DETEKSI KEMIRIPAN DOKUMEN TEKS DENGAN METODE N-GRAM BERBASIS PANJANG KATA

Fatma Indriani[1], Irwan Budiman[2]
Program Studi Ilmu Komputer, Fakultas MIPA
Universitas Lambung Mangkurat[1,2]
f.indriani@unlam.ac.id[1], irwan.budiman@unlam.ac.id[2]

ABSTRACT

Various methods to detect plagiarism of text documents have been researched and developed. One type of method for detecting literal plagiarism is lexical-based, mostly character-based n-gram and word-based n-gram. N-gram based on word length has the advantage of smaller space and computing time, because it only encodes word length (number of letters per word). This paper discusses the application of the word length based n-gram word for representing documents and using the Dice coefficient to measure the similarity between the n-grams. The system is tested on a corpus of Indonesian language documents containing literal plagiarism, and succeeded in detecting all pairs of documents that contain plagiarism. From the test results, the optimal value of n is n> = 6 and the limit for Dice Coefficient is t <0.3.

***Keywords**: text similarity, plagiarism detection, word length n-gram, Dice-Coefficient, Indonesian language*

ABSTRAK

Ada berbagai metode deteksi plagiarisme dokumen teks yang sudah diteliti dan dikembangkan. Salah satu metode bagi penjiplakan yang tidak terlalu banyak melakukan perubahan dari teks asli (disebut plagiarisme literal), adalah metode deteksi berbasis leksikal, pada umumnya teknik n-gram berbasis karakter dan n-gram berbasis kata. Teknik n-gram berbasis karakter dan n-gram berbasis kata memakan waktu yang lama dengan bertambahnya jumlah dokumen, sehingga muncul berbagai variasi teknik n-gram, seperti n-gram berbasis stop word (kata-kata umum) dan n-gram berbasis panjang kata. Teknik n-gram berbasis panjang kata memiliki keunggulan yaitu membutuhkan ruang dan waktu komputasi yang lebih kecil, karena hanya meng-encode panjang kata (jumlah huruf per kata). Pada makalah ini dibahas penerapan metode *n-gram* berbasis panjang kata untuk merepresentasikan dokumen serta menggunakan Dice Coefficient untuk mengukur kemiripan antar n-gram. Sistem diujicobakan pada korpus dokumen berbahasa Indonesia yang mengandung plagiarisme literal, dan berhasil mendeteksi semua pasangan dokumen yang mengandung plagiarisme. Dari hasil pengujian, didapat bahwa nilai *n* yang optimal adalah n >= 6 dan batas Dice Coefficient *t*< 0.3.

Kata Kunci: kesamaan teks, deteksi plagiarisme, *n-gram* panjang kata, *Dice-Coefficient,* bahasa Indonesia

PENDAHULUAN

Salah satu kecurangan akademik yang dilakukan mahasiswa yaitu menyalin tugas hasil karya temannya untuk diakui sebagai karya sendiri. Atau kasus lain yang sejenis, yaitu dua atau lebih mahasiswa bekerja sama dalam menyelesaikan tugas perorangan, dan masing-masing mengumpulkan dokumen tugas yang sama atau hampir sama. Jika terjadi kecurangan, idealnya semua kasus terdeteksi dan pihak yang terlibat diberi sanksi.

Kemiripan karya ini bisa dideteksi secara manual oleh dosen pemeriksa tugas. Namun, semakin banyak dokumen yang diperiksa, akan semakin memperlambat pemeriksaan. Ini karena tingkat memori manusia yang terbatas yang akan sulit mengingat dan membandingkan dokumen yang banyak. Selain dari segi waktu, kelemahan manusia ini juga berpotensi berdampak pada tingkat keakurasian dan keobjektifan proses pemeriksaan.

Oleh karena itu, dibutuhkan suatu sistem yang bisa melakukan deteksi kemiripan dokumen teks dengan cepat. Sistem ini bisa mendeteksi pasangan atau kelompok dokumen yang memiliki banyak kesamaan, untuk dipisahkan dari dokumen-dokumen yang original. Kumpulan dokumen yang mirip ini berpotensi sebagai hasil penjiplakan, dan selanjutnya ditindaklanjuti oleh pengajar dengan pemeriksaan secara manual.

Menurut Ridhatillah (2003), plagiarisme adalah tindakan penyalahgunaan, pencurian/perampasan, penerbitan, pernyataan, atau menyatakan sebagai milik sendiri sebuah pikiran, ide, tulisan, atau ciptaan yang sebenarnya milik orang lain. Ada berbagai metode deteksi plagiarisme dokumen teks yang sudah diteliti dan dikembangkan.

Salah satu metode yang cocok untuk jenis penjiplakan yang tidak terlalu banyak melakukan perubahan dari teks asli (disebut plagiarisme literal), adalah metode deteksi berbasis leksikal (Alzahrani, dkk, 2012), pada umumnya teknik *n-gram* berbasis karakter dan *n-gram* berbasis kata. Teknik *n-gram* berbasis karakter untuk deteksi plagiarisme dokumen berbahasa Indonesia pernah diteliti oleh Lisangan (2013).

Teknik *n-gram* berbasis karakter dan *n-gram* berbasis kata memakan waktu yang lama dengan bertambahnya jumlah dokumen, sehingga muncul berbagai variasi teknik *n-gram*, di antaranya *n-gram* berbasis *stop word* (kata-kata umum) (Stamatatos, 2011) dan *n-gram* berbasis panjang kata (Barrón-Cedeño, 2010).

Teknik *n-gram* berbasis panjang kata memiliki keunggulan yaitu membutuhkan ruang dan waktu komputasi yang lebih kecil, karena hanya meng-*encode* panjang kata (jumlah huruf per kata). Kelebihan ini akan memudahkan pemeriksa tugas dalam menyortir mana dokumen yang dicurigai sebagai jiplakan, mana yang bukan jiplakan, secepat mungkin.

Makalah ini membahas penerapan metode n-gram panjang kata untuk mendeteksi kemiripan teks pada dokumen berbahasa Indonesia. Selain itu, dibahas kinerjanya berdasar nilai *precision* dan *recall*, serta berapa nilai n dan batas kesamaan (*threshold)* yang optimal untuk mendeteksi kemiripan tersebut.

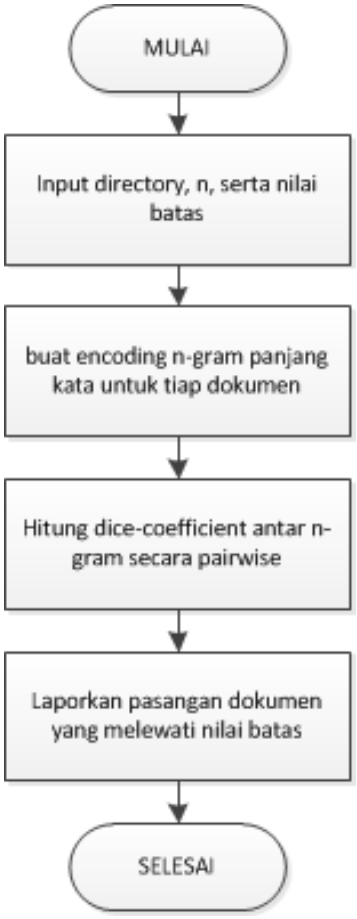
METODE PENELITIAN

Pada bagian ini dijelaskan tahap-tahap penelitian:

1. Pembangunan korpus dokumen

   Korpus dokumen berfungsi untuk mengetes kualitas sistem pendeteksi plagiarisme yang akan dibuat. Korpus dokumen uji dibangun dari kumpulan artikel tugas mahasiswa dan artikel di internet sebanyak 100 dokumen, terdiri dari dua tipe: Tipe (a), 25 pasang dokumen yang mengandung plagiarisme buatan, dan Tipe (b) yaitu 50 dokumen unik. Panjang tiap dokumen berkisar antara 250-750 kata. Tingkat plagiarisme pada dokumen Tipe (a) antara 20%-80% berupa plagiarisme literal, namun tanpa plagiarisme cerdas.

2. Perancangan dan implementasi sistem



Gambar 1. Proses deteksi dengan n-gram panjang kata

Proses utama pada sistem ada empat langkah (Gambar 1), yaitu:

a) Menerima input. Pada tahap ini sistem membaca directory dan mengambil daftar file dokumen pada directory tersebut

b) Membaca dan meng-*encode* tiap dokumen ke dalam bentuk *n*-gram panjang kata. Pada tahap ini, tiap file akan dibaca, diekstraksi isi teksnya, semua tanda baca dibuang, dan dikonversi menjadi *n*-gram panjang kata.

Berikut ini adalah tahapan encodinguntuk *n*-gram panjang kata (Barrón-Cedeño, 2010) beserta contoh (Tabel 1).

Tabel 1. Contoh pemrosesan *input*

Langkah	Hasil
***Input* teks asli**	Observatorium Bosscha akan dipindah ke Kupang, NTT
Pra-pemrosesan:penghapusan tanda baca, karakter non-huruf diubah menjadi karakter blank	Observatorium Bosscha akan dipindah ke Kupang NTT
***Encoding*:**pengubahan tiap kata w menjadi **minimum(|*w*|,9)**	9 7 4 8 2 6 3
Pembentukan n-gram	97, 74, 48, 82, 26, 63 (bigram) 974, 748, 482, 826, 263 (trigram) dst.

c) Menghitung kesamaan antar dokumen secara pairwise menggunakan Dice-Coefficient. Rumus nilai Dice-Coeffient dari himpunan A dan B adalah:

$$S(A,B) = \frac{2|A \cap B|}{|A| + |B|}$$

dimana:

$|A \cap B|$ adalah jumlah elemen dari irisan n-gram A dan B.
$|A|$ adalah jumlah elemen dari n-gram A
$|B|$ adalah jumlah elemen dari n-gram B

Pada tahap ini, tiap pasang n-gram (merepresentasikan dokumen) akan dihitung nilai Dice-Coefficientnya. Semakin tinggi Nilai Dice Coefficient, maka tingkat kemiripan semakin tinggi. Misal dua buah dokumen A dan B sebagai berikut:
A = ”Observatorium Bosscha akan dipindah ke Kupang, NTT”
B = “Observatorium Bosscha rencananya akan dipindah ke Kupang, NTT”
memiliki bigram panjang-kata masing-masing {97, 74, 48, 82, 26, 63} dan {97, 79, 94, 48, 82, 26, 63}. Maka,

$$S(A,B) = \frac{2|A \cap B|}{|A| + |B|} = \frac{10}{13} = 0{,}769$$

d) Pemrosesan output. Pada tahap ini, pasangan dokumen akan dipilah berdasar nilai Dice-Coefficientnya. Pasangan dokumen yang nilai Dice nya melebihi batas yang telah ditentukan di awal, akan dilaporkan kepada pengguna sebagai daftar pasangan dokumen yang mencurigakan. Daftar ini disampaikan dalam bentuk file teks.

3. Pengujian

Ada dua variabel yang akan diteliti pada tahap pengujian, yaitu jumlah token yang digunakan untuk membuat n-gram (*n*) serta batas minimum dari nilai Dice Coefficient (*t*). Untuk *n,* nilai yang akan diujikan adalah *n* = {1, 2, 3, ..., 9}. Sedangkan nilai t yang diujikan *t* = {0,1, 0,15, 0,2, ..., 0.9} Pengujian akan mengkombinasikan nilai *n* dan *t*. Hasil deteksi sisten dinilai dengan nilai *Precision,Recall,* dan *F1 Score*. *Precision* yaitu, dari daftar pasangan dokumen yang dicurigai (dilaporkan) oleh sistem, berapa persen yang benar-benar

mengandung plagiarisme. *Recall* yaitu dari 25 pasangan dokumen yang mengandung plagiarisme yang ada di korpus uji, berapa persen yang berhasil dilaporkan oleh sistem. Idealnya nilai *Precision* adalah 1, dan nilai *Recall* adalah 1. Nilai *Precision* dan *Recall* pada umumnya bertolak belakang, sehingga digunakan pula nilai *F1 Score* , yang merupakan gabungan dari *Precision* dan *Recall*.

$$\text{Pr}\,ecision = \frac{TruePositive}{TruePositive + FalsePositive}$$

$$\text{Re}\,call = \frac{TruePositive}{TruePositive + FalseNegative}$$

$$F1_{Score} = 2\frac{\text{Pr}ecision \cdot \text{Re}call}{\text{Pr}ecision + \text{Re}call}$$

HASIL DAN PEMBAHASAN

Berikut ini adalah hasil perhitungan Precision, Recall, dan F1 Score untuk pengujian berbagai nilai *n* dan *t* (
Tabel 2, Tabel ,

Tabel **4**).

Tabel 2 Nilai Precision untuk pengujian berbagai nilai n dan t

t	***n***							
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0.1	0.005	0.005	0.005	0.510	1	1	1	1
0.15	0.005	0.005	0.006	1	1	1	1	1
0.2	0.005	0.005	0.008	1	1	1	1	1
0.25	0.005	0.005	0.018	1	1	1	1	1
0.3	0.005	0.005	0.061	1	1	1	1	1
0.35	0.005	0.005	0.253	1	1	1	1	1
0.4	0.005	0.005	0.719	1	1	1	1	1
0.45	0.005	0.005	0.955	1	1	1	1	1
0.5	0.005	0.006	1	1	1	1	1	1
0.55	0.005	0.007	1	1	1	1	1	1
0.6	0.005	0.009	1	1	1	1	1	1
0.65	0.005	0.014	1	1	1	1	1	1
0.7	0.005	0.030	1	1	1	1	1	1
0.75	0.005	0.106	1	1	1	1	1	1
0.8	0.006	0.65	1	1	1	1	1	1
0.85	0.007	1	1	1	1	1	1	1
0.9	0.015	1	1	-	-	-	-	-

ISSN 2541-5662 (Cetak)
ISSN 2541-5670 (Online)

Tabel 3 Nilai Recall untuk pengujian berbagai nilai n dan t

t	*n*							
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0.1	1	1	1	1	1	1	1	1
0.15	1	1	1	1	1	1	1	1
0.2	1	1	1	1	1	1	1	1
0.25	1	1	1	1	1	1	1	1
0.3	1	1	1	0.96	0.84	0.84	0.84	0.84
0.35	1	1	1	0.8	0.8	0.8	0.8	0.8
0.4	1	1	0.92	0.76	0.72	0.68	0.68	0.68
0.45	1	1	0.84	0.68	0.6	0.6	0.56	0.56
0.5	1	1	0.76	0.52	0.4	0.4	0.4	0.4
0.55	1	1	0.6	0.4	0.36	0.36	0.36	0.36
0.6	1	0.96	0.44	0.32	0.28	0.28	0.28	0.28
0.65	1	0.92	0.4	0.28	0.28	0.28	0.28	0.28
0.7	1	0.88	0.32	0.16	0.08	0.08	0.08	0.08
0.75	1	0.68	0.16	0.08	0.08	0.08	0.08	0.08
0.8	1	0.52	0.08	0.08	0.08	0.08	0.08	0.08
0.85	0.96	0.32	0.08	0.08	0.08	0.08	0.08	0.08
0.9	0.92	0.12	0.04	0	0	0	0	0

Tabel 4 Nilai F1 Score untuk pengujian berbagai nilai n dan t

t	*n*							
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0.1	0.01	0.01	0.01	0.68	1	1	1	1
0.15	0.01	0.01	0.01	1	1	1	1	1
0.2	0.01	0.01	0.02	1	1	1	1	1
0.25	0.01	0.01	0.04	1	1	1	1	1
0.3	0.01	0.01	0.11	0.98	0.91	0.91	0.91	0.91
0.35	0.01	0.01	0.40	0.89	0.89	0.89	0.89	0.89
0.4	0.01	0.01	0.81	0.86	0.84	0.81	0.81	0.81
0.45	0.01	0.01	0.89	0.81	0.75	0.75	0.72	0.72
0.5	0.01	0.01	0.86	0.68	0.57	0.57	0.57	0.57
0.55	0.01	0.01	0.75	0.57	0.53	0.53	0.53	0.53
0.6	0.01	0.02	0.61	0.48	0.44	0.44	0.44	0.44
0.65	0.01	0.03	0.57	0.44	0.44	0.44	0.44	0.44
0.7	0.01	0.06	0.48	0.28	0.15	0.15	0.15	0.15
0.75	0.01	0.18	0.28	0.15	0.15	0.15	0.15	0.15
0.8	0.01	0.58	0.15	0.15	0.15	0.15	0.15	0.15

0.85	0.01	0.48	0.15	0.15	0.15	0.15	0.15	0.15
0.9	0.03	0.21	0.08	-	-	-	-	-

Dari pengujian, ternyata untuk *n* sangat rendah (*n*=2 dan *n*=3) nilai Dice Coefficient (similaritas) senantiasa tinggi, baik diantara pasangan dokumen mengandung plagiasi, maupun dokumen yang unik (tidak ada plagiasi). Dengan demikian tidak bisa dibedakan antara dokumen plagiasi dan non-plagiasi. Ini karena kombinasi panjang frase 2-kata kurang bervariasi. Untuk kasus n=2, hanya ada 9x9 = 81 bigram yang mungkin, sehingga cukup tinggi kemungkinan suatu frasa 2-kata memiliki panjang-2-kata yang sama dengan 2-kata yang lain dan tidak berhubungan (Tabel).

Tabel 5 Jumlah n-gram unik untuk beberapa nilai n

n	**Jumlah n-gram unik**
2	81
3	729
4	6561
5	59049
6	531441
7	4782969
8	43046721
9	387420489

Namun dengan semakin tinggi nilai *n* (*n*=4 dan n=5), perhitungan Dice Coefficient semakin mampu untuk membedakan pasangan dokumen yang mirip dengan pasangan dokumen yang saling unik (tidak mirip). Sistem semakin bisa mengenali kesamaan teks pada dokumen karena dokumen yang mirip memiliki nilai Dice tinggi, sedangkan dokumen unik memiliki nilai Dice rendah. Nilai *n* optimal yang optimal adalah mulal *n*=6 ke atas.

Dari
Tabel , nilai *precision* yang sempurna didapat pada *n*>=5. Precision bagus juga didapat pada *n*=3 dan *n*=4 namun dengan threshold yang tinggi. Dari Tabel , nilai recall yang sempurna didapat pada *threshold* yang rendah (*t*< 0,3). Hal ini karena korpus uji mengandung pasangan dokumen yang mengandung plagiarisme level rendah (20%), sehingga tidak terdeteksi jika *threshold* tinggi.

Dari

Tabel **4** yang menunjukkan nilai F1 Score, maka nilai *n* dan *t* yang memaksimalkan precision dan recall yaitu *n* >= 6 dan nilai *t* < 0,3.

KESIMPULAN

Dari penelitian ini maka didapat bahwa:

1. Metode n-gram berbasis panjang kata dapat digunakan untuk mendeteksi kesamaan teks berbahasa Indonesia
2. Hasil pengujian kinerja menunjukkan bahwa didapatkan nilai *Recall*, *Precision*, dan *F1 Score* sebesar 100%, tergantung nilai *n* dan batas Dice Coefficient (*t*) yang digunakan.

3. Nilai *n* dan *t* yang memaksimalkan *precision* dan *recall* yaitu *n* >= 6 dan nilai *t* < 0,3.

Sistem yang telah dibuat ini berfokus untuk mendeteksi kesamaan teks, sehingga cocok untuk mendeteksi plagiarisme literal/*copy-paste*. Diperlukan penelitian lebih lanjut untuk plagiarisme yang lebih canggih, khususnya plagiarisme dengan substitusi kata/frase, dan plagiarisme dengan menukar klausa kalimat. Selain itu, baru diteliti deteksi kesamaan teks pada koleksi dokumen yang terbatas (kumpulan tugas mahasiswa untuk mengecek potensi *copy paste* antar mahasiswa). Diperlukan penelitian lanjut metode ini untuk mendeteksi plagiarisme antara suatu dokumen dengan koleksi dokumen yang sangat besar, misal memeriksa suatu tugas atau makalah dibandingkan dengan koleksi dokumen/website di internet.

DAFTAR PUSTAKA

Alzahrani, S. M., Salim, N., & Abraham, A. (2012). “Understanding plagiarism linguistic patterns, textual features, and detection methods”. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, *42*(2), 133–149. doi:10.1109/TSMCC.2011.2134847

Barrón-Cedeño, A., Basile, C., Esposti, M. D., & Rosso, P. (2010). ”Word length n-grams for text re-use detection”. *Lecture Notes in Computer Science* (Vol. 6008 LNCS, pp. 687–699). doi:10.1007/978-3-642-12116-6_58

Dice, Lee R. (1945). "Measures of the Amount of Ecologic Association Between Species". *Ecology 26* (3): 297–302. doi:10.2307/1932409

Lisangan, E. A., (2013). ”Implementasi n-Gram technique dalam deteksi plagiarism pada tugas mahasiswa”. *TEMATIKA, Journal of Informatics and Information Systems*. Vol 1, No 2

N-gram. (n.d.). Di *Wikipedia*. Diakses 12 Maret 2016, dari http://en.wikipedia.org/wiki/N-gram

Ridhatillah, Ardini.(2003).”Dealing with Plagiarism in the Information System Research Community: A Look at Factors That Drive Plagiarism and Ways to Address Them”.*MIS Quarterly*; Vol. 27, No. 4, p. 511-532/December 2003.

Stamatatos, E. (2011). ”Plagiarism detection using stopword n-grams”. *Journal of the American Society for Information Science and Technology*, 62(12), 2512–2527. doi:10.1002/asi.21630