HASIL CEK_Tsani Elvia Nita, Lisna Zahrotun_udul kerja praktik, text mining, Manhattan Distance Similarity

by Tsani Elvia Nita, Lisna Zahrotun Penerapan Metode Single Linkage

Submission date: 20-Sep-2022 12:53PM (UTC+0700)
Submission ID: 1904324235
File name: Penerapan_Metode_Single_Linkage_dengan_Manhattan_Distance.pdf (1.23M)
Word count: 2853
Character count: 16289

Jurnal Riset Sains dan Teknologi
Volume 5 No. 1 Maret 2021, 39 - 44
e-ISSN: 2549-9750
p-ISSN: 2579-9118

Penerapan Metode Single Linkage dengan Manhattan Distance Similarity dalam Mengelompokkan Trens Topik Kerja Praktik

Application of the Single Linkage Method with Manhattan Distance Similarity in Grouping Trens of Practical Work Topics

Tsani Elvia Nita[1], Lisna Zahrotun[2*]
[1,2] Teknik Informatika, Fakultas Teknologi Industri, Universitas Ahmad Dahlan
Jl. Ringroad Selatan, Tamanan, Kec. Banguntapan, Bantul, Daerah Istimewa Yogyakarta 55191, Indonesia
email: elvianitatsani@gmail.com, *lisna.zahrotun@tif.uad.ac.id

DOI;
10.30595/jrst.v5i1.9083

Histori Artikel:

Diajukan:
24/11/2020

Diterima:
05/03/2021

Diterbitkan:
27/03/2021

ABSTRAK

Data laporan judul kerja praktik (KP) biasanya hanya terkumpul di perpustakaan dan jarang dipubilkasikan ke mahasiswa, hal ini menyebabkan kesulitan bagi mahasiswa yang akan mengkasesnya. Berdasarkan permasalahan tersebut, maka dibuatlah suatu program pada penlitian ini untuk pengelompokkan Trend Topik. Metode yang digunakan dalam penelitian ini adalah *Manhattan Distance Similariy* dan *Single Linkage*. Sebelum masuk tahapan *text mining*, perlu dilakukan perancangan diantaranya perancangan basis data dan antar muka (*interface*). Tahapan dan *text mining* adalah mengumpulkan data (*collect data*), penguraian teks (text mining), penyaringan teks (*text filtering*), pembobotan kata (calculate term count), similarity, pengelompokan, dan pengujian. Hasil dari penelitian ini adalah program yang dapat mengolah data judul KP menjadi pola kelompok Trend Topik KP. Dari 905 data yang di dapatkan, terbentuk 7 kelompok yaitu Sistem Informasi, Multimedia, Jaringan, Web, Kewirausahaan, Magang, dan Pelatihan. Tetapi dari hasil pengujian *Purity Test* didapatkan nilai sebesar 0,267, yang artinya Manhattan Distance Similarity dan Single *Linkage* kurang cocok untuk mengelompokkan Judul KP.

Kata Kunci: judul kerja praktik, *text mining, Manhattan Distance Similarity*

ABSTRACT

Data on practical work titles (KP) reports are usually only collected in libraries and are rarely published to students, this causes difficulties for students who will access them. Based on these problems, a program in this study was made to group Trend Topics. The method used in this research is Manhattan Distance Similariy and Single Linkage. Before entering the text mining stage, it is necessary to design including database design and interfaces. The stages and text mining are collecting data (collect data), text mining, text filtering, word weighting (calculate term count), similarity, grouping, and testing. The result of this research is a program that can process data of KP titles into patterns of group KP Topic Trends. From the 905 data obtained, 7 groups were formed, namely Information Systems, Multimedia, Networking, Web, Entrepreneurship, Internships, and Training. But from the Purity Test test results obtained a value of 0.267, which means that Manhattan Distance Similarity and Single Linkage are not suitable for grouping KP Title

Keywords: practical work titles, text mining, Manhattan Distance Similarity.

1. PENDAHULUAN

Salah satu syarat kelulusan di Program Studi Teknik Informatika (TIF) Universitas Ahmad Dahlan (UAD) adalah telah menyelesaikan Kerja Praktik (KP). KP merupakan mata kuliah wajib yang mulai ditawarkan di semester 5 pada kurikulum baru 2016 ini. KP dapat diambil secara individu oleh mahasiswa pada semester genap maupun semester ganjil. Tahapan untuk dapat melaksanakan KP sendiri dibagikan melalui *website* TIF. Saat tahap bimbingan, mahasiswa menggunakan Kartu Bimbingan sebagai dokumen evaluasi. Kartu Bimbingan didapatkan dari website TIF. Kartu Bimbingan ada 7 jenis yakni Sistem Informasi, Multimedia, Jaringan, Web, Magang, Kewirausahaan, dan Pelatihan.

Seringkali kendala yang dialami oleh mahasiswa yang akan mengambil matakuliah KP adalah kesulitan mencari topik KP. Referensi mahasiswa dalam menyusun laporan adalah dari bertanya-tanya kepada angkatan atas yang telah selesai KP. Hal itupun dilakukan oleh mahasisa yang aktif saja, karena mahasiswa belum mengetahui adanya publikasi judul laporan KP terdahulu pada *website* TIF. Data judul laporan KP terdahulu yang dipublikasi pada *website* TIF juga kurang *up-to-date* dibandingkan dengan data yang ada pada Kantor Tata Usaha (TU) Fakultas Teknologi Industri (FTI).

Selama ini, dokumentasi yang dilakukan baru sekedar menyimpan dan mempublikasikan laporan KP saja. Padahal, jika judul KP dikelompokkan dapat mempermudah dalam mengetahui trend topik KP yang dapat menjadi salah satu referensi mahasiswa dalam menentukan topik apa yang sesuai dengan keahliannya. Dapat juga membantu Koordinator KP dan Program Studi mengolah data judul serta bahan pertimbangan dalam pengadaan referensi tempat KP serta perkembangan matakuliah dan kurikulum.

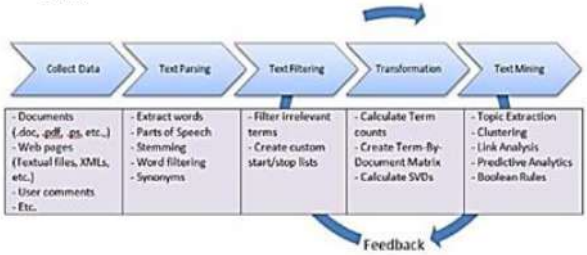
Beberapa penyelesaian tentang data text adalah pengelompokkan judul kerja praktek menggunakan metode *Shared Nearest Neighbour* (Zahrotun, 2017), rancang bangun aplikasi pengelompokkan judul penellitian dosen menggunakan metode *Shared Nearest Neighbour* (Zahrotun and Mushlihudin, 2017), perbandingan *Jaccard dan Cosine Similarity* (Zahrotun, 2016), Pengelompokkan judul skripsi menggunakan metode K-Means (Zahrotun, Putri and Khusna, 2018),

Dengan demikian, dari permasalahan KP yang selama ini ada, maka dibuat *Text Mining* Menggunakan Metode *Single Linkage* dengan *Manhattan Distance Similarity* untuk Pengelompokan Trend Topik Kerja Praktik. Diharapkan dengan penelitian ini dapat melakukan pengelompokan trend topik KP lebih baik dari pada penelitian sebelumnya. Sehingga, dapat mengolah data judul KP menjadi pola kelompok trend topik KP untuk bahan pertimbangan dalam pengadaan tempat KP serta perkembangan matakuliah dan kurikulum.

2. METODE PENELITIAN

2.1 *Text Mining*

Text mining melibatkan penerapan algoritma data mining tradisional seperti pengelompokan, klasifikasi. *Text mining* merupakan proses berulang yang melibatkan pengulangan analisis menggunakan pengaturan yang berbeda dan menggunakan atau mengecualikan syarat tertentu untuk hasil yang lebih baik. Hasil dari langkah ini bisa berupa kumpulan dokumen, daftar topik jangka panjang atau multi-istilah, atau aturan yang menjawab masalah klasifikasi. Langkah-langkah *text mining* menurut (Chakraborty, Pagolu and Garla, 2013) ditunjukkan pada Gambar 1. *Text Mining Process Flow*



Gambar 1. *Text mining process flow*

2.1.1 Mengumpulkan Data (*Collect Data*)

Langkah pertama dalam proyek penelitian penambangan teks adalah mengumpulkan data tekstual yang diperlukan untuk digali informasi yang berkualitas dari data tesktual tersebut.

2.1.2 Penguraian Text (*Text Parsing*)

Penguraian data teks dimulai dengan mengambil urutan karakter (seperti urutan kalimat dalam dokumen teks) dan memecahnya menjadi token (unit, di mana satu unit berada adalah sebuah kata, angka, atau tanda baca) proses ini dinamakan tokenisasi. Proses tokenisasi termasuk ke dalam *extract words*. Setelah tokens ditemukan dalam dokumen,

selanjutnya adalah normalisasi untuk menghilangkan kekompleksan kata, atau disebut stemming. Untuk tujuan komputasi, *stemming* berguna untuk mengurangi semua variasi kata-kata yang mirip.

Salah satu *library* yang bisa digunakan dalam melakukan proses *stemming* bahasa indonesia adalah menggunakan *Library Python Sastrawi dan function Stemmer Factory*. *Library* ini merupakan pengembangan dari *Library* PHP Sastrawi dimana *library* tersebut menerapkan algoritma Algoritma Nazief dan Adriani.

2.1.3 Penyaringan teks (*Text Filtering*)

Korpus dalam beribu-ribu dokumen mungkin akan banyak hal yang tidak relevan, baik untuk membedakan dokumen dari satu sama lain atau untuk meringkas dokumen. Menelusuri terms secara manual untuk menghilangkan terms yang tidak relevan adalah salah satu hal yang paling sering memakan waktu dan tugas subjektif dari semua langkah *text mining*.

2.1.4 Pembobotan Kata (*Calculate Terms count*)

Pembobotan kata dilakukan dengan cara menghitung *Term Frequency-Inverse Document Frequency* (tf-idf) (Janani and Vijayarani, 2016). Tf-idf adalah statistik numerik yang mengungkapkan bahwa seberapa penting sebuah kata dalam sebuah dokumen di koleksi. Tf - IDF sering digunakan sebagai faktor pembobotan dalam pencarian informasi dan penambangan teks. Pembobotan dijabarkan dalam persamaan 1 (Kao and Poteet, 2006).

$$\boldsymbol{tfidf}\ (\boldsymbol{t_i}, \boldsymbol{d_j}) = \boldsymbol{tf}\ (\boldsymbol{t_i}, \boldsymbol{d_j}) \times \log(\frac{\boldsymbol{N}}{\boldsymbol{N}\ (\boldsymbol{t_i})})$$

Dimana :

$\boldsymbol{tfidf}\ (\boldsymbol{t_i}, \boldsymbol{d_j})$ = bobot kata / term $\boldsymbol{t_i}$ terhadap dokumen d_j

$\boldsymbol{tf}\ (\boldsymbol{t_i}, \boldsymbol{d_j})$ = jumlah kemunculan kata / term t_i dalam dokumen d_j

N = Jumlah semua dokumen

N (t_i) = Jumlah dokumen yang mengandung kata / term t_i

2.2 *Manhattan Distance Similarity*

Manhattan distance adalah metode pengukuran jarak dua vektor (Muzzammil, Ginardi and Purwitasari, 2016). Nilai yang digunakan adalah nilai absolut dari masing-masing fitur yang dihitung selisihnya. Berikut adalah persamaan Manhattan distance dapat dilihat pada (2).

$$\boldsymbol{ManhDis}(\boldsymbol{p}, \boldsymbol{q}) = \| \boldsymbol{p} - \boldsymbol{q} \| = \sum_{i=1}^{n} \| \boldsymbol{p} - \boldsymbol{q} \| \qquad (2)$$

Dimana p dan q adalah bobot token di masing-masing dokumen

2.3 Pengelompokkan *Single Linkage*

Ada dua pendekatan utama pengelompokkan yaitu pengelompokkan dengan pendekatan partisi dan pendekatan hierarki. Pengelompokkan dengan pendekatan partisi (partition-based clustering) mengelompokkan data dengan memilah-milah data yang dianalisa ke dalam cluster-cluster yang ada. Contohnya adalah Fuzzy, K-Means, K-Medoids, juga K-Harmonics. Pengelompokkan dengan pendekatan hierarki (*hierarchical clustering*) mengelompokkan data dengan membuat suatu hierarki berupa dendrogram dimana data yang mirip akan ditempatkan pada hirarki yang berdekatan dan yang tidak pada hirarki yang berjauhan.

Metode pengelompokkan hierarki (*hierarchical clustering*) terdiri atas dua bagian, yaitu metode *agglomerative* (penyatuan) dan *divisive* (penyebaran). Dalam metode *agglomerative* dikenal beberapa metode untuk membentuk cluster, yaitu metode *single*, *complete*, *average*, dan *ward linkage*. Prinsip kerja dari pengelompokan *Hierarchical Clustering* dilakukan secara bertahap. Dan disetiap iterasi dari pengelompokan Hierarchical Clustering hanya ada satu pemilihan penggabungan suatu dokumen terhadap dokumen lainnya (Handoyo *et al.*, 2014).

Single Link disebut juga minimum link dimana similaritas dari dua kelpompok didasarkan pada dua titik paling dekat dari dua kelompok yang berbeda.

3. HASIL DAN PEMBAHASAN

3.1 *Text Mining*

Data yang diambil adalah judul Laporan Kerja Praktik sejumalh 905 judul, dimana setiap judul dianggap sebagai satu dokumen. Setelah mendapatkan data maka akan dilakukan tahap preprocesing text.

3.2 Penguraian Text (*Text Parsing*)

Dalam penguraian text ini terdiri dari beberapa tahap diantaranya :

1. Tokenisasi
   Dalam proses ini dilakukan pemecahan kalimat menjadi beberapa kata. Contoh

pemecahan kalimat ditunjukkan dalam Tabel 1.

Tabel 1. Tabel tokenisasi

ID	T1	T2	T3	T4	T5	T6
D1	sistem	informasi	klinik	gigi	berbasis	web

2. Stemming
   Dalam proses ini dilakukan pencarian akar kata dari token. Contoh hasil proses stemming ditunjukkan dalam Tabel 2

Tabel 2. Tabel Stemming

ID	T1	T2	T3	T4	T5	T6
D1	sistem	informasi	klinik	gigi	basis	web

3.3 Filtering

Setelah dilakukan stemming maka dilakukan proses filtering. Dalam penelitian ini teknik filtering yang digunakan adalah stop word. Contoh proses filtering ditunjukkan dalam Tabel 3.

Tabel 3. Tabel Filtering

ID	T1	T2	T3	T4	T5	T6
D1	sistem	informasi	klinik	gigi	basis	web

3.4 Pembobotan Kata

Langkah selanjutnya adalah pembobotan kata, dimana dalam pembobotan ini menggunakan rumus Tf-Idf dan hasil pembobotan 10 data dapat dilihat pada Tabel 4.

3.5 Implemetasi

Implementasi merupakan interface dari aplikasi yang dibuat. Tampilan awal dalam penelitian ini ditunjukkan dalam Gambar 1

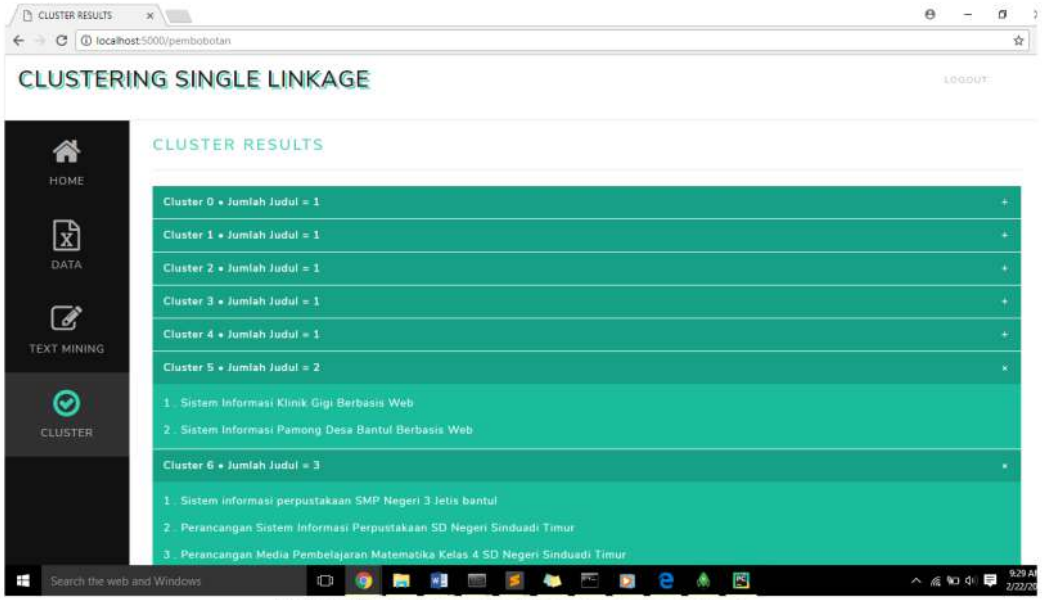
Gambar 1. Load Data

Untuk tampilan preprocessing ditunjukkan dalam Gambar 2 dan tampilan hasil akhir clustering ditunjukkan dalam Gambar 3.

Gambar 2. Preprocesing text

Tabel 4. Tabel Pembobotan Kata

	term	Tf										df	D / df	idf	w = tf x idf								
		D0	D1	D2	D3	D4	D5	D6	D7	D8	D9				D0	D1	D2	D3	D4	D5	D6	D7	D8	D9
1	sistem	1	1	1			1	1		1	1	7	1.43	0.15	0.15	0.15	0.15	0.00	0.00	0.15	0.15	0.00	0.15	0.15
2	informasi	1	1	1			1	1		1	1	7	1.43	0.15	0.15	0.15	0.15	0.00	0.00	0.15	0.15	0.00	0.15	0.15
3	poin	1										1	10	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	kirim	1										1	10	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	agen	1										1	10	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
...	...											...	...	...	...	...	...	...	...	...	...	...	...	...
43	desa									1		1	10	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
44	3										1	1	10	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
45	jetis										1	1	10	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
		9	6	8	9	10	10	6	7	7	8	80			6.23	4.01	4.20	6.36	6.64	7.71	3.23	5.44	3.75	4.63



Gambar 3. Implementasi hasil pengelompokan

3.2.8 Pengujian Akurasi

Pengujian akurasi menggunakan *Purity Test* dilakukan untuk mengetahui baik buruknya kelompok yang dihasilkan dari proses pengelompokkan dengan metode *Single Linkage*. Jika hasil *Puriity Test* semakin mendekati 1, maka semakin baik kualitas kelompoknya. Sedangkan, jika hasil *Puriity Test* semakin mendekati 1, maka semakin baik kualitas kelompoknya. Hasil dari *Purity Test* dapat dilihat pada Tabel 9 : Hasil Purity Test.

No	Kelompok	Asli	Single Linkage	Sesuai
1	SISTEM INFORMASI	168	1	1
2	MULTIMEDIA	192	1	1
3	JARINGAN	27	1	1
4	WEB	213	1	1
5	KEWIRAUSAHAAN	19	1	1
6	MAGANG	50	1	1
7	PELATIHAN	236	899	236
		905	905	242

$$r = \frac{1}{n}\sum_{i=1}^{k} a_i$$

Pengujian akurasi menggunakan *Purity Test* dilakukan untuk mengetahui baik buruknya kelompok yang dihasilkan dari proses pengelompokkan dengan metode *Single Linkage*. Jika hasil *Puriity Test* semakin mendekati 1, maka semakin baik kualitas kelompoknya. Sedangkan, jika hasil Puriity Test semakin mendekati 1, maka semakin baik kualitas kelompoknya. Hasil dari Purity Test dapat dilihat pada Tabel 9 : Hasil *Purity Test*.

Dimana:

r : tingkat akurasi pengelompokan

k : jumlah cluster

a_i : objek yang muncul di *cluster Ci* dan pada label *class* yang sesuai.

$$r = \frac{1}{905}\,(1+1+1+1+\ 1+1+236)$$

$$r = \frac{1}{905}\,(242)$$

$$r = 0.267$$

Data yang diuji adalah semua data bersih dari tahun 2012 hingga 2017 yaitu sejumlah 905 data. Jumlah kelompok (cluster) sesuai dengan jenis kartu bimbingan Kerja Praktik yaitu sejumlah 7 (Sistem Informasi, Multimedia, Jaringan, Web, Kewirausahaan, Magang, dan Pelatihan). Pengelompokan data asli dilakukan secara manual agar dapat dibandingkan dengan hasil dari pengelompokkan menggunakan metode Single Linkage. Dari perhitungan diatas didapatkan hasil *Purity Test* adalah sebesar 0.267. Hasil tersebut tergolong rendah karena *Purity test*

memiliki range nilai antara 0 – 1. Semakin mendekati 1 hasil kelompoknya semakin baik.

4. KESIMPULAN

Berdasarkan hasil penelitian *"Text Mining* Menggunakan Metode *Single Linkage* dengan *Manhattan Distance Similarity* untuk Pengelompokan Trend Topik Kerja Praktik" dapat ditarik kesimpulan sebagai berikut :

1. Uji *Purity Test* yang dilakukan pada aplikasi "*Text Mining* Menggunakan Metode *Single Linkage* dengan *Manhattan Distance Similarity* untuk Pengelompokan Trend Topik Kerja Praktik" menggunakan 905 data dan dibagi menjadi 7 kelompok sesuai pada jenis Kartu Bimbingan KP, menunjukan hasil sebesar 0.267. Sehingga, Metode pengelompokkan *Single Linkage* di penelitian ini menghasilkan kelompok yang kurang bagus.
2. Hasil *Uji Purity Test* tergolong rendah karena data yang digunakan adalah data teks berupa judul laporan, dimana judul laporan ditulis dengan kata-kata yang berbeda antara judul satu dengan judul yang lainnya, beda kata menjadikan bobotnya juga berbeda, bobot yang berbeda menjadikan nilai similaritas berbeda, padahal untuk menjadi satu kelompok suatu judul harus memiliki nilai similaritas yang hampir sama. Lalu dari pemilihan metode, Single Linkage merupakan pengelompokan dengan pendekatan hierarki (*hierarchical clustering*) mengelompokan data dengan membuat suatu hierarki berupa dendrogram dimana data yang mirip akan ditempatkan pada hirarki yang berdekatan dan yang tidak pada hierarki yang berjauhan.
3. Metode pengelompokkan *Single Linkage* dengan *Manhattan Distance Similarity* di penelitian ini menghasilkan kelompok yang kurang bagus dan kurang cocok untuk pengelompokkan Trend Topik Kerja Praktik.

DAFTAR PUSTAKA

Chakraborty, G., Pagolu, M. and Garla, S. (2013) *PREVIEW: Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS*, *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS*.

Handoyo, R. *et al.* (2014) "Perbandingan Metode Clustering Mengggunakan metode Single Linkage dan K-Means Pada Pengelompokkan Dokumen," *JSM STMIK Mikroskil*, 15(2), pp. 73–82.

Janani, R. and Vijayarani, S. (2016) "Text Mining Research : A Survey," *International Journal of innovative Research in Computer and Communication Engineering*, 4(4), pp. 6564–6571. doi: 10.15680/IJIRCCE.2016.

Kao, A. and Poteet, S. R. (2006) *Natural Language Processing and Text Mining*. USA: Springer.

Muzzammil, R. R., Ginardi, R. V. hari and Purwitasari, D. (2016) "Modul Klasifikasi Aduan dengan Pendekatan Kemiripan Teks pada Aplikasi Perangkat Bergerak Suara Warga (Surga) Kota Kediri," *Jurnal Teknik ITS*, 5(1), pp. 52–57.

Zahrotun, L. (2016) "Comparison Jaccard Similarity , Cosine Similarity and Combined Both of the Data Clustering With Shared Nearest Neighbor Method," *Computer Engineering and Applications*, 5(1), pp. 11–18.

Zahrotun, L. (2017) "Text Mining for Internship Titles Clustering Using Shared Nearest-Neighbor Method," *Computer Engineering and Applications*, 6(3).

Zahrotun, L. and Mushlihudin (2017) "Rancang Bangun Aplikasi Text Mining dalam Mengelompokkan Judul Penelitian Dosen Menggunakan Metode Shared Nearest Neighbor dan Euclidean Similarity," *Jurnal Ilmu Teknik elektro Komputer dan informatika (JITEKI)*, 3(2), pp. 91–99.

Zahrotun, L., Putri, N. hutami and Khusna, A. N. (2018) "The Implementation of K-Means Clustering Method in Classifying Undergraduate Thesisi Titles," in *12th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*. Yogyakarta: IEEE.

HASIL CEK_Tsani Elvia Nita, Lisna Zahrotun_udul kerja praktik, text mining, Manhattan Distance Similarity

ORIGINALITY REPORT

2%	0%	5%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Reni Dwi Astuti, Devi Meilina Khoirun Nisa. "Penghambat Niat dan Perilaku Masyarakat Aktif dalam Kegiatan Bank Sampah", JRST (Jurnal Riset Sains dan Teknologi), 2021 Publication	2%

Exclude quotes	On	Exclude matches	< 2%
Exclude bibliography	On		

Jurnal Riset Sains dan Teknologi
Volume 5 No. 1 Maret 2021, 39 - 44
e-ISSN: 2549-9750
p-ISSN: 2579-9118

Penerapan Metode Single Linkage dengan Manhattan Distance Similarity dalam Mengelompokkan Trens Topik Kerja Praktik

Application of the Single Linkage Method with Manhattan Distance Similarity in Grouping Trens of Practical Work Topics

Tsani Elvia Nita[1], Lisna Zahrotun[2*]
[1,2] Teknik Informatika, Fakultas Teknologi Industri, Universitas Ahmad Dahlan
Jl. Ringroad Selatan, Tamanan, Kec. Banguntapan, Bantul, Daerah Istimewa Yogyakarta 55191, Indonesia
email: elvianitatsani@gmail.com, *lisna.zahrotun@tif.uad.ac.id

DOI;
10.30595/jrst.v5i1.9083

Histori Artikel:

Diajukan:
24/11/2020

Diterima:
05/03/2021

Diterbitkan:
27/03/2021

ABSTRAK

Data laporan judul kerja praktik (KP) biasanya hanya terkumpul di perpustakaan dan jarang dipubilkasikan ke mahasiswa, hal ini menyebabkan kesulitan bagi mahasiswa yang akan mengkasesnya. Berdasarkan permasalahan tersebut, maka dibuatlah suatu program pada penlitian ini untuk pengelompokkan Trend Topik. Metode yang digunakan dalam penelitian ini adalah *Manhattan Distance Similariy* dan *Single Linkage*. Sebelum masuk tahapan *text mining*, perlu dilakukan perancangan diantaranya perancangan basis data dan antar muka (*interface*). Tahapan dan *text mining* adalah mengumpulkan data (*collect data*), penguraian teks (text mining), penyaringan teks (*text filtering*), pembobotan kata (calculate term count), similarity, pengelompokan, dan pengujian. Hasil dari penelitian ini adalah program yang dapat mengolah data judul KP menjadi pola kelompok Trend Topik KP. Dari 905 data yang di dapatkan, terbentuk 7 kelompok yaitu Sistem Informasi, Multimedia, Jaringan, Web, Kewirausahaan, Magang, dan Pelatihan. Tetapi dari hasil pengujian *Purity Test* didapatkan nilai sebesar 0,267, yang artinya Manhattan Distance Similarity dan Single *Linkage* kurang cocok untuk mengelompokkan Judul KP.

Kata Kunci: judul kerja praktik, *text mining, Manhattan Distance Similarity*

ABSTRACT

Data on practical work titles (KP) reports are usually only collected in libraries and are rarely published to students, this causes difficulties for students who will access them. Based on these problems, a program in this study was made to group Trend Topics. The method used in this research is Manhattan Distance Similariy and Single Linkage. Before entering the text mining stage, it is necessary to design including database design and interfaces. The stages and text mining are collecting data (collect data), text mining, text filtering, word weighting (calculate term count), similarity, grouping, and testing. The result of this research is a program that can process data of KP titles into patterns of group KP Topic Trends. From the 905 data obtained, 7 groups were formed, namely Information Systems, Multimedia, Networking, Web, Entrepreneurship, Internships, and Training. But from the Purity Test test results obtained a value of 0.267, which means that Manhattan Distance Similarity and Single Linkage are not suitable for grouping KP Title

Keywords: practical work titles, text mining, Manhattan Distance Similarity.

1. PENDAHULUAN

Salah satu syarat kelulusan di Program Studi Teknik Informatika (TIF) Universitas Ahmad Dahlan (UAD) adalah telah menyelesaikan Kerja Praktik (KP). KP merupakan mata kuliah wajib yang mulai ditawarkan di semester 5 pada kurikulum baru 2016 ini. KP dapat diambil secara individu oleh mahasiswa pada semester genap maupun semester ganjil. Tahapan untuk dapat melaksanakan KP sendiri dibagikan melalui *website* TIF. Saat tahap bimbingan, mahasiswa menggunakan Kartu Bimbingan sebagai dokumen evaluasi. Kartu Bimbingan didapatkan dari website TIF. Kartu Bimbingan ada 7 jenis yakni Sistem Informasi, Multimedia, Jaringan, Web, Magang, Kewirausahaan, dan Pelatihan.

Seringkali kendala yang dialami oleh mahasiswa yang akan mengambil matakuliah KP adalah kesulitan mencari topik KP. Referensi mahasiswa dalam menyusun laporan adalah dari bertanya-tanya kepada angkatan atas yang telah selesai KP. Hal itupun dilakukan oleh mahasisa yang aktif saja, karena mahasiswa belum mengetahui adanya publikasi judul laporan KP terdahulu pada *website* TIF. Data judul laporan KP terdahulu yang dipublikasi pada *website* TIF juga kurang *up-to-date* dibandingkan dengan data yang ada pada Kantor Tata Usaha (TU) Fakultas Teknologi Industri (FTI).

Selama ini, dokumentasi yang dilakukan baru sekedar menyimpan dan mempublikasikan laporan KP saja. Padahal, jika judul KP dikelompokkan dapat mempermudah dalam mengetahui trend topik KP yang dapat menjadi salah satu referensi mahasiswa dalam menentukan topik apa yang sesuai dengan keahliannya. Dapat juga membantu Koordinator KP dan Program Studi mengolah data judul serta bahan pertimbangan dalam pengadaan referensi tempat KP serta perkembangan matakuliah dan kurikulum.

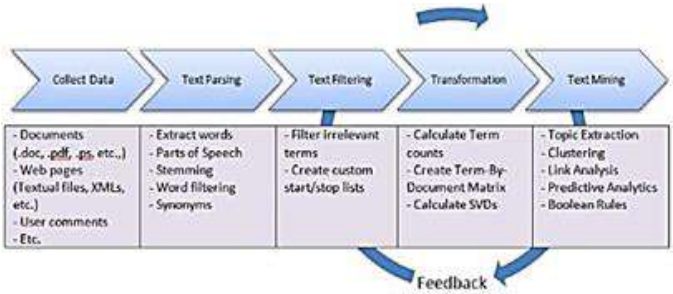
Beberapa penyelesaian tentang data text adalah pengelompokkan judul kerja praktek menggunakan metode *Shared Nearest Neighbour* (Zahrotun, 2017), rancang bangun aplikasi pengelompokkan judul penellitian dosen menggunakan metode *Shared Nearest Neighbour* (Zahrotun and Mushlihudin, 2017), perbandingan *Jaccard dan Cosine Similarity* (Zahrotun, 2016), Pengelompokkan judul skripsi menggunakan metode K-Means (Zahrotun, Putri and Khusna, 2018),

Dengan demikian, dari permasalahan KP yang selama ini ada, maka dibuat *Text Mining* Menggunakan Metode *Single Linkage* dengan *Manhattan Distance Similarity* untuk Pengelompokan Trend Topik Kerja Praktik. Diharapkan dengan penelitian ini dapat melakukan pengelompokan trend topik KP lebih baik dari pada penelitian sebelumnya. Sehingga, dapat mengolah data judul KP menjadi pola kelompok trend topik KP untuk bahan pertimbangan dalam pengadaan tempat KP serta perkembangan matakuliah dan kurikulum.

2. METODE PENELITIAN

2.1 *Text Mining*

Text mining melibatkan penerapan algoritma data mining tradisional seperti pengelompokan, klasifikasi. *Text mining* merupakan proses berulang yang melibatkan pengulangan analisis menggunakan pengaturan yang berbeda dan menggunakan atau mengecualikan syarat tertentu untuk hasil yang lebih baik. Hasil dari langkah ini bisa berupa kumpulan dokumen, daftar topik jangka panjang atau multi-istilah, atau aturan yang menjawab masalah klasifikasi. Langkah-langkah *text mining* menurut (Chakraborty, Pagolu and Garla, 2013) ditunjukkan pada Gambar 1. *Text Mining Process Flow*



Gambar 1. *Text mining process flow*

2.1.1 Mengumpulkan Data (*Collect Data*)

Langkah pertama dalam proyek penelitian penambangan teks adalah mengumpulkan data tekstual yang diperlukan untuk digali informasi yang berkualitas dari data tesktual tersebut.

2.1.2 Penguraian Text (*Text Parsing*)

Penguraian data teks dimulai dengan mengambil urutan karakter (seperti urutan kalimat dalam dokumen teks) dan memecahnya menjadi token (unit, di mana satu unit berada adalah sebuah kata, angka, atau tanda baca) proses ini dinamakan tokenisasi. Proses tokenisasi termasuk ke dalam *extract words*. Setelah tokens ditemukan dalam dokumen,

selanjutnya adalah normalisasi untuk menghilangkan kekompleksan kata, atau disebut stemming. Untuk tujuan komputasi, *stemming* berguna untuk mengurangi semua variasi kata-kata yang mirip.

Salah satu *library* yang bisa digunakan dalam melakukan proses *stemming* bahasa indonesia adalah menggunakan *Library Python Sastrawi dan function Stemmer Factory*. *Library* ini merupakan pengembangan dari *Library* PHP Sastrawi dimana *library* tersebut menerapkan algoritma Algoritma Nazief dan Adriani.

2.1.3 Penyaringan teks (*Text Filtering*)

Korpus dalam beribu-ribu dokumen mungkin akan banyak hal yang tidak relevan, baik untuk membedakan dokumen dari satu sama lain atau untuk meringkas dokumen. Menelusuri terms secara manual untuk menghilangkan terms yang tidak relevan adalah salah satu hal yang paling sering memakan waktu dan tugas subjektif dari semua langkah *text mining*.

2.1.4 Pembobotan Kata (*Calculate Terms count*)

Pembobotan kata dilakukan dengan cara menghitung *Term Frequency-Inverse Document Frequency* (tf-idf) (Janani and Vijayarani, 2016). Tf-idf adalah statistik numerik yang mengungkapkan bahwa seberapa penting sebuah kata dalam sebuah dokumen di koleksi. Tf - IDF sering digunakan sebagai faktor pembobotan dalam pencarian informasi dan penambangan teks. Pembobotan dijabarkan dalam persamaan 1 (Kao and Poteet, 2006).

$$\boldsymbol{tfidf}\ (\boldsymbol{t_i}, \boldsymbol{d_j}) = \boldsymbol{tf}\ (\boldsymbol{t_i}, \boldsymbol{d_j}) \times \mathbf{log}(\frac{\boldsymbol{N}}{\boldsymbol{N}\ (\boldsymbol{t_i})})$$

Dimana :

$\boldsymbol{tfidf}\ (\boldsymbol{t_i}, \boldsymbol{d_j})$ = bobot kata / term $\boldsymbol{t_i}$ terhadap dokumen d_j

$\boldsymbol{tf}\ (\boldsymbol{t_i}, \boldsymbol{d_j})$ = jumlah kemunculan kata / term t_i dalam dokumen d_j

N = Jumlah semua dokumen

N (t_i) = Jumlah dokumen yang mengandung kata / term t_i

2.2 *Manhattan Distance Similarity*

Manhattan distance adalah metode pengukuran jarak dua vektor (Muzzammil, Ginardi and Purwitasari, 2016). Nilai yang digunakan adalah nilai absolut dari masing-masing fitur yang dihitung selisihnya. Berikut adalah persamaan Manhattan distance dapat dilihat pada (2).

$$\boldsymbol{ManhDis}(\boldsymbol{p}, \boldsymbol{q}) = \| \boldsymbol{p} - \boldsymbol{q} \| = \sum_{i=1}^{n} \| \boldsymbol{p} - \boldsymbol{q} \| \qquad (2)$$

Dimana p dan q adalah bobot token di masing-masing dokumen

2.3 Pengelompokkan *Single Linkage*

Ada dua pendekatan utama pengelompokkan yaitu pengelompokkan dengan pendekatan partisi dan pendekatan hierarki. Pengelompokkan dengan pendekatan partisi (partition-based clustering) mengelompokkan data dengan memilah-milah data yang dianalisa ke dalam cluster-cluster yang ada. Contohnya adalah Fuzzy, K-Means, K-Medoids, juga K-Harmonics. Pengelompokkan dengan pendekatan hierarki (*hierarchical clustering*) mengelompokkan data dengan membuat suatu hierarki berupa dendrogram dimana data yang mirip akan ditempatkan pada hirarki yang berdekatan dan yang tidak pada hirarki yang berjauhan.

Metode pengelompokkan hierarki (*hierarchical clustering*) terdiri atas dua bagian, yaitu metode *agglomerative* (penyatuan) dan *divisive* (penyebaran). Dalam metode *agglomerative* dikenal beberapa metode untuk membentuk cluster, yaitu metode *single*, *complete*, *average*, dan *ward linkage*. Prinsip kerja dari pengelompokan *Hierarchical Clustering* dilakukan secara bertahap. Dan disetiap iterasi dari pengelompokan Hierarchical Clustering hanya ada satu pemilihan penggabungan suatu dokumen terhadap dokumen lainnya (Handoyo *et al.*, 2014).

Single Link disebut juga minimum link dimana similaritas dari dua kelpompok didasarkan pada dua titik paling dekat dari dua kelompok yang berbeda.

3. HASIL DAN PEMBAHASAN

3.1 *Text Mining*

Data yang diambil adalah judul Laporan Kerja Praktik sejumalh 905 judul, dimana setiap judul dianggap sebagai satu dokumen. Setelah mendapatkan data maka akan dilakukan tahap preprocesing text.

3.2 Penguraian Text (*Text Parsing*)

Dalam penguraian text ini terdiri dari beberapa tahap diantaranya :

1. Tokenisasi

   Dalam proses ini dilakukan pemecahan kalimat menjadi beberapa kata. Contoh

pemecahan kalimat ditunjukkan dalam Tabel 1.

Tabel 1. Tabel tokenisasi

ID	T1	T2	T3	T4	T5	T6
D1	sistem	informasi	klinik	gigi	berbasis	web

2. Stemming
   Dalam proses ini dilakukan pencarian akar kata dari token. Contoh hasil proses stemming ditunjukkan dalam Tabel 2

Tabel 2. Tabel Stemming

ID	T1	T2	T3	T4	T5	T6
D1	sistem	informasi	klinik	gigi	basis	web

3.3 Filtering

Setelah dilakukan stemming maka dilakukan proses filtering. Dalam penelitian ini teknik filtering yang digunakan adalah stop word. Contoh proses filtering ditunjukkan dalam Tabel 3.

Tabel 3. Tabel Filtering

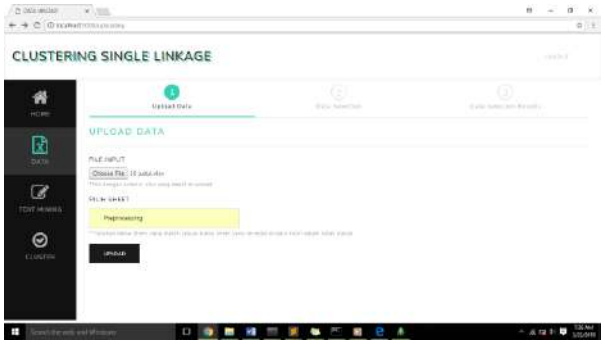
ID	T1	T2	T3	T4	T5	T6
D1	sistem	informasi	klinik	gigi	basis	web

3.4 Pembobotan Kata

Langkah selanjutnya adalah pembobotan kata, dimana dalam pembobotan ini menggunakan rumus Tf-Idf dan hasil pembobotan 10 data dapat dilihat pada Tabel 4.

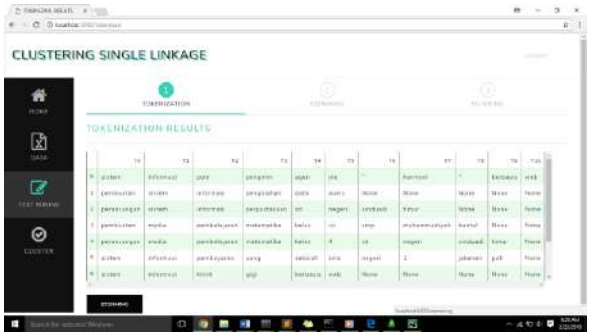
3.5 Implemetasi

Implementasi merupakan interface dari aplikasi yang dibuat. Tampilan awal dalam penelitian ini ditunjukkan dalam Gambar 1



Gambar 1. Load Data

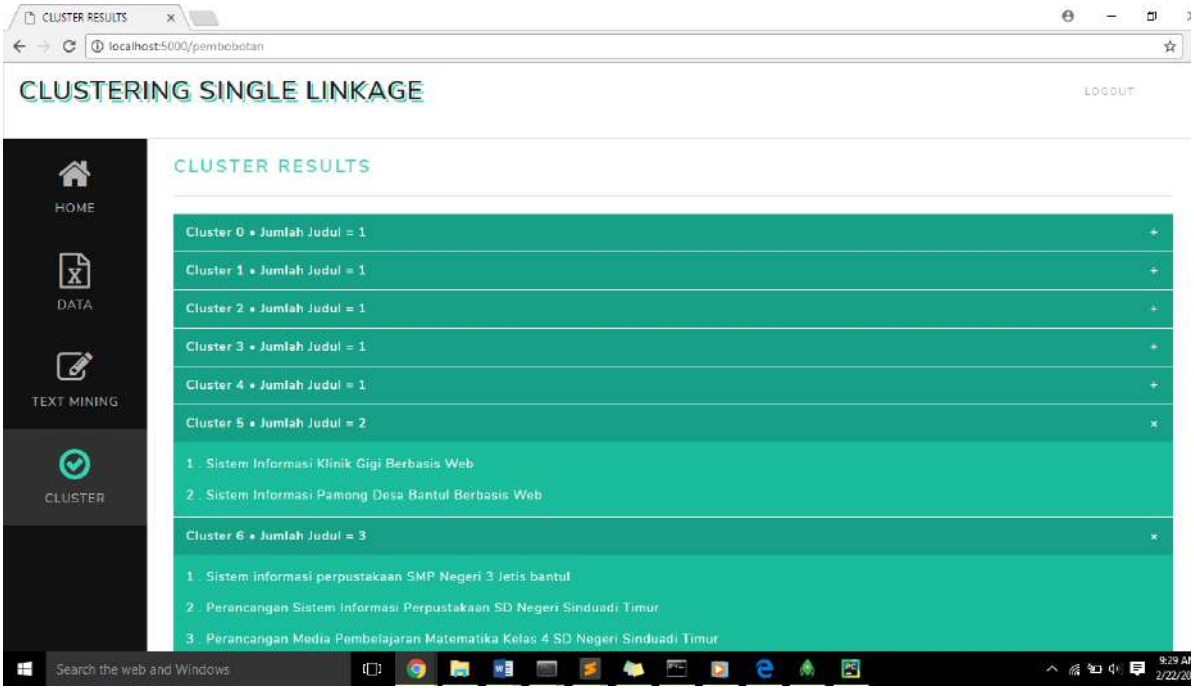
Untuk tampilan preprocessing ditunjukkan dalam Gambar 2 dan tampilan hasil akhir clustering ditunjukkan dalam Gambar 3.



Gambar 2. Preprocesing text

Tabel 4. Tabel Pembobotan Kata

	term	Tf										df	D / df	idf	w = tf x idf								
		D0	D1	D2	D3	D4	D5	D6	D7	D8	D9				D0	D1	D2	D3	D4	D5	D6	D7	D8	D9
1	sistem	1	1	1			1	1		1	1	7	1.43	0.15	0.15	0.15	0.15	0.00	0.00	0.15	0.15	0.00	0.15	0.15
2	informasi	1	1	1			1	1		1	1	7	1.43	0.15	0.15	0.15	0.15	0.00	0.00	0.15	0.15	0.00	0.15	0.15
3	poin	1										1	10	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	kirim	1										1	10	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	agen	1										1	10	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
...	...											...	...	...	...	...	...	...	...	...	...	...	...	...
43	desa									1		1	10	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
44	3										1	1	10	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
45	jetis										1	1	10	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
		9	6	8	9	10	10	6	7	7	8	80			6.23	4.01	4.20	6.36	6.64	7.71	3.23	5.44	3.75	4.63



Gambar 3. Implementasi hasil pengelompokan

3.2.8 Pengujian Akurasi

Pengujian akurasi menggunakan *Purity Test* dilakukan untuk mengetahui baik buruknya kelompok yang dihasilkan dari proses pengelompokkan dengan metode *Single Linkage*. Jika hasil *Puriity Test* semakin mendekati 1, maka semakin baik kualitas kelompoknya. Sedangkan, jika hasil *Puriity Test* semakin mendekati 1, maka semakin baik kualitas kelompoknya. Hasil dari *Purity Test* dapat dilihat pada Tabel 9 : Hasil Purity Test.

No	Kelompok	Asli	Single Linkage	Sesuai
1	SISTEM INFORMASI	168	1	1
2	MULTIMEDIA	192	1	1
3	JARINGAN	27	1	1
4	WEB	213	1	1
5	KEWIRAUSAHAAN	19	1	1
6	MAGANG	50	1	1
7	PELATIHAN	236	899	236
		905	905	242

$$r = \frac{1}{n}\sum_{i=1}^{k} a_i$$

Pengujian akurasi menggunakan *Purity Test* dilakukan untuk mengetahui baik buruknya kelompok yang dihasilkan dari proses pengelompokkan dengan metode *Single Linkage*. Jika hasil *Puriity Test* semakin mendekati 1, maka semakin baik kualitas kelompoknya. Sedangkan, jika hasil Puriity Test semakin mendekati 1, maka semakin baik kualitas kelompoknya. Hasil dari Purity Test dapat dilihat pada Tabel 9 : Hasil *Purity Test*.

Dimana:

r : tingkat akurasi pengelompokan

k : jumlah cluster

a_i : objek yang muncul di *cluster Ci* dan pada label *class* yang sesuai.

$$r = \frac{1}{905}(1+1+1+1+1+1+236)$$

$$r = \frac{1}{905}(242)$$

$$r = 0.267$$

Data yang diuji adalah semua data bersih dari tahun 2012 hingga 2017 yaitu sejumlah 905 data. Jumlah kelompok (cluster) sesuai dengan jenis kartu bimbingan Kerja Praktik yaitu sejumlah 7 (Sistem Informasi, Multimedia, Jaringan, Web, Kewirausahaan, Magang, dan Pelatihan). Pengelompokan data asli dilakukan secara manual agar dapat dibandingkan dengan hasil dari pengelompokkan menggunakan metode Single Linkage. Dari perhitungan diatas didapatkan hasil *Purity Test* adalah sebesar 0.267. Hasil tersebut tergolong rendah karena *Purity test*

memiliki range nilai antara 0 – 1. Semakin mendekati 1 hasil kelompoknya semakin baik.

4. KESIMPULAN

Berdasarkan hasil penelitian *"Text Mining* Menggunakan Metode *Single Linkage* dengan *Manhattan Distance Similarity* untuk Pengelompokan Trend Topik Kerja Praktik" dapat ditarik kesimpulan sebagai berikut :

1. Uji *Purity Test* yang dilakukan pada aplikasi "*Text Mining* Menggunakan Metode *Single Linkage* dengan *Manhattan Distance Similarity* untuk Pengelompokan Trend Topik Kerja Praktik" menggunakan 905 data dan dibagi menjadi 7 kelompok sesuai pada jenis Kartu Bimbingan KP, menunjukan hasil sebesar 0.267. Sehingga, Metode pengelompokkan *Single Linkage* di penelitian ini menghasilkan kelompok yang kurang bagus.
2. Hasil *Uji Purity Test* tergolong rendah karena data yang digunakan adalah data teks berupa judul laporan, dimana judul laporan ditulis dengan kata-kata yang berbeda antara judul satu dengan judul yang lainnya, beda kata menjadikan bobotnya juga berbeda, bobot yang berbeda menjadikan nilai similaritas berbeda, padahal untuk menjadi satu kelompok suatu judul harus memiliki nilai similaritas yang hampir sama. Lalu dari pemilihan metode, Single Linkage merupakan pengelompokan dengan pendekatan hierarki (*hierarchical clustering*) mengelompokan data dengan membuat suatu hierarki berupa dendrogram dimana data yang mirip akan ditempatkan pada hirarki yang berdekatan dan yang tidak pada hierarki yang berjauhan.
3. Metode pengelompokkan *Single Linkage* dengan *Manhattan Distance Similarity* di penelitian ini menghasilkan kelompok yang kurang bagus dan kurang cocok untuk pengelompokkan Trend Topik Kerja Praktik.

DAFTAR PUSTAKA

Chakraborty, G., Pagolu, M. and Garla, S. (2013) *PREVIEW: Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS*, *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS*.

Handoyo, R. *et al.* (2014) "Perbandingan Metode Clustering Mengggunakan metode Single Linkage dan K-Means Pada Pengelompokkan Dokumen," *JSM STMIK Mikroskil*, 15(2), pp. 73–82.

Janani, R. and Vijayarani, S. (2016) "Text Mining Research : A Survey," *International Journal of innovative Research in Computer and Communication Engineering*, 4(4), pp. 6564–6571. doi: 10.15680/IJIRCCE.2016.

Kao, A. and Poteet, S. R. (2006) *Natural Language Processing and Text Mining*. USA: Springer.

Muzzammil, R. R., Ginardi, R. V. hari and Purwitasari, D. (2016) "Modul Klasifikasi Aduan dengan Pendekatan Kemiripan Teks pada Aplikasi Perangkat Bergerak Suara Warga (Surga) Kota Kediri," *Jurnal Teknik ITS*, 5(1), pp. 52–57.

Zahrotun, L. (2016) "Comparison Jaccard Similarity , Cosine Similarity and Combined Both of the Data Clustering With Shared Nearest Neighbor Method," *Computer Engineering and Applications*, 5(1), pp. 11–18.

Zahrotun, L. (2017) "Text Mining for Internship Titles Clustering Using Shared Nearest-Neighbor Method," *Computer Engineering and Applications*, 6(3).

Zahrotun, L. and Mushlihudin (2017) "Rancang Bangun Aplikasi Text Mining dalam Mengelompokkan Judul Penelitian Dosen Menggunakan Metode Shared Nearest Neighbor dan Euclidean Similarity," *Jurnal Ilmu Teknik elektro Komputer dan informatika (JITEKI)*, 3(2), pp. 91–99.

Zahrotun, L., Putri, N. hutami and Khusna, A. N. (2018) "The Implementation of K-Means Clustering Method in Classifying Undergraduate Thesisi Titles," in *12th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*. Yogyakarta: IEEE.