APLIKASI ARSIP DIGITAL DENGAN MENGIMPLEMENTASIKAN OCR DAN METODE TEXT CLASSIFICATION TF-IDF BERBASIS WEB

Uray Adri Hidayat
Mahasiswa Program Studi Teknik Informatika Fakultas Teknik Universitas Wahidiyah
ure.uray@gmail.com

Khamid
Dosen Teknik Informatika Fakultas Teknik Universitas Wahidiyah
khamid@uniwa.ac.id

Abstrak

Penelitian ini dilatarbelakangi oleh sulitnya melakukan pencarian dokumen dalam jumlah banyak pada saat dibutuhkan dalam waktu yang cepat. Aplikasi ini dibuat menggunakan Framework Laravel dengan bantuan Tesseract OCR dan MongoDB sebagai Database. Dengan kemudahan yang diberikan aplikasi pengguna hanya tinggal mengupload dokumen kemudian aplikasi secara otomatis memberikan indeks terhadap dokument tersebut dengan menggunakan teknik Text Classification dan Reversal Indexing sehingga dapat dilakukan pencarian kembali secara cepat dan tepat.

Kata Kunci: *Mesin Pencari, Text Classification, TF-IDF, OCR, Reversal Indexing.*

PENDAHULUAN

Sekarang di abad ke 20 adalah era dimana informasi berkembang dengan sangat cepat dan teknologi sudah mulai dipergunakan di segala bidang termasuk salah satunya adalah bidang perkantoran yang dimana segala bentuk dokumen surat telah terkomputerisasi. Bahkan mungkin setiap instansi setidaknya memiliki sebuah komputer yang digunakan untuk membuat dokumen surat baik surat-surat kantor ataupun karya tulis.

Dokumen yang telah dibuat ini nantinya di cetak sebagai bukti yang digunakan untuk berbagai keperluan misalnya permohonan persetujuan dan lain-lain. Termasuk juga dalam bidang akademis yang dimana setiap mahasiswa di universitas pastinya membuat dokumen-dokumen karya tulis ilmiah. Dan dengan berjalannya waktu maka jumlah dokumen-dokumen ini pasti akan menyentuh angka yang membuat proses pengarsipan dokumen surat menjadi tidak efisien jika dilakukan dengan pengarsipan manual.

Karena banyaknya dokumen yang dicetak tersebut menimbulkan sebuah permasalahan baru yakni bagaimana agar data dokumen tersebut tetap aman sehingga tidak terjadi kehilangan arsip dan bagaimana cara untuk menemukan kembali dokumen yang dikehendaki jika jumlah arsip telah cukup banyak. Apabila user memiliki manajemen yang baik dalam menyimpan arsip mungkin bukan menjadi masalah besar untuk menemukan kembali dokumen yang telah disimpan namun tetap membutuhkan waktu jika hanya harus mencari 1 dokumen diantara ratusan atau bahkan ribuan dokumen yang ada.

Sekarang telah banyak sistem pengarsipan dokumen berbasis digital yang dimana sistem pengarsipan ini sangat membantu dalam penyimpanan dan pencarian dokumen. Dan juga dengan adanya sistem pengarsipan digital proses penyimpanan dokumen menjadi mudah, efiesien dan hemat tempat. Namun jika suatu instansi belum pernah melakukan penyimpanan dokumen yang berbasis digital sedangkan arsip dokumen yang dimiliki sudah sangat banyak. Tentu akan menjadi pekerjaan yang merepotkan dalam menginputkan data dokumen. Karena didalam sistem terkadang pengguna masih harus melakukan proses identifikasi baik judul maupun isi dokumen saat melakukan proses input data.

Dari munculnya permasalahan tersebut penulis mencoba untuk melakukan penelitian tentang aplikasi manajemen arsip menggunakan OCR (*Optical Character Recognition*) untuk melakukan pembacaan isi dokumen dari hasil scan dan otomatis melakukan identifikasi tentang isi dokumen tersebut agar waktu dalam proses input dokumen menjadi efisien sehingga user tidak perlu melakukan input data untuk judul dan isi dokumen dan dalam proses pencarian kembali dokumen tersebut menjadi lebih mudah dan cepat dengan menggunakan metode pencarian seperti *search engine* yang menggunakan *reversal indexing* yang di-*support* dengan metode *Text Classification TF-IDF*.

Dan juga dalam aplikasi manajemen arsip ini nanti penulis akan menggunakan aplikasi berbasis web agar jika sewaktu – waktu aplikasi ini akan digunakan dalam cangkupan yang lebih luas maka aplikasi dapat ditempatkan pada sebuah *server* agar user dari manapun dapat mengakses aplikasi secara langsung dimanapun berada.

Aplikasi manajemen arsip digital yang akan dirancang oleh penulis sebisa mungkin menerapkan konsep *mobile friendly*, agar user dapat mengakses aplikasi dari segala jenis device. Untuk teknologi yang

akan digunakan penulis dalam perancangan aplikasi ini menggunakan *Laravel* PHP *Framework* dalam konstruksi aplikasi, MongoDB sebagai database engine dan Tessract OCR sebagai modul dalam pemrosesan citra pada dokumen yang berupa gambar.

METODE

Tahapan dalam penelitian yang dilakukan adalah sebagai berikut :

Pengumpulan Data

- **Observasi**
  Observasi yaitu mengumpulkan informasi dengan pengamatan atau peninjauan langsung terhadap objek penelitian.
- **Studi Pustaka**
  Studi pustaka dapat diartikan sebagai suatu langkah untuk memperoleh informasi dari penelitian terdahulu yang harus dikerjakan, tanpa memperdulikan apakah sebuah penelitian menggunakan data primer atau data sekunder, apakah penelitian tersebut menggunakan penelitian lapangan, laboratorium ataupun didalam museum. Yang dimaksud dengan studi kepustakaan adalah segala usaha yang dilakukan oleh peneliti untuk menghimpun informasi yang relevan dengan topik dan masalah yang akan atau sedang diteliti.

Tahapan Penelitian

Adapun tahapan penelitian yang akan dilakukan penyusun dalam pengembangan aplikasi menggunakan siklus pengembangan aplikasi (*Software Development Life Cycle*) Air Terjun (*Waterfall*) dengan langkah-langkah sebagai berikut:

- **Analisa Kebutuhan**
  Meliputi Analisa Sistem, Analisa Kebutuhan Perangkat Lunak, Analisa Kebutuhan Perangkat Keras
- **Desan Sistem**
  Desain perangkat lunak adalah proses multi langkah yang fokus pada de- sain pembuatanprogram perangkat lunak termasuk struktur data, arsitektur perangkat lunak, representasi antar muka, dan prosedur pengodean. Tahap ini mentranslasi kebutuhan perangkat lunak dari tahap analisis kebutuhan ke representasi desain agar dapat diimplementasikan menjadi program pa- da tahap selanjutnya. Desain perangkat lunak yang dihasilkan pada tahap ini juga perlu didokumentasikan.
- **Penulisan Kode Program**
  Desain harus ditranslasikan kedalam program perangkat lunak. Hasil dari tahap ini adalah program komputer sesuai dengan desain yang telah dibuat pada tahap desain.
- **Pengujian Program**
  Pengujian fokus pada perangkat lunak secara dari segi *logic* dan fungsional dan memastikan bahwa semua bagian sudah diuji. Hal ini dilakukan untuk meminimalisir kesalahan (*error*) dan memastikan keluaran yang dihasilkan sesuai dengan yang diinginkan.
- **Pendukung dan Pemeliharaan**
  Tidak menutup kemungkinan sebuah perangkat lunak mengalami perubahan ketika sudah diterapkan. Perubahan bisa terjadi karena adanya kesalahan yang muncul dan tidak terdeteksi saat pengujian atau perangkat lunak harus beradaptasi dengan lingkungan yang baru. Tahap pendukung atau pemeliharaan dapat mengulangi proses pengembangan mulai dari analisis spesifikasi untuk perubahan perangkat lunak yang sudah ada, tapi tidak untuk membuat perangkat lunak baru.

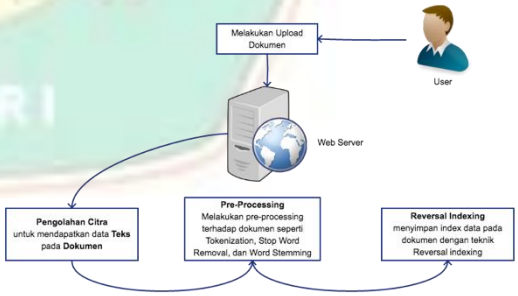
Alat dan Bahan

Adapun alat dan bahan yang digunakan pada penelitian ini yaitu Macbook Pro (13-*inch*, *Early* 2011) 2.3 GHz Intel Core i5 sistem operasi MacOS Sierra (10.12.6) dengan RAM 16 GB yang ter-*install* perangkat lunak Tesseract OCR, MAMP, MongoDB dan peramban *web* (*browser*) Google Chrome sebagai program untuk mensimulasikan penelitian yang dilaksanakan.

HASIL DAN PEMBAHASAN

Proses Upload Dokumen

Pada aplikasi arsip digital yang dibangun dilakukan proses seperti yang terlihat pada gambar dibawah.



Gambar 1. Proses Pada Aplikasi

Proses paling awal yang dilakukan oleh aplikasi adalah dengan menyiapkan dokumen yang diupload oleh user dan dilakukannya pembacaan kata oleh OCR maka untuk selanjutnya dilakukan *pre-processing.* Yakni mengolah dokumen dengan menghilangkan *stop word* dan melakukan *word stemming.*

Setelah dokumen melewati tahapan *pre-processing*, tahapan selanjutnya adalah melakukan *reversal indexing*. Konsep dari *reversal indexing* adalah proses index yang dibalik dari proses index pada umumnya. Jika pada umumnya proses indeks dilakukan dengan mencari kata pada 1 dokumen. Maka reversal indexing menyimpan nilai dokumen apa saja yang terdapat dalam 1 kata.

Proses Pencarian Dokumen

Fungsi *Search* adalah fungsi yang penting didalam aplikasi arsip digital dikarenakan kecepatan dan ketepatan dalam menemukan kembali dokumen yang telah disimpan dengan banyak dokumen yang mungkin memiliki kesamaan merupakan *point* yang ingin dicapai. Pada penelitian ini tahapan dalam pencarian dokumen dilakukan berdasarkan *query* dari *keyword* yang dimasukan pada inputan form pencarian. Dari *query* tersebut kemudian dipecah menjadi *term* yang akan dibandingkan dengan data yang telah diproses sebelumnya. Dikarenakan pada proses *reversal indexing* telah didapatkan bobot per kata dari setiap dokumen yang juga merupakan penerapan dari *Raw Term Frequency*, maka bisa dilanjutkan ke perhitungan *Inverse Document Frequency* (IDF) dari masing-masing *keyword* yang sesuai dengan *term* pada *query*. Rumus untuk melakukan perhitungan IDF adalah:

$$IDFj = log(D/dfj)$$

Dimana D adalah jumlah semua dokumen dalam koleksi sedangkan dfj adalah jumlah dokumen yang mengandung *term*.

Sebagai contoh penulis akan menggunakan *query* "surat pernyataan" sebagai inputan pada *form search*. Dengan *query* tersebut didapatkan 2 *term* yakni "surat" dan "pernyataan". Dari hasil *upload* data dokumen di *database* terdapat 5 jenis dokumen dengan jenis yang berbeda. 2 Diantaranya merupakan Surat Perintah Kerja dan Surat Keputusan. Pada Surat Perintah Kerja diketahui hanya terdapat 1 *term* yang cocok yakni "surat" dengan bobot TF 1. Dan pada dokumen kedua yakni Surat Pernyataan, didapatkan setiap *term* pada *query* dengan nilai bobot TF untuk "surat" adalah 1 dan "pernyataan" bernilai 2. Untuk ketiga dokumen lain tidak termasuk dalam koleksi *query* ini karena sama sekali tidak mengandung *keyword* untuk *query* tersebut.

Setelah didapatkan nilai TF untuk masing-masing dokumen maka nilai IDF dapat ditentukan dengan rumus perhitungan diatas.

PENUTUP

Simpulan

Dari hasil pengujian dari aplikasi yang dibangun pada tugas akhir ini dapat diambil beberapa kesimpulan sebagai berikut :

1. Dengan banyaknya dokumen yang telah dibuat maka dibutuhkan aplikasi untuk mempermudah manajemen dokumen-dokumen tersebut.
2. Aplikasi ini dapat membantu penyimpanan dokumen dalam bentuk digital yang dinilai dapat mengurangi tingkat kehilangan data yang bersifat penting.
3. Dengan fitur OCR pada aplikasi sangat membantu dalam menentukan informasi yang terdapat didalam dokumen. Dan proses menyimpan informasi menjadi lebih cepat.
4. Serta dilengkapi dengan algoritma pencarian yang cocok untuk menemukan kembali dokumen yang relevan dengan *query*.

Saran

Tentunya aplikasi ini sangat jauh dari kata sempurna, maka dapat diberikan saran-saran yang sekiranya dapat digunakan sebagai acuan dalam pengembangan lanjutan aplikasi. Saran-saran tersebut antara lain

1. Kedepannya diharapkan aplikasi dapat ditambahkan sistem untuk menganalisa dokumen dan menentukan sentimen yang terdapat dalam dokumen.
2. Diharapkan aplikasi dapat menerapkan konsep clustering ketika dokumen yang di-upload ke aplikasi menjadi sekian banyak.
3. Diharapkan juga dalam pengembangan kedepannya dapat ditambahkan algoritma machine learning dalam proses pencarian dokumen sehingga aplikasi dapat belajar seiring dengan bertambahnya jumlah data.

DAFTAR PUSTAKA

Agusta, Y. (2007). *K-Means - Penerapan, Permasalahan dan Metode Terkait.* Jurnal Sistem dan Informatika.

Arif, A. Y. (2018, Desember 15). *Pengertian Indeks adalah: Jenis, Ciri dan Pembagian*. Retrieved Maret 20, 2019, from Rocket Manajemen: https://rocketmanajemen.com/definisi-indeks/

Basuki, A. (2005). *Metode Numerik dan Algoritma Komputasi.* Yogyakarta: Andi.

Darma Putra. (2010). *Pengolahan Citra Digital.* Yogyakarta: ANDI.

Fathansyah. (2018). *Basis Data.* Bandung: Penerbit Informatika.

Hermawati, F. A. (2013). *Data Mining.* Yogyakarta: Andi.

Korde, V., & Mahender, C. N. (2012, March). TEXT CLASSIFICATION AND CLASSIFIERS: A SURVEY. *International Journal of Artificial Intelligence & Applications, 3*, 2.

Kusumadewi, S. (2003). *Artificial Intelligence (Teknik dan Aplikasinya). Edisi Pertama.* Jakarta: Graha Ilmu.

Larose, D. T. (2005). *Discovering Knowledge in Data : An Introduction to Data Mining.* John Willey & Sons, Inc.

MacQueen, J. B. (1967). *Some Methods for classification and Analisys of Multivariate Observation, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probabilty.* California: University of California Press.

Melita, R., Amrizal, V., Suseno, H. B., & Dirjam, T. (2018, Oktober). Penerapan Metode Term Frequency Inverse Document Frequency (TF-IDF) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (STUDI KASUS: SYARAH UMDATIL AHKAM). *JURNAL TEKNIK INFORMATIKA, 11*(2).

newbieputrab13. (n.d.). *Pengertian Framework: Scribd*. Retrieved Maret 20, 2019, from Scribd: https://www.scribd.com/doc/52982287/PENGERTIAN-FRAMEWORK

Phangtriastu, M. R. (2017, Juli 03). *Binus University Article*. Retrieved from Binus University Website: http://mti.binus.ac.id/2017/07/03/optical-character-recognition-ocr/

Prasetyo, E. (2012). *Data Mining Konsep dan Aplikasi Menggunakan Matlab.* Yogyakarta: Andi.

Rao, V., Sastry, A., Chakracarthy, A., & Kalyanchakravarthi, P. (2016). Optical Character Recognition Technique Algorithms. *Journal of Theoretical and Applied Information Technology*, 275-282.

Sutojo, & Siswanto. (2004). *Membangun Citra Perusahaan.* Jakarta: Damar Mulia Pustaka.

Sutojo, T., Mulyanto, E., & Suhartono, V. (2011). *Kecerdasan Buatan, Edisi 1.* Yogyakarta: Andi.

Sutoyo, T., & dkk. (2009). *Teori Pengolahan Citra Digital.* Yogyakarta: Andi.

Tan, & dkk. (2006). *Tahapan Knowledge Discovery in Database.* Retrieved from http://informatika.web.id/category/data-mining/

Whitten, J. L., & Bentley, L. D. (2007). *System Analysis and Design Methods - 7th Edition.* New York: McGraw-Hill.