UNIVERSITAS
MERCU BUANA

PENERAPAN ALGORITMA *COSINE SIMILARITY* UNTUK PENGKLASIFIKASIAN OTOMATIS DOKUMEN KEPEGAWAIAN DENGAN TEKNIK *OCR* DI KEMENTERIAN DALAM NEGERI

TUGAS AKHIR

JONATHAN EKA RATMOKO
41519310035

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS MERCU BUANA
JAKARTA
2021

PENERAPAN ALGORITMA *COSINE SIMILARITY* UNTUK PENGKLASIFIKASIAN OTOMATIS DOKUMEN KEPEGAWAIAN DENGAN TEKNIK *OCR* DI KEMENTERIAN DALAM NEGERI

Tugas Akhir

Diajukan Untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer

Oleh:
JONATHAN EKA RATMOKO
41519310035

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS MERCU BUANA
JAKARTA
2021

LEMBAR PERNYATAAN ORISINALITAS

Yang bertanda tangan dibawah ini:

NIM	:	41519310035
Nama	:	Jonathan Eka Ratmoko
Judul Tugas Akhir	:	Penerapan Algoritma *Cosine Similarity* Untuk Pengklasifikasian Otomatis Dokumen Kepegawaian Dengan Teknik *OCR* Di Kementerian Dalam Negeri

Menyatakan bahwa Laporan Tugas Akhir saya adalah hasil karya sendiri dan bukan plagiat. Apabila ternyata ditemukan didalam laporan Tugas Akhir saya terdapat unsur plagiat, maka saya siap untuk mendapatkan sanksi akademik yang terkait dengan hal tersebut.

Jakarta, 14 Juli 2021

Jonathan Eka Ratmoko

SURAT PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR

Sebagai mahasiswa Universitas Mercu Buana, saya yang bertanda tangan di bawah ini :

Nama Mahasiswa	:	Jonathan Eka Ratmoko
NIM	:	41519310035
Judul Tugas Akhir	:	Penerapan Algoritma *Cosine Similarity* Untuk Pengklasifikasian Otomatis Dokumen Kepegawaian Dengan Teknik *OCR* Di Kementerian Dalam Negeri

Dengan ini memberikan izin dan menyetujui untuk memberikan kepada Universitas Mercu Buana **Hak Bebas Royalti Noneksklusif** (*None-exclusive Royalty Free Right*) atas karya ilmiah saya yang berjudul diatas beserta perangkat yang ada (jika diperlukan).

Dengan Hak Bebas Royalti/Noneksklusif ini Universitas Mercu Buana berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat dan mempublikasikan tugas akhir saya.

Selain itu, demi pengembangan ilmu pengetahuan di lingkungan Universitas Mercu Buana, saya memberikan izin kepada Peneliti di Lab Riset Fakultas Ilmu Komputer, Universitas Mercu Buana untuk menggunakan dan mengembangkan hasil riset yang ada dalam tugas akhir untuk kepentingan riset dan publikasi selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Jakarta, 14 Juli 2021

Jonathan Eka Ratmoko

SURAT PERNYATAAN LUARAN TUGAS AKHIR

Sebagai mahasiswa Universitas Mercu Buana, saya yang bertanda tangan di bawah ini :

Nama Mahasiswa	:	Jonathan Eka Ratmoko
NIM	:	41519310035
Judul Tugas Akhir	:	Penerapan Algoritma *Cosine Similarity* Untuk Pengklasifikasian Otomatis Dokumen Kepegawaian Dengan Teknik *OCR* Di Kementerian Dalam Negeri

Menyatakan bahwa :

1. Luaran Tugas Akhir saya adalah sebagai berikut :

No	Luaran	Jenis		Status	
1	Publikasi Ilmiah	Jurnal Nasional Tidak Terakreditasi		Diajukan	V
		Jurnal Nasional Terakreditasi	V		
		Jurnal International Tidak Bereputasi		Diterima	
		Jurnal International Bereputasi			
	Disubmit/dipublikasikan di :	Nama Jurnal	: Jurnal Teknologi Informasi dan Ilmu Komputer		
		ISSN	: 2355-7699		
		Link Jurnal	: https://jtiik.ub.ac.id/index.php/jtiik		
		Link File Jurnal Jika Sudah di Publish	:		

2. Bersedia untuk menyelesaikan seluruh proses publikasi artikel mulai dari submit, revisi artikel sampai dengan dinyatakan dapat diterbitkan pada jurnal yang dituju.
3. Diminta untuk melampirkan scan KTP dan Surat Pernyataan (Lihat Lampiran Dokumen HKI), untuk kepentingan pendaftaran HKI apabila diperlukan

Demikian pernyataan ini saya buat dengan sebenarnya.

Mengetahui
Dosen Pembimbing TA

Sri Dianing Asri, ST, M. Kom

Jakarta, 14 Juli 2021

Jonathan Eka Ratmoko

LEMBAR PERSETUJUAN

Nama Mahasiswa	:	Jonathan Eka Ratmoko
NIM	:	41519310035
Judul Tugas Akhir	:	Penerapan Algoritma *Cosine Similarity* Untuk Pengklasifikasian Otomatis Dokumen Kepegawaian Dengan Teknik *OCR* Di Kementerian Dalam Negeri

Tugas Akhir ini telah diperiksa dan disetujui

Jakarta, 13 Juli 2021

Menyetujui,

(Sri Dianing Asri, ST, M. Kom)
Dosen Pembimbing

LEMBAR PERSETUJUAN PENGUJI

NIM	:	41519310035
Nama	:	Jonathan Eka Ratmoko
Judul Tugas Akhir	:	Penerapan Algoritma *Cosine Similarity* Untuk Pengklasifikasian Otomatis Dokumen Kepegawaian Dengan Teknik *OCR* Di Kementerian Dalam Negeri

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 13 Agustus 2021

(Wawan Gunawan, S. Kom., MT)

LEMBAR PERSETUJUAN PENGUJI

NIM	:	41519310035
Nama	:	Jonathan Eka Ratmoko
Judul Tugas Akhir	:	Penerapan Algoritma *Cosine Similarity* Untuk Pengklasifikasian Otomatis Dokumen Kepegawaian Dengan Teknik *OCR* Di Kementerian Dalam Negeri

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 13 Agustus 2021

(Sukma Wardhana, S.Kom,M.Kom)

LEMBAR PERSETUJUAN PENGUJI

NIM	:	41519310035
Nama	:	Jonathan Eka Ratmoko
Judul Tugas Akhir	:	Penerapan Algoritma *Cosine Similarity* Untuk Pengklasifikasian Otomatis Dokumen Kepegawaian Dengan Teknik *OCR* Di Kementerian Dalam Negeri

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 13 Agustus 2021

(Giri Purnama, S.Pd., M.Kom)

LEMBAR PENGESAHAN

NIM	:	41517310035
Nama	:	Jonathan Eka Ratmoko
Judul Tugas Akhir	:	Penerapan Algoritma *Cosine Similarity* Untuk Pengklasifikasian Otomatis Dokumen Kepegawaian Dengan Teknik *OCR* Di Kementerian Dalam Negeri

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 13 Agustus 2021

Menyetujui,

(Sri Dianing Asri, ST, M.Kom)
Dosen Pembimbing

Mengetahui,

(Wawan Gunawan, S.Kom., MT)	(Herry Derajad Wijaya, S.Kom., MM)
Koord. Tugas Akhir Teknik Informatika	Ka. Prodi Teknik Informatika

ABSTRAK

Nama	:	Jonathan Eka Ratmoko
NIM	:	41519310035
Pembimbing TA	:	Sri Dianing Asri, ST, M. Kom
Judul	:	Penerapan Algoritma *Cosine Similarity* Untuk Pengklasifikasian Otomatis Dokumen Kepegawaian Dengan Teknik *OCR* Di Kementerian Dalam Negeri

Penanganan yang baik dalam bidang kearsipan tentu akan mendukung jalannya kegiatan administrasi pada setiap organisasi. Permasalahan di Kementerian Dalam Negeri adalah pengarsipan dibidang kepegawaian masih secara manual dan belum tertata berdasarkan jenis kategori dokumennya, dengan sistem yang masih manual dapat memperlambat kinerja PNS dalam pengarsipan dokumen-dokumen kepegawaian. *Text mining* merupakan salah satu cara penanganan masalah tersebut dengan memanfaatkan teknologi *OCR* untuk mendapatkan kata-kata yang ada di dalam gambar, algoritma *text mining* yang tepat untuk pengklasifikasian dokumen adalah *cosine similarity.* Pengimplementasian sistem tersebut menggunakan *python*, *library tesseract* berfungsi untuk penerapan *OCR,* sedangkan proses *cosine similarity* dilakukan oleh *library sklearn*. Hasil dari *OCR* menunjukan terdapat karakter-karakter yang tidak terdapat di dokumen sk_cpns.jpg juga ikut terkestraksi, sedangkan hasil yang didapatkan oleh algoritma *cosine similarity* bahwa dokumen tersebut memiliki nilai kemiripan 0.22209373 dengan persentase 22.20%, yang berarti bahwa dokumen sk_cpns.jpg merupakan kategori dari Surat Keputusan Calon Pegawai Negeri Sipil.

Kata kunci:
dokumen, klasifikasi, *OCR*, *Cosine Similarity*

ABSTRACT

Name	:	Jonathan Eka Ratmoko
Student Number	:	41519310035
Counsellor	:	Sri Dianing Asri, ST, M. Kom
Title	:	Application Of Cosine Similarity Algorithm For Automatic Classification Of Employee Documents Using Ocr Techniques In Kementerian Dalam Negeri

Good handling in the field of archives will certainly support the course of administrative activities in each organization. The problem at the Ministry of Home Affairs is that the civil servant archiving is still manual and has not been organized based on the type of document category, with a manual system that can slow down the performance of civil servants in archiving personnel documents. Text Mining is one way to handle this problem by utilizing OCR technology to get the words in the image, the right text mining algorithm for classifying documents is Cosine Similarity. The implementation of the system uses python, the tesseract library functions for the implementation of OCR, while the cosine similarity process is carried out by the sklearn library. The results of the OCR show that characters that are not in the sk_cpns.jpg document are also extracted. The results obtained by the Cosine Similarity algorithm are that the document has a similarity value of 0.22209373 with a percentage of 22.20%, which means that the sk_cpns.jpg document is a category of the Prospective Civil Servant Decree.

Key words:
document, classification, *OCR, Cosine Similarity*

KATA PENGANTAR

Puji syukur kita panjatkan atas kehadirat Allah SWT atas segala rahmat dan nikmat-Nya yang telah diberikan, sehingga penulis dapat menyelesaikan Tugas Akhir sebagai persyaratan guna memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Penulis menyadari bahwa tanpa bantuan dan bimbingan dari banyak pihak. Oleh karena itu, penulis mengucapkan terima kasih kepada:..

1. Bapak Herry Derajat W., S.Kom, MM selaku kepala program studi Teknik Informatika.
2. Ibu Sri Dianing Asri, ST, M.Kom. selaku Dosen Pembimbing.
3. Staff Kepegawaian Kementerian Dalam Negeri yang telah membantu dan mengarahkan dalam proses penilitian.
4. Teristimewa kepada keluarga besar penulis, Ayahanda tercinta Djoko Permono, Ibunda tersayang Grace Ratna Liesdyani, semua Saudara dan Saudari penulis yang telah memberikan kasih sayang dan juga dukungan secara moril maupun materil serta doa yang tulus kepada penulis.
5. Nurul Adhita Kusumawardani yang telah membantu dan memberi dukungan kepada penulis dalam penyelesaian Tugas Akhir ini.
6. Semua rekan dan teman yang tidak dapat penulis sebutkan satu persatu yang mana banyak memberi doa dan dukungan kepada penulis.

Akhir kata, penulis berharap Tugas Akhir ini dapat bermanfaat bagi rekan – rekan mahasiswa dan juga para pembaca sekalian untuk menjadi pembelajaran ataupun refrensi. Semoga Allah SWT selalu memberikan rahmat dan nikmat-Nya kepada kita semua.

Jakarta, 15 Juli 2021
Penulis

DAFTAR ISI

Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik…

Vol. x, No. x, April 2014, hlm. x-x p-ISSN: 2355-7699

Akreditasi KEMENRISTEKDIKTI, No. 30/E/KPT/2018 e-ISSN: 2528-6579

NASKAH JURNAL

PENERAPAN ALGORITMA *COSINE SIMILARITY* UNTUK PENGKLASIFIKASIAN OTOMATIS DOKUMEN KEPEGAWAIAN DENGAN TEKNIK *OCR* DI KEMENTERIAN DALAM NEGERI

Jonathan Eka R[*1], Sri Dianing Asri, ST, M.Kom[*2]

[1]Jonathan Eka Ratmoko
[2]Sri Dianing Asri, ST, M.Kom
Email: [1] 41519310035@mercubuana.ac.id, [2]dianing.asri@gmail.com
[*]Penulis Korespondensi

(Naskah masuk: 12 Agustus 2021, diterima untuk diterbitkan: dd mmm yyyy)

Abstrak

Penanganan yang baik dalam bidang kearsipan tentu akan mendukung jalannya kegiatan administrasi pada setiap organisasi. Permasalahan di Kementerian Dalam Negeri adalah pengarsipan dibidang kepegawaian masih secara manual dan belum tertata berdasarkan jenis kategori dokumennya, dengan sistem yang masih manual dapat memperlambat kinerja PNS dalam pengarsipan dokumen-dokumen kepegawaian. *Text mining* merupakan salah satu cara penanganan masalah tersebut dengan memanfaatkan teknologi *OCR* untuk mendapatkan kata-kata yang ada di dalam gambar, algoritma *text mining* yang tepat untuk pengklasifikasian dokumen adalah *cosine similarity*. Pengimplementasian sistem tersebut menggunakan *python*, *library tesseract* berfungsi untuk penerapan *OCR,* sedangkan proses *cosine similarity* dilakukan oleh *library sklearn*. Hasil dari *OCR* menunjukan terdapat karakter-karakter yang tidak terdapat di dokumen sk_cpns.jpg juga ikut terkestraksi, sedangkan hasil yang didapatkan oleh algoritma *cosine similarity* bahwa dokumen tersebut memiliki nilai kemiripan 0.22209373 dengan persentase 22.20%, yang berarti bahwa dokumen sk_cpns.jpg merupakan kategori dari Surat Keputusan Calon Pegawai Negeri Sipil.
Kata kunci: *dokumen, klasifikasi, OCR, Cosine Similarity*

APPLICATION OF COSINE SIMILARITY ALGORITHM FOR AUTOMATIC CLASSIFICATION OF EMPLOYEE DOCUMENTS USING OCR TECHNIQUES IN KEMENTERIAN DALAM NEGERI

Abstract

Good handling in the field of archives will certainly support the course of administrative activities in each organization. The problem at the Ministry of Home Affairs is that the civil servant archiving is still manual and has not been organized based on the type of document category, with a manual system that can slow down the performance of civil servants in archiving personnel documents. Text Mining is one way to handle this problem by utilizing OCR technology to get the words in the image, the right text mining algorithm for classifying documents is Cosine Similarity. The implementation of the system uses python, the tesseract library functions for the implementation of OCR, while the cosine similarity process is carried out by the sklearn library. The results of the OCR show that characters that are not in the sk_cpns.jpg document are also extracted. The results obtained by the Cosine Similarity algorithm are that the document has a similarity value of 0.22209373 with a percentage of 22.20%, which means that the sk_cpns.jpg document is a category of the Prospective Civil Servant Decree.
Keywords: *document, classification, OCR, Cosine Similarity*

1. PENDAHULUAN

Perkembangan teknologi memiliki dampak yang sangat signifikan dalam kehidupan sehari-hari, mulai dari kegiatan yang sederhana hingga kegiatan yang membutuhkan tingkat ketelitian yang tinggi. Kegiatan yang umum dilakukan oleh sebuah instansi adalah kegiatan pengarsipan dokumen, baik dokumen dalam bentuk fisik maupun elektronik. Umumnya kegiatan pengarsipan melibatkan dokumen dengan jumlah yang cukup besar, sehingga diperlukan suatu metode yang praktis dan efisien dalam pengelolaanya.

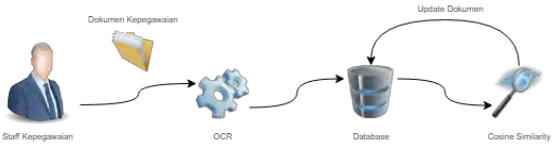
Pengklasifikasian dokumen elektronik dengan jumlah yang banyak diperlukan agar data yang terkumpul dapat diproses menjadi informasi yang tepat. Pengklasifikasian dokumen dilakukan dalam upaya memisahkan atau mengelompokkan dokumen berdasarkan ciri-ciri atau kategori tertentu. Dengan banyaknya dokumen, proses pengklasifikasian tidak mungkin dilakukan secara manual karena memerlukan banyak waktu dan tenaga. Salah satu metode yang dapat digunakan adalah dengan pengklasifikasian secara otomatis dengan text mining. Banyak metode text mining yang digunakan dalam mengklasifikasikan dokumen atau teks, salah satunya adalah algoritma Consine Similarity.

Staff kepegawaian di lingkungan Kementerian Dalam Negeri memerlukan waktu yang lama untuk pengarsipan dokumen kepegawaian ke dalam aplikasi, dimana para staff harus scan dokumen terlebih dahulu, kemudian diunggah ke aplikasi dan diklasifikasi secara manual di aplikasi. *Python* merupakan salah satu media yang dapat mengimplementasikan suatu penelitian ke dalam sistem, dengan banyaknya *library* yang tersedia di *python* memudahkan proses tersebut. *Library tesseract* untuk menjalankan *OCR* sedangkan untuk mengimplementasikan algoritma *cosine similarity* menggunakan *library sklearn.*

Maka dari itu, penulis mengambil penelitian tugas akhir bagaimana cara kerja algoritma Consine Similarity untuk pengklasifikasian otomatis dokumen kepegawaian dengan bantuan teknik OCR di lingkungan Kementerian Dalam Negeri menggunakan *library tesseract* dan *library sklearn* pada bahasa pemograman *python.*

2. METODE PENELITIAN

Pada penelitian ini, diusulkan model sistem seperti Gambar 1. Tahapan yang digunakan meliputi persiapan dokumen kepegawaian, mengekstrak gambar ke dalan bentuk text dan mengklasifikasikan dokumen kepegawaian berdasarkan jenis dokumen.



Gambar 1. Alur kerja *OCR* dan *Cosine Similarity*

Dokumen kepegawaian yang digunakan merupakan dokumen pribadi atau tidak dapat dipublikasikan, dokumen yang diberikan hanya berjumlah 7 data berformatkan .jpg, .jpeg dan png. Setiap dokumen memiliki kategorinya masing-masing, akan tetapi dokumen tersebut belum dikategorikan oleh staff kepegawaian. Maka dari itu, proses ekstraksi gambar menggunakan *OCR* dan proses pengklasifikasian pada penelitian ini menggunakan algoritma *Cosine Similarity.*

Proses ekstraksi gambar menggunakan *OCR* pada penelitian ini, sebagai berikut:

a. *File input*
   File input berupa dokumen kepegawaian dengan format *.png atau *.jpg.
b. *Preprocessing*
   Preprocessing merupakan suatu proses untuk menghilangkan bagian-bagian yang tidak diperlukan pada gambar *input* untuk proses selanjutnya.
c. Segmentasi
   Segmentasi adalah proses memisahkan area pengamatan (*region*) pada setiap karakter yang dideteksi.
d. Normalisasi
   Normalisasi merupakan proses untuk merubah dimensi *region* setiap karakter dan ketebalan karakter.
e. Ekstraksi ciri
   Ekstraksi ciri adalah proses untuk mengambil ciri-ciri tertentu dari karakter yang diamati.

f. *Recognition*
Recognition merupakan proses untuk mengenali karakter yang diamati dengan cara membandingkan ciri-ciri karakter yang diperoleh dengan ciri-ciri karakter yang ada pada basis data.

Dan cara kerja algoritma *Cosine Similarity* dalam mengklasifikasikan dokumen, sebagai berikut:

a. *Case folding*
Tahap ini merupakan tahap merubah huruf dari kapital menjadi huruf kecil.
b. *Tokenizing*
Tokenizing adalah proses memecah dokumen menjadi kumpulan kata. *Tokenization* dapat dilakukan dengan menghilangkan tanda baca dan memisahkannya per-spasi. Tahapan ini juga menghilangkan karakter-karakter tertentu seperti tanda baca dan mengubah semua *token* ke bentuk huruf kecil (*lower case*).
c. *Stopwords/Filtering*
Stopwords removal merupakan proses penghilangan kata tidak penting pada deskripsi melalui pengecekan kata-kata hasil *parsing* deskripsi apakah termasuk di dalam daftar kata tidak penting (*stoplist*) atau tidak. Jika termasuk di dalam *stoplist* maka kata-kata tersebut akan di-*remove* dari deskripsi sehingga kata-kata yang tersisa di dalam deskripsi dianggap sebagai kata-kata penting atau keywords.

Metode penelitian yang diterapkan pada penelitian ini adalah dengan pengembangan metode waterfall. Metode waterfall merupakan model pengembangan sistem informasi yang sistematik dan sekuensial. Metode Waterfall pada penelitian memiliki tahapan-tahapan sebagai berikut:

a. *Requirements analysis and definition*
Pada tahap ini penulis melakukan penelitian selama 3 bulan untuk mendapatkan informasi tentang dokumen kepegawaian dengan cara diskusi *online.*
b. *System and software design*
Informasi yang telah diperoleh kemudian dirancang menjadi suatu sistem yang dapat diimplementasikan ke aplikasi induk agar dapat mempermudah *staff* kepegawaian dalam mengklasifikasikan dokumen kepegawaian.
c. *Implementation and unit testing*
Di tahap ini, penulis mulai menerapkan *OCR* dan *Cosine Similarity* ke dalam bahasa pemograman untuk klasifikasi dokumen kepegawaian berdasarkan jenis dokumennya.
d. *Integration and system testing*
Setelah sistem sudah diimplementasikan kemudian dilakukan testing dan aplikasi yang penulis sudah bangun akan diberikan ke *staff IT* kepegawaian untuk diintegrasikan dengan aplikasi induk.
e. *Operation and maintenance*
Tahap terakhir ini dilakukan oleh *staff IT* kepegawaian untuk *maintenance* sistem setiap bulannya.

3. HASIL DAN PEMBAHASAN

3.1. Proses Implementasi Teknik *OCR*

Teknik ini digunakan untuk mendapatkan kata pada gambar, gambar yang digunakan dalam penelitian ini adalah dokumen kepegawaian yang telah di-*scan* menggunakan *scanner.* Gambar yang telah di-*scan* harus berformat .jpg, jpeg dan .png agar *OCR* dapat membacanya.

Dalam proses ekstraksi dokumen, data pengujian yang digunakan sebanyak 10 dokumen. Salah satu contoh hasil dari pengujian *OCR* terhadap dokumen kepegawaian pada Tabel 1.

Tabel 1. Hasil ekstraksi *OCR* sk_cpns.jpg

Nama Dokumen	Query
sk_cpns.jpg	DEPARTEMEN DALAM NEGERI REPUBLIK INDONESIA KEPUTUSAN MENTERI DALAM NEGERI NOMOR : 811.133.83 MENTERI DALAM NEGERI Menimbang > bahwa datam rangka pengisian formasi yang lowong di lingkungan Departemen Dalam Negeri untuk Tahun Anggaran 2007 dipandang perlu mengangkat yang namanya tersebut di bawah ini menjadi Calon Pegawai Negeri Sipil dalam masa percobaan. Mengingat > 1. Undang-undang Nomer 8 Tahun 1974 jo. Undang-undang Nomor 43 Tahun 1999; 2. Peraturan Pemerintah Nomor 7 Tahun 1977 jo. PP No. 66 Tahun 2005; 3. Peraturan Pemerintah Nomor 54 Seka 2003 tentang Perubahan Atas PP No. 97 Tahun 2000; 4, Peraturan Pemerintah Nomor

	i1 Tahun 2002 Teritang Perubahan Atas PP Nomor 98 Tahun 2000; 5. Peraturan Pemerintah Nomor @ Tahun 2003 Tentang Perubahan Atas PP Nomor 96 _ Tahun 2000; 6. Keputusan Kepala Badan Kepegawaian Negara Nomor 11 Tahun 2002 tanggal 17 Juni 2002. Memperhatikan : Penetapan Sdr/i. , _oleh Kepala Badan Kepegawaian Negara pada tanggal 28 Desember 2007 MEMUTUSKAN Memperhatikan : PERTAMA > Perhitungan mulai tanggal 1 Januari 2008 mengangkat sebagai Calon Pegawai Negeri Sipil. Nama : N,I.P. : ~- Tempat/tanggal lahir : Jakarta, 14-03-1973 - Jenis kelamin : LAKI-LAKT - Pendidikan : S1. Manajemen Informatika Thn. 1999 .. Golongan ruarg : Ila Masa kerja golongan : Otahun 0 hulan - Gaji pokok : Rp. 960,480,- - Jabatan : Pranata Computer - Unit kerja : Direktorat Jenderal Otonomi Daerah ~ Instansi induk : Departemen Dalam Negeri. - KEDUA : Diatas gaji pokok tersebut, kepada yang bersangkutan diberikan penghasilan lain yang sah sesuai dengan peraturan perundang-undangan yang berlaku. - KETIGA : Apabila dikemudian hari ternyata terdapat kekeliruan dalam keputusan ini akan diadakan perbaikan dan perhitungan kembali sebagaimana mestinya. ASLI —_Keputusan ini diberikan kepada yang bersangkutan untuk dipergunakan sebagaimana mestinya. Ditetapkan di Jakarta Pada tengga: 21 Januari 2008 7 SWANTO WV Fee 8 Utama Muda SNIP, 010 137 064 Tembusan : i. Kepala Badan Kepegawaian Negara Up. Deputi Bidang Informasi Kepegawaian; 2. Dirjen Anggaran Departemen Keuangan; 3. Kepala Kantor Perbendaharaan Pembayaran dan-Kas Negara IV Jakarta; 4. Kepala Bagian Perencanaan Biro Kepegawaian Depdagri di Jakarta; 7x

Hasil ekstraksi *OCR* menunjukan bahwa sudah dapat mengekstraks teks dari dokumen kepegawaian yang berformat .jpg. Namun, hasil menunjukan terdapat karakter-karakter yang tidak terdapat di dokumen tersebut juga ikut terkestraksi. Banyaknya huruf-huruf atau karakter yang tidak terdapat di kartu nama tapi juga ikut diekstraksi disebabkan oleh adanya figure atau grafik di dalam kartu nama tersebut yang dapat menghasilkan teks atau karater kacau pada bagian tersebut.

3.2. Proses Implementasi *Cosine Similarity*

Proses ini merupakan proses dimana dokumen yang sudah melewati tahap *OCR* akan diklasifikasikan otomatis oleh sistem menggunakan algoritma *cosine similarity*. Dalam tahap ini data-data yang digunakan sebagai D1 s/d D14 ditunjukkan pada Tabel 2.

Tabel 2. Data jenis kategori

Data	Kategori
D1	Surat Keputusan Calon Pegawai Negeri Sipil
D2	Surat Keputusan Pegawai Negeri Sipil
D3	Surat Keputusan NIP 18
D4	Surat Keputusan Gaji Berkala
D5	Surat Keputusan Riwayat Jabatan
D6	Impassing
D7	Kartu Pegawai
D8	TASPEN
D9	Kartu Istri/Kartu Suami
D10	Asuransi Kesehatan/BPJS
D11	Surat Nikah/Akta Cerai
D12	Penghargaan
D13	Surat Mutasi
D14	Sertifikat Pendidikan dan Pelatihan

Langkah perhitungan menggunakan algoritma *cosine similarity* terdapat 10 tahapan, yaitu:

1. Ditentukan terlebih dahulu masing-masing *query*, yaitu *query* dari jawaban (D), query dari *key* jawaban (Q) dan gabungan keduanya (*Queries*)
2. Ketiga *query* tersebut dihilangkan *stoplist* atau simbol-simbol yang tidak mempengaruhi penilaian, seperti tanda titik, tanda koma, tanda seru, dan sebagainya
3. Ketiga query tersebut dihilangkan stopwords atau kata-kata umum yang lazim digunakan dalam suatu *query*, seperti "dan", "jika", "di", "namun", "tetapi", dan sebagainya
4. Dihitung nilai *term frequency query* jawaban dan *query key* jawaban terhadap *queries*. Jadi perhitungan term di *query* jawaban dan *query key* jawaban merujuk pada *term* yang terdapat dalam *queries*
5. Dihitung nilai *document frequency* (n) atau banyaknya file (N) yang memiliki suatu *term* untuk tiap term dalam *queries*
6. Dihitung nilai *inverse document frequency* dengan rumus Log (n/df)+1

Tabel 3. Perhitungan tf & idf

Term	tf										df	idf
	Q	D1	D2	D3	D4	D5	D6	D7	…	D14		

keputusan	1	1	1	1	1	1	0	0		0	6	1,3979400009
calon	1	1	0	0	0	0	0	0		0	2	1,8750612633
pegawai	1	1	1	0	0	0	0	1		0	4	1,5740312688
nip	1	0	0	1	0	0	0	0		0	2	1,8750612633
jabatan	1	0	0	0	0	1	0	0		0	2	1,8750612633
kartu	0	0	0	0	0	0	0	1		0	1	2,1760912599
l8	0	0	0	1	0	0	0	0		0	1	2,1760912599
negeri	1	1	1	0	0	0	0	0		0	3	1,6989700004
sipil	1	1	1	0	0	0	0	0		0	3	1,6989700004
gaji	1	0	0	0	1	0	0	0		0	2	1,8750612633
berkala	0	0	0	0	1	0	0	0		0	1	2,1760912599

7. Dikalikan nilai *term frequency* dengan nilai *inverse document frequency* tiap *term* dalam Q maupun D.

Tabel 4. Perhitungan tf/idf = tf * idf

Term	**tf**									**D14**
	Q	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**		
keputusan	1,954236268	1,954236268	1,954236268	1,954236268	1,954236268	1,954236268	0	0		0
calon	3,515854741	3,515854741	0	0	0	0	0	0		0
pegawai	2,477574432	2,477574432	2,477574432	0	0	0	0	2,477574432		0
nip	3,515854741	0	0	3,515854741	0	0	0	0	..	0
jabatan	3,515854741	0	0	0	0	3,515854741	0	0		0
kartu	0	0	0	0	0	0	0	4,735373168		0
l8	0	0	0	4,735373	0	0	0	0		0

				168						
negeri	2,886499076	2,886499076	2,886499076	0	0	0	0	0		0
sipil	2,886499076	2,886499076	2,886499076	0	0	0	0	0		0
gaji	3,515854741	0	0	0	3,515854741	0	0	0		0
berkala	0	0	0	0	4,735373168	0	0	0		0
lon	1585474 41	1585474 41								
pegawai	2,477574432	2,477574432	2,477574432	0	0	0	0	2,477574432		0
nip	3,515854741	0	0	3,515854741	0	0	0	0		0
jabatan	3,515854741	0	0	0	0	3,515854741	0	0		0
kartu	0	0	0	0	0	0	0	4,735373168		0
18	0	0	0	4,735373168	0	0	0	0		0
negeri	2,886499076	2,886499076	2,886499076	0	0	0	0	0		0
si	2,8	2,8	2,8	0	0	0	0	0		0

8. Dihitung hasil perkalian skalar masing-masing *query* jawaban terhadap *query key* jawaban. Hasil perkalian dari setiap jawaban dengan *query* dijumlahkan
9. Dihitung hasil perkalian vektor tiap query key jawaban dan query jawaban

Tabel 5. Perhitungan SQRT

Term	tf									
	Q	D1	D2	D3	D4	D5	D6	D7	..	D14
keputusan	1,954236268	1,954236268	1,954236268	1,954236268	1,954236268	1,954236268	0	0	..	0
ca	3,5	3,5	0	0	0	0	0	0		0

pil	8649907676	8649907676	8649907676							
gaji	3,515854741	0	0	0	3,515854741	0	0	0		0
berkala	0	0	0	0	4,735373168	0	0	0		0
Total	**24,268227822**	**13,720663589**	**10,204808085**	**10,205464418**	**10,205464418**	**5,470091009**	**0**	**7,212947676**		**0**
SQRT	**4,926279308**	**3,704141411**	**3,194496651**	**3,194594922**	**3,194594922**	**2,338822569**	**0**	**2,685693132**		**0**

10. Dihitung nilai *cosine similarity* (nilai vektor beda antara D terhadap Q) dengan rumus :

$$sim(d,q) = \sum_{k=1}^{t}(weight_{ik} weight_{qk}) \Big/ \sqrt{\sum_{k=1}^{t}(weight_{ik}^{2} weight_{qk}^{2})}$$

Tabel 6. Hasil *Cosine Similarity*

Data	Kategori	Nilai Kemiripan	Persentase
D1	Surat Keputusan Calon Pegawai Negeri Sipil	0.22209373	22.20%
D2	Surat Keputusan Pegawai Negeri Sipil	0.21288006	21.28%
D3	Surat Keputusan NIP 18	0.05100153	5.10%
D4	Surat Keputusan Gaji Berkala	0.08500255	8.50%
D5	Surat Keputusan Riwayat Jabatan	0.06800204	6.80%
D6	Impassing	0	-
D7	Kartu Pegawai	0.0480847	4.81%
D8	TASPEN	0	-
D9	Kartu Istri/Kartu Suami	0	-
D10	Asuransi Kesehatan/BPJS	0	-
D11	Surat Nikah/Akta Cerai	0	-
D12	Penghargaan	0	-
D13	Surat Mutasi	0	-
D14	Sertifikat Pendidikan dan Pelatihan	0	-

Hasil akhir yang didapatkan dari contoh hasil pengujian tersebut, bahwa dokumen sk_cpns.jpg merupakan jenis kategori surat keputusan calon pegawai negeri sipil berdasarkan nilai kemiripan yang paling tinggi. Nilai kemiripan dari kategori surat keputusan calon pegawai negeri sipil adalah 0.22209373 dengan persentase kemiripan 22.20% dan terdapat 8 kategori yang tidak memiliki nilai kemiripan dengan dokumen sk_cpns.jpg.

4. KESIMPULAN DAN SARAN

Berdasarkan hasil dan pembahasan yang telah dijabarkan sebelumnya, maka kesimpulan yang dapat diambil adalah sebagai berikut:

a. Data yang diambil oleh *OCR* berupa teks yang berada di dalam gambar, sedangkan untuk pengklasifikasian otomatis dokumen memiliki 14 jenis kategori dokumen. Berdasarkan hasil pengujian *OCR* yang dilakukan pada dokumen sk_cpns.jpg, menunjukan bahwa gambar sudah dapat mengekstraks teks. Namun, hasil menunjukan terdapat karakter-karakter yang tidak terdapat di dokumen tersebut juga ikut terkestraksi.
b. Algoritma *cosine similarity* pada aplikasi ini hanya berlaku jika terdapat kemiripan kata, apabila teks yang sudah diekstrak oleh *OCR* tidak memiliki kemiripan kata sama sekali dengan jenis kategori dokumen maka nilai kemiripan yang

akan didapatkan adalah 0.0 dan jenis kategorinya adalah lainnya.

c. Hasil yang didapatkan dari algoritma *Cosine Similarity* dari dokumen sk_cpns.jpg memiliki nilai kemiripan yang paling tinggi berjumlah 0.22209373 dengan persentase kemiripan 22.20%, dokumen yang dapat diklasifikasikan dengan kategori Surat Keputusan Calon Pegawai Negeri Sipil adalah sk_cpns.jpg.

Penelitian yang penulis lakukan masih dapat dikembangkan lebih lanjut, beberapa saran untuk mengembangkan penelitian ini adalah sebagai berikut:

a. Untuk saat ini teknik *OCR* harus menggunakan file yang berformatkan .jpg, .jpeg dan .png. Penulis berharap teknik *OCR* dapat dilakukan menggunakan file yang berformat .pdf tanpa harus mengkonversikannya terlebih dahulu.
b. Pada penelitian ini, penulis hanya menggunakan satu contoh saja dikarenakan data yang digunakan bersifat pribadi dan tidak dapat dikonsumsi oleh publik seperti surat keputusan gaji berkala, KTP dan surat keputusan jabatan. Untuk penelitian selanjutnya diharapkan menggunakan data yang bersifat umum.

DAFTAR PUSTAKA

Utami, Tri Budi. 2020. Jurnal Ilmu Teknik dan Komputer. Universitas Mercu Buana.

Hatta A, Ahmad. Rancang Bangun Sistem Pengelolaan Dokumen-Dokumen Penting Menggunakan Text Mining. Surabaya: Politeknik Elektronika Negeri Surabaya.

Susandi, D. dan Sholahudin, U. 2016. Pemanfaatan Vector Space Model pada Penerapan Algoritma Nazief Adriani, KNN dan Fungsi Similarity Cosine untuk Pembobotan IDF dan WIDF pada Prototipe Sistem Klasifikasi Teks Bahasa Indonesia. Jurnal Teknologi Informasi Volume 3, Nomor 1.

Nurdiana, O., Jumadi., dan Nursantika, D. 2016. Perbandingan Metode Cosine Similarity dengan Metode Jaccard Similarity pada Aplikasi Pencarian Terjemahan Al-Qur'an dalam Bahasa Indonesia. Jurnal Online Informatika Volume 1, Nomor 1.

Zulhida Putri, Dewinta. 2018. Konversi Citra Kartu Nama Ke Teks Menggunakan Teknik Ocr Dan Jaro-Winkler Distance. Universitas Bengkulu.

Firdaus. 2019. Implementasi Cosine Similarity Untuk Peningkatan Akurasi Pengukuran Kesamaan Dokumen Pada Klasifikasi Dokumen Berita Dengan K Nearest Neighbour. Makassar: STMIK AKBA.

Nurdiana, Ogie. 2016. Perbandingan Metode Cosine Similarity Dengan Metode Jaccard Similarity Pada Aplikasi Pencarian Terjemah Al-Qur'an Dalam Bahasa Indonesia. Bandung: Universitas Islam Negeri Sunan Gunung Djati Bandung.

Wahyu Irianandа, Syahroni. 2018. Identifikasi Kemiripan Teks Menggunakan Class Indexing Based Dan Cosine Similarity Untuk Klasifikasi Dokumen Pengaduan. Universitas Brawijaya.

Pakpahan, Rolina. 2019. Perancangan Aplikasi Pendeteksian Kemiripan Dokumen Teks Menggunakan Algoritma Cosine Similarity Berbasis Android. ISSN 2339-210X. Majalah Ilmiah INTI, Volume 6, Nomor 3.

KERTAS KERJA

Ringkasan

Kertas kerja ini merupakan material kelengkapan artikel jurnal dengan judul "Penerapan Algoritma *Cosine Similarity* Untuk Pengklasifikasian Otomatis Dokumen Kepegawaian Dengan Teknik *OCR* Di Kementerian Dalam Negeri*"*. Kertas kerja berisi semua material hasil penelitan Tugas Akhir yang tidak dimuat/atau disertakan di artikel jurnal. Di dalam kertas kerja ini disajikan: literature review, dataset yang digunakan, source code, dan hasil eksperimen secara keseluruhan.

Pada bagian I literature review, di dalam literature review ini disajikan hasil review atas literature yang terkait dengan penelitian yaitu: konsep data mining, *OCR* dan *cosine similarity*. Literatur membantu penulis dalam mencari informasi yang dibutuhkan. membantu memperkuat informasi hasil dari suatu analisis atau hipotesa dan juga memberi tambahan informasi. Bagian II adalah analisis dan perancangan, pada bagian analisis dan perancangan di uraikan tahapan rancangan pengolahan data, bagaimana rancangan dari awal hingga selesai. Bagian III adalah source code, pada bagian ini penulis memaparkan proses pengolahan data yaitu parameter masukannya apa, dan keluaran yang dihasilkan pengklasifikasian otomatis menggunakan aplikasi yang telah dibangun. Bagian IV adalah dataset, bagian ini memaparkan dataset yang diolah/digunakan yaitu data dokumen kepegawaian Kementerian Dalam Negeri. Kemudian bagian V yaitu hasil eksperimen secara keseluruhan termasuk dalam pengujian menggunakan aplikasi yang sudah dibangun.