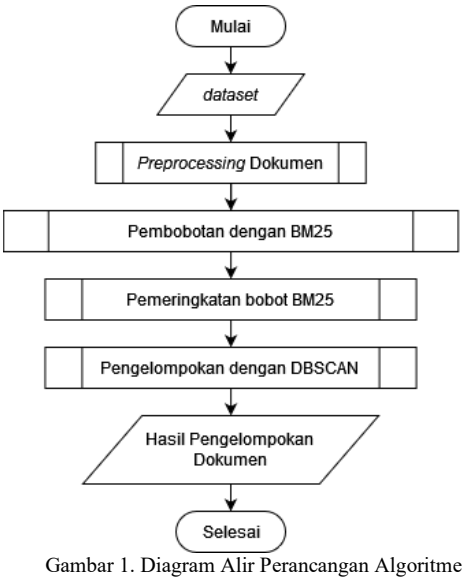
Pembobotan dan Pemeringkatan Dokumen

Pembobotan dan pemeringkatan dokumen pada penelitian ini dilakukan dengan metode BM25. Pembobotan kata dilakukan terhadap setiap kata atau *term* yang terdapat pada masing masing sampel. Tahapan berikutnya dilakukan pemeringkatan dokumen dengan menjumlahkan bobot *term* yang terdapat pada kunci untuk setiap dokumen. Dokumen yang diperingkatkan adalah dokumen yang memiliki jumlah bobot *term* pada kata kunci tidak sama dengan nol, yang berarti hanya dokumen yang tidak mengandung satupun *term* yang terdapat pada kata kunci yang tidak akan diperingkatkan. Dokumen yang diperoleh dari hasil pemeringkatan diperoleh bobot dokumennya dengan menjumlahkan seluruh *term* yang terdapat pada masing masing sampel dokumen. Bobot dari dokumen tersebut kemudian akan dijadikan acuan dalam proses pengelompokan dengan metode DBSCAN.

Implementasi Algoritme

Tahapan dimulai dengan memasukan *input* yang berupa *dataset* yang berisi sampel judul dan abstrak skripsi. Tahapan dilanjutkan dengan *preprocessing* dokumen dengan menerapkan *text preprocessing* untuk keseluruhan sampel. Tahapan selanjutnya adalah pembobotan kata pada setiap sampel dengan BM25, dan dilanjutkan dengan memasukan kata kunci pencarian guna melakukan pemeringkatan

sesuai dengan kata kunci. Bobot dari dokumen yang dikembalikan melalui hasil pemeringkatan kemudian dikelompokan dengan metode *Density Based Spatial Clustering of Applications with Noise* (DBSCAN) untuk memperoleh hasil pengelompokan dokumen hasil pencarian berupa label *cluster* untuk setiap dokumen. Diagram alir dari implementasi algoritme ditampilkan pada Gambar 1.



Gambar 1. Diagram Alir Perancangan Algoritme

3. PENGUJIAN

Penelitian ini akan menguji parameter nilai $epsilon$ (ε) dan $MinPts$ pada metode pengelompokan DBSCAN dan nilai $k1$ serta b pada metode pembobotan BM25. Nilai parameter terbaik untuk masing – masing parameter ditentukan oleh hasil pengelompokan yaitu kualitas dari *cluster* yang diukur menggunakan rata – rata *silhouette coefficient.* Pada penelitian ini digunakan 4 skenario kata kunci pencarian yang berbeda, dan pengujian parameter dilakukan untuk masing masing skenario. Skenario ke-1 menggunakan seluruh sampel, sedangkan skenario ke-2 hingga ke-4 menggunakan sampel hasil dari pemeringkatan dengan kata kunci yang secara berurutan antara lain adalah "Pengembangan sistem informasi manajemen pada lingkungan belajar", "Analisis Pengalaman Pengguna pada *Website E-commerce"* dan "Implementasi K-Means pada analisis sentimen". Masing – masing skenario akan menguji parameter dengan menggunakan sampel hasil pencarian yang dihasilkan sesuai dengan kata kunci yang secara berurutan sebanyak 1426, 1209, 609 dan 743 sampel. Tahapan pengujian pada bagian ini hanya akan menampilkan pengujian pada skenario pertama dari keseluruhan skenario dengan menggunakan sebanyak 1426 sampel.

Pengujian Nilai Epsilon dan MinPts Pada Skenario Pertama

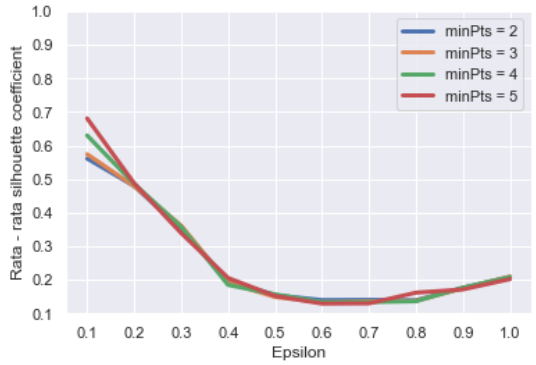
Pengujian nilai $epsilon$ (ε) dan $MinPts$ bertujuan untuk memperoleh nilai parameter $epsilon$ (ε) dan $MinPts$ terbaik pada skenario pertama. Pada pengujian ini dokumen akan melalui tahapan pembobotan dengan nilai terendah dari $k1$ dan b yaitu $k1$=1,2 dan b=0,5. Penelitian yang dilakukan oleh Jambak dan Efendi (2021), dalam menguji pengklasteran dengan DBSCAN untuk mengelompokan data berdimensi tinggi dan rendah dilakukan dengan menggunakan kondisi nilai parameter $epsilon$ (ε) pada rentang nilai 0,1 hingga 1 dan kondisi nilai parameter $MinPts$ pada rentang 2 hingga 5. Pada penelitian ini pengujian metode pengelompokan DSBCAN juga dilakukan dengan menguji nilai $epsilon$ (ε) pada rentang 0,1 hingga 1 dan parameter $MinPts$ pada rentang nilai 2 hingga 5. Hasil pengujian untuk nilai parameter $epsilon$ (ε) dan $MinPts$ terbaik dapat dilihat pada Tabel 1.

Tabel 1. Hasil Pengujian Epsilon (ε) dan MinPts pada Skenario Pertama

Epsilon (ε)	MinPts	Rata – rata *Silhouette coefficient*	Jumlah *Cluster* yang Terbentuk
0.1	2	0.561031	211
0.2	2	0.479302	86
0.3	2	0.351442	54
0.4	2	0.198174	44
0.5	2	0.154400	26
0.6	2	0.139645	22
0.7	2	0.140241	22
0.8	2	0.139020	18
0.9	2	0.176321	17
1	2	0.208807	14
0.1	3	0.574014	133
0.2	3	0.478766	61
0.3	3	0.361726	38
0.4	3	0.191596	33
0.5	3	0.148856	18
0.6	3	0.132787	14
0.7	3	0.136009	16
0.8	3	0.137593	15
0.9	3	0.174755	13
1	3	0.209659	10
0.1	4	0.629942	111
0.2	4	0.485888	58
0.3	4	0.357995	27
0.4	4	0.184777	27
0.5	4	0.156662	17
0.6	4	0.133740	12
0.7	4	0.133684	12
0.8	4	0.136198	11
0.9	4	0.176641	11
1	4	0.207864	9
0.1	**5**	**0.680644**	**92**
0.2	5	0.487534	53
0.3	5	0.339321	22
0.4	5	0.205694	19
0.5	5	0.151886	14
0.6	5	0.129204	11
0.7	5	0.129887	11
0.8	5	0.161763	7
0.9	5	0.171326	8
1	5	0.202132	7

Berdasarkan Tabel 1 dapat diambil kesimpulan bahwa nilai $epsilon$ (ε) dan $MinPts$ menghasilkan kualitas *cluster* terbaik pada skenario pertama ketika $epsilon$ (ε) bernilai sebesar 0.1 dan $MinPts$ sebesar 5. Hal tersebut dibuktikan dengan nilai rata – rata

silhouette coefficient sebesar 0,681 yang merupakan nilai tertinggi dibandingkan dengan pengujian nilai *epsilon* (ε) dan *MinPts* yang lain. Guna mempermudah proses analisis, hasil pada Tabel 1 dapat ditampilkan dalam bentuk grafik garis yang dapat dilihat pada Gambar 2.



Gambar 2. Grafik Rata – Rata *Silhouette coefficient* Terhadap Penggunaan Nilai Epsilon (ε) dan MinPts pada Skenario Pertama

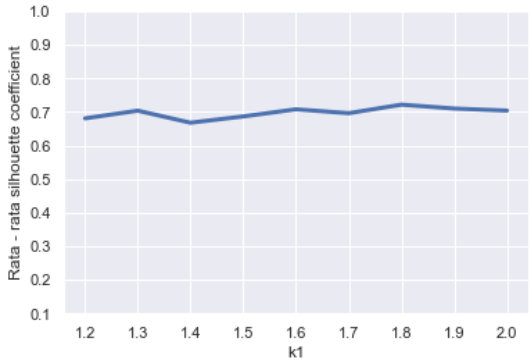
Pengujian Nilai k1 Pada Skenario Pertama

Pengujian nilai $k1$ bertujuan untuk memperoleh nilai parameter k $k1$ 1 terbaik pada skenario pertama. Pada pengujian ini dokumen akan melalui tahapan pembobotan dengan nilai terendah dari b yaitu b =0,5 dan dokumen akan dikelompokan dengan menggunakan nilai *epsilon* (ε) dan *MinPts* terbaik yang diperoleh dari pengujian sebelumnya. Menurut penelitian Ghawi dan Pfeffer (2019), nilai parameter $k1$ berguna untuk menentukan tingkat saturasi terhadap frekuensi *term*. Menurut penelitian Robertson dan Zaragoza (2009), dari berbagai percobaan yang telah dilakukan menunjukan bahwa nilai yang disarankan untuk parameter pada BM25 adalah dan 1,2 hingga 2 untuk $k1$. Penelitian yang dilakukan oleh Hesay, Indriati, dan Adinugroho (2021), menguji nilai parameter $k1$ pada rentang nilai 1,2 hingga 2 dengan peningkatan nilai sebesar 0,1 untuk setiap pengujiannya. Pada penelitian ini juga akan menggunakan variasi nilai $k1$ dengan rentang 1,2 hingga 2 dengan kenaikan nilai sebesar 0,1 pada setiap pengujian. Hasil pengujian untuk nilai parameter $k1$ terbaik dapat dilihat pada Tabel 2.

Tabel 2. Hasil Pengujian k1 pada Skenario Pertama

k1	Rata – rata *Silhouette coefficient*	Jumlah *Cluster* yang Terbentuk
1.2	0.680644	92
1.3	0.703563	84
1.4	0.667689	85
1.5	0.686418	75
1.6	0.707588	88
1.7	0.696063	89
1.8	**0.721516**	**93**
1.9	0.709866	89
2	0.703750	88

Berdasarkan Tabel 2 dapat diambil kesimpulan bahwa nilai $k1$ menghasilkan kualitas *cluster* terbaik pada skenario pertama ketika $k1$ bernilai sebesar 1,8. Hal tersebut dibuktikan dengan nilai rata – rata *silhouette coefficient* sebesar 0,722 yang merupakan nilai tertinggi dibandingkan dengan pengujian nilai $k1$ yang lain. Guna mempermudah proses analisis, hasil pada Tabel 2 dapat ditampilkan dalam bentuk grafik garis yang dapat dilihat pada Gambar 3.



Gambar 3. Grafik Rata – Rata *Silhouette coefficient* Terhadap Penggunaan Nilai k1 pada Skenario Pertama

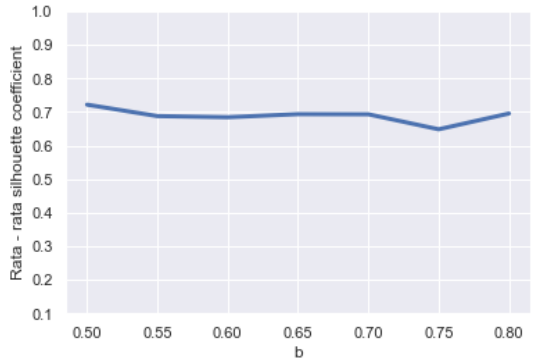
Pengujian Nilai b Pada Skenario Pertama

Pengujian nilai b bertujuan untuk memperoleh nilai parameter b terbaik pada skenario pertama. Pada pengujian ini dokumen akan melalui tahapan pembobotan dengan nilai terbaik dari $k1$ dan dokumen akan dikelompokan dengan menggunakan nilai terbaik dari *epsilon* (ε) dan *MinPts* yang ketiga nilai parameter tersebut diperoleh dari pengujian sebelumnya. Menurut penelitian Ghawi dan Pfeffer (2019), parameter b digunakan untuk mengontrol normalisasi ukuran dokumen. Menurut Robertson dan Zaragoza (2009), dari berbagai percobaan yang telah dilakukan menunjukan bahwa nilai yang disarankan untuk parameter b pada BM25 adalah 0,5 hingga 0,8. Penelitian yang dilakukan oleh Hesay, Indriati, dan Adinugroho (2021) menguji parameter b pada rentang nilai 0,5 hingga 0,8 dengan kenaikan sebesar 0,05 pada setiap pengujiannya. Pada penelitian ini juga akan menggunakan variasi nilai b dengan rentang 0,5 hingga 0,8 dengan kenaikan nilai sebesar 0,05 pada setiap pengujiannya Hasil pengujian untuk nilai parameter b terbaik dapat dilihat pada Tabel 3.

Tabel 3. Hasil Pengujian b pada Skenario Pertama

b	Rata – rata *Silhouette coefficient*	Jumlah *Cluster* yang Terbentuk
0.5	**0.721516**	**93**
0.55	0.687452	83
0.6	0.683994	89
0.65	0.693234	93
0.7	0.692812	89
0.75	0.648116	86
0.8	0.695163	89

Berdasarkan Tabel 3 dapat diambil kesimpulan bahwa nilai b menghasilkan kualitas *cluster* terbaik pada skenario pertama ketika b bernilai sebesar 0,5. Hal tersebut dibuktikan dengan nilai rata – rata *silhouette coefficient* sebesar 0,722 yang merupakan nilai tertinggi dibandingkan dengan pengujian nilai b yang lain. Guna mempermudah proses analisis, hasil pada Tabel 3 dapat ditampilkan dalam bentuk grafik garis yang dapat dilihat pada Gambar 4.



Gambar 4. Grafik Rata – Rata *Silhouette coefficient* Terhadap Penggunaan Nilai b pada Skenario Pertama

4. HASIL DAN PEMBAHASAN

Analisis pengujian terdiri dari analisis pengujian nilai $epsilon\ (\varepsilon)$ dan $MinPts$ sebagai nilai parameter pada metode pengelompokan DBSCAN dan nilai $k1$ dan b sebagai nilai parameter pada metode pembobotan BM25. Berdasarkan hasil evaluasi untuk setiap skenario pengujian, dapat diambil kesimpulan bahwa seluruh variasi nilai parameter, baik pada parameter pengelompokan DBSCAN yaitu $epsilon\ (\varepsilon)$ dan $MinPts$, maupun pada parameter pembobotan BM25 yaitu $k1$ dan b berpengaruh terhadap hasil pengelompokan. Pada hasil pengujian ditemukan bahwa parameter terbaik yang didasarkan pada hasil rata – rata *silhouette coefficient* tertinggi yang digunakan pada pengelompokan dengan metode DBSCAN dan pembobotan BM25 untuk masing masing skenario pengujian dapat dilihat pada Tabel 4.

Tabel 4. Perolehan Nilai Parameter Terbaik Untuk Setiap Skenario Pengujian

Skenario	Eps	$MinPts$	$k1$	b	Rata – rata *Silhouette coefficient*
1	0,1	5	1,8	0,5	0,721516
2	0,1	5	1,9	0,5	0,762294
3	0,1	5	1,4	0,55	0,944634
4	0,1	5	1,8	0,65	0,906822

Pada keseluruhan skenario pengujian parameter $Epsilon\ (\varepsilon)$ dan $MinPts$ dapat diambil kesimpulan bahwa nilai rata – rata *silhouette coefficient* dipengaruhi secara signifikan oleh parameter $Epsilon\ (\varepsilon)$ dan $MinPts$. Penggunaan nilai $Epsilon\ (\varepsilon)$ yang semakin kecil akan meningkatkan kohesi *cluster*, sehingga rata-rata *silhouette coefficient* yang dihasilkan cenderung meningkat. Namun semakin bertambahnya nilai $MinPts$ tidak memastikan nilai rata - rata *silhouette coefficient* semakin baik. Selain itu jumlah *cluster* yang terbentuk dipengaruhi oleh parameter $Epsilon\ (\varepsilon)$ dan $MinPts$. Penggunaan nilai $MinPts$ yang semakin kecil akan meningkatkan jumlah *cluster* yang terbentuk. Namun semakin bertambahnya nilai $Epsilon\ (\varepsilon)$ cenderung menurunkan jumlah *cluster* yang terbentuk.

Pada keseluruhan skenario pengujian variasi nilai parameter $k1$, dapat diambil kesimpulan bahwa nilai rata – rata *silhouette coefficient* yang diperoleh dipengaruhi oleh nilai parameter $k1$. Namun setiap variasi nilai $k1$ memiliki pengaruh berupa peningkatan maupun penurunan nilai rata – rata *silhouette coefficient* yang berbeda beda untuk masing masing kasus skenario pengujian. Selain itu selisih peningkatan serta penurunan nilai rata – rata *silhouette coefficient* pada setiap nilai variasi parameter $k1$ yang ada bernilai kecil, sehingga pola yang terbentuk pada grafik setiap skenario pengujian tidak terlalu signifikan. Penggunaan nilai $k1$ yang semakin besar tidak memastikan nilai rata - rata *silhouette coefficient* semakin meningkat atau menurun. Selain memiliki pengaruh pada nilai rata – rata *silhouette coefficient,* nilai parameter $k1$ juga mempengaruhi jumlah *cluster* yang terbentuk dari proses pengelompokan. Namun sama halnya seperti pada pengaruhnya pada nilai rata – rata *silhouette coefficient,* setiap variasi nilai $k1$ memiliki pengaruh berupa peningkatan maupun penurunan jumlah *cluster* yang berbeda beda untuk masing masing kasus skenario pengujian. Penggunaan nilai $k1$ yang semakin besar tidak memastikan jumlah *cluster* yang terbentuk semakin besar atau kecil.

Pada keseluruhan skenario pengujian variasi nilai parameter b dapat diambil kesimpulan bahwa nilai rata – rata *silhouette coefficient* yang diperoleh dipengaruhi oleh nilai parameter b . Namun setiap variasi nilai b memiliki pengaruh berupa peningkatan maupun penurunan nilai rata – rata *silhouette coefficient* yang berbeda beda untuk masing masing kasus skenario pengujian. Selain itu selisih peningkatan serta penurunan nilai rata – rata *silhouette coefficient* pada setiap nilai variasi parameter b yang ada bernilai kecil, sehingga pola yang terbentuk pada grafik setiap skenario pengujian tidak terlalu signifikan. Penggunaan nilai b yang semakin besar tidak memastikan nilai rata - rata *silhouette coefficient* semakin meningkat atau menurun. Selain memiliki pengaruh pada nilai rata – rata *silhouette coefficient,* nilai parameter b juga mempengaruhi jumlah *cluster* yang terbentuk dari proses pengelompokan. Namun sama halnya seperti pada pengaruhnya pada nilai rata – rata *silhouette coefficient,* setiap variasi nilai b memiliki pengaruh berupa peningkatan maupun penurunan jumlah *cluster* yang berbeda beda untuk masing masing kasus skenario pengujian. Penggunaan nilai b yang

semakin besar tidak memastikan jumlah *cluster* yang terbentuk semakin besar atau kecil.

5. KESIMPULAN DAN SARAN

Penentuan parameter $k1$ dan b pada metode BM25 cukup mempengaruhi kualitas pengelompokan dengan metode DBSCAN yang dibuktikan dengan nilai parameter terbaik $k1$ dan b yang diperoleh bukanlah nilai terendah seperti yang digunakan pada pengujian sebagai nilai bawaan. Walaupun nilai b=0.5 merupakan parameter terbaik pada skenario pertama dan kedua. Hasil rata – rata *silhouette coefficient* untuk skenario pertama, skenario kedua, skenario ketiga, dan skenario keempat yang telah diperoleh secara berurutan adalah 0,722516, 0,762, 0,945, 0,907. Parameter terbaik yang diperoleh berdasarkan rata – rata *silhouette coefficient* tertinggi adalah $k1$=1,8, b=0,5, $Epsilon\ (\varepsilon)$=0,1 dan $minPts$=5 pada skenario pertama. $k1$=1,9, b=0,5, $Epsilon\ (\varepsilon)$=0,1 dan $minPts$=5 pada skenario kedua. $k1$=1,4, b=0,55, $Epsilon\ (\varepsilon)$=0,1 dan $minPts$=5 pada skenario ketiga dan $k1$=1,8, b=0,65, $Epsilon\ (\varepsilon)$=0,1 dan $minPts$=5 pada skenario keempat.

Terdapat beberapa saran yang dapat digunakan untuk membantu pengembangan penelitian serupa yang akan datang, diantaranya:

1. Menambahkan tahapan pemrosesan angka dalam tahapan *preprocessing* guna memproses bilangan, terutama bilangan rasional desimal yang tertulis pada bagian abstrak skripsi.
2. Terdapat beberapa kata yang mengalami kesalahan pengetikan pada bagian judul maupun abstrak pada dokumen yang digunakan. Pada penelitian selanjutnya dapat dilakukan proses perbaikan atau penghapusan pada kata yang mengalami kesalahan pengetikan.
3. Memperbanyak skenario pengujian dengan menggunakan variasi kata kunci pencarian dengan beragam tingkat umum dan spesifik.

DAFTAR PUSTAKA

AYU, P. D. W., 2015. Perbandingan Kinerja Fuzzy C-Means dan DBSCAN Dalam Segmentasi Citra USG Kepala Janin. Jurnal Sistem dan Informatika, 9(2), pp. 79-85.

B, D. W. & HETAMI, A., 2015. Perancangan Information Retrieval (IR) Untuk Pencarian Ide Pokok Teks Artikel Berbahasa Inggris dengn Pembobotan Vector Space Model. Jurnal Ilmiah Teknologi Informasi Asia, 9(1), pp. 53-59.

BIRANT, D. & KUT, A., 2007. ST-DBSCAN: An algorithm for clustering spatial–temporal data. Data & Knowledge Engineering, 60(1), pp. 208-221.

DEVI, N. M. A. S., PUTRA, I. K. G. D. & SUKARSA, I. M., 2015. Implementasi Metode Clustering DBSCAN pada Proses Pengambilan Keputusan. Lontar Komputer, 6(3), pp. 185-191.

GHAWI, R. & PFEFFER, J., 2019. Efficient Hyperparameter Tuning with Grid Search for Text Categorization using KNN Approach with BM25 Similarity. Open Computer Science, 9(1), pp. 160-180.

HADI, S., 2017. Pemeriksaan Keabsahan data penelitian kualitatif pada skripsi. Jurnal Ilmu Pendidikan, 22(1), pp. 74-79.

HASANAH, N., 2017. Sistem Pencarian Skripsi Berbasis Information Retrieval di FASTIKOM UNSIQ. Jurnal Penelitian dan Pengabdian Kepada Masyarakat UNSIQ, 4(1), pp. 105-113.

HERMAWAN, L. & ISMIATI, M. B., 2020. Pembelajaran Text Preprocessing berbasis Simulator Untuk Mata Kuliah Information Retrieval. Jurnal Transformatika, 17(2), pp. 188-199.

HESAY, I. K., INDRIATI & ADINUGROHO, S., 2021. Analisis Sentimen Ulasan Pengunjung Simpang Lima Gumul Kediri menggunakan Metode BM25 dan Neighbor-Weighted K-Nearest Neighbor. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 5(7), pp. 3160-3169.

IŞIK, M. & DAĞ, H., 2020. The impact of text preprocessing on the prediction of review ratings. Turkish Journal of Electrical Engineering and Computer Sciences, 28(3), pp. 1405-1421.

ISNARWATY, D. P. & IRHAMAH, 2019. Text Clustering pada Akun TWITTER Layanan Ekspedisi JNE, J&T, dan Pos Indonesia Menggunakan Metode Density-Based Spatial Clustering of Applications with Noise (DBSCAN) dan K-Means. Jurnal Sains dan Seni ITS, 8(2), pp. D137-D144.

JAMBAK, M. I. & EFENDI, R., 2021. Pengaruh Reduksi Dimensi Terhadap Metode Pengklasteran Berbasis Centroid dan Metode Pengklasteran Berbasis Density Dalam Pengklasteran Dokumen Teks. Indonesian Journal of Business Intelligence (IJUBI), 4(2), pp. 53-62.

KARAMI, A. & JOHANSSON, R., 2014. Choosing DBSCAN Parameters Automatically using Differential Evolution. International Journal of Computer Applications, 91(7), pp. 1-11.

NISHOM, M., 2019. Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square.

Jurnal Informatika: Jurnal Pengembangan IT (JPIT), 4(1), pp. 20-24.

PRAMUDITA, W., TOMASOUW, B. P., LELEURY, Z. A. & RIJOLY, M. E., 2021. Perancangan Sistem Deteksi Plagiarisme Skripsi (Judul Dan Abstrak) Berbasis Matlab Menggunakan Algoritma Winnowing. Tensor: Pure and Applied Mathematics Journal, 2(2), pp. 67-76.

PRIANDONO, I. R., HAKIMAH, M. & ROZI, N. F., 2020. Implementasi Vector Space Model Dengan Pembobotan Berbasis Kelas Pada Mesin Pencari Dokumen Skripsi. Jurnal Informatika: Jurnal pengembangan IT (JPIT), 5(2), pp. 54-58.

RACHMAN, D. A. C., GOEJANTORO, R. & AMIJAYA, F. D. T., 2020. Implementasi Text Mining Pengelompokkan Dokumen Skripsi Menggunakan Metode K-Means Clustering. Jurnal EKSPONENSIAL, 11(2), pp. 167-174.

RAMADHANA, CUT, B. & HUSNA, J., 2019. Rancangan Bangun Sistem E-Repository Skripsi Mahasiswa Berbasis Qr (Quick Response) Code. Kandidat: Jurnal Riset dan Inovasi Pendidikan, 1(1), pp. 9-14.

ROBERTSON, S. & ZARAGOZA, H., 2009. The Probabilistic Relevance Framework: BM25 and Beyond. [pdf] City University of London Staff Personal Pages. Tersedia di: <https://www.staff.city.ac.uk/~sbrp622/papers/foundations_bm25_review.pdf> [Diakses 17 November 2022]

SAKARIANA, M. I. D., INDRIATI & DEWI, C., 2020. Analisis Sentimen Pemindahan Ibu Kota Indonesia Dengan Pembobotan Term BM25 Dan Klasifikasi Neighbor Weighted K-Nearest Neighbor. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 4(3), pp. 748-755.

SARI, H., GINTING, G. L., ZEBUA, T. & MESRAN, 2021. Penerapan Algoritma Text Mining dan TF-IDF Untuk Pengelompokan Topik Skripsi Pada Aplikasi Repository STMIK Budi Darma. TIN: Terapan Informatika Nusantara, 2(7), pp. 414-432.

SIMANJUNTAK, K. P. & KHAIRA, U., 2021. Pengelompokkan Titik Api di Provinsi Jambi dengan Algoritma Agglomerative Hierarchical Clustering. MALCOM: Indonesian Journal of Machine Learning and Computer Science, 1(1), pp. 7-16.

STRUYF, A., HUBERT, M. & ROUSSEEUW, P. J., 1997. Clustering in an Object-Oriented Environment. Journal of Statistical Software, 1(4), pp. 1-30.

SUGIYAMTO, SURARSO, B. & SUGIHARTO, A., 2014. Analisa Performa Metode Cosine dan Jacard Pada Pengujian Kesamaan Dokumen. Jurnal Masyarakat Informatika, 5(10), pp. 1-8.

SUHARDI, ET AL., 2021. Implementasi Information Retrieval System untuk Klasifikasi Berita Offline di Indonesia Menggunakan Metode Extended Boolean. CERMIN: Jurnal Penelitian, 5(1), pp. 124-137.

TINEGA, G. A., MWANGI, W. & RIMIRU, R., 2018. Text Mining in Digital Libraries using OKAPI BM25 Model. International Journal of Computer Applications Technology and Research, 7(10), pp. 398-406.

ZHANG, J., GAO, J., ZHOU, M. & WANG, J., 2001. Improving the Effectiveness of Information Retrieval with Clustering and Fusion. Computational Linguistics and Chinese Language Processing, 6(1), pp. 109-125.

Halaman ini sengaja dikosongkan