Automated Data Extractor for Evaluation Form Data Extraction

Automated Data Extractor untuk Ekstraksi Data Formulir Evaluasi

Taukhid W. Broto[1*], Yessi N. Ulfia[2], Dian E. H. Purnomo[3]

[1),2), 3)] *Politeknik Industri Furnitur dan Pengolahan Kayu*

*Coresponding author

E-mail addresses: taukhidwb@poltek-furnitur.ac.id

Abstract. *This paper presents an application program for automated PDF data extraction tool which is implemented in Python, designed to automate the extraction of structured data from PDF documents, specifically peer evaluation forms. The tool features an ease of use, allowing users to upload PDF files, process the data extraction from standardized PDF forms, which traditionally would require manual data entry or complex optical character recognition (OCR) systems, and view extracted data in a tabular formats. This paper discusses the methodology, results, and potential future enhancements of the tool.*
Keywords: *data extraction, evaluation form, pdf*

Abstrak. *Makalah ini menyajikan pembahasan tentang program aplikasi sebagai alat ekstraksi data PDF otomatis yang diimplementasikan menggunakan bahasa pemrograman Python, dirancang untuk mengotomatiskan ekstraksi data terstruktur dari dokumen PDF khususnya formulir evaluasi rekan kerja. Aplikasi ini memiliki fitur kemudahan penggunaan, memungkinkan pengguna untuk mengunggah file PDF, memproses ekstraksi data dari formulir PDF standar yang secara tradisional memerlukan entri data manual atau sistem pengenalan karakter optik (OCR) yang kompleks, dan memperoleh tampilan data hasil ekstraksi dalam format tabel. Makalah ini memuat pembahasan tentang metodologi, hasil, dan potensi peningkatan alat di masa depan.*
Kata kunci: *ekstraksi data, formulir evaluasi, pdf*

PENDAHULUAN

Tantangan dalam mengekstraksi data terstruktur secara efisien dari dokumen PDF sangat signifikan, terutama ketika formulir standar seperti evaluasi rekan kerja yang diisi secara mandiri oleh anggota tim dapat memuat isian yang berbeda antara satu dengan lainnya [1]. Metode tradisional sering kali memerlukan entri data manual atau sistem pengenalan karakter optik (OCR) yang kompleks. OCR adalah sistem yang mempunyai fungsi mengidentifikasi sebuah karakter huruf atau angka [2]. Penggunaan metode OCR bertujuan untuk maerubah suatu citra yang terdiri dari tulisan dan background, menjadi sebuah tulisan atau angka[3][4]. Makalah ini memperkenalkan program aplikasi berbasis bahasa pemrograman Python yang mengotomatiskan proses dimaksud, menargetkan informasi spesifik dari dokumen seperti nama dan peringkat (*rating*) berdasarkan kriteria yang telah ditentukan sebelumnya. Program aplikasi ini akan sangat berguna dalam pengelolaan kegiatan pendidikan maupun kegiatan profesional lainnya di mana penilaian di antara sesama rekan kerja merupakan hal yang umum.

Skrip aplikasi yang dibuat merupakan alat ekstraksi data dari dokumen berformat PDF yang ditulis dalam bahasa pemrograman Python. Aplikasi ini bertujuan untuk menjawab kebutuhan ekstraksi data terstruktur yang efisien dari dokumen PDF, terutama fokusnya adalah pada formulir evaluasi sejawat. Program aplikasi ini dirancang demi kemudahan penggunaan, memungkinkan bagi pengguna untuk mengunggah file PDF, memprosesnya, melihat data yang diekstraksi dalam format tabel, dan untuk pengembangan di masa mendatang diharapkan dapat mengekspor hasilnya ke dalam bentuk file CSV atau Excel.

Masalah utama yang hendak diselesaikan melalui skrip program aplikasi ini adalah otomatisasi ekstraksi data dari formulir PDF standar, yang secara tradisional memerlukan entri data manual atau sistem pengenalan karakter optik (OCR) yang kompleks.

METODE PENELITIAN

Ekstraktor Data PDF dikembangkan menggunakan metode Software Development Life Cycle (SDLC) dengan pendekatan Rapid Application Development (RAD). Metode ini memungkinkan fleksibilitas dan peningkatan berkelanjutan selama proses pengembangan. Menurut Simarmata (2010:39), SDLC mengacu pada model dan proses yang digunakan untuk

mengembangkan sistem perangkat lunak dan menguraikan proses, yaitu pengembang menerima perpindahan dari permasalahan ke solusi.

Pengembangan rekayasa sistem informasi (system development) dan atau perangkat lunak (software engineering) dapat berarti menyusun sistem atau perangkat lunak yang benar – benar baru atau yang lebih sering terjadi menyempurnakan yang sebelumnya (Nugroho, 2010:2)[10]. Pendekatan RAD diperkenalkan untuk mengatasi penunndaan dalam jangka waktu lama yang dialami pengguna aplikasi ketika pengembangan dilakukan menggunakan pendekatan struktural. Pengembangan aplikasi pemrograman secara tradisional sering menyita waktu lama ketika terjadi perubahan yang mendasar terkait persyaratan dan kebutuhan seiring berjalannya waktu pengembangan. Proyek dimulai dengan fase perencanaan dan diikuti oleh fase perancangan, di mana persyaratan utama diidentifikasi: kebutuhan untuk mengekstrak data spesifik dari formulir PDF, dan menyajikannya dengan cara yang ramah bagi pengguna.

Tahap pengembangan kemudian berlanjut dalam beberapa iterasi, dengan setiap siklus menambahkan fitur baru dan menyempurnakan fitur yang sudah ada. Logika inti skrip mengikuti algoritma yang jelas:

1. **PDF Parsing**: Skrip memanfaatkan PyPDF2 *library* [7] untuk mengekstrak teks dari file PDF. PDF parsing melibatkan pemahaman terhadap struktur file PDF dan proses ekstraksi teks terbaca darinnya.
2. **Regular Expressions (Regex)**: Regular expressions [6] dimanfaatkan untuk mengidentifikasi dan mengekstrak pola teks spesifik dari konten PDF yang diekstrak. Teknik ini sangat penting untuk menemukan informasi yang relevan di antara teks yang tidak terstruktur [8].
3. **Multithreading**: Skrip mengimplementasikan proses yang terpisah untuk pemrosesan data.
4. **Data Manipulation**: *Library* Pandas [5] digunakan untuk menangani data yang diekstrak, dan selebihnya digunakan untuk mengekspor ke format file lainnya.

Sepanjang proses ini, penanganan kesalahan memastikan bahwa setiap masalah ditangkap dan dilaporkan, sementara pengelogan melacak peristiwa penting untuk ditinjau nanti. Iterasi terbaru menambahkan pengukuran kinerja, menghitung kecepatan dan akurasi pemrosesan, dan mengumpulkan umpan balik pengguna tentang kegunaan.

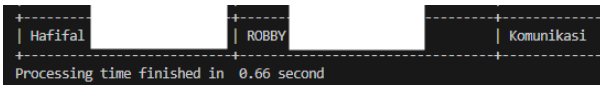
Pengujian dilakukan setelah setiap siklus pengembangan, dengan pengujian unit untuk masing-masing komponen dan pengujian integrasi untuk seluruh sistem. Umpan balik pengguna dimasukkan ke dalam iterasi berikutnya, yang mengarah pada peningkatan tata letak.

Pendekatan SDLC ini, dikombinasikan dengan alur logika skrip yang jelas, menghasilkan alat yang secara efektif mengotomatiskan proses ekstraksi data sambil memberikan wawasan berharga tentang kinerjanya sendiri. Sifat berulang dari pengembangan memungkinkan penyempurnaan berkelanjutan, memastikan bahwa produk akhir sangat sesuai dengan kebutuhan dan harapan pengguna.

Untuk memastikan keandalan dan memfasilitasi debugging, pengembangan aplikasi ini menggabungkan mekanisme penanganan kesalahan dan pencatatan yang komprehensif. Setiap pengecualian yang ditemui selama proses ditangkap, dicatat ke file dengan stempel waktu terperinci dan deskripsi kesalahan, dan dilaporkan kembali kepada pengguna. Pendekatan ini tidak hanya meningkatkan pengalaman pengguna dengan memberikan umpan balik yang jelas tentang masalah apa pun, tetapi juga membantu penyempurnaan alat yang berkelanjutan.

Program yang dikembangan juga direncakan untuk dapat mencakup pengukuran kinerja aplikasi dalam bentuk komponen terintegrasi. Skrip ini didesain untuk dapat melacak waktu pemrosesan untuk setiap file PDF serta ditampilkan hasilnya.

Keberadaan metrik pengukuran ini pada akhirnya menjadi sarana pemantauan secara *real-time* tentang efektivitas dan efisiensi alat. Dengan menggabungkan proses teknis ini dengan fitur yang berpusat pada pengguna, pendekatan yang dipilih ini dapat menghasilkan proses yang efisien, dan ramah pengguna untuk mengotomatiskan tugas intensif kerja tradisional untuk mengekstraksi data terstruktur dari formulir PDF.



Gambar 1. Durasi pemrosesan data.

HASIL PENELITIAN DAN PEMBAHASAN

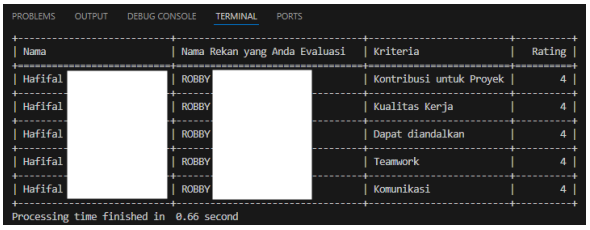
Sementara Enhanced PDF Data Extractor pada dasarnya adalah alat daripada skrip analisis data, aplikasinya menghasilkan dua hasil penting yang secara signifikan merampingkan proses penanganan data evaluasi rekan:

1. ***Successful extraction of peer evaluation data from PDF forms***: Fungsi utama dari alat ini adalah untuk mengotomatiskan ekstraksi informasi spesifik dari formulir evaluasi PDF standar. Proses ini, yang secara tradisional memerlukan entri data manual atau sistem OCR yang kompleks, dicapai melalui pencocokan pola yang canggih menggunakan ekspresi reguler. Alat ini memindai konten PDF dengan cermat, mengidentifikasi dan menangkap poin data utama seperti nama evaluator, nama orang yang dievaluasi, dan peringkat untuk kriteria yang telah ditentukan sebelumnya. Proses ekstraksi ini dirancang agar kuat, mampu menangani variasi tata letak formulir atau sedikit inkonsistensi dalam entri data. Keberhasilan pelaksanaan langkah ini mengubah dokumen PDF statis yang berpotensi sulit diproses menjadi data yang dinamis dan dapat dimanipulasi. Hasil ini sangat berharga dalam skenario yang melibatkan sejumlah besar formulir evaluasi, di mana ekstraksi data manual akan sangat memakan waktu dan rentan terhadap kesalahan manusia. Dengan mengotomatiskan proses ini, alat ini tidak hanya menghemat waktu yang signifikan tetapi juga mengurangi kemungkinan kesalahan transkripsi, memastikan tingkat integritas data yang lebih tinggi.
2. ***Structured presentation of the data in a tabular format***: Setelah data evaluasi rekan diekstraksi, alat ini mengatur dan menyajikan informasi ini

dalam format tabel yang jelas dan terstruktur dalam antarmuka pengguna grafis. Presentasi ini mengubah data mentah dan tidak terstruktur dari PDF menjadi format logis dan mudah dibaca yang memfasilitasi pemahaman dan analisis cepat. Struktur tabel biasanya mencakup kolom untuk nama evaluator, nama orang yang dievaluasi, kriteria evaluasi, dan peringkat yang sesuai. Format terstruktur ini melayani banyak tujuan: memungkinkan pengguna untuk memverifikasi keakuratan data yang diekstraksi dengan cepat, memberikan gambaran langsung tentang hasil evaluasi, dan menyediakan dasar bagi manipulasi atau analisis data lebih lanjut. Dengan menyajikan data secara terstruktur ini, alat ini menjembatani kesenjangan antara konten PDF mentah dan informasi yang dapat ditindaklanjuti, secara signifikan meningkatkan utilitas data evaluasi rekan anggota tim.

A. Ekstraksi Dokumen Penilaian Rekan Anggota Tim

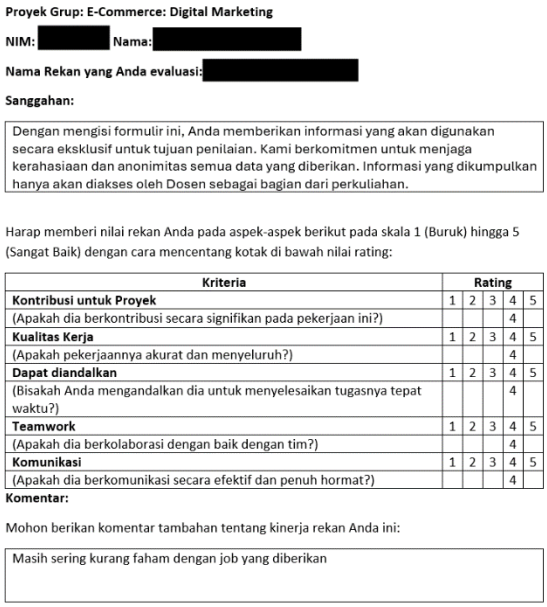
Ekstraktor Data PDF menggunakan metodologi multi-langkah yang canggih untuk mengekstrak, memproses, dan menganalisis data dari dokumen PDF secara efisien. Pada intinya, proses ini memanfaatkan kombinasi penguraian PDF, pencocokan pola teks, multithreading, dan teknik penataan data. Perjalanan dimulai ketika pengguna mengunggah file PDF melalui antarmuka pengguna grafis. Setelah memulai tugas pemrosesan, aplikasi menelurkan utas terpisah (DataExtractorThread) untuk menangani pekerjaan intensif komputasi, memastikan GUI utama tetap responsif. Thread ini pertama-tama menggunakan PyPDF2 *library* untuk mengekstrak semua konten tekstual dari PDF dengan cermat, menavigasi setiap halaman dan mengkompilasi teks menjadi satu string yang komprehensif.



Gambar 2. Hasil ekstraksi data

B. Tampilan Data Secara Terstruktur untuk Pengolahan Lebih Lanjut

Setelah teks mentah diperoleh, metodologi bergeser ke pengenalan pola yang canggih. Memanfaatkan kekuatan ekspresi reguler, skrip mencari pola tertentu dalam teks yang sesuai dengan informasi penting seperti nama evaluator, nama orang yang dievaluasi, dan peringkat untuk kriteria yang telah ditentukan sebelumnya. Langkah pencocokan pola ini sangat rumit, karena harus cukup kuat untuk menangani variasi pemformatan sambil tetap menangkap data yang diperlukan secara akurat. Informasi yang diekstraksi kemudian disusun dengan hati-hati ke dalam daftar daftar, di mana setiap daftar dalam mewakili deretan data yang berisi nama evaluator, nama orang yang dievaluasi, kriteria tertentu, dan peringkat yang sesuai.



Proyek Grup: E-Commerce: Digital Marketing

NIM: [redacted] **Nama:** [redacted]

Nama Rekan yang Anda evaluasi: [redacted]

Sanggahan:

Dengan mengisi formulir ini, Anda memberikan informasi yang akan digunakan secara eksklusif untuk tujuan penilaian. Kami berkomitmen untuk menjaga kerahasiaan dan anonimitas semua data yang diberikan. Informasi yang dikumpulkan hanya akan diakses oleh Dosen sebagai bagian dari perkuliahan.

Harap memberi nilai rekan Anda pada aspek-aspek berikut pada skala 1 (Buruk) hingga 5 (Sangat Baik) dengan cara mencentang kotak di bawah nilai rating:

Kriteria	Rating				
Kontribusi untuk Proyek	1	2	3	4	5
(Apakah dia berkontribusi secara signifikan pada pekerjaan ini?)				4	
Kualitas Kerja	1	2	3	4	5
(Apakah pekerjaannya akurat dan menyeluruh?)				4	
Dapat diandalkan	1	2	3	4	5
(Bisakah Anda mengandalkan dia untuk menyelesaikan tugasnya tepat waktu?)				4	
Teamwork	1	2	3	4	5
(Apakah dia berkolaborasi dengan baik dengan tim?)				4	
Komunikasi	1	2	3	4	5
(Apakah dia berkomunikasi secara efektif dan penuh hormat?)				4	

Komentar:

Mohon berikan komentar tambahan tentang kinerja rekan Anda ini:

Masih sering kurang faham dengan job yang diberikan

Gambar 3. Formulir Peer Evaluation

KESIMPULAN

Skrip Enhanced PDF Data Extractor menghadirkan solusi yang kuat untuk mengotomatiskan ekstraksi data terstruktur dari formulir PDF standar. Dengan menggabungkan penguraian PDF, pencocokan pola teks, dan antarmuka yang ramah pengguna, diharapkan akan secara signifikan mengurangi waktu dan upaya yang diperlukan untuk menyusun data evaluasi rekan sejawat. Pendekatan multithreaded memastikan bahwa aplikasi tetap responsif, bahkan saat memproses PDF besar atau kompleks.

Kemampuan alat untuk mengekspor data ke format umum seperti CSV dan Excel meningkatkan kegunaannya, memungkinkan integrasi tanpa batas dengan alat dan alur kerja analisis data lainnya. Secara keseluruhan, skrip ini menunjukkan pendekatan yang efektif untuk memecahkan masalah spesifik mengekstraksi data evaluasi rekan sejawat dari formulir PDF, dengan aplikasi potensial dalam berbagai konteks pendidikan dan profesional.

REFERENSI

[1] Chawla, N. V. (2005). Data Mining for Imbalanced Datasets: An Overview. In Data Mining and Knowledge Discovery Handbook (pp. 853-867). Springer, Boston, MA.

[2] Syahri Muharom " Penerapan Metode Hough Line Transform Untuk Mendeteksi Ruangan Menggunakan Kamera" Jurnal IPTEK Vol. 21. No. 1, Mei 2017

[3] Parul Shah, Sunil Karamchandani, Taskeen Nadkar, Nikita Gulecha, Kaushik Koli, Ketan Lad “OCR-based Chassis-Number Recognition using Artificial Neural Networks” IEEE, ICVES 2009.

[4] Muharom, S. (2019). Pengenalan Nomor Ruangan Menggunakan Kamera Berbasis OCR Dan

Template Matching. Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi, 4(1), 27-32.
[5] Shaw, Z. A (2024). Learn Python the Hard Way. Addison-Wesley Profesional.
[6] Shaw, Z. A (2018). Learn Python 3 the Hard Way. Addison-Wesley Profesional.
[7] Sweigart, A. (2019). Automate The Boring Stuff With Python. O'reilly Media.
[8] Miner, G., Elder IV, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. Academic Press.
[9] Davis, S., Fendall, J. (2012). Software design and development: the HSC course. Douglas Park, N. S. W., Parramatta Education Center.
[10] Sofyan, A. A., Puspitorini, P., & Yulianto, M. A. (2016). Aplikasi Media Informasi Sekolah Berbasis SMS Gateway Dengan Metode SDLC (System Development Life Cycle). Jurnal Sisfotek Global, 6(2), 297726.

Conflict of Interest Statement:
The author declares that the research was conducted in the absence of any commercial or financial relation- ships that could be construed as a potential conflict of interest.

Article History:
Received: 12 September 2024 | Accepted: 25 Oktober 2024| Published: 30 November 2024