The Digitalisation of Financial Documents

Student: Bilț Florin Alexandru
Profesor: Sabo Cosmin Nicolae
Departamentul de informatica
Universitatea Tehnică din Cluj Napoca

Centrul Universitar Nord din Baia Mare
Baia Mare, România

***Abstract*— The Digitalisation of Financial Documents** is the results of a study which aims to extract content from scanned documents using the artificial intelligence as well as its implementation in a web platform with we can create a good management but also a simple accessibility, wherever and whenever, just a click away.

Keywords— management, inteligenta artificiala, management, accesibilitate

I. INTRODUCERE

Aplicația pe care aș dori să v-o prezint astăzi este o platformă web, cu ajutorul căreia, realizarea evidențelor tuturor documentelor dar și transformarea acestora din format fizic în format digital este mult mai simplă și mai eficientă.

Principalul motiv pentru care mi-am ales această temă este necesitatea de a avea acest soft la îndemână, acest soft care să îmi poată oferi oricând și oriunde accesul la toate documentele importante.

De asemenea, lipsa pe piață a unui asemenea produs, precum și cererea mare venită din partea marilor companii, m-au făcut să realizez ca nu sunt singurul care are o problemă de acest tip. Astfel, aceste motive au definitivat alegerea mea în ceea ce privește tema prezentată.

II. TESSERACT (SOFTWARE)

A. *Descriere*

Tesseract este un motor optic de recunoaștere a caracterelor pentru diverse sisteme de operare. Este un software gratuit , lansat sub Licența Apache. Dezvoltat inițial de Hewlett-Packard ca software privat în anii '80, mai târziu a fost lansat ca open source în 2005, iar dezvoltarea a fost sponsorizată de Google din 2006.

În 2006, Tesseract a fost considerat unul dintre cele mai precise motoare OCR cu sursă deschisă, disponibile atunci.

B. *Istoric*

Motorul Tesseract a fost inițial dezvoltat la laboratoarele Hewlett Packard din Bristol, Anglia și Greeley, Colorado, între 1985 și 1994, cu alte modificări făcute în 1996 către port c Windows și unele migrații de la C la C ++ în 1998.

O parte a codului a fost scris în C , iar apoi alte câteva au fost scrise în C ++. De atunci, tot codul a fost convertit în cel puțin compilat cu un compilator C ++.

S-a lucrat foarte puțin în deceniul următor. A fost apoi lansat ca sursă deschisă în 2005 de către Hewlett Packard și Universitatea din Nevada, Las Vegas (UNLV).

C. *Caracteristici*

- Tesseract a fost în primele trei motoare OCR în ceea ce privește precizia de caractere în 1995. Este disponibil pentru Linux , pentru Windows și Mac OS X . Cu toate acestea, datorită resurselor limitate, acesta este testat riguros de dezvoltatori sub Windows și Ubuntu .

- Tesseract, până la și inclusiv versiunea 2, ar putea accepta doar imagini TIFF cu text simplu dintr-o coloană ca date de intrare. Aceste versiuni timpurii nu includeau analiza machetei și, prin urmare, introducerea de text, imagini sau ecuații cu mai multe coloane a produs rezultate eronate . Începând cu versiunea 3.00, Tesseract acceptă formatarea textului de ieșire, informații de poziție hOCR și analiză de aspect al paginii. Suportul pentru o serie de noi formate de imagini a fost adăugat folosind biblioteca Leptonica . Tesseract poate detecta dacă textul este monospațiat sau distanțat proporțional.

- Versiunile inițiale ale Tesseract ar putea recunoaște doar textul în limba engleză. Tesseract v2 a adăugat șase limbi occidentale suplimentare (franceză, italiană, germană, spaniolă, portugheză braziliană, olandeză). Versiunea 3 a extins în mod semnificativ suportul lingvistic pentru a include limbi ideografice (chineză și japoneză, arabă, ebraică), precum și multe alte scripturi. Noi limbi includ arabă, bulgară, catalană, chineză (simplificată și tradițională), croată, cehă, daneză, germană (Frakturscript), greacă, finlandeză, ebraică, hindi, maghiară, indoneziană, japoneză, coreeană, letonă, lituaniană, norvegiană, poloneză, portugheză, română, rusă, sârbă, slovacă (script standard și Fraktur), slovene, suedeză, tagalog, tamil , thailandeză, turcă, ucraineană și vietnameză. V3.04, lansat în iulie 2015, a adăugat 39 de combinații de limbă / script adiționale, aducând numărul total de limbi de asistență la peste 100. Noi coduri lingvistice incluse: amh (amharic), asm (Assamese), aze_cyrl (Azerbaidjan în chirilic script), bod (tibetan), bos (bosniacă), ceb (cebuano), cym (galez), dzo (dzongkha), fas (persan), gle (irlandez), guj (gujaian), pălărie (cretă haitiană și haitiană), iku (Inuktitut), jav (javanez), kat (georgian), kat_old (veche georgiana), kaz (kazah), khm (Central Khmer), kir (kirgiz), kur (kurdu), lao (Lao), lat (lat.), mar (Marathi),[10]

- În plus, Tesseract poate fi instruit pentru a lucra în alte limbi dar si pentru a îmbunătății rezultatele . Atenție la antrenarea excesivă care dupa un anumit prag, va duce la scăderea performanțelor. Tesseract poate procesa text de la dreapta la stânga, cum ar fi araba sau ebraica, multe scripturi Indic, precum și CJK destul de bine .
- Tesseract este potrivit pentru a fi folosit ca backend și poate fi utilizat pentru sarcini OCR mai complicate, inclusiv analiza machetei folosind un frontend, cum ar fi OCRopus .

Ieșirea Tesseract va avea o calitate foarte slabă dacă imaginile de intrare nu sunt preprocesate pentru a se potrivi cu aceasta: Imaginile (în special capturile de ecran) trebuie să fie reduse astfel încât înălțimea x a textului să fie de cel puțin 20 de pixeli, orice rotație sau oblic trebuie corectată. sau nu va fi recunoscut niciun text, modificările de luminozitate de frecvență scăzută trebuie să fie filtrate în trepte mari sau etapa de binarizare a lui Tesseract va distruge o mare parte a paginii, iar marginile întunecate trebuie eliminate manual sau vor fi interpretate greșit ca si caractere.

D. *Version 4*

Versiunea 4 adaugă motor și model OCR bazat pe LSTM pentru multe limbi și scripturi suplimentare, aducând totalul la 116 limbi.

În plus, scripturile pentru 37 de limbi sunt acceptate, astfel încât este posibil să recunoașteți o limbă folosind scriptul în care este scris.

III. ÎMBUNĂTĂȚIREA CALITĂȚII PRODUCȚIEI

Există o varietate de motive pentru care nu puteți obține o calitate bună de la Tesseract. Este important să rețineți că, cu excepția cazului în care utilizați un font foarte neobișnuit sau un limbaj nou, Tesseract are șanse minime în a vă ajuta fără o preprocesare.

Procesarea imaginii:

- Rescaling
- Binarisation
- Înlăturarea zgomotului
- Dilatație / eroziune
- Rotire
- Border
- Transparență / canal Alpha
- Instrumente / Biblioteci

A. *Procesarea imaginii*

Tesseract efectuează diverse operații de procesare a imaginilor pe plan intern (folosind biblioteca Leptonica) înainte de a efectua OCR-ul propriu-zis. În general, face o treabă foarte bună în acest sens, dar inevitabil vor exista cazuri în care nu este suficient de bun, ceea ce poate duce la o reducere semnificativă a preciziei.

Puteți vedea modul în care Tesseract a procesat imaginea folosind setând variabila de configurare tessedit_write_images setata la true(sau folosind configfile get.images) atunci când rulați Tesseract. Dacă tessinput.tif (fișierul rezultat) pare problematic, încercați unele dintre aceste operațiuni de procesare a imaginii înainte de a trece imaginea la Tesseract.

B. *Invertirea imaginilor*

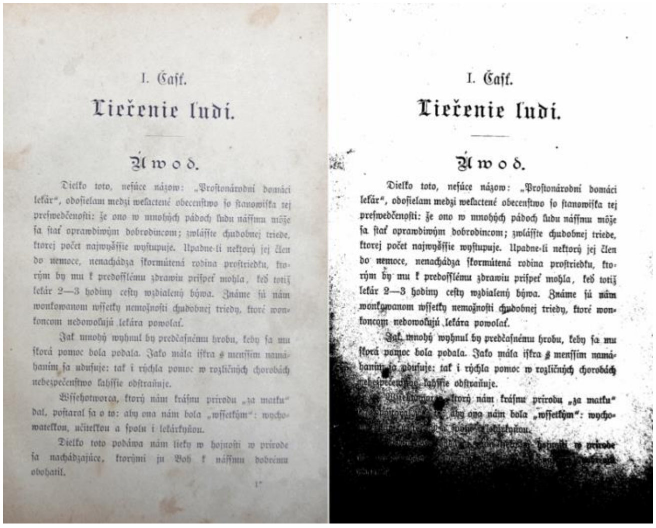
În timp ce tesseract versiunea 3.05 (și mai veche) se ocupă cu o imagine inversată (fundal închis și text deschis) fără probleme, pentru versiunea 4.x folosiți text întunecat pe fundal deschis.

C. *Rescaling*

Tesseract funcționează cel mai bine la imaginile care au un DPI de cel puțin 300 dpi, astfel încât poate fi benefică redimensionarea imaginilor. Cu cât dpi-ul este mai mare, cu atât rezultatele vor fi mai bune dar și timpul de procesare va creste.

„Willus Dotkom” a făcut un test interesant pentru rezoluția optimă a imaginii, cu sugestia pentru înălțimea optimă a literelor majuscule în pixeli.

D. *Binarisation*


I. Časť.

Liečenie ľudí.

Úvod.

Dielko toto, nesúce názow: „Prostonárodní domáci lekár“, odosielam medzi weľactené obecenstwo so stanowiska tej preswedčenosti: že ono w mnohých pádoch ľudu nášmu môže sa stať oprawdiwým dobrodincom; zwlášte chudobnej triede, ktorej počet najwyšše wystupuje. Upadne-li niektorý jej člen do nemoce, nenachádza skormútená rodina prostriedku, ktorým by mu k predošlému zdrawiu prispeť mohla, keď totiž lekár 2—3 hodiny cesty wzdialený býwa. Známe sú nám wonkowanom wšetky nemožnosti chudobnej triedy, ktoré wonkoncom nedowoľujú lekára powolať.

Jak mnohý wyhnul by predčasnému hrobu, keby sa mu skorá pomoc bola podala. Jako malá iskra s menším namáhaním sa udusuje: tak i rýchla pomoc w rozličných chorobách nebezpečenstwo ľahšie odstraňuje.

Wšehotworca, ktorý nám krásnu prírodu „za matku“ dal, postaral sa o to: aby ona nám bola „wšetkým“: wychowateľkou, učiteľkou a spolu i lekárkyňou.

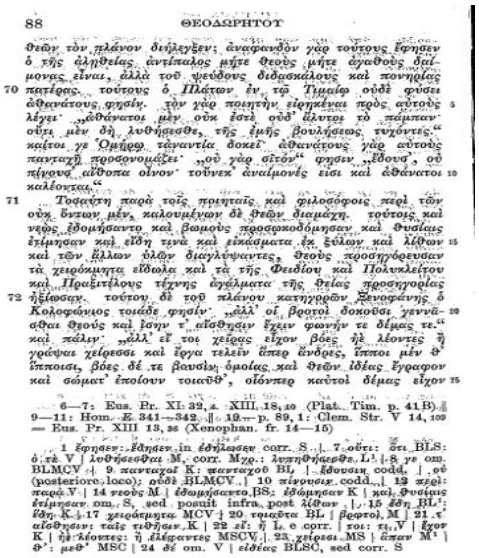
Dielko toto podáwa nám lieky w hojnosti w prírode sa nachádzajúce, ktorými ju Boh k nášmu dobrému obohatil.

1*

Aceasta transformă o imagine în alb și negru. Tesseract face acest lucru intern (algoritmul Otsu), dar rezultatul poate fi suboptim, în special dacă fundalul paginii este întunecat neuniform.

Dacă nu se poate să remediați acest lucru oferind o imagine de intrare mai bună, puteți încerca un algoritm diferit cum ar fi ImageJ (java) sau OpenCV Image Thresholding (python) sau OpenCV Image Thresholding (python) or scikit-image Thresholding .

E. *Înlăturarea zgomotului(Noise)*

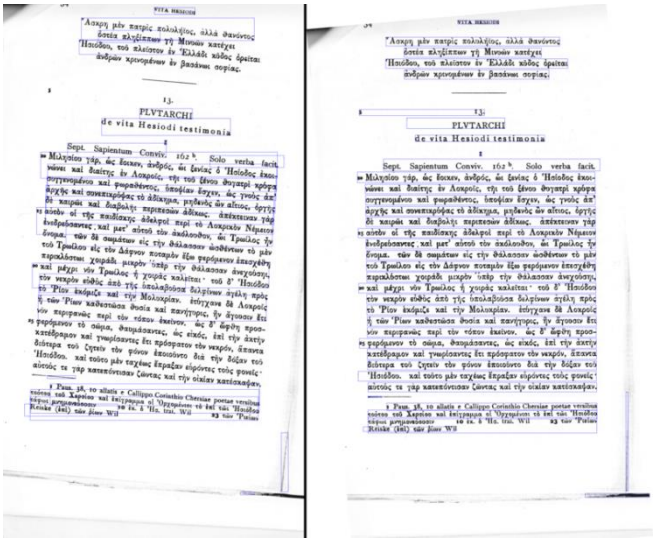


Zgomotul(Noise) este o variație aleatoare de luminozitate sau culoare într-o imagine, care poate face textul imaginii mai dificil de citit. Anumite tipuri de zgomot nu pot fi înlăturate de Tesseract în etapa de binarizare, ceea ce poate determina scăderea ratelor de precizie.

F. *Dilatarea și eroziunea*

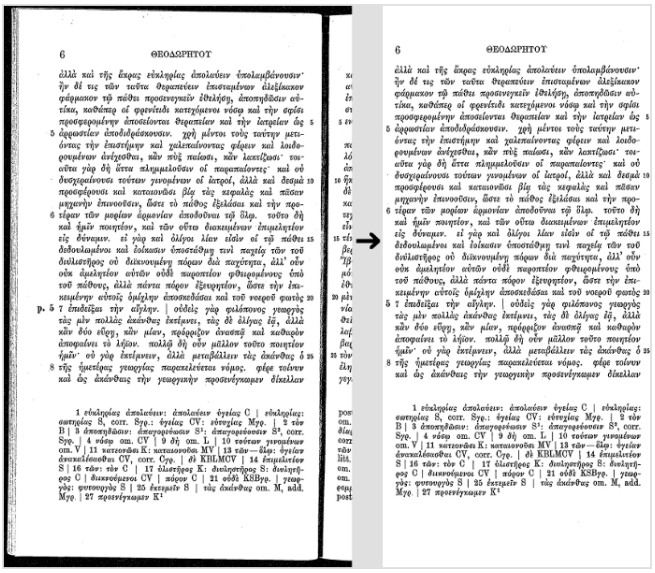
Caractere "îndrăznețe" sau caractere subțiri pot afecta recunoașterea caracterelor. Multe programe de procesare a imaginilor permit Dilatarea și eroziunea marginilor caracterelor pe un fundal comun să se dilate sau să crească în dimensiune (Dilatare) sau să se micșoreze (Eroziune).

E. *Rotire / deschewing*



Calitatea segmentării liniei Tesseract se reduce semnificativ dacă o pagină este prea slabă, ceea ce are un impact sever asupra calității OCR. Pentru a aborda acest lucru, rotiți imaginea, astfel încât liniile de text să fie orizontale.

G. *Borders*



Paginile scanate au adesea margini întunecate în jurul lor. Acestea pot fi intelese în mod eronat ca si caractere suplimentare, mai ales dacă acestea variază ca formă și gradare

H. *Transparență / canal Alpha*

Unele formate de imagine (de ex. Png) pot avea un canal alfa pentru a oferi o caracteristică de transparență.

Tesseract 3.0x se așteaptă ca utilizatorii să elimine canalul alfa din imagine înainte de a utiliza imaginea din tesseract. Acest lucru se poate face de exemplu cu comanda ImageMagick:

convert input.png -alpha off output.png

Tesseract 4.00 elimină canalul alfa cu funcția leptonica pixRemoveAlpha () : elimină componenta alfa amestecând-o cu un fundal alb. În unele cazuri (de exemplu, OCR a subtitrărilor de filme) acest lucru poate duce la probleme, astfel încât utilizatorii ar trebui să elimine canalul alfa (sau să preproceseze imaginea prin inversarea culorilor imaginii) de unul singur.

I. *Instrumente / Biblioteci*

Leptonica

OpenCV

ScanTailor Advanced

ImageMagick

unpaper

ImageJ

IV. SECURITATEA

Înainte de a implementa Tessaract într-o aplicație care poate intra în producție, este foarte important să cunoaștem câteva elemente de securitate mai ales dacă vorbim de clienți a căror date sunt foarte importante pentru o companie. Sunt multe metode de securitate, dar am să vă prezint câteva dintre cele mai importante:

Criptarea datelor

Criptarea – este un proces de codificare a informației astfel încât să poată fi înțeleasă doar de persoanele autorizate. Totodată, reprezintă și o metodă simplă de protejare a datelor cu caracter sensibil. În cazul în care o persoana neautorizată reușește să ajungă la baza de date, datele colectate nu îi vor fi utille atâta timp cât nu știe și metoda de decriptare, astfel reușim să protejăm datele clienților .

Token

Token-ul este un concept care presupune că fiecare persoană autentificată va primi un token care reprezintă o serie de date codificate astfel încât să se poată demonstra autenticitatea persoanei . Fiecare token este unic și are o semnatură specială, ceea ce crește nivelul de securitate.

Totuși este bine de știut că acest token nu ar trebui sa fie valabil mereu, și este foarte important ca acesta să expire dupa un anumit interval de timp (cateva ore).

Refresh Token

Este un alt concept care presupune ca la fiecare acțiune pe care utilizatoul o face (dupa validarea token-ului) acesta va primi un nou token cu nou valabilitate de timp astfel se evita deconectarea forțată a utilizatorului și riscul de nefinisare a muncii depuse. Astfel se creează un token dinamic.

Nivele de securitate

Este foarte importat ca înainte de fiecare acțiune pe care utilizatorul dorește să o facă, să se verifice autenticitatea lui, aceste verificări pot fi puse atât in partea de frontend cât și în partea de backend.

V. BIBLIOGRAFIE

[1] Google (2008). „tesseract-ocr”. Preluat 2016-03-08 .

[2] "Comunicări - tesseract-ocr / tesseract" . Preluat 5 ianuarie 2020 - prin GitHub .

[3] Kay, Anthony (iulie 2007). "Tesseract: un motor de recunoaștere a caracterului optic cu sursă deschisă". Jurnalul Linux . Preluat 28 septembrie 2011.

[4] Vincent, Luc (august 2006). „Anunțarea TCR-ului Tesseract”. Arhivat dela originalpe 26 octombrie 2006. Preluat 26.06.2008 .

[5] Canonical Ltd. (februarie 2011). „OCR”. Preluat 2011-02-11 .

[6] Anunțarea Tesseract OCR- blogul oficial Google

[7] Willis, Nathan (septembrie 2006). „Motorul Tesseract OCR al Google este un salt cuantic înainte” . Preluat 2008-07-18 .

[8] Rice Ștefan V., Frank R. Jenkins și Thomas A. Nartker A patra probă anuală de precizie OCR , expervision.com, preluată la 21 mai 2013

[9] Proiect Tesseract (februarie 2011). "Problema 263: patch pentru a activa ieșirea hOCR" . Arhivat de la original pe 13 noiembrie 2012 . Preluat 26 februarie 2011 .

[10] "langdata - Date de instruire sursă pentru Tesseract pentru multe limbi" . Preluat 6 noiembrie 2016 .

[11] "Instruirea rețelelor LSTM pe 100 de limbi și rezultatele testelor" (PDF) . Preluat 18 martie 2018 .

[12] Anunțarea sistemului OCRopus OCR Open Source (Thomas Breuel, OCRopus Project Leader).

[13] "FAQ - tesseract-ocr - Întrebări frecvente - Un motor OCR care a fost dezvoltat la HP Labs între 1985 și 1995 ... și acum la Google. - Google Project Hosting" . Preluat 2014-05-30 .

[14] "ImproveQuality - tesseract-ocr - Sfaturi pentru îmbunătățirea calității producției dvs. - Un motor OCR care a fost dezvoltat la HP Labs între 1985 și 1995 ... și acum la Google. - Google Project Hosting" . 2014-01-27 . Preluat 2014-05-30 .

[15] "TESSERACT (1) Pagina manuală" . Preluat 15 martie 2018 .

[16] Cod Google - Tesseract Readme

[17] "3rdParty - tesseract-ocr - GUI și alte proiecte folosind Tesseract OCR" . github.com . Preluat 2017-03-30 .

[18] "OCRFeeder" . Wiki GNOME . Preluat 12 ianuarie 2019 .

[19] OnDemand, HPE Haven. "OCR Document".

[20] OnDemand, HPE Haven. "undefined".

[21] Schantz, Herbert F. (1982). The history of OCR, optical character recognition. [Manchester Center, Vt.]: Recognition Technologies Users Association. ISBN 9780943072012.

[22] Dhavale, Sunita Vikrant (March 10, 2017). Advanced Image-Based Spam Detection and Filtering Techniques. Hershey, PA: IGI Global. p. 91. ISBN 9781683180142. Retrieved September 27, 2019.

[23] d'Albe, E. E. F. (July 1, 1914). "On a Type-Reading Optophone". Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences. 90 (619): 373–375. Bibcode:1914RSPSA..90..373D. doi:10.1098/rspa.1914.0061.

[24] "The History of OCR". Data Processing Magazine. 12: 46. 1970.

[25] "Extracting text from images using OCR on Android". June 27, 2015.

[26] "[Tutorial] OCR on Google Glass". October 23, 2014.

[27] Qing-An Zeng (October 28, 2015). Wireless Communications, Networking and Applications: Proceedings of WCNA 2014. Springer. ISBN 978-81-322-2580-5.

[28] "[javascript] Using OCR and Entity Extraction for LinkedIn Company Lookup". July 22, 2014.

[29] "How To Crack Captchas". andrewt.net. June 28, 2006. Retrieved June 16, 2013.

[30] "Breaking a Visual CAPTCHA". Cs.sfu.ca. December 10, 2002. Retrieved June 16, 2013.

[31] John Resig (January 23, 2009). "John Resig – OCR and Neural Nets in JavaScript". Ejohn.org. Retrieved June 16, 2013.

[32] Tappert, C. C.; Suen, C. Y.; Wakahara, T. (1990). "The state of the art in online handwriting recognition". IEEE Transactions on Pattern Analysis and Machine Intelligence. 12 (8): 787. doi:10.1109/34.57669.

[33] "Optical Character Recognition (OCR) – How it works". Nicomsoft.com. Retrieved June 16, 2013.

[34] Sezgin, Mehmet; Sankur, Bulent (2004). "Survey over image thresholding techniques and quantitative performance evaluation" (PDF). Journal of Electronic Imaging. 13 (1): 146. Bibcode:2004JEI....13..146S. doi:10.1117/1.1631315. Retrieved May 2, 2015.

[35] Gupta, Maya R.; Jacobson, Nathaniel P.; Garcia, Eric K. (2007). "OCR binarisation and image pre-processing for searching historical documents" (PDF). Pattern Recognition. 40 (2): 389. doi:10.1016/j.patcog.2006.04.043. Retrieved May 2, 2015.

[36] Trier, Oeivind Due; Jain, Anil K. (1995). "Goal-directed evaluation of binarisation methods" (PDF). IEEE Transactions on Pattern Analysis and Machine Intelligence. 17 (12): 1191–1201. doi:10.1109/34.476511. Retrieved May 2, 2015.

[37] Milyaev, Sergey; Barinova, Olga; Novikova, Tatiana; Kohli, Pushmeet; Lempitsky, Victor (2013). "Image binarisation for end-to-end text understanding in natural images" (PDF). Document Analysis and Recognition (ICDAR) 2013. 12th International Conference on. Retrieved May 2, 2015.

[38] Pati, P.B.; Ramakrishnan, A.G. (May 29, 1987). "Word Level Multi-script Identification". Pattern Recognition Letters. 29 (9): 1218–1229. doi:10.1016/j.patrec.2008.01.027.

[39] "Basic OCR in OpenCV | Damiles". Blog.damiles.com. November 20, 2008. Retrieved June 16, 2013.

[40] Ray Smith (2007). "An Overview of the Tesseract OCR Engine" (PDF). Retrieved May 23, 2013.

[41] "OCR Introduction". Dataid.com. Retrieved June 16, 2013.

[42] "How OCR Software Works". OCRWizard. Retrieved June 16, 2013.

[43] "The basic pattern recognition and classification with openCV | Damiles". Blog.damiles.com. November 14, 2008. Retrieved June 16, 2013.

[44] "How does OCR document scanning work?". Explain that Stuff. January 30, 2012. Retrieved June 16, 2013.

[45] "How to optimize results from the OCR API when extracting text from an image? - Haven OnDemand Developer Community".

[46] Fehr, Tiff, How We Sped Through 900 Pages of Cohen Documents in Under 10 Minutes, Times Insider, The New York Times, March 26, 2019

[47] "Train Your Tesseract". Train Your Tesseract. September 20, 2018. Retrieved September 20, 2018.

[48] "What is the point of an online interactive OCR text editor? - Fenno-Ugrica". February 21, 2014.

[49] Riedl, C.; Zanibbi, R.; Hearst, M. A.; Zhu, S.; Menietti, M.; Crusan, J.; Metelsky, I.; Lakhani, K. (February 20, 2016). "Detecting Figures and Part Labels in Patents: Competition-Based Development of Image Processing Algorithms". International Journal on Document Analysis and Recognition. 19 (2): 155. arXiv:1410.6751. doi:10.1007/s10032-016-0260-8.

[50] "Code and Data to evaluate OCR accuracy, originally from UNLV/ISRI". Google Code Archive.

[51] Holley, Rose (April 2009). "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs". D-Lib Magazine. Retrieved January 5, 2014.

[52] Suen, C.Y.; Plamondon, R.; Tappert, A.; Thomassen, A.; Ward, J.R.; Yamamoto, K. (May 29, 1987). Future Challenges in Handwriting and Computer Applications. 3rd International Symposium on Handwriting and Computer Applications, Montreal, May 29, 1987. Retrieved October 3, 2008.

[53] Sarantos Kapidakis, Cezary Mazurek, Marcin Werla (2015). Research and Advanced Technology for Digital Libraries. Springer. p. 257. ISBN 9783319245928. Retrieved April 3, 2018.

[54] https://en.wikipedia.org/wiki/Optical_character_recognition

[55] https://en.wikipedia.org/wiki/Tesseract_(software)

[56] ^ SetThings (6 septembrie 2009). „Dezvoltarea web”. SetThings.com. Accesat în 12 ianuarie 2018.

[57] ^ „Retail e-commerce sales CAGR forecast in selected countries from 2016 to 2021”. Statista. octombrie 2016. Accesat în 1 ianuarie 2018.

[58] ^ SetThings (10 martie 2009). „Industria dezvoltării web”. SetThings.com. Accesat în 12 ianuarie 2018.

[59] ^ Bureau of Labor Statistics, U.S. Department of Labor. „Information Security Analysts, Web Developers, and Computer Network Architects”. Occupational Outlook Handbook, 2012-13 Edition. Accesat în 17 ianuarie 2013.

[60] „LEMP Stack (Linux, Nginx, MySQL, PHP)”. lemp.io. Accesat în 15 octombrie 2014.

[61] Kawa, Arkadiusz (2017). „Fulfillment Service in E-Commerce Logistics”. Logforum. 13 (4): 429–438. doi:10.17270/J.LOG.2017.4.4.

[62] SetThings (15 martie 2016). „Modalități de comunicare a cunoașterii”. SetThings.com. Accesat în 12 ianuarie 2018.

[63] Davoust, Emmanuel. "A hundred years of science at the Pic du Midi Observatory". arXiv:astro-ph/9707201

[64] Engineering Web Applications", by Sven Casteleyn, Florian Daniel, Peter Dolog and Maristella Matera, Springer, 2009, ISBN: 978-3-540-92200-1

[65] SetThings (15 martie 2016). „Considerente de securitate în dezvoltarea web”. SetThings.com. Accesat în 1 august 2014.

[66] „Information Security: A Growing Need of Businesses and Industries Worldwide”. University of Alabama at Birmingham Business Program. Accesat în 20 noiembrie 2014.