

Dokumen Pemanding	Nilai Kemiripan	Dokumen pemanding	Nilai kemiripan
Abstrak 32	0.06918988	Abstrak 90	0.01961846
Abstrak 33	0.56484648	Abstrak 91	0.14907149
Abstrak 34	0.12248006	Abstrak 92	0.05695438
Abstrak 35	0.20853131	Abstrak 93	0.08996868
Abstrak 36	0.20321736	Abstrak 94	0.18929787
Abstrak 37	1	Abstrak 95	0.08612819
Abstrak 38	0.05765132	Abstrak 96	0.10208029
Abstrak 39	0.52166193	Abstrak 97	0.12013396
Abstrak 40	0.05518797	Abstrak 98	0.07204142
Abstrak 41	0.11289404	Abstrak 99	0.0454027
Abstrak 42	0.09784821	Abstrak 100	0.13592193
Abstrak 43	0.12173014	Abstrak 101	0.07949741
Abstrak 44	0.07610631	Abstrak 102	0.08950764
Abstrak 45	0.3787092	Abstrak 103	0.09404702
Abstrak 46	0.1142238	Abstrak 104	0.08130033
Abstrak 47	0.2566103	Abstrak 105	0.03692271
Abstrak 48	0.10479726	Abstrak 106	0.06769954
Abstrak 49	0.05095993	Abstrak 107	0.06885338
Abstrak 50	0.04392516	Abstrak 108	0.08777579
Abstrak 51	0.22777396	Abstrak 109	0.13189072
Abstrak 52	0.07249316	Abstrak 110	0.05934871
Abstrak 53	0.0616631	Abstrak 111	0.09832263
Abstrak 54	0.09090472	Abstrak 112	0.06003592
Abstrak 55	0.08920618	Abstrak 113	0.22045239
Abstrak 56	0.09169365	Abstrak 114	0.06745582
Abstrak 57	0.44973581	Abstrak 115	0.12717072
Abstrak 58	0.09966345	Abstrak 116	0.10648255

Berdasarkan tabel 4.3 merupakan hasil kemiripan dokumen uji dengan semua dokumen pembanding yaitu sebanyak 116 dokumen abstrak. Memiliki nilai 1 jika dokumen uji dan dokumen pembanding memiliki kesamaan. Semakin rendah nilai kemiripan maka dokumen tersebut tidak berplagiat dengan dokumen pembanding.

4.3.2 Metode *Word2vec* dan Cosine Similarity

percobaan selanjutnya yaitu menggunakan *Word2vec* untuk mengubah kata-kata menjadi bentuk *vector* dengan tujuan untuk mencari nilai kedekatan antar kata.

Penulis menggunakan Gensim untuk implementasi *Word2vec*. Gensim merupakan open-source model ruang vektor dan topic modeling, yang diimplementasikan dalam Bahasa pemrograman Python. Langkah pertama yang penulis buat yaitu membuat wrapper class yaitu membaca setiap *corpus*, *word vector*, *Cosine Similarity*. Penulis membuat dokumen uji sebagai model dan menghilangkan kata-kata stopwords terlebih dahulu. Penulis menggunakan abstrak 33 sebagai percobaan. Untuk data pembanding menggunakan seluruh abstrak sebagai corpus atau kamus bahasa Indonesia. Berdasarkan percobaan hasil vektor dokumen pembanding sebanyak 116 abstrak dapat diketahui. Berikut merupakan vektor dari dokumen uji dan dokumen pembanding.

[6.4605789e-04	-9.8795583e-04	-1.6786055e-04	8.6065271e-04
-1.0657071e-03	-8.0747210e-04	-6.7189737e-04	-1.5671735e-04
3.1455970e-04	-1.6626263e-04	-1.0366469e-03	-3.6189471e-05
4.8287632e-04	-7.0258748e-04	-3.7939640e-04	-6.6730950e-04
-9.9935173e-04	2.1032250e-04	2.8884114e-04	6.7331456e-04
3.8648158e-04	5.4567080e-04	2.8482359e-04	-3.3652212e-04
4.0950271e-04	-3.8750772e-04	-8.4294828e-05	-6.5942644e-04
-8.4716530e-04	6.2747582e-05	6.1152963e-04	-9.4357226e-04
-2.7130614e-04	6.8225234e-04	-7.0055079e-04	-1.3118732e-04
-5.1276671e-04	2.2719115e-04	2.9873684e-06	-1.9587431e-04
2.6846157e-05	-1.3454105e-03	6.1151409e-04	1.3446531e-03
-1.2331762e-04	1.3055446e-03	9.8450785e-04	-1.3894533e-03
-6.6434266e-05	-4.6188003e-04	-8.1398216e-04	1.4866308e-04
-8.3038372e-05	-4.0379024e-04	-1.4635023e-03	-6.0109940e-04
-2.9034985e-04	9.5342554e-04	6.3497777e-04	-2.5496684e-04
3.9609364e-04	1.2204911e-03	-1.2583763e-03	4.5424947e-04
-9.0273324e-04	4.2306425e-04	5.2422460e-04	2.5525904e-05
-2.7251532e-04	-9.8369841e-04	6.3935883e-04	-2.6353574e-04
-5.4667139e-04	1.4697940e-03	5.2317703e-04	-3.4383123e-04
5.7293434e-04	-8.0324837e-04	-1.2237944e-03	1.8983742e-04
5.8821915e-04	-3.4955738e-04	-3.3313129e-04	-1.3440380e-04
4.1452993e-04	2.6510027e-04	3.7606878e-04	7.9698261e-04
3.7641768e-04	2.1470188e-04	2.7438405e-04	-4.2632848e-05
-1.5154136e-03	-1.0912312e-03	-3.0114426e-04	-1.0885895e-03
1.3130957e-03	4.1631702e-04	-2.8079201e-04	3.3564647e-04]

Setelah diketahui nilai vektor dari dokumen uji dan dokumen pembanding maka akan dihitung kemiripan dokumen menggunakan *Cosine Similarity*. Hasil

kemiripan dari sebesar 0.8538808. Tabel 4.4 merupakan nilai kemiripan dari dokumen uji dengan 116 astrak menggunakan metode *Word2vec*.

Tabel 4. 4 Nilai Kemiripan Dokumen Uji dengan Dokumen Pembanding dengan *Word2vec*

Dokumen Pembanding	Nilai Kemiripan	Dokumen Pembanding	Nilai Kemiripan
Abstrak 1	0.22264175	Abstrak 59	0.48094505
Abstrak 2	0.1371604	Abstrak 60	0.26394352
Abstrak 3	0.19711533	Abstrak 61	0.37786728
Abstrak 4	0.030326221	Abstrak 62	0.21892473
Abstrak 5	0.030326221	Abstrak 63	0.3255793
Abstrak 6	0.20269045	Abstrak 64	0.22778605
Abstrak 7	0.22121407	Abstrak 65	0.32405353
Abstrak 8	0.32464153	Abstrak 66	0.054884713
Abstrak 9	0.1591025	Abstrak 67	0.35012674
Abstrak 10	0.27428207	Abstrak 68	0.052798368
Abstrak 11	0.23478605	Abstrak 69	0.5316477
Abstrak 12	0.21604858	Abstrak 70	0.27392897
Abstrak 13	0.27917206	Abstrak 71	0.07851979
Abstrak 14	0.17544876	Abstrak 72	0.59820443
Abstrak 15	0.27753904	Abstrak 73	0.014364033
Abstrak 16	0.06526552	Abstrak 74	0.26684174
Abstrak 17	0.18365003	Abstrak 75	0.39069095
Abstrak 18	0.40248156	Abstrak 76	0.10698622
Abstrak 19	0.45778444	Abstrak 77	0.3081009
Abstrak 20	0.24866031	Abstrak 78	0.014909746
Abstrak 21	0.05562416	Abstrak 79	0.0673433
Abstrak 22	0.21290481	Abstrak 80	0.2075908
Abstrak 23	0.272257	Abstrak 81	0.16604188
Abstrak 24	0.20632786	Abstrak 82	0.3673765
Abstrak 25	0.2638586	Abstrak 83	0.18059812
Abstrak 26	0.33374706	Abstrak 84	0.40624055
Abstrak 27	0.245693	Abstrak 85	0.07344479
Abstrak 28	0.155023	Abstrak 86	0.337499
Abstrak 29	0.17255001	Abstrak 87	0.51006514
Abstrak 30	0.3630265	Abstrak 88	0.23633797
Abstrak 31	0.30094576	Abstrak 89	0.5133403
Abstrak 32	0.19714086	Abstrak 90	0.10799556
Abstrak 33	0.8538808	Abstrak 91	0.46947724

Dokumen Pembanding	Nilai Kemiripan	Dokumen pembanding	Nilai kemiripan
Abstrak 34	0.361169	Abstrak 92	0.3180679
Abstrak 35	0.3334012	Abstrak 93	0.23035413
Abstrak 36	0.653275	Abstrak 94	0.33945078
Abstrak 37	1	Abstrak 95	0.22371814
Abstrak 38	0.079147026	Abstrak 96	0.2018097
Abstrak 39	0.7952894	Abstrak 97	0.29611626
Abstrak 40	0.27249074	Abstrak 98	0.21718523
Abstrak 41	0.35934255	Abstrak 99	0.37035003
Abstrak 42	0.12408732	Abstrak 100	0.3431035
Abstrak 43	0.30626315	Abstrak 101	0.121295236
Abstrak 44	0.33298886	Abstrak 102	0.13309595
Abstrak 45	0.7617192	Abstrak 103	0.40465915
Abstrak 46	0.20407476	Abstrak 104	0.07607996
Abstrak 47	0.5086409	Abstrak 105	0.06869649
Abstrak 48	0.23851185	Abstrak 106	0.22570054
Abstrak 49	0.021405412	Abstrak 107	0.26656616
Abstrak 50	0.18489933	Abstrak 108	0.30313465
Abstrak 51	0.4590556	Abstrak 109	0.3915175
Abstrak 52	0.13239165	Abstrak 110	0.13512716
Abstrak 53	0.06245791	Abstrak 111	0.40531063
Abstrak 54	0.22661968	Abstrak 112	0.08269849
Abstrak 55	0.27087072	Abstrak 113	0.5593587
Abstrak 56	0.12848037	Abstrak 114	0.27770603
Abstrak 57	0.6732309	Abstrak 115	0.24340634
Abstrak 58	0.30504596	Abstrak 116	0.31784543

Berdasarkan tabel 4.4 diketahui bahwa nilai kemiripan dokumen uji abstrak 37 dengan dokumen pembanding yang sama memperoleh nilai kemiripan 1 dan mirip. Untuk nilai kemiripan 0 maka dokumen uji tidak mirip dengan dokumen pembanding.

4.4 Analisis

Penelitian ini dilakukan 2 percobaan yaitu menggunakan *Word2vec Cosine Similarity* dan TF-IDF. Percobaan menggunakan abstrak 37 sebagai dokumen uji (Q) dan abstrak 33 sebagai dokumen pembanding sekenario penelitian pada tabel 5.4 jika dilakukan *stemming* dan tidak dilakukan *stemming* pada masing-masing percobaan.

Tabel 4. 5 Perbandingan Nilai *Cosine Similarity* tanpa *Stemming* dan dengan *Stemming*

Metode vektor	<i>Cosine Similarity</i> (<i>Stemming</i>)	<i>Cosine Similarity</i> (tidak <i>Stemming</i>)
TF-IDF	0.56484648	0.51942993
<i>WORD2VEC</i>	0.8538808	0.82454777

Berdasarkan tabel 4.5 Tahap *preprocessing stemming* sangat berpengaruh terhadap hasil kemiripan dari dokumen uji. *stemming* merupakan proses untuk menghilangkan semua imbuhan dari kata, baik itu awalan, sisipan, maupun akhiran. Menggunakan *stemming* nilai yang didapatkan lebih besar karena kata memiliki imbuhan yang berbeda tetapi pada saat dilakukan proses *stemming* menjadi satu kata yang sama, sedangkan kata berbeda imbuhan dapat memiliki arti yang berbeda. Hasil kemiripan *Word2vec* lebih tinggi dibandingkan dengan TD-IDF karena beberapa kemungkinan salah satunya yaitu besarnya ukuran korpus data, apabila ukuran korpus semakin besar, maka semakin banyak kosa kata yang dimiliki korpus tersebut, sehingga akan semakin baik hasil nilai kemiripan yang dihasilkan oleh sistem. Nilai kemiripan yang dihasilkan oleh sistem dipengaruhi oleh kosa kata yang dimiliki oleh korpus serta kemungkinan pasangan kata yang sering muncul Bersama pada korpus tersebut. Penulis juga melakukan percobaan dengan software aplikasi plagiarism detector yang ada di perpustakaan hasil dari dokumen uji memiliki nilai kemiripan sebanyak 31%.

Plagiarism Detector v. 1092 - Originality Report:

Analyzed document: 1/25/2019 2:26:25 PM

"ABSTRAK37.pdf"

Licensed to: Heru Priyanto_License14

Relation chart:



Distribution graph:



Comparison Preset: Rewrite. Detected language: Indonesian

Top sources of plagiarism:

→	% 25	wrds: 57	E:\dataset\ABSTRAK33.pdf
→	% 22	wrds: 52	E:\dataset\ABSTRAK39.pdf
→	% 20	wrds: 44	E:\dataset\ABSTRAK67.pdf

[Show other Sources:]

Processed resources details:

Gambar 4. 1 Hasil Plagiarisme Detector

Pada gambar 4.1 menggunakan aplikasi plagiatarism detector penulis mencoba mengetahui nilai kemiripan dokumen uji yaitu abstrak 37 dengan dokumen pembanding sebanyak 116 abstrak dan sudah penulis pindah ke komputer perpustakaan. Hasil yang didapatkan yaitu nilai kemiripan dokumen 37 sebesar 31% berplagiat dengan dokumen 33 dan 39. Dokumen yang berplagiat. Sistem tersebut menampilkan dokumen yang memiliki kemiripan dan memiliki kesamaan dengan yang penulis buat yaitu abstrak 33 dan 39 tetapi nilai yang didapatkan jauh berbeda yaitu sebesar 0.254846. Beberapa faktor yang mempengaruhi perbedaan nilai similarity yaitu jika menggunakan TF-IDF kemiripan suatu kalimat dihitung dengan berapa jumlah kata atau *term* yang sering muncul. Sedangkan aplikasi menghitung berdasarkan paragraph. Menggunakan *Word2vec* lebih besar karena membandingkan nilai vektor pada setiap kata yang ada didalam kalimat.

Percobaan selanjutnya yaitu penulis mencoba merubah salah satu kata didalam abstrak 37 yaitu kata “delay” menjadi kata “delali”. Kemudian akan di deteksi kemiripan dengan abstrak 37. Hasil yang didapatkan yaitu 0.997