Perhitungan Kemiripan Dokumen Bahasa Indonesia Menggunakan Metode Cosine Similarity (Studi Kasus : Abstrak Tugas Akhir Fakultas Informatika Universitas Telkom)

Calculating Similarity Of Document In Bahasa Indonesia With Cosine Similarity Method

[1]Irfan Pahlevi. [2] Moch. Arif Bijaksana, Ir., M.Tech,

[1,2,]Prodi Sl Teknik Informatika, Fakultas Informatika, Universitas Telkom
[1]irfan.pahlevi93@gmail.com, [2]arifbijaksana@telkomuniversity.ac.id.

Abstrak

Perkembangan teknologi sekarang ini dapat memudahkan kegiatan yang biasanya harus dilakukan oleh manusia. Dengan banyaknya informasi di internet maka manusia akan dengan mudah mengakses dan mengolah informasi dari manapun. Seperti melakukan pengecekan tingkat kemiripan dokumen yang membutuhkan waktu lama dan tenaga yang besar, maka dapat digantikan oleh aplikasi yang menerapkan *text mining*.

***Text Mining* adalah suatu bahasan yang menganalis dan mengelola teks menjadi sebuah informasi yang dapat diolah untuk tujuan tertentu. Pada *Text Mining* dikenal istilah *preprocessing* yang terdiri dari *case folding, tokenizing, filtering, dan stemming*. Hal ini yang dilakukan sebelum proses pembobotan *tf-idf* dan perhitungan *cosine similarity.* Metode yang digunakan adalah Cosine Similarity. Cosine Similarity merupakan metode perhitungan jarak antara vektor A dan B yang menghasilkan sudut kosinus x diantara kedua vektor tersebut. Nilai sudut kosinus antara dua vektor menentukan kesamaan dua buah objek yang dibandingkan dimana nilai terkecil adalah 0 dan nilai terbesar adalah 1. Nilai 0 menandakan bahwa kedua abstrak yang dibandingkan tidak mirip sama sekali dan semakin mendekati nilai 1 berarti tingkat kemiripannya semakin besar**

Dengan didapatkannya hasil sebesar 0.5729 pada perhitungan koefisien korelasi Pearson, maka dapat dinyatakan bahwa korelasi antara perhitungan menggunakan TF-IDF dan metode Cosine Similarity dengan penilaian secara manual memiliki nilai korelasi linear positif karena berada diantara 0 sampai 1.

Kata kunci : ***text mining, kemiripan, tugas akhir, Cosine Similarity, Telkom University***

Abstract

The current technological developments can facilitate activities that must be performed by humans. With the a lot of information on the internet, so people will easily access and process information from anywhere. Such as checking the degree of similarity of documents that take a long time and a great power, it can be replaced by applying text mining.

Text Mining is a discussion that analyze and manage text into an information that can be processed for a specific purpose. In the Text Mining be known about preprocessing, consisting of folding case, tokenizing, filtering, and stemming. This is done before the tf-idf weighting and calculation cosine similarity. The method used is Cosine Similarity. Cosine Similarity is a method of calculating the distance between vectors A and B which produces x cosine angle between the two vectors. Cosine value of the angle between two vectors determine the similarity of two objects are compared where the smallest value is 0 and the largest value is 1. A value of 0 indicates that the two abstracts were compared are not similar at all and getting closer to the value of 1 means that the greater the degree of similarity.

Attaining a yield of 0.5729 at Pearson correlation coefficient calculation, it can be stated that the correlation between the calculations using TF-IDF and Cosine Similarity method with manual assessment has a positive linear correlation value because it is located between 0 and 1.

Keywords: text mining, similarity, final project, Cosine Similarity, Telkom University

1. Pendahuluan

Teknologi tercipta untuk membantu pekerjaan manusia agar lebih cepat dan tepat. Dengan berkembangnya teknologi informasi maka pengolahan data akan lebih mudah untuk dilakukan. Banyaknya data di internet maka dapat diolah untuk dijadikan informasi yang bermanfaat bagi orang lain. Pekerjaan yang biasanya dilakukan secara manual oleh manusia, kini sudah dapat digantikan oleh teknologi informasi. Dengan memanfaatkan teknologi informasi, maka dapat mempercepat waktu dan mengurangi tenaga yang dibutuhkan dibanding dilakukan oleh manusia. Selain memakan waktu yang

cukup lama untuk membaca semua dokumen, kelemahan dari pengecekan manual adalah terkadang nilai yang diberikan masih bersifat subjektif dan tidak berimbang.

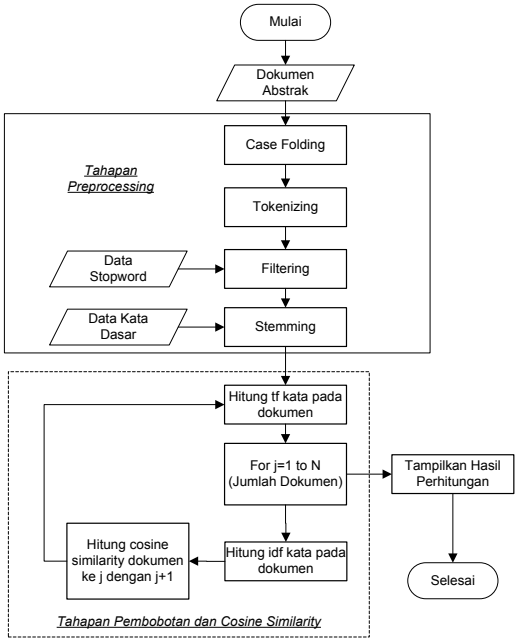
Menurut Kamus Besar Bahasa Indonesia, kemiripan adalah hampir sama atau serupa. Dalam tugas akhir ini akan dibahas mengenai cara untuk mengetahui tingkat kemiripan suatu tugas akhir dengan tugas akhir lainnya dilihat dari judul dan abstraknya. Hal ini dilakukan karena abstrak adalah rangkuman dari seluruh bab yang dikerjakan di tugas akhir dan di dalam abstrak terdapat kata kunci yang bersifat unik dan mencakup seluruh pembahasan di tugas akhir terkait.

Text Mining adalah suatu bahasan yang menganalis dan mengelola teks menjadi sebuah informasi yang dapat diolah untuk tujuan tertentu. Pada *Text Mining* dikenal istilah *case folding, tokenizing, filtering, dan stemming*. Hal ini yang dilakukan pertama kali terhadap data abstrak sebelum dilakukan pembobotan *tf-idf* dan perhitungan *cosine similarity.*

2. Dasar Teori dan Perancangan

Sistem akan dibangun menggunakan bahasa Java kemudian aplikasi yang tersedia menerima inputan berupa data abstrak dari tugas akhir mahasiswa dalam format .csv. Data abstrak yang sudah ada dilakukan proses *preprocessing* yaitu *case folding, tokenizing, filtering,* dan *stemming.* Kemudian dilanjutkan pembobotan term menggunakan *tf-idf* dan perhitungan menggunakan *cosine similarity.* Pada bagian ini akan ditampilkan alur proses dari sistem yang akan dibangun.

2.1 Proses



Gambar 1. Rancangan Umum

Berdasarkan gambar 1 Rancangan Umum menyatakan bahwa rancangan sistem dari awal hingga akhir adalah:

a. **Dokumen Abstrak**

Dokumen abstrak maksudnya adalah dataset yang sudah ada yaitu abstrak tugas akhir dalam format .csv

b. ***Preprocessing***

Tahap *preprocessing* pada gambar di atas terdiri dari :

i. *Case Folding*

Pada tahapan ini adalah mengubah seluruh huruf kapital menjadi huruf kecil untuk kemudahan pengolahan data.

ii. *Tokenizing*

Pada tahapan ini adalah isi abstrak yang berupa kalimat dan paragraf dipecah menjadi susunan per kata.

iii. *Filtering*

Pemisahan kata yang akan digunakan dan kata yang tidak akan digunakan. Untuk yang akan digunakan biasa disebut *wordlist* sedangkan yang tidak digunakan biasa disebut *stoplist*, termasuk membuang *string* yang dianggap sebagai tanda baca seperti titik, koma, atau tanda kutip.

iv. *Stemming*

Mengubah kata yang sudah melewati *tokenizing* dan *filtering* menjadi kata dasar contohnya penggunaan menjadi guna. *Stemming* ini dilakukan dengan mengimplementasikan algoritma *Stemming Nazief – Adriani.*

Berikut adalah peraturan dari *Stemming* Nazief – Adriani :

1. Algoritma *stemming* Nazief dan Adriani dikembangkan berdasarkan aturan morfologi Bahasa Indonesia yang mengelompokkan imbuhan menjadi awalan (*prefix*), sisipan (*infix*), akhiran (*suffix*) dan gabungan awalan akhiran (*confixes*). Algoritma ini menggunakan kamus kata dasar dan mendukung *recoding*, yakni penyusunan kembali kata-kata yang mengalami proses *stemming* berlebih. Aturan morfologi Bahasa Indonesia mengelompokkan imbuhan ke dalam beberapa kategori sebagai berikut [11] : *Inflection suffixes* yakni kelompok akhiran yang tidak merubah bentuk kata dasar. Sebagai contoh, kata "duduk" yang diberikan akhiran "-lah" akan menjadi "duduklah".

Kelompok ini dapat dibagi menjadi dua:

i. *Particle* (P) atau partikel yakni termaksud di dalamnya "-lah", "kah", "tah" dan "pun".

ii. *Possessive pronoun* (PP) atau kata genti kepunyaan, termaksud di dalamnya "-ku","-mu" dan "- nya".

2. *Derivation suffixes* (DS) yakni kumpulan akhiran asli Bahasa Indonesia yang secara langsung ditambahkan pada kata dasar yaitu akhiran "-i", "-kan", dan "-an".

3. *Derivation prefixes* (DP) yakni kumpulan awalan yang dapat langsung diberikan pada kata dasar murni, atau pada kata dasar yang sudah mendapatkan penambahan sampai dengan 2 awalan. Termaksud di dalamnya adalah:

i. Awalan yang dapat bermorfologi ("me-", "be-", "pe-" dan "te").

ii. Awalan yang tidak bermorfologi ("di-", "ke-" dan "se-").

[DP+ [DP+ [DP+]]] Kata Dasar [[+DS] [+PP]

Gambar 2 Model Kata dalam Bahasa Indonesia

Keterangan : *DP : Derivation prefixes DS : Derivation suffixes PP : Possessive pronoun*

c. Pembobotan TF-IDF

Jika suatu kata atau *term* muncul di dokumen, maka nilai vektornya adalah bukan-nol (nonzero). Beberapa cara berbeda untuk menghitung nilai tersebut, dikenal sebagai bobot *term*. Salah satu formula yang terkenal yaitu pembobotan *term frequency-inverse document frequency (tf-idf).*

Terdapat tiga bentuk pembobotan yang dapat diberikan pada suatu kata dalam matriks yaitu pembobotan lokal, pembobotan global, dan normalisasi. Pembobotan lokal diberikan pada suatu kata berdasarkan jumlah kemunculannya pada satu dokumen. Pembobotan global diberikan pada suatu kata atas kemunculannya di semua dokumen, sedangkan normalisasi adalah sebuah cara untuk menormalkan panjang vektor dokumen sehingga vektor tersebut independen terhadap panjangnya. Setiap kata mendapatkan bobot dari tiga bentuk pembobotan ini sehingga totalnya adalah perkalian dari tiga pembobotan tersebut (Robertson, 2005).

d. Perhitungan *Cosine Similarity*

Metode *cosine similarity* adalah metode untuk menghitung kesamaan dari dua dokumen. Penentuan kesesuaian dokumen dengan *query* dipandang sebagai pengukuran (*similarity measure*) antara vektor dokumen (A) dengan vektor query (B).

2.2. Implementasi Metode *Cosine Similarity*

Pendeteksi kemiripan dokumen teks pada tugas akhir ini menggunakan algoritma *Stemming Nazief & Adriani* dan metode *Cosine Similarity*. Algoritma *Nazief & Adriani* dimanfaat untuk mengubah kata menjadi bentuk kata dasar. Metode *Cosine Similarity* digunakan untuk menghitung

nilai kemiripan antar dokumen teks. Algoritma *Nazief & Adriani* pada penelitian ini merupakan algoritma yang digunakan untuk mempengaruhi nilai yang dihasilkan oleh metode *Cosine Similarity*.

Dijelaskan di bawah ini perhitungan kemiripan antara dua kalimat . Kalimat no.1 yaitu, "Komputer membantu pekerjaan manusia menjadi lebih mudah". kalimat no.2 yaitu, "Pekerjaan manusia manjadi lebih mudah bila dibantu oleh komputer".

Tabel 1 Contoh *Cosine Similarity*

No	Kata	Kalimat 1	Kalimat 2
1	Bila	0	1
2	Dibantu	0	1
3	Computer	1	1
4	Lebih	1	1
5	Manusia	1	1
6	Membantu	1	0
7	Menjadi	1	1
8	Mudah	1	1
9	Oleh	0	1
10	Pekerjaan	1	1

Dari tabel di atas akan dibentuk vektor menjadi dibawah ini :

Kalimat 1 : {0,0,1,1,1,1,1,1,0,1}

Kalimat 2 : {1,1,1,1,1,0,1,1,1,1}

Dengan menggunakan rumus *Cosine Similarity* maka didapatkan hasil yaitu 0,7599 dan persentase kemiripan yaitu 75,99

3. Pengujian

3.1 Tujuan Pengujian

Pada tahap ini akan dibahas mengenai analisis dari hasil pengujian sesuai dengan sistem yang telah dibuat. Pengujian dilakukan bertujuan untuk mengetahui performansi algoritma *Cosine Similarity* dengan pembobotan TF-IDF dalam menghitung kemiripan abstrak tugas akhir.

3.2 Data Uji

Data Uji yang digunakan berasal dari dataset yang didapat dari awal dan digunakan keseluruhan bagian abstrak tanpa dikurangi sama sekali. Data uji sebanyak 341 abstrak mahasiswa Fakultas Informatika Universitas Telkom. Selanjutnya dilakukan perhitungan kemiripan tiap satu abstrak dengan abstrak yang lain dengan keterkaitan 1 ke N dengan N sebanyak jumlah data abstrak kurang 1.

3.3 Pengujian

Hasil yang didapat dari perhitungan sistem yang telah dibuat selanjutnya dibandingkan dengan perhitungan kemiripan secara manual dan selanjutnya dicari nilai korelasi antar dua nilai yang sudah didapat baik secara manual maupun hasil perhitungan sistem. Nilai dari korelasi tersebut akan berada pada *range* -1 sampai 1 dengan keterangan mendekati 1 berarti kedua variabel memiliki nilai korelasi linear positif yang kuat dan mendekati -1 berarti kedua variabel memiliki nilai korelasi linear negatif yang kuat..

3.4 Penilaian Secara Manual

Penilaian secara manual dilakukan dengan cara membaca seluruh isi dokumen dan membandingkan dengan dokumen pembanding lalu memberikan skor dengan skala nol sampai satu. Dengan penilaian nol adalah tidak mirip sama sekali dan apabila mendekati satu berarti kedua dokumen memiliki kemiripan kata yang tinggi.

Skala nilai awal adalah nol sampai lima, lalu dinormalisasi menjadi nol sampai satu. Contohnya 3 pada nilai awal, setelah menjadi 0.6 setelah dinormalisasi.

3.4 Hasil Penilaian Secara Manual

Berikut adalah tabel perhitungan kemiripan abstrak secara *random* dan manual :

Tabel 2 Penilaian Manual

no	Abstrak	Abstrak	Nilai

	Pembanding 1	Pembanding 2	
1	Dokumen 1	Dokumen 12	0.6
2	Dokumen 30	Dokumen 41	0.2
3	Dokumen 42	Dokumen 196	0.6
4	Dokumen 200	Dokumen 65	1
5	Dokumen 166	Dokumen 224	0.6
6	Dokumen 299	Dokumen 146	0
7	Dokumen 135	Dokumen 136	0
8	Dokumen 113	Dokumen 66	0.4
9	Dokumen 29	Dokumen 172	0.2
10	Dokumen 179	Dokumen 268	0
11	Dokumen 270	Dokumen 272	0
12	Dokumen 282	Dokumen 300	0.8
13	Dokumen 304	Dokumen 68	0.2
14	Dokumen 69	Dokumen 142	0.2
15	Dokumen 153	Dokumen 173	0.4
16	Dokumen 189	Dokumen 140	0.8
17	Dokumen 144	Dokumen 227	0.2
18	Dokumen 235	Dokumen 105	0.2
19	Dokumen 63	Dokumen 267	0.6
20	Dokumen 268	Dokumen 280	0.6
21	Dokumen 281	Dokumen 302	0
22	Dokumen 305	Dokumen 164	0.8
23	Dokumen 161	Dokumen 269	0.2
24	Dokumen 271	Dokumen 281	0
25	Dokumen 127	Dokumen 313	0.4
26	Dokumen 314	Dokumen 97	0.2
27	Dokumen 99	Dokumen 130	0.2
28	Dokumen 208	Dokumen 135	0.2
29	Dokumen 229	Dokumen 91	0
30	Dokumen 94	Dokumen 143	0.2

3.5 Hasil Perhitungan Sistem

Berikut adalah tabel perhitungan kemiripan abstrak secara sistematik menggunakan pembobotan TF-IDF dan *cosine similarity* :

Tabel 3 Perhitungan Sistem

no	Abstrak Pembanding 1	Abstrak Pembanding 2	Nilai
1	Dokumen 1	Dokumen 12	0,28
2	Dokumen 30	Dokumen 41	0,11
3	Dokumen 42	Dokumen 196	0,28

4	Dokumen 200	Dokumen 65	0,30
5	Dokumen 166	Dokumen 224	0,11
6	Dokumen 299	Dokumen 146	0,07
7	Dokumen 135	Dokumen 136	0,04
8	Dokumen 113	Dokumen 66	0,25
9	Dokumen 29	Dokumen 172	0,20
10	Dokumen 179	Dokumen 268	0,19
11	Dokumen 270	Dokumen 272	0,02
12	Dokumen 282	Dokumen 300	0,41
13	Dokumen 304	Dokumen 68	0,07
14	Dokumen 69	Dokumen 142	0,07
15	Dokumen 153	Dokumen 173	0,54
16	Dokumen 189	Dokumen 140	0,67
17	Dokumen 144	Dokumen 227	0,05
18	Dokumen 235	Dokumen 105	0,05
19	Dokumen 63	Dokumen 267	0,37
20	Dokumen 268	Dokumen 280	0,90
21	Dokumen 281	Dokumen 302	0,34
22	Dokumen 305	Dokumen 164	0,22
23	Dokumen 161	Dokumen 269	0,05
24	Dokumen 271	Dokumen 281	0,27
25	Dokumen 127	Dokumen 313	0,46
26	Dokumen 314	Dokumen 97	0,29
27	Dokumen 99	Dokumen 130	0,10
28	Dokumen 208	Dokumen 135	0,36
29	Dokumen 229	Dokumen 91	0,09
30	Dokumen 94	Dokumen 143	0,20

4.6 Perhitungan Korelasi

Berikut adalah tabel perhitungan koefisien korelasi antara perhitungan kemiripan melalui sistem yang dibangun dan perhitungan manual oleh manusia :

Tabel 4 Perhitungan Korelasi

No.	X	Y	x2	y2	Xy
1	0.6	0,28	0.36	0.08	0.168
2	0.2	0,11	0.04	0.0121	0.022
3	0.6	0,28	0.36	0.0784	0.168
4	1	0,30	1	0.09	0.3
5	0.6	0,11	0.36	0.0121	0.066
6	0	0,07	0	0.5776	0
7	0	0,04	0	0.1849	0

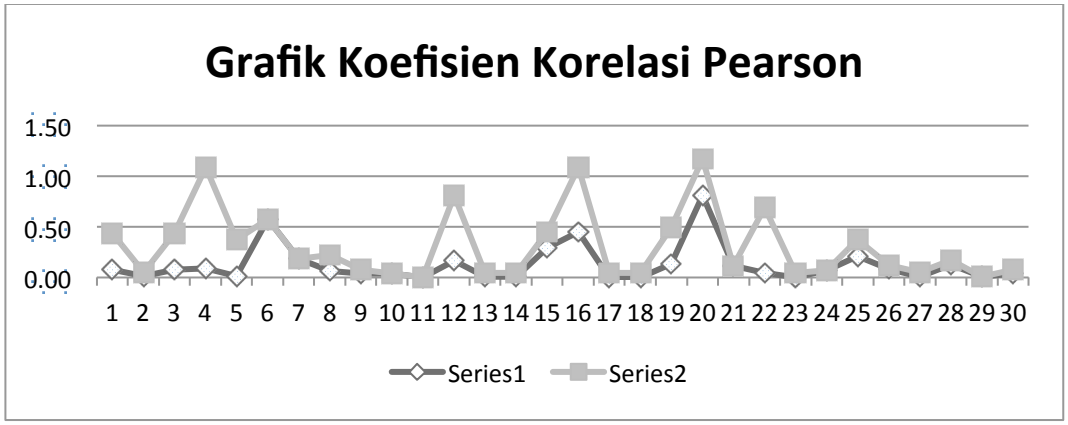
8	0.4	0,25	0.64	0.0625	0.1
9	0.2	0,20	0.04	0.04	0.04
10	0	0,19	0.04	0.0361	0
11	0	0,02	0.16	0.004	0
12	0.8	0,41	0.64	0.1681	0.328
13	0.2	0,07	0.04	0.0049	0.014
14	0.2	0,07	0.04	0.004	0.014
15	0.4	0,54	0.16	0.291	0.216
16	0.8	0,67	0.64	0.448	0.536
17	0.2	0,05	0.04	0.025	0.01
18	0.2	0,05	0.04	0.002	0.01
19	0.6	0,37	0.36	0.136	0.22
20	0.6	0,90	0.36	0.81	0.54
21	0	0,34	0	0.115	0
22	0.8	0,22	0.64	0.048	0.176
23	0.2	0,05	0.04	0.002	0.01
24	0	0,27	0	0.072	0
25	0.4	0,46	0.16	0.211	0.184
26	0.2	0,29	0.04	0.084	0.058
27	0.2	0,10	0.04	0.01	0.02
28	0.2	0,36	0.04	0.129	0.072
29	0	0,09	0	0.008	0
30	0.2	0,20	0.04	0.04	0.04
Σ	8.44	9.8	3.77	5.64	3.31

Keterangan :

x : nilai perbandingan secara manual

y : nilai perbandingan melalui sistem

Dari tabel di atas maka bentuk grafik perbandingan untuk nilai perhitungan kemiripan dokumen berdasarkan penilaian manual dan perhitungan sistem adalah sebagai berikut :



Grafik 1 Grafik Korelasi

4.6.1 Perhitungan Koefisien Korelasi Pearson

Dengan didapatkannya hasil sebesar 0.5729 pada perhitungan koefisien korelasi Pearson, maka dapat dinyatakan bahwa korelasi antara perhitungan menggunakan TF-IDF dan metode Cosine Similarity dengan penilaian secara manual memiliki nilai korelasi linear positif karena berada diantara 0 sampai 1.

4. Kesimpulan dan Saran

Berdasarkan pengujian serta analisis yang telah dilakukan pada tugas akhir ini, dapat diambil beberapa kesimpulan, yaitu :

1. Dengan menerapkan text mining maka dapat dilihat tingkat kemiripan dokumen Bahasa Indonesia dalam hal ini abstrak tugas akhir.
2. Secara perhitungan korelasi koefisien perbandingan antara penilaian manual dan perhitungan sistem untuk menentukan kemiripan abstrak tugas akhir, didapatkan bahwa secara rata – rata nilainya tidak terlampau jauh yang berarti validasi aplikasi sudah baik.
3. Dengan menerapkan text mining, maka untuk melakukan perbandingan dokumen yang sangat banyak dapat dilakukan dengan cepat tanpa harus membaca keseluruhan dokumen.

Pengembangan dari tugas akhir ini dapat dilakukan :

1. Penilaian manual sebaiknya dilakukan oleh lebih dari 5 orang agar penilaian yang dilakukan bersifat lebih objektif dan bervariatif sehingga akan didapatkan nilai tengahnya.
2. Data yang diolah sebaiknya lebih banyak dan lebih beragam supaya kinerja sistem dapat lebih baik.

Daftar Pustaka :

[1] Anna Huang. Similarity Measures for Text Document Clustering.

[2] Ardiansyah, Adryan. Pengembangan Aplikasi Pendeteksi Plagiarisme Menggunakan Metode Latent Semantic Analysis (LSA), Studi Kasus Plagiarisme Karya Ilmiah Berbahasa Indonesia), 2011

[3] Azhar Firdaus, Ernawati, Arie Vatresia. Aplikasi Pendeteksi Kemiripan Pada Dokumen Teks Menggunakan Algoritma Nazief & Adriani Dan Metode Cosine Similarity. 2014

[4] Bhanu Prasad A., Venkata Gopala. Space and Cosine Simililarity measures for Text Document Clustering. 2013

[5] Kemal Ade Sekarwati, Lintang Yuniar Banowosari, I Made Wiryana, Djati Kerami. Pengukuran Kemiripan Dokumen Dengan Menggunakan Tools GENSIM. 2015

[6] Nova Khairunnisa, Dadang Syarif, Ardianto Wibowo. Aplikasi Pendeteksi Plagiat dengan Menggunakan Metode Latent Semantic Analysis (Studi Kasus : Laporan TA PCR).

[7] Nur Rosyid M, Entin Martiana, Damitha Vidyastana, *Penentuan Kemiripan Topik Proyek Akhir Berdasarkan Abstrak Pada Jurusan Teknik Informatika Menggunakan Metode Single Linkage Hierarchical*.

[8] O. Maimon and L. Rokach, Data Mining and Knowledge Discovery Handbook. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2005..

[9] Oxford Advanced Learner's dictionary 5th edition 1995

[10] Radiant Victor Imbar, Adelia, Mewati Ayub, Alexander Rehatta. Implementasi Cosine Similarity dan Algoritma Smith-Waterman untuk Mendeteksi Kemiripan Teks. 2014

[11] Saleh, Muhammad. 2012. Menentukan Kemiripan Topik Tugas Akhir Berdasarkan Deskripsi Pada Jurusan Teknik Informatika Menggunakan Metode Inner Product dan Single Linkage Hierarchical. Skripsi Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Malang

[12] Tahitoe, Andita Dwiyoga(tanpa tahun). "Modifikasi Enhanced Confix Stripping Stemmer Untuk Bahasa Indonesia dengan Metode Corpus Based Stemming". Institut Teknologi Sepuluh Nopember, 1-15

[13] Winda Ayu Irianto, Drs. Achmad Ridok, M.Kom., Lailil Muflikhah, S.Kom., M.Sc. Penentuan Tingkat Plagiarisme Dokumen Penelitian Menggunakan Centroid Linkage Hierarchical Method (CLHM).

[14] X. Wu and V. Kumar, eds., The Top Ten Algorithms in Data Mining.Chapman and Hall, 2009.