Evaluacija OCR sustava

Sumpor, Josipa

SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI
Ak. god. 2021./2022.

Josipa Sumpor

Evaluacija OCR sustava

Završni rad

Mentor: dr. sc. Hrvoje Stančić, red. prof.
Neposredni voditelj: dr. sc. Željko Trbušić

Zagreb, rujan, 2022

Izjava o akademskoj čestitosti

Izjavljujem i svojim potpisom potvrđujem da je ovaj rad rezultat mog vlastitog rada koji se temelji na istraživanjima te objavljenoj i citiranoj literaturi. Izjavljujem da nijedan dio rada nije napisan na nedozvoljen način, odnosno da je prepisan iz necitiranog rada, te da nijedan dio rada ne krši bilo čija autorska prava. Također izjavljujem da nijedan dio rada nije korišten za bilo koji drugi rad u bilo kojoj drugoj visokoškolskoj, znanstvenoj ili obrazovnoj ustanovi.

Sadržaj

1. Uvod

Arhivi, ali i druge kulturne institucije zahvaćene su procesom digitalizacije. Digitalizacijom se nastoji povećati dostupnost gradiva te olakšati pristup gradivu kroz unaprijed osmišljen tok rada. Kako bi papirnato gradivo koje je digitalizirano moglo biti dostupno za korištenje, nije ga dovoljno skenirati i pohraniti na neki sustav. Slika gradiva nije pretraživa te, neposredno nakon skeniranja, ne govori ništa o samom gradivu. U teoriji, postavljanjem velikog broja takvih slika na neki sustav bilo bi nemoguće navigirati kroz njih te pronaći odgovarajuće gradivo. Iz tog se razloga digitaliziranoj inačici gradiva dodjeljuju metapodaci prije samog pohranjivanja u sustav. Metapodaci se mogu dodijeliti ručnim opisom gradiva, no moguće je provesti i automatizirani proces generiranja metapodataka primjenom optičkog prepoznavanja znakova (engl. *Optical character recognition, OCR*). Primjenom OCR-a nad digitaliziranim gradivom, tekst slike postaje pretraživ i dostupan za korištenje.

U praksi, provođenje OCR-a i dobivanje kvalitetnih rezultata predstavlja izazov. Cijeli proces provođenja optičkog prepoznavanja znakova provodi se kroz nekoliko faza, a odluke koje se donesu tijekom provođenja tih faza znatno utječu na kvalitetu rezultata odnosno točnost prepoznatog teksta. Proces provođenja optičkog prepoznavanja znakova uključuje i odabir programa čija pouzdanost prepoznavanja znakova može značajno varirati. Na točnost, također, utječe i kvaliteta digitalizacije. Metodama obrade slike prilagođenim OCR programima nastoji se povećati točnost rezultata provedenog OCR-a. Pouzdana istraživanja koja se temelje na testiranju različitih OCR programa, ali i primjeni metoda za obradu slika za uspješnije rezultate optički prepoznatog teksta od velike su važnosti jer, u mnoštvu programskih rješenja koja su dostupna danas, postoji velika vjerojatnost za odabir lošijeg rješenja što bi u konačnici moglo rezultirati ispodprosječnim rezultatima.

2. Optičko prepoznavanje znakova

Optičko prepoznavanje znakova je proces u kojem se, putem softvera te namjene, tiskani tekst pretvara u računalno kodirani tekst čija je glavna karakteristika mogućnost daljnje računalne obrade. Optičko prepoznavanje znakova često se pogrešno poistovjećuje s optičkim prepoznavanjem rukopisa. Optičko prepoznavanje znakova provodi se nad tiskanim tekstom čiji su pojedinačni znakovi dosljednih oblika, dok se pojedinačni znakovi unutar rukopisnog teksta razlikuju prema veličini i obliku. Iz tog razloga, provođenje procesa optičkog prepoznavanja rukopisa puno je zahtjevniji proces te nije istovjetan procesu optičkog prepoznavanja znakova.

Postoje komercijalni OCR programi koji se plaćaju i nekomercijalni OCR programi koji su besplatni.Primjeri komercijalnih programa za optičko prepoznavanje znakova su Google Cloud Vision, Microsoft Azure Computer Vision, Abbyy FineReader itd. Primjeri nekomercijalnih programa otvorenog koda za optičko prepoznavanje tekstova su Calamri, Kraken, SwiftOCR, Tesseract i sl. OCR programi razlikuju se prema dodatnim opcijama poput mogućnosti za obradu slike, jezicima i pismima nad kojima je moguće provesti prepoznavanje znakova, podržanim operacijskim sustavima, razini točnosti i jednostavnosti korištenja.

Današnji OCR sustavi temelje se na neuronskim mrežama. Sustavi se treniraju na uzorcima istih slova. „Strojno učenje se izvodi na način da se stroju prikazuju primjeri znakova svih klasa. Na temelju ovih primjera OCR sustav gradi prototip ili opis svake klase znakova“.[1] Ovakvo treniranje prepoznavanja uzoraka istih slova pospješuje točnost prepoznavanja kod drugačijeg oblika istog slova ili kod oštećenih slova.

[1] Eikvil, Line. Optical Character Recognition. Dec. 1993. Dostupno na: https://www.academia.edu/6214026/OCR_Optical_Character_Recognition_OCR_-Optical_Character_Recognition (2.8.2022.)

3. Povijest OCR-a

Prve ideje vezane uz razvoj OCR-a javljaju se već početkom 19. stoljeća. „Tehnologija optičkog prepoznavanja znakova izumljena je početkom 1800-ih, kada je patentirana kao pomagalo za čitanje za slijepe.“[2] Prva imena koja vežemo uz OCR su C.R. Carey i P.G. Nipkow. „Godine 1870. C. R. Carey patentirao je sustav za prijenos slike pomoću fotoćelija, a 1890. godine, P.G. Nipkow izumio je sekvencijalno skeniranje OCR-om.“[3]

„U početku je optičko prepoznavanje znakova (OCR) razvijeno kako bi tekst mogli „čitati“ oni s poteškoćama u čitanju, zadatak koji je vršio optofon Edmunda Edwarda Fourniera d'Albea (1910.) koji je pretvarao znakove u zvukove“.[4] Pojavom prvih računala i modernizacijom obrade podataka, javljala se sve veća potreba za primjenom OCR-a u obliku u kojem nam je danas najpoznatiji. Imena koja se vežu uz prve OCR programe su Gustav Tauschek i Paul Handel. Gustav Tauschek je 1929. godine dobio patent za OCR u Njemačkoj, dok je Paull Handel dobio američki patent za OCR u SAD-u *1933. godine.* Gustav Tauschek zaslužan je za prvu komercijalnu instalaciju OCR sustava za tvrtku Reader's Digest. Takav stroj mogao je pretvoriti strojopisne tekstove o prodaji u bušene kartice. Raymond Kurzeweil je 1976. godine izumio stroj koji je pretvarao tekst u zvučni zapis te ga je nazvao Kurzweil Reading Machine. „Stroj obrađuje običan tiskani materijal– slike, knjige, pisma, izvješća, memorandume itd., s najčešćim stilovima i tipovima veličina, a dobiveni rezultat je sintetički glas na engleskom jeziku.“[5]

Krajem 1980-ih, izrađuje se sve više OCR programa za osobna računala. Abbyy Finerader, Omnipage, Tesseract, Google Document OCR samo su neki od brojnih koji su se počeli pojavljivati. Jedan od prvih i najpoznatijih OCR programa je OmniPage OCR program za optičko prepoznavanje znakova. Razvila ga je tvrtka Caere Corporation. Ruska tvrtka Abbyy, koja je zaslužna za jedan od najpoznatijih OCR programa današnjice

[2] Awel, M. A., Abidi, A. I. Review on Optical Character Recognition. *International Research Journal of Engineering and Technology* 6, br. 6 (2019).
[3] Ibid.
[4] Romein C.A., Kemman, M., Birkholz. J. M., Baker, J., Gruijter, M., Merono-Penuela, A., Ries, R., Ros, R., Scagliola, S. State of the Field: Digital History. *The Journal of the Historical Association History* 365, br. 105: 197-376 (2020). Dostupno na: https://onlinelibrary.wiley.com/doi/10.1111/1468-229X.12969 (2.8.2022.)
[5] Kleiner, A., Kurzweil, R. C. A Description of the Kurzweil Reading Machine and a Status Report on its Testing and Dissemination. Dostupno na: https://www.rehab.research.va.gov/jour/77/14/1/kleiner.pdf (2.8.2022)

- Abbby FineReader, osnovana je 1989. godine. Prvi OCR program tvrtke Abbyy javnosti je bio dostupan 1993. godine. Nekomercijalni OCR program Tesseract, kojeg je razvila tvrtka Hewlett-Packard postaje dostupan kao sustav otvorenog koda 2005. godine. Google, u sklopu Google Drive-a, nudi opciju besplatnog optičkog prepoznavanja znakova *Google Document* od 2015. godine.

S popularizacijom osobnih računala, danas se na tržištu nude brojni OCR računalni programi, ali i za mobilne aplikacije. Uz to, primjena OCR-a nije ograničena samo na optičko prepoznavanje dokumenata već se danas primjenjuje i za prepoznavanje registarskih tablica, za prepoznavanje raznih podataka na čekovima, putovnicama, računima itd.

Prema L. Eikvilu, glavnom istraživaču na Odjelu za statističku analizu, analizu slike i prepoznavanje uzoraka (SAMBA) Norveškog računalnog centra, povijest razvoja OCR-a moguće je podijeliti u tri faze tj. generacije OCR-a.

Prva generacija OCR-a traje od 1960. do 1965. godine. „Ova generacija OCR strojeva uglavnom je karakterizirana s ograničenim oblicima slova koje su mogli pročitati.“[6] Komercijalizacijom strojeva, pojavio se velik broj različitih fontova koje su stvarale problem kod optičkog prepoznavanja znakova što je dovelo do daljnjeg razvoja strojeva za optičko prepoznavanje znakova.

Druga generacija OCR-a, trajala je između 1965. i 1975. godine. Ovu generaciju obilježava veliki pomak u vidu tehnoloških mogućnosti prepoznavanja znakova. Osim što su strojevi mogli prepoznati strojne zapise, neki su strojevi mogli prepoznati i rukopis. „No, u ovom periodu prepoznavanje rukopisa je bilo ograničeno na brojeve, nekoliko slova i simbola.“[7] Prvi i najpoznatiji takav stroj iz tog vremena je IBM 1287.[8] „IBM 1287

[6] Eikvil, L. Optical Character Recognition“, n. dj., str. 9.
[7] Ibid.
[8] Ibid.

predstavljen je na 'svjetskom sajmu' u New Yorku.“[9] Adrian Frutiger je 1968. godine razvio font koji je nazvao OCR-B. „Proglašen je međunarodnim standardom za optičko prepoznavanje kod elektroničkih računala i uključuje brojke, velika i mala slova i određene povezane simbole.“[10]

Nakon Druge generacije OCR-a slijedi Treća generacija OCR-a čime započinje treća faza razvoja koja traje do 1985. godine. „Na tržištu su se počeli pojavljivati sofisticiraniji OCR strojevi, no jednostavniji OCR uređaji su i dalje bili vrlo korisni.“[11] Jednostavniji OCR uređaji bili su posebno korisni za strojopisne tekstove, no pojavom prvog osobnog računala 1971. godine započela je nova era u kojoj se značajno povećala potreba za razvojem OCR sustava. OCR strojevi su do završetka Treće generacije OCR-a bili iznimno skupi i nisu bili namijenjeni osobnoj uporabi, no završetkom te faze i zahvaljujući tehnološkom napretku, postupno započinje komercijalizacija OCR sustava te postaju cjenovno prihvatljiviji i dostupniji široj javnosti.

[9] Memon, J., Sami, M., Khan, R.A., Uddin, M. Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR). *IEEE Access*, Vol. 8 (2020). Dostupno na: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9151144 (2.8.2022)
[10] Frutiger, A., Delamarre, N., & Gurtler, A. (Collaborators). (1967). OCR-B: A standardized character for optical recognition. *Journal of Typographic Research, 1*(2), 137–146.
[11] Eikvil, L. Optical Character Recognition“, n. dj., str. 11.

4. Preporuke za provođenje optičkog prepoznavanja znakova

Smjernice za digitalizaciju kulturne baštine ključne su smjernice za provođenje optičkog prepoznavanja znakova kojima Ministarstvo kulture potiče institucije na korištenje OCR tehnologija prilikom digitalizacije jer se time omogućuje brzo i efikasno pretraživanje sadržaja u digitaliziranom obliku. „S ciljem uspostave i poticanja sustavnog i ujednačenog pristupa digitalizaciji građe u kulturnim ustanovama i privatnim zbirkama te stvaranja jedinstvenih i usklađenih normi u postupku digitalizacije, u okviru projekta „e-Kultura – Digitalizacija kulturne baštine“ izrađene su smjernice za digitalizaciju kulturne baštine.“[12] Smjernice za digitalizaciju kulturne baštine objavljene su 11. listopada 2021. godine. Smjernice se, između ostalog, dotiču postupaka koji prethode provođenju optičkog prepoznavanja znakova te postupaka vezanih uz odabir formata nakon provedenog optičkog prepoznavanja znakova nad gradivom.

Prema Smjernicama, slika koja će biti optički prepoznata treba biti u TIFF formatu. TIFF format razvio je Aldus 1986. godine i kratica je za *tagged image file format. „*Tagged se odnosi na kompliciranu strukturu formata. Nakon početnog zaglavlja podataka datoteke slijede „komadići“ podataka koji se nazivaju „tagovi“ i koji prenose informacije o slici programu koji prikazuje datoteku.“[13] Zbog toga što TIFF podržava veliki raspon standardiziranih slikovnih formata, a pod tim se podrazumijeva i široki spektar veličina slika, rezolucija i dubine boja, te zbog kompresije bez gubitka koja osigurava da slika ne gubi kvalitetu prilikom uređivanja ili spremanja, TIFF se koristi kao standardni format u arhivima.

Nadalje, preporučuje se 300 ppi (do 600 ppi max za sitan tekst). PPI (engl. *pixels per inch*) je kratica za piksele po inču. Odnosi se na gustoću piksela u digitalnoj slici. Često se koristi i termin DPI (engl*. dots per inch*) koji je kratica za točke po inču. Odnosi se na raspon točaka na postojećoj (isprintanoj) slici.

[12] Smjernice za digitalizaciju kulturne baštine. Dostupno na: https://min-kulture.gov.hr/vijesti-8/objavljene-smjernice-za-digitalizaciju-kulturne-bastine/21484 (2.8.2022.)
[13] Wiggins, R. H., Christian Davidson, H., Ric Harnsberger, H., Lauman, J. R., Goede, P. A. Image File Formats: Past, Present, Future. *RadioGraphics* 21, br. *3* (2001). Dostupno na: https://pubs.rsna.org/doi/full/10.1148/radiographics.21.3.g01ma25789 (2.8.2022.)

Preporučuje se 8-bitna siva skala (1 bit b/w za kvalitetan tisak) koja sadrži 256 nijansi sive boje. Takve slike razlikuju se od bitonalnih crno-bijelih slika koje sadrže samo dvije nijanse – crnu i bijelu.

Nakon optičkog prepoznavanja računalno kodirani tekst je potrebno odabrati i spremiti u jedan od tri preporučena formata: obični tekst (engl. *plain text*) koji je kodiran UTF-8 sustavom kodiranja znakova, PDF/A-1a ili XML TEI format. UTF-8 je sustav kodiranja koji koristi nizove od 8 bajtova. Za razliku od ASCII, podržava i one jezike čiji se znakovi ne nalaze unutar engleske abecede. Standard ISO 19005 definirao je formate datoteke PDF/A-1b i PDF/A-1a. „PDF/A-1b osnovna je razina i definira minimalni skup zahtjeva za pouzdanu reprodukciju vizualnog izgleda dokumenta, dok PDF/A-1a proširuje razinu B kako bi dodatno uključila značajke za poboljšanje pristupačnosti dokumenta i pouzdanog izdvajanja teksta.“[14] TEI (engl. *Text Encoding Initiative*) je inicijativa koja je pokrenuta 1987. godine i kojom su razvijene smjernice kojima su propisane oznake koje se koriste u TEI dokumentima kako bi takvi dokumenti bili strojno čitljivi.

[14] PDF Association. Dostupno na: https://www.pdfa.org/resource/iso-19005-pdfa/ (2.8.2022.)

5. Uloga OCR-a u digitalizaciji

Postoje brojne predrasude oko koncepta digitalizacije koji je izvan kruga stručnjaka koji se bave samim procesom nerijetko definiran ugrubo kao proces skeniranja gradiva. Takva definicija je pogrešna jer digitalizacija obuhvaća različite vrsta gradiva, od papirnatog gradiva, pa sve do zvučnih i video zapisa te fizičkih 3D objekata. Nadalje, kod same digitalizacije papirnatog gradiva, skeniranje predstavlja tek jedan od početnih koraka u cjelokupnom procesu, a prethode mu koraci odabira i pripreme gradiva. Nakon što je određeno gradivo pretvoreno u digitalno gradivo, potrebna je dodatna obrada provođenjem optičkog prepoznavanja znakova i dodjeljivanjem metapodataka. „Rezultat OCR-a omogućuje korištenje naprednih tehnologija poput sintetiziranja teksta u govor, analize velikih količina podataka (engl. *big data*) i dubinske analize teksta bez potrebe za ručnim prepisivanjem ili ručnom izradom metapodataka.“[15] Potom započinje proces uspostavljanja zaštite nad gradivom u vidu dodavanja digitalnih potpisa, certifikata, vodenih žigova i sl. Pohranjivanjem digitaliziranog gradiva na neki sustav proces digitalizacije ne završava nego tek započinje. Zbog sve bržeg razvitka računalnih tehnologija, dolazi do ubrzanog zastarijevanja formata i računalno-programskih okolina što posljedično utječe i na gradivo koje je digitalizirano jer postaje neupotrebljivo. Stoga je najzahtjevniji izazov u procesu digitalizacije dugoročni proces očuvanja gradiva. Kao rješenje osmišljen je OAIS referentni model koji „opisuje potrebne komponente i usluge potrebne za razvoj i održavanje arhiva, kako bi se podržao dugoročni pristup i razumijevanje informacija u arhivima“[16]. Moguće je implementirati sustav za optičko prepoznavanje znakova tijekom prihvata gradiva u arhivske informacijske sustave. Primjer je digitalna knjižnica Hathi gdje je „cjelokupni sustav izgrađen u skladu s OAIS referentnim modelom“.[17]

[15] Trbušić, Ž. Mogućnosti implementacije sustava za optičko prepoznavanje znakova tijekom prihvata gradiva u arhivske informacijske sustave. // *Radovi 52. savjetovanja hrvatskih arhivista* Zagreb/Šibenik. Hrvatsko arhivističko društvo, 2020. Str. 215-235. Dostupno na: https://www.researchgate.net/publication/349723490_Mogucnosti_implementacije_sustava_za_opticko_prepoznavanje_znakova_tijekom_prihvata_gradiva_u_arhivske_informacijske_sustave (2.8.2022.)

[16] Lee, Christopher A. Open Archival Information System (OAIS) Reference Model. Encyclopedia of Library and Information Sciences, Third Edition. (2010). Dostupno na: https://ils.unc.edu/callee/p4020-lee.pdf (2.8.2022.)

[17] Trbušić, Ž. Mogućnosti implementacije sustava za optičko prepoznavanje znakova tijekom prihvata gradiva u arhivske informacijske sustave, n.dj.

6. Provođenje procesa optičkog prepoznavanja znakova

Provođenje procesa optičkog prepoznavanja znakova obuhvaća više faza u kojima je potrebno donijeti određene odluke koje će odrediti uspješnost tijeka samog procesa, a samim time odredit će i uspješnost cjelokupnog procesa digitalizacije. Prema Ž. Trbušiću, „proces optičkog prepoznavanja znakova korištenjem suvremenih tehnoloških dostignuća sastoji se od četiriju koraka: stvaranje slike teksta i prethodna obrada, optičko prepoznavanje znakova, naknadna obrada optički prepoznatog teksta i kontrola kvalitete“.[18] Prije stvaranja slike teksta tj. snimanja gradiva, potrebno je odabirati gradivo. Prema Smjernicama za odabir građe za digitalizaciju, „na odabir građe utječe niz okolnosti kao što su, primjerice, zadaće ustanove, autorsko pravo nad građom, vrsta i stanje izvornika koji se digitaliziraju te skupina korisnika kojima je projekt namijenjen i sl.“[19] Nakon što je utvrđeno koje će gradivo biti digitalizirano, prije snimanja gradiva potrebno je pripremiti gradivo za digitalizaciju. Priprema gradiva obuhvaća razne postupke poput utvrđivanja potpunosti gradiva, utvrđivanja redoslijeda, fizičke obrade gradiva poput uklanjanja spojnica i sl. Nakon što je gradivo odabrano i pripremljeno, započinje proces snimanja gradiva tj. stvaranja slike teksta.

6.1. Stvaranje slike teksta i prethodna obrada

Kako bi se stvorila slika teksta, gradivo se snima fotoaparatom ili skenerom. Najčešće se bira skener. Ovisno o gradivu, odabir će se vršiti između koračnih i protočnih skenera, a u obzir će se uzimati karakteristike skenera, poput brzine, razlučivosti (rezolucije), dinamičkog raspona, polja skeniranja, veznih uređaja, softvera za skeniranje i opsega za skeniranje. Potom će se gradivo snimiti te pohraniti u određeni format.

Prije provođenja optičkog prepoznavanja znakova provest će se prethodna obrada slike kojom se nastoji povećati točnost izlaznog teksta. Postoje brojne metode kojima se može povećati točnost. Neki primjeri metoda su uklanjanje nakošenosti, uklanjanje šuma binarizacija, stanjivanje. Metodom uklanjanja nakošenosti moguće je poravnati dokument koji je digitaliziran nakošeno, rotirajući ga u smjeru kazaljke na satu ili suprotno od smjera

[18] Trbušić, Ž. Mogućnosti implementacije sustava za optičko prepoznavanje znakova tijekom prihvata gradiva u arhivske informacijske sustave, n.dj.
[19] Smjernice za odabir građe za digitalizaciju. Dostupno na: https://bib.irb.hr/datoteka/590089.smjernice_odabir.pdf (2.8.2022.)

kazaljke na satu kako bi nastao okomit ili vodoravan dokument. Metodom uklanjanja šuma zagladit će se područja u kojima su vidljive određene smetnje, pritom ne utječući na tekst čime je moguće povećati točnost optičkog prepoznavanja znakova. Metodom binarizacije provodi se pretvorba slike u boji u crnu i bijelu boju čime se jasno razgraničava tekst od njegove pozadine. Metodom stanjivanja ujednačava se širina poteza između znakova. Primjena metoda je opcionalna.

Autor Ž. Trbušić navodi kako prilikom ovog koraka treba razmotriti dva često obrnuto proporcionalna elementa – vrijeme i kvalitetu. Pritom se treba primjenjivati pravilo "kvaliteta nad kvantitetom". „Ako je primarni cilj što veća točnost optički prepoznatog teksta, potrebno je izraditi pilot-testiranje koje će na manjemu uzorku identificirati optimalnu kvalitetu skeniranog teksta, ali i uključiti ispitivanje različitih kompresijskih algoritama i formata zapisa koji se susreću u procesu stvaranja slike"."[20] Na temelju rezultata pilot-projekta, odabire se optimalno rješenje s obzirom na zahtjeve projekta digitalizacije.

6.2. Optičko prepoznavanje znakova

Nakon što je gradivo snimljeno, moguće je provesti optičko prepoznavanje znakova. Potrebno je razmotriti odabir softvera za optičko prepoznavanje znakova koji su na raspolaganju. „Dvije glavne kategorije softvera kojima se koristi u suvremenom okruženju su: komercijalni sustavi i sustavi otvorenog koda."[21] Osim tog kriterija, u obzir treba uzeti koji operativni sustavi podržavaju odabrani OCR sustav, koje jezike i pisma podržava OCR sustav, koje tipove slova prepoznaje, u koje se izlazne formate optički prepoznati tekst može pohraniti, dodatne opcije koje se mogu odabrati vezane uz obradu slike te pouzdanost samog OCR programa.

[20] Trbušić, Ž. Mogućnosti implementacije sustava za optičko prepoznavanje znakova tijekom prihvata gradiva u arhivske informacijske sustave, n.dj.
[21] Ibid.

6.3. Naknadna obrada prepoznatog teksta

Prema autoru Ž. Trbušiću, „ciljevi naknadne obrade optički prepoznatog teksta su učiniti dobiveni prepoznati tekst istovjetnim originalnom predlošku, ekstrakcija metapodataka i stvaranje novog, efikasnog sustava prepoznavanja*“*. Naknadna obrada prepoznatog teksta se odnosi na korekciju grešaka nad izlaznim tekstom, nakon provedenog optičkog prepoznavanja znakova nad izvornim dokumentom. Greške, u vidu krivo prepoznatih znakova, uobičajene su i točnost prepoznavanja iznimno rijetko je stopostotna. Nerijetko je moguće naići i na gramatičke i pravopisne greške ili, pak, tiskarske pogreške u originalnom dokumentu, no te se greške ne smiju prepravljati kako bi se sačuvala vjerodostojnost dokumenta.

„Ručno ispravljanje može obaviti operater ili zaposlenik arhiva, ali [se može provesti] izradom sustava za korisničko označivanje, koji podrazumijeva “skupinu usluga koja korisnicima omogućuje da po vlastitom nahođenju odabranim informacijskim izvorima dodjeljuju oznake (tags), i to ad hoc i bez konzultiranja nekog postojećeg uređenog ili kontroliranog rječnika”.[22] „Automatsko ispravljanje ostvaruje se korištenjem računalnog rječnika i sustava za analizu teksta bez kontrole operatera.“[23] Nakon što je izlazni tekst naknadno obrađen, može se isporučiti u željenom formatu.

6.4. Kontrola kvalitete

Kontrola kvalitete može se provoditi usporedno sa sve tri prethodne faze. Kontrola kvalitete odnosi se na “pilot-testiranje”, tj. proces testiranja neke od faza na manjem uzorku gradiva te pronalazak optimalnog rješenja kojim bi se postigla što veća točnost u kombinaciji sa što manje utrošenog vremena i financijskih sredstava.

[22] Špirenec, S., Ivanjko, T. Korisničko označavanje tekstualnih i vizualnih informacija: što mogu očekivati AKM ustanove? U: *16. seminar Arhivi , knjižnice, muzeji: mogućnosti suradnje u okruženju globalne informacijske strukture: zbornik radova.* Hrvatsko knjižničarsko društvo: Zagreb, 2013.

[23] Trbušić, Ž. Mogućnosti implementacije sustava za optičko prepoznavanje znakova tijekom prihvata gradiva u arhivske informacijske sustave, n.dj.

7. Prednosti i nedostaci OCR sustava

Primjena optičkog prepoznavanja teksta nad digitaliziranim gradivom čest je postupak u arhivskoj praksi. Iako postupak ima mnogo prednosti, postoje i nedostaci, koji bi se drugim riječima mogli nazvati izazovima, s kojima se arhivisti suočavaju korištenjem OCR programa. Tijekom istraživanja, uočene su određene prednosti, ali i izazovi vezani uz OCR tehnologiju.

7.1. Prednosti OCR sustava

Optički prepoznati tekst je lako pretraživ. S obzirom na to da je svrha digitalizacije lakši pristup i korištenje gradiva, izrazito je bitno da gradivo bude pretraživo. U pretraživanju se može koristiti i samo dio neke riječi što je osobito pogodno kod pretrage onih riječi koje nisu u potpunosti točno optički prepoznate ili kod onih riječi koje sadrže tiskarske greške. Mogućnošću strojnog pretraživanja smanjuje se i vrijeme koje je potrebno utrošiti za pronalazak neke riječi unutar teksta.

Izlazni tekst koji se dobije provođenjem optičkog prepoznavanja znakova moguće spremiti u više različitih formata za pohranu teksta poput DOC, DOCX, TXT itd. Ovisno o korištenom programu, ali i namjeni, korisnik može birati spremanje u jedan ili više različitih formata.

Optičko prepoznavanje znakova provedeno nad nepretraživim formatima drastično smanjuje trošak koji bi obuhvaćao profesionalnog daktilografa koji bi ručno prepisivao tekst. Daktilografsko prepisivanje je dugotrajan i skup proces, a provođenje optičkog prepoznavanja može biti relativno brz proces i jeftiniji u odnosu na daktilografske usluge.

7.2. Nedostaci OCR sustava

Najveći nedostatak optičkog prepoznavanja teksta je netočnost izlaznog teksta. Postoje razni postupci kojima se nastoji poboljšati točnost izlaznog teksta, no taj problem je i dalje aktualan. „Konsenzus je da dobra stopa točnosti OCR-a iznosi između 98%-99% uspješno prepoznatih znakova.“[24] Na točnost izlaznog teksta utječe font, slični znakovi, fizičko stanje dokumenta, kvaliteta slike i sl. Prema Veileiu, glavne poteškoće na koje se nailazi u različitim dokumentima mogu se klasificirati kao varijacije oblika (zbog serifa i stilskih varijacija); deformacije uzrokovane isprekidanim znakovima, zamrljanim znakovima i mrljama; varijacije u razmacima (zbog subskripta i supSerskripta, kosih i promjenjivih razmaka; mješavine teksta i grafike).[25]

S obzirom na nedostatak točnosti, izlazni tekst je potrebno dodatno urediti. Uređivanje podrazumijeva zamjenu pogrešno prepoznatih znakova, brisanje nepostojećih znakova, dodavanje onih znakova koji nedostaju kao i umetanje i brisanje razmaka, prijeloma teksta i sl. Uređivanje može uključivati i grafičku obradu teksta poput dodavanja stilova, odabir vrste i veličine fonta i sl. Ovisno o količini gradiva, uređivanje može oduzeti puno vremena s obzirom na to da se obavlja ručno.

Cjelokupni proces optičkog prepoznavanja znakova odvija se u više faza te samim time oduzima dosta vremena. Prije njegova provođenja, proces je potrebno provesti nad manjim testnim uzorkom gradiva kako bi se moglo odlučiti o optimalnim rješenjima vezanim uz odabir programa, način snimanja gradiva, postavke skenera, odabir izlaznog formata i sl. O takvim odlukama ovisi trajanje cjelokupnog procesa digitalizacije, ali i uspješnost jer je svaki projekt digitalizacije ograničen budžetom i raspoloživim stručnim osobljem za provođenje digitalizacije. Iz tog razloga, provođenje takvog procesa nije isplativo nad manjim gradivom.

[24] Stančić, H., Trbušić, Ž., Optimisation.of archival processes involving digitisation of typewritten documents. *Aslib Journal of Information Management.* Vol. 72, br. 4, 2020.
[25] Eikvil, L. Optical Character Recognition“, n. dj., str. 28.

8. Provedeno istraživanje

Provedeno istraživanje obuhvaćalo je odabir uzoraka nad kojima se provelo optičko prepoznavanje znakova te odabir repozitorija iz kojeg su ekstrahirani uzorci. Nad odabranim uzorcima provedeno je optičko prepoznavanje znakova programima Abbyy FineReader i Tesseract. Istraživanje je uključivalo i provedbu evaluacije te su se za potrebe evaluacije izradile datoteke istovjetne izvornim uzorcima.

8.1. Odabir repozitorija

Uzorci su ekstrahirani iz digitalnog repozitorija HathiTrust. „HathiTrust je neprofitna suradnja akademskih i istraživačkih knjižnica koje čuvaju više od 17 milijuna digitaliziranih jedinica.“[26] Repozitorij je osnovan 2008. godine. Digitalizaciju jedinica provodi Google, Internet Archive, Microsoft te druge inicijative. Digitalizirane jedinice obuhvaćaju djela s javnom domenom, otvorenim pristupom i Creative Commons licencom. „Misija HathiTrust-a je doprinijeti istraživanju, učenju i općem dobru zajedničkim prikupljanjem, organiziranjem, očuvanjem, komuniciranjem i dijeljenjem zapisa ljudskog znanja.“[27]

„Reviziju sustava obavili su 2008.godine DCC-a (Digital Curation Centre) i DPE-a (Digital Preservation Europe) korištenjem DRAMBORA (Digital Repository Audit Method Based on Risk Assessment) revizijskog standarda, a godine 2007. (objavljeno 2011.) knjižnica je uspješno certificirana TRAC (Trust-worthy Respositories Audit & Certification) standardom.“[28] Kriteriji vrednovanja za TRAC su organizacijska struktura u kojoj je Hathi na skali od 1 do 5 zadovoljio ocjenom 2, upravljanje digitalnim objektima u kojoj je Hathi zadovoljio ocjenom 3 te tehnologije, tehnička infrastruktura i sigurnost u kojoj je Hathi zadovoljio ocjenom 4.

Na službenim stranicama HathiTrust repozitorija, navedeno je kako je „HathiTrust stvoren prema okviru za otvoreni arhivski informacijski sustav (OAIS)“[29]. Nadalje, navedeno je kako je i „u skladu sa Standardom za kodiranje i prijenosom metapodataka (METS)“.[30]

---

[26] HathiTrust: About. Dostupno na: https://www.hathitrust.org/about (2.8.2022.)
[27] HathiTrust: Mission Goals. Dostupno na: https://www.hathitrust.org/mission_goals (2.8.2022.)
[28] Center for Research Libraries: HathiTrust Audit Report 2011, https://www.crl.edu/reports/hathitrust-audit-report-2011 (2.8.2022.)
[29] HathiTrust: Digital object specifications. Dostupno na: https://www.hathitrust.org/digital_object_specifications (2.8.2022.)
[30] HathiTrust: Digital object specifications. Dostupno na: https://www.hathitrust.org/digital_object_specifications (2.8.2022.)

8.2. Ekstrakcija testnog uzorka iz repozitorija

Za testni uzorak odabrane su dvije knjige zbirki radova Instituta za oceanografiju[31]. Jedna knjiga potječe iz 1966. godine, a druga iz 1967. godine. Radovi unutar knjige su tiskani na različite načine, no odabrani su strojopisni radovi po uzoru na testne uzorke strojopisnih zapisnika Društva hrvatskih književnika iz časopisa *Aslib Journal of Information Management,* u članku pod nazivom *Optimisation of archival processes involving digitisation of typewritten documents* i čiji su autori H. Stančić i Ž. Trbušić.[32] Članci Društva hrvatskih književnika potječu iz istog vremenskog razdoblja te su, također, strojopisni radovi. Obje su odabrane knjige dostupne u varijanti koju je digitalizirao Google i koja je binarizirana te onoj koju je digitalizirao Internet Archive i koja je nebinarizirana. Testni uzorci preuzeti su u obje varijante. Odabrane uzorke je moguće preuzeti u 3 različita formata: Ebook (PDF), Text (.txt) i Image (.jpeg). Odabrani uzorci preuzeti su u .txt i .jpeg formatu. Pri dnu stranice nalazi se alatna traka pomoću koje je moguće upravljati i navigirati odabranim uzorkom. Unutar nje nalazi se opcija prikaza VIEW gdje je moguće odabrati prikaz uzorka kao slike opcijom JPEG ili optički prepoznate datoteke (.txt) opcijom Plain Text. Testni uzorci preuzeti su u .jpeg i .txt formatu. Iz zbirke radova Instituta za oceanografiju koja potječe iz 1966. godine, odabrano je 50[33] testnih uzoraka, a iz zbirke radova Instituta za oceanografiju koja potječe iz 1967.godine odabrano je 38[34] testnih uzoraka.

S obzirom na to da su preuzete JPEG datoteke sadržavale vodene žigove koje bi OCR sustavi prepoznali kao dio teksta, testne je uzorke bilo potrebno urediti. JPEG uzorci su odrezani 125 pixela od dna. Na par stranica došlo je i do odrezivanja teksta izvornika, no obje varijante su i dalje ostale identične.

8.3. Izrada datoteka istovjetnih izvorniku

Datoteka istovjetna izvorniku odnosno temeljni tekst (engl. *ground truth)* je potpuna i točna transkripcija svakog znaka i riječi sukladno izvorniku. Koristi se za provjeru točnosti

[31] Collected reprints Institute for Oceanography
[32] Stančić, H., Trbušić, Ž., Optimisation.of archival processes involving digitisation of typewritten documents. *Aslib Journal of Information Management.* Vol. 72, br. 4, 2020.
[33] 25 stranica za varijantu Google, 25 stranica za varijantu Internet Arhive
[34] 19 stranica za varijantu Google, 19 stranica za varijantu Internet Archive

izlaznih datoteka OCR sustava. Datoteku istovjetnu izvorniku moguće je izraditi na dva načina: direktnim prepisivanjem teksta sa slike ili korištenjem izlaznih datoteka OCR sustava te naknadnim, ručnim ispravljanjem greški. Spremaju se kao tekstualna datoteka (.txt) te se koriste zajedno s ekstrahiranim OCR-om u alatima za evaluaciju.

Datoteke su izrađene korištenjem izlaznih datoteka te ručnim ispravljanjem. U uzorcima su postojali određeni simboli, tj. posebni znakovi koje nije bilo moguće unijeti u datoteku .txt formata te se umjesto njih koristila tilda kako ju alati za evaluaciju ne bi brojali kao pogrešku.

9. Programi korišteni u istraživanju

Programi koji su korišteni u istraživanju su Abbyy FineReader, Tesseract, ImageMagick, ISRI evaluacijski alati te programski kod napisan u Pythonu. Abbyy FineReader i Tesseract su korišteni za provođenje optičkog prepoznavanja znakova. ImageMagick je korišten za binarizaciju slika. ISRI evaluacijski alati su korišteni za evaluaciju točnosti izlaznog teksta dobivenog prethodno provedenim optičkim prepoznavanjem znakova. Programski kod je napisan u svrhu ekstrakcije podataka iz izlaznih datoteka dobivenih ISRI evaluacijom radi lakše i brže analize evaluacije.

9.1. Abbyy FineReader 15

Abbyy FineReader jedan je od najpoznatijih komercijalnih OCR programa današnjice. Na licencu za korištenje programa za Windows operacijski sustav moguće se pretplatiti mjesečno ili godišnje te godišnje za Mac operacijski sustav. Moguće je odabrati verziju za osobnu upotrebu (standardna) ili verziju za korporacije (korporativnu) ukoliko se radi o licenci za Windows operacijski sustav, dok za Mac operacijski sustav postoji samo jedna prilagođena licenca. Detaljan popis razlika između verzije za osobnu upotrebu i verzije za korporacije i plan plaćanja dostupan je na njihovoj službenoj stranici.[35] Moguće ga je i besplatno isprobati u periodu od 7 dana. Tijekom tog perioda primjena OCR-a moguća je na sveukupno 100 stranica te se nudi mogućnost uređivanja PDF dokumenata, komentiranja i uspoređivanja.

Ovaj alat omogućuje „digitalizaciju, dohvaćanje, uređivanje, zaštitu, dijeljenje i surađivanje na raznim vrstama dokumenata u istom radnom tijeku“.[36] Prilikom pokretanja ovog programa, izdvojena je mogućnost pregleda i uređivanja PDF formata te mogućnost konvertiranja gdje se nudi opcija konvertiranja u formate: PDF, Microsoft Word i Microsoft Excel. Dodatno se nudi i opcija konvertiranja u druge formate. Podržava 193 jezika te nije dostupan za Linux.

Abbyy FineReader PDF izrazito je popularan uredski alat koji je višestruko nagrađivan od kojih je najrecentnija nagrada ona iz 2022. godine kada je proglašen drugim najboljim uredskim alatom na listi najboljih globalnih prodavača *G2 annual Best Office Product.*

[35] Abbyy FineReader PDF. Dostupno na: https://pdf.abbyy.com/pricing/ (20.9.2022.)
[36] Abbyy FineReader PDF. Dostpuno na: https://pdf.abbyy.com/finereader-pdf/ (2.8.2022.)

Lista *G2 annual Best Office Product* rangira 50 najboljih proizvoda na temelju autentičnih, pravovremenih recenzija stvarnih korisnika.[37]

Provođenje optičkog prepoznavanja znakova korištenjem programa Abbyy FineReader vrlo je jednostavno. Na alatnoj traci nalazi se gumb otvori (engl. *open*) te se nakon toga odabire slika koju se želi optički prepoznati. Program prepoznaje dijelove stranice npr. tekst, tablicu, sliku te ih dijeli linijama. Samu stranicu moguće je i ručno podijeliti na tekst, tablicu, sliku i sl. S desne strane se pojavljuje prozor u kojem je prikazan optički prepoznati tekst. Takav tekst je moguće spremiti odabirom na gumb spremi te odabirom vrste formata u kojem ga želimo pohraniti.

9.2. Tesseract

Tesseract je jedan od poznatijih nekomercijalnih OCR programa otvorenog koda. „Tesseract je izvorno razvijen u laboratorijima Hewlett-Packard Laboratories Bristol i Hewlett Packard Co, Greeley Colorado između 1985. i 1994., s nekim promjenama napravljenim 1996. godine za prijenos na Windows i za programiranje u C++ 1998. godine.“[38] Krajem 2005. godine, HP je objavio Tesseract kao sustav otvorenog koda.[39] Dostupan je pod licencom Apache 2.0. Podržava preko 100 jezika, no prethodno je potrebno instalirati jezik koji će se koristiti. Također, moguće je i trenirati sustav u svrhu prepoznavanja jezika koji nije podržan. Trenutno je dostupan za Linux i Windows operativne sustave. Program je moguće preuzeti s GitHuba[40]. Provođenje optičkog prepoznavanja znakova provodi se putem komandne linije

U istraživanju provedenom na Odsjeku za Inženjering Sveučilišta Southern Maine, uočeno je da program ima brojna ograničenja. „Tijekom istraživanja otkriveno je da su

[37] Abbyy FineReader PDF. Dostpuno na: https://pdf.abbyy.com/blog/finereader-pdf-earns-2-ranking-on-g2-2022-best-software-awards-for-office-products/ (2.8.2022.)
[38] Github: Tesseract. Dostupno na: https://github.com/tesseract-ocr/tesseract (2.8.2022.)
[39] Patel, C. I., Patel, D. Optical Character Recognition by Open source OCR Tool Tesseract: A Case Study. *International Journal of Computer Applications.* Vol 55, br. 10 (2012), str. 50. Dostupno na: https://www.researchgate.net/profile/Chirag-Patel-12/publication/235956427_Optical_Character_Recognition_by_Open_source_OCR_Tool_Tesseract_A_Case_Study/links/00463516fa43a64739000000/Optical-Character-Recognition-by-Open-source-OCR-Tool-Tesseract-A-Case-Study.pdf (2.8.2022.)
[40] GitHub. Dostupno na (20.9.2022.): https://github.com/UB-Mannheim/tesseract/wiki (20.9.2022.):

određeni fontovi bolje prihvaćeni u odnosu na druge te da veličina fonta, razmak i kvaliteta slike igraju ulogu u tome koliko dobro Tesseract prepoznaje tekst.“[41]

Za provođenje optičkog prepoznavanja znakova korištenjem Tesseract OCR programa potrebno je otvoriti odredišnu mapu u kojoj je instalacijski paket instaliran. Unutar odredišne mape nalazi se konzola (engl. *Console*), pomoću koje se provodi optičko prepoznavanje znakova. Unutar odredišne mape nalaze se i dokumentacija (engl. *Documentation*), često postavljena pitanja (engl. *FAQ*), početna stranica (engl. *Homepage*), „Pročitaj me“ dokument (engl. *ReadMe*) te deinstalacijski paket (engl. *Uninstall*).
Otvaranjem konzole, moguće je unijeti naredbe.

Naredbe koje su korištene u ovom istraživanju su sljedeće:

a) *cd DesktopOCR/TesseractTest*
b) *tesseract xxxxxxx.tiff test1*

Naredba *cd* kratica je za *change directory* te služi za postavljanje radnog direktorija. Radni direktorij je onaj u kojem se nalaze slike nad kojima će se provesti optičko prepoznavanje znakova.
Pomoću naredbe *tesseract xxxxxxx.tiff test1* pokreće se optičko prepoznavanje znakova. Umjesto *xxxxxxx* treba navesti naziv slike, zajedno s ekstenzijom. *Test1* odnosi se na ime koje će Tesseract dodijeliti izlaznoj tekstualnoj datoteci koja će sadržavati optički prepoznati tekst u TXT formatu. Moguće je i postaviti i drugačiji format izlazne datoteke korištenjem naredbi.

9.3. ImageMagick

ImageMagick je nekomercijalan program kojim se obrađuju slike. Podržava velik broj slikovnih formata. Pogodan je za korištenje na raznim operacijskim sustavima, poput Windows, Linux, iOS, Android OS itd. „ImageMagick može promijeniti veličinu,

[41] Suitter, J. Acuracy of Optical Character Recognition sofware Google Tesseract. Dostupno na: https://digitalcommons.usm.maine.edu/cgi/viewcontent.cgi?article=1042&context=thinking_matters (2.8.2022.)

preokrenuti, zrcaliti, rotirati, izobličiti i transformirati slike, prilagoditi boje slike, primijeniti razne specijalne efekte ili crtati tekst, linije, poligone, elipse i Bézierove krivulje.“[42]
Poput Tesseracta, pokreće se pomoću komandne linije. Detaljan popis naredbi može se pronaći na službenoj stranici, na kartici *Annotated List of Command-line Options*[43]. Naredbe se pokreću ključnom riječi *magick* te odabranom naredbom s određenim parametrima.

Imagemagick korišten je za binarizaciju slika. Naredba koja je korištena je -threshold. Granična vrijednost (engl. *Threshold*) je postavljena na 50%.
U svrhu automatizacije procesa, naredba se koristila u sklopu datoteke s .bat ekstenzijom koja se pokretala iz CMD-a. Datoteka se sastojala od nekoliko naredbi:

a) *FOR %%Y IN ("*.jpg") DO magick "%%Y" -threshold 50%% "%%Y_b.jpg"*

b) *FOR %%Y IN ("*_b.jpg") DO rename "%%Y" "??????????_b.jpg"*

c) *FOR %%Y IN ("*_b.jpg") DO tesseract "%%Y" "%%Y_tesseract"*

d) *FOR %%Y IN ("*_tesseract.txt") DO rename "%%Y" "???????????_tesseract.txt"*

Komandom linijom *a*, postavljena je granična vrijednost na 50% te je na sliku dodan nastavak _b.jpg, kako bi se slika imenski razlikovala od slike nad kojom je provedena naredba, tj. kako ne bi zamijenila već postojeću, nego bila spremljena uz nebinariziranu sliku u odredišnoj mapi. To je ključno jer će se kasnije uspoređivati razina točnosti izlaznog teksta koji je optički prepoznat i s binariziranim i nebinariziranim datotekama.
Komandnom linijom *a* dobivene su datoteke koje su kao dio svog imena sadržavale .jpg i na koju je nadodana ekstenzija .jpg. Stoga, korištena je komandna linija *b* kojom su se datoteke preimenovale tako što bi dio imena bio zamijenjen s *_b*.

[42] ImageMagick. Dostupno na: https://imagemagick.org/index.php (2.8.2022.)
[43] ImageMagick. Dostupno na: https://imagemagick.org/script/command-line-options.php (20.9.2022.)

Komandnom linijom *c* pokrenuto je optičko prepoznavanje znakova Tesseractom nad binariziranim datotekama.
Komandnom linijom *d* izlaznim datotekama dodijeljen je dodatak na ime *_tesseract.txt*.

9.4. ISRI evaluacijski alati

ISRI evaluacijske alate razvio je Institut za informacijske znanosti Las Vegas, Sveučilišta u Nevadi 1996. godine.[44] Unatoč svojoj starosti, alati su i danas korisni za zadatke evaluacije novih OCR modela.[45] Za uspješnu evaluaciju potrebno je izlaznu datoteku optički prepoznatog teksta usporediti s datotekama istovjetnim izvorniku. Provođenjem evaluacije ISRI evaluacijskim alatima, dobiva se datoteka tekstualnog formata (.txt) koja, između ostalog, sadrži broj znakova u izlaznoj datoteci optički prepoznatog teksta, broj pogrešaka u odnosu na znakove datoteka istovjetnih izvorniku te već izračunat postotak točnosti na temelju tih podataka. Ovi podaci su ključni za daljnju analizu evaluacije OCR alata Abbyy FineReader i Tesseract.

9.5. Programski kod za lakšu obradu podataka

Kako podaci ne bi bili ručno upisivani u Excel datoteku, radi lakše obrade korišten je programski kod napisan u Pythonu pomoću kojeg su podaci iz izlazne datoteke dobivene provođenjem evaluacije ISRI evaluacijskim alatima zapisani u datoteku .csv formata. Podaci koji su korišteni za analizu su broj znakova u izlaznoj datoteci evaluacije optički prepoznatog teksta, broj pogrešaka u odnosu na znakove datoteka istovjetnih izvorniku te postotak točnosti na temelju znakova. Format .csv je odabran jer je otvaranjem tog formata u Excelu, ali po potrebi i u drugim programima, lako uvesti podatke koje sadrži te je time znatno olakšana daljnja obrada i analiza podataka. Programski kod je u cijelosti dostupan kao Prilog 1. Primjer izlazne datoteke dobivene evaluacijom dostupne su kao Prilog 2.

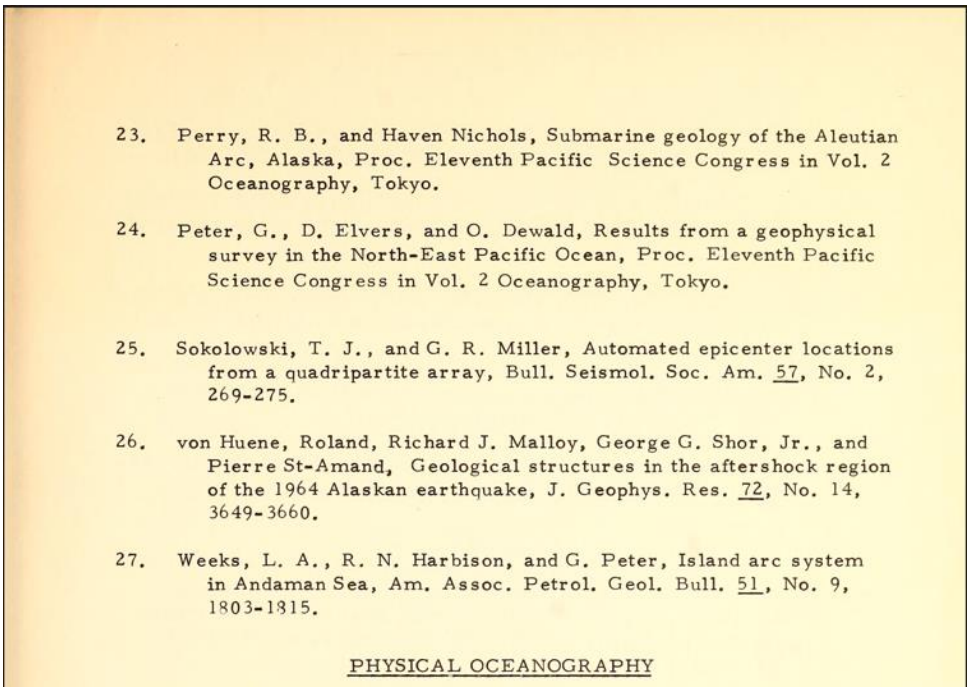
[44] Rice, S.V., Nartker, T.A. 1996. The ISRI analytic tools for OCR evaluation. *Information Science Research Institute* (1996). Dostupno na: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.216.9427&rep=rep1&type=pdf (2.8.2022.)
[45] Santos, E.A. OCR evaluation tools for the 21st century. *National Research Council Canada.* Dostupno na: https://aclanthology.org/W19-6004.pdf (2.8.2022.)

10. Čimbenici koji utječu na točnost OCR-a

Prema N. Andersonu i G. Muhlbergeru[46], čimbenici koji utječu na točnost OCR-a, mogu se podijeliti na negativne faktore koji proizlaze iz izvornog materijala (požutjeli papir, zgužvani papir, prijenos tinte, vidljivost tinte na obostrano ispisanom tankom papiru, lošem ispisu, korištenje više od jedne boje tinte, anotacije, manjak leksičkih podataka) te na negativne faktore koji proizlaze iz procesa snimanja slike (uski uvez, neobrezane slike, iskrivljenost slike).

10.1. Negativni faktori koji proizlaze iz izvornog materijala na odabranom uzorku gradiva

Na neobrađenim, nebinariziranim uzorcima Internet Archive-a, uočljivo je da papir nije klasične bijele boje. Nije jasno da li je papir izvorno takvog materijala ili je blagu promjenu u nijansi boje poprimio s vremenom, no na određenim uzorcima vidljivo je da je na rubovima nekih stranica došlo do promjene u boji. „Slika 1“



23. Perry, R. B., and Haven Nichols, Submarine geology of the Aleutian Arc, Alaska, Proc. Eleventh Pacific Science Congress in Vol. 2 Oceanography, Tokyo.

24. Peter, G., D. Elvers, and O. Dewald, Results from a geophysical survey in the North-East Pacific Ocean, Proc. Eleventh Pacific Science Congress in Vol. 2 Oceanography, Tokyo.

25. Sokolowski, T. J., and G. R. Miller, Automated epicenter locations from a quadripartite array, Bull. Seismol. Soc. Am. 57, No. 2, 269-275.

26. von Huene, Roland, Richard J. Malloy, George G. Shor, Jr., and Pierre St-Amand, Geological structures in the aftershock region of the 1964 Alaskan earthquake, J. Geophys. Res. 72, No. 14, 3649-3660.

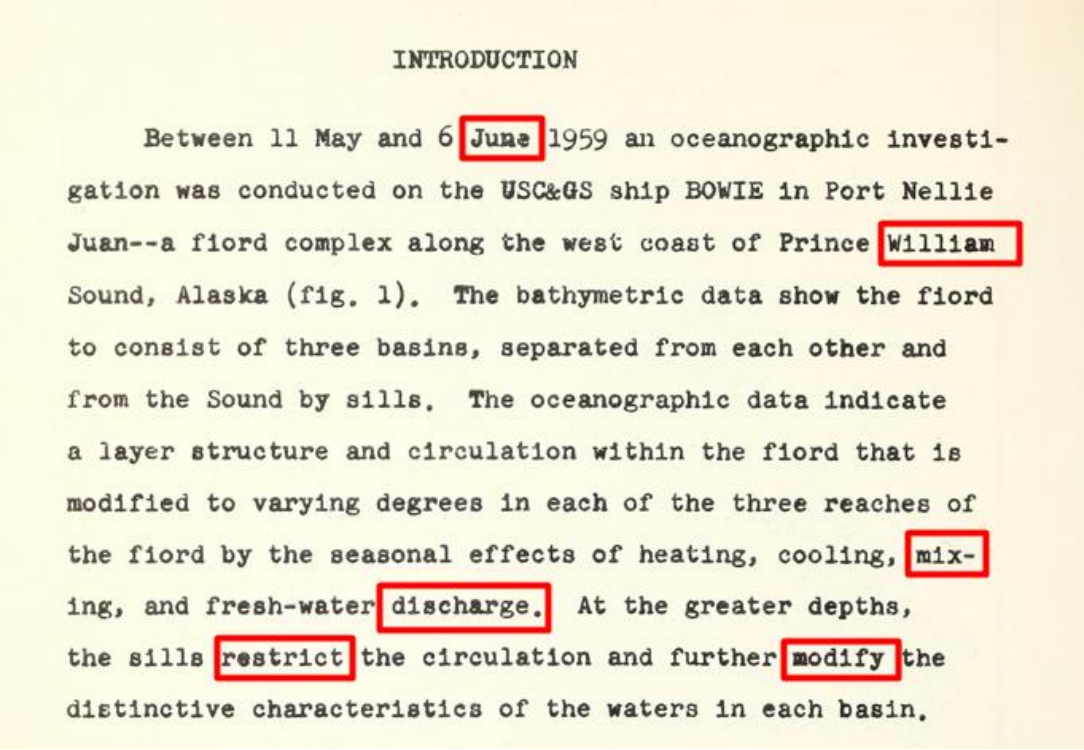
27. Weeks, L. A., R. N. Harbison, and G. Peter, Island arc system in Andaman Sea, Am. Assoc. Petrol. Geol. Bull. 51, No. 9, 1803-1815.

PHYSICAL OCEANOGRAPHY

Slika 1. Uzorak 1967_IA_03

[46] Anderson, N., Muhlberger, G.. Optical Character Recognition. Dostupno na: https://www.digitisation.eu/download/website-files/BPG/OpticalCharacterRecognition-IBPG_01.pdf (20.9.2022.)

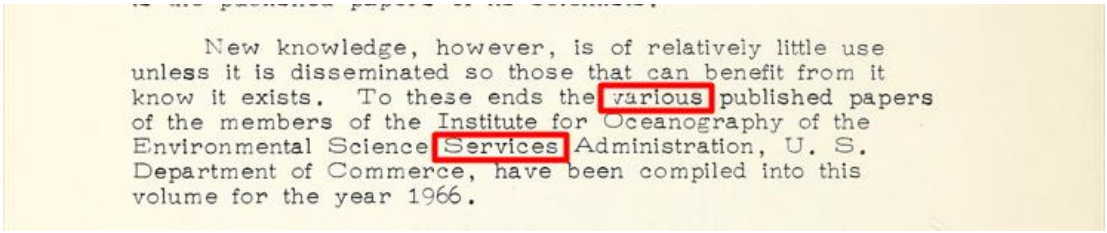
Papir uzoraka je poprilično očuvan te ne sadrži mrlje, niti promjene na papiru koje bi utjecale na jasnoću pojedinačnih znakova. U vezi s tim, vidljivo je da papir nije bio izložen vlazi te ne postoje uzorci čiji su znakovi nejednaki zbog zgužvanosti papira. S obzirom na to da je papir neproziran, ne postoji vidljivost tinte na obostrano isprintanim stranicama i, također, nije došlo do međusobnog prijenosa tinte s jedne stranice na drugu, koji se najčešće događa zbog toga što se stranice preklope prije nego što se tinta na njima osuši. Problem kod ispisa uključuje zamućene (Slika 2) prelomljene (Slika 3), izblijedjele znakove (Slika 4) te znakove koji su neujednačeno ispisani po pitanju širine samih linija znakova u odnosu na druge (Slika 5).



INTRODUCTION

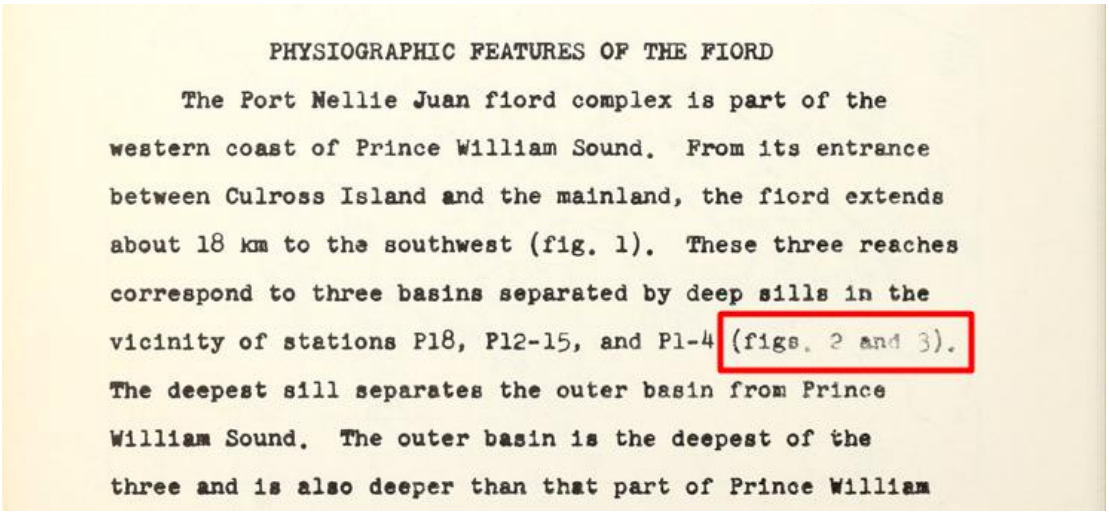
Between 11 May and 6 June 1959 an oceanographic investigation was conducted on the USC&GS ship BOWIE in Port Nellie Juan--a fiord complex along the west coast of Prince William Sound, Alaska (fig. 1). The bathymetric data show the fiord to consist of three basins, separated from each other and from the Sound by sills. The oceanographic data indicate a layer structure and circulation within the fiord that is modified to varying degrees in each of the three reaches of the fiord by the seasonal effects of heating, cooling, mixing, and fresh-water discharge. At the greater depths, the sills restrict the circulation and further modify the distinctive characteristics of the waters in each basin.

Slika 2. Uzorak 1966_IA_18



New knowledge, however, is of relatively little use unless it is disseminated so those that can benefit from it know it exists. To these ends the various published papers of the members of the Institute for Oceanography of the Environmental Science Services Administration, U. S. Department of Commerce, have been compiled into this volume for the year 1966.

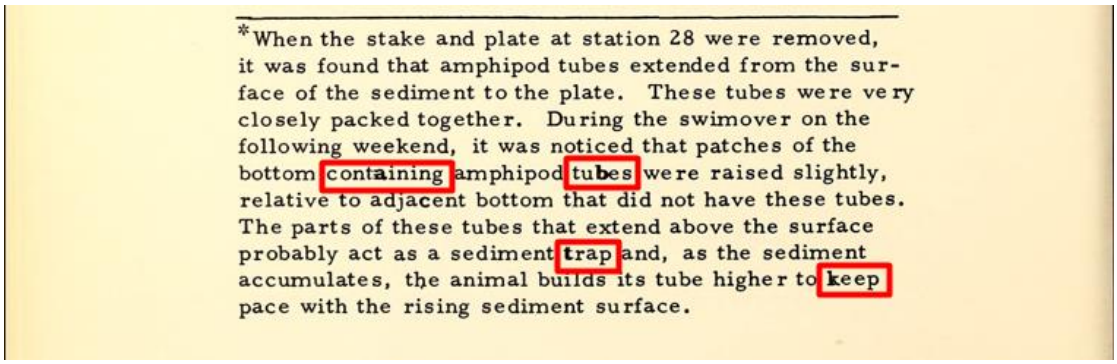
Slika 3. Uzorak 1966_IA_01



PHYSIOGRAPHIC FEATURES OF THE FIORD

The Port Nellie Juan fiord complex is part of the western coast of Prince William Sound. From its entrance between Culross Island and the mainland, the fiord extends about 18 km to the southwest (fig. 1). These three reaches correspond to three basins separated by deep sills in the vicinity of stations P18, P12-15, and P1-4 (figs. 2 and 3). The deepest sill separates the outer basin from Prince William Sound. The outer basin is the deepest of the three and is also deeper than that part of Prince William

Slika 4. Uzorak 1966_IA_18 b



*When the stake and plate at station 28 were removed, it was found that amphipod tubes extended from the surface of the sediment to the plate. These tubes were very closely packed together. During the swimover on the following weekend, it was noticed that patches of the bottom containing amphipod tubes were raised slightly, relative to adjacent bottom that did not have these tubes. The parts of these tubes that extend above the surface probably act as a sediment trap and, as the sediment accumulates, the animal builds its tube higher to keep pace with the rising sediment surface.

Slika 5. Uzorak 1967_IA_16

Prema N. Andersonu i G. Muhlbergeru, u anotacije spadaju bilješke i crteži korisnika, također i knjižnični štambilji i vodeni žigovi.[47] U odabranim uzorcima ne postoje bilješke i crteži korisnika, no postojali su vodeni žigovi koji su prije provođenja optičkog prepoznavanja znakova odstranjeni obrezivanjem slike. Prema N. Andersonu i G. Muhlbergeru manjak leksičkih podataka se događa kad OCR mehanizam nema pristup relevantnim jezičnim podacima za dokument. [48] Oba OCR programa, Abbyy FineReader i Tesseract podržavaju engleski jezik te pristup leksičkim podacima nije bio problem.

10.2. Negativni faktori koji proizlaze iz procesa snimanja slike

Prema N. Andersonu i G. Muhlbergeru, ponekad može doći do geometrijskog izobličenja zbog zategnutosti uveza ili nemogućnosti određenog skenera da snimi knjige koje se ne mogu otvoriti više od 60°.[49] Takva vrsta geometrijskog izobličenja nije uočena kod odabranih uzoraka. Također, uzorci na kojima je vidljivo da su rubovi promijenili boju tijekom vremena i koje je moguće usporediti s binariziranim uzrocima Google-a, nisu zahtijevali dodatnu obradu jer je boja ujednačena i jasno odudara od boje znakova. Iskrivljenost slika nije uočena u uzorcima.

[47] Digitisation EU. https://www.digitisation.eu/download/website-files/BPG/OpticalCharacterRecognition-IBPG_01.pdf str 11.
[48] Anderson, N., Muhlberger, G.. Optical Character Recognition. Dostupno na: https://www.digitisation.eu/download/website-files/BPG/OpticalCharacterRecognition-IBPG_01.pdf (2.8.2022.)
[49] Ibid.

11. Proces evaluacije

Ovo istraživanje napravljeno je po uzoru na istraživanje čiji su rezultati objavljeni u časopisu *Aslib Journal of Information Management* pod nazivom *Optimisation of archival processes involving digitisation of typewritten documents* i čiji su autori H. Stančić i Ž. Trbušić[50]. U tom istraživanju, optičko prepoznavanje znakova provedeno je nad strojopisnim zapisima *Društva hrvatskih književnika*. Programi koji su bili korišteni su Abbyy FineReader i Tesseract. Evaluacija rezultata provedena je s ISRI evaluacijskim alatima. Pretpostavke su bile da će točnost OCR rezultata biti veća s poboljšanjem kvalitete digitalizacije te primjenom metode binarizacije. Teze su odbačene.

U ovom radu, testiranje točnosti provedeno je nad datotekama JPEG i TXT formata dostupnim na digitalnom repozitoriju HathiTrust. Pretpostavka je da TXT datoteke sadrže optički prepoznati tekst dostupnih JPEG datoteka. Datoteke .txt formata evaluirane su ISRI evaluacijskim alatima kako bi se točnost optički prepoznatog teksta repozitorija HathiTrust mogla usporediti s rezultatima postignutima provođenjem optičkog prepoznavanja znakova programima Abbyy FineReader i Tesseract. Pritom treba napomenuti kako je u repozitoriju HathiTrust dostupna varijanta koju je digitalizirao Google i koja je binarizirana te varijanta koju je digitalizirao Internet Archive i koja je nebinarizirana. Stoga, ne postoji binarizirana varijanta Internet Archive-a nad kojom se mogla provesti evaluacija rezultata ISRI alatima. Rezultati evaluacije za svaku sliku zasebno dostupni su kao Prilog 3.

Evaluacija rezultata je, potom, provedena nad novostvorenim datotekama .txt formata koje su rezultat optičkog prepoznavanja znakova programom Abbyy FineReader i to nad datotekama .jpeg formata koja su binarizirana varijanta Google-a, nad datotekama .jpeg formata koja su nebinarizirana varijanta inicijative Internet Archive te nad binariziranom varijantom inicijative Internet Arhive čija je binarizacija provedena programom ImageMagick u sklopu ovog istraživanja. Rezultati evaluacije za svaku sliku zasebno priloženi su kao Prilog 4.

[50] Stančić, H., Trbušić, Ž., Optimisation.of archival processes involving digitisation of typewritten documents. *Aslib Journal of Information Management.* Vol. 72, br. 4, 2020.

Evaluacija rezultata je, naposljetku, provedena nad novostvorenim datotekama .txt formata koje su rezultat optičkog prepoznavanja znakova programom Tesseract i to nad datotekama .jpeg formata koja su binarizirana varijanta Google-a, nad datotekama .jpeg formata koja su nebinarizirana varijanta inicijative Internet Archive te nad binariziranom varijantom inicijative Internet Arhive čija je binarizacija provedena programom ImageMagick u sklopu ovog istraživanja. Rezultati evaluacije za svaku sliku zasebno priloženi su kao Prilog 5.

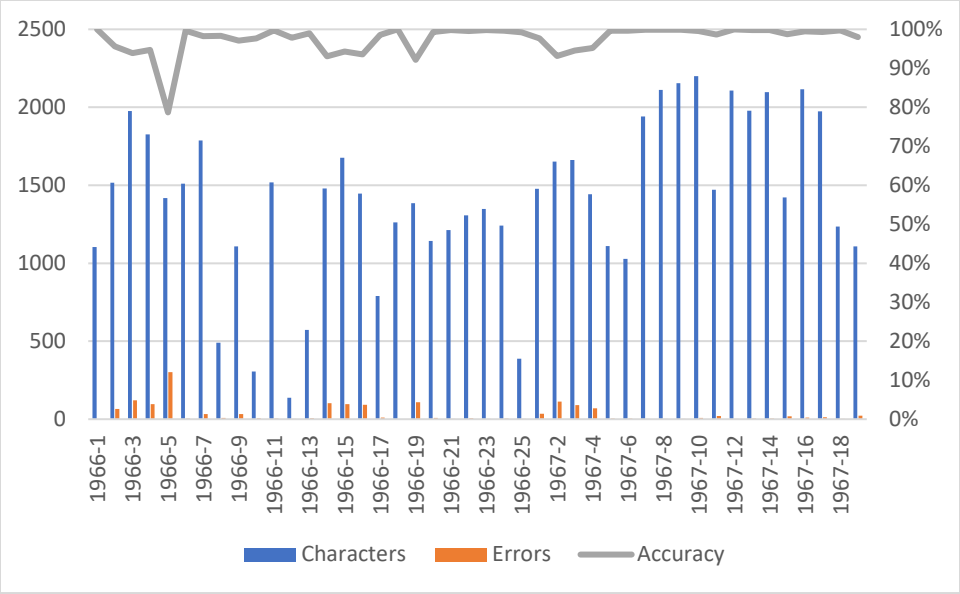
U analizi evaluacije, rezultati su prikazani u obliku grafikona. Na vodoravnoj osi nalaze se oznake koje predstavljaju nazive uzoraka. Broj ispred crtice odnosi se na godinu, a broj iza crtice odnosi se na broj stranice uzorka. Plavi stupac predstavlja broj znakova u uzorku, narančasti stupac predstavlja broj pogrešaka u optički prepoznatom tekstu, dok siva krivulja predstavlja točnost. Na okomitoj osi s lijeve strane nalazi se broj znakova u uzorku, a s desne strane postotak koji se odnosi na točnost.

11.1. Evaluacija optički prepoznatog teksta gradiva Zaklade HathiTrust

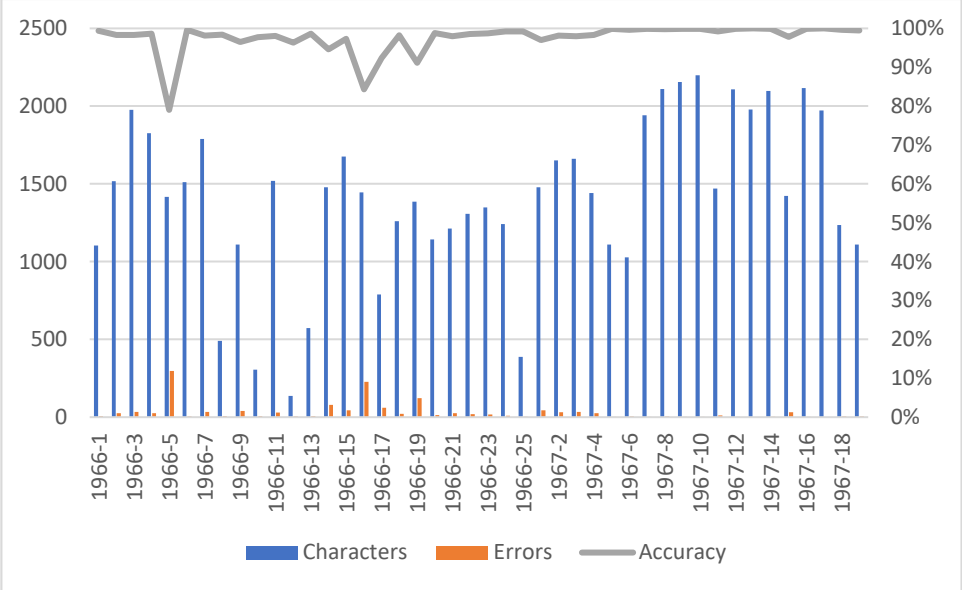
Na službenoj stranici Zaklade HathiTrust navedeno je da je za prihvat gradiva potrebno dostaviti optički prepoznati tekst gradiva. Navedeno je da je „datoteka u .txt formatu obavezna za svaku stranicu, osim ako se radi o rukopisu ili jeziku nad kojim nije moguće provesti OCR“.[51] Osim toga, navedeno je i da svaka stranica mora biti kodirana s UTF-8 te da nazivi moraju biti identični nazivu slike, izuzev ekstenzije. Nije naveden prag stope točnosti optički prepoznatog teksta. S obzirom na navedeno, nije moguće identificirati koji su OCR programi korišteni za optičko prepoznavanje odabranih uzoraka.

Iz dostupnih rezultata vidljivo je da je više od polovice optički prepoznatih datoteka koje je digitalizirao Google (Grafikon 1), ali i više od polovice optički prepoznatih datoteka koje je digitalizirao Internet Archive (Grafikon 2) zadovoljilo pravilo o zadovoljavajućim rezultatima OCR-a koji iznosi od 98%.

[51] HathiTrust: Submission package. Dostupno na: https://www.hathitrust.org/submission-package-requirements-digitized-content-submitted-to-hathitrust



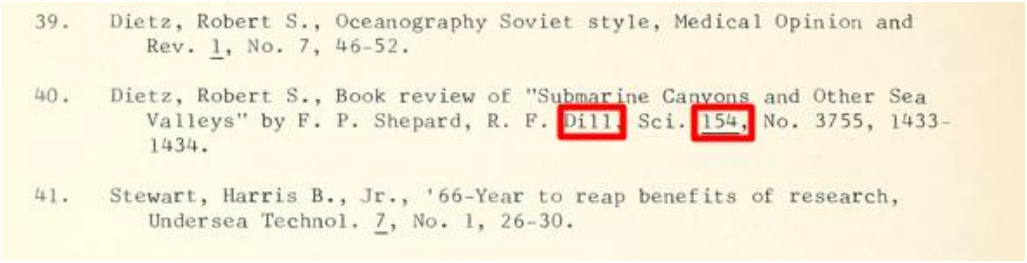
Grafikon 1. Google HathiTrust 1966.-1967.



Grafikon 2. IA HathiTrusts 1966.-1967.

U nekim slučajevima stopa točnosti iznosila je 100%. Stopa točnosti je, za neke uzorke bila znatno lošija. Primjer takvih uzoraka su Google i IA *1966-5* (Slika 6) čija je stopa točnosti iznosila 78,69% i 79,04%. Pregledom samog uzorka uočeno je kako se unutar

slike nalazi jako puno brojeva koji su vrlo vjerojatno utjecali na konačni rezultat. To posebice vrijedi za broj 1 koji izrazito nalikuje slovu l.


39. Dietz, Robert S., Oceanography Soviet style, Medical Opinion and Rev. 1, No. 7, 46-52.

40. Dietz, Robert S., Book review of "Submarine Canyons and Other Sea Valleys" by F. P. Shepard, R. F. Dill, Sci. 154, No. 3755, 1433-1434.

41. Stewart, Harris B., Jr., '66-Year to reap benefits of research, Undersea Technol. 7, No. 1, 26-30.

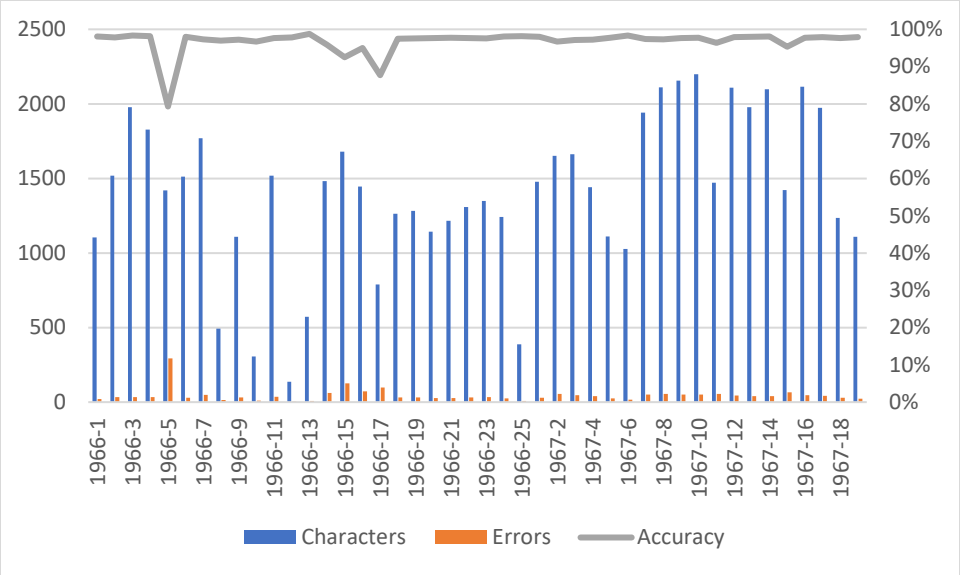
Slika 6. IA 1966-1967 – broj 1 i slovo l

Prosječna točnost OCR-a koje je proveo Google iznosi 97,55%, dok prosječna točnost OCR-a koje je proveo Internet Archive iznosi 97,58%. Razlika u stopi točnosti je minimalna te samim time i zanemariva, a stopa točnosti je, iako ne zadovoljava pravilo, visoka.

11.2. Evaluacija OCR sustava Abbyy FineReader

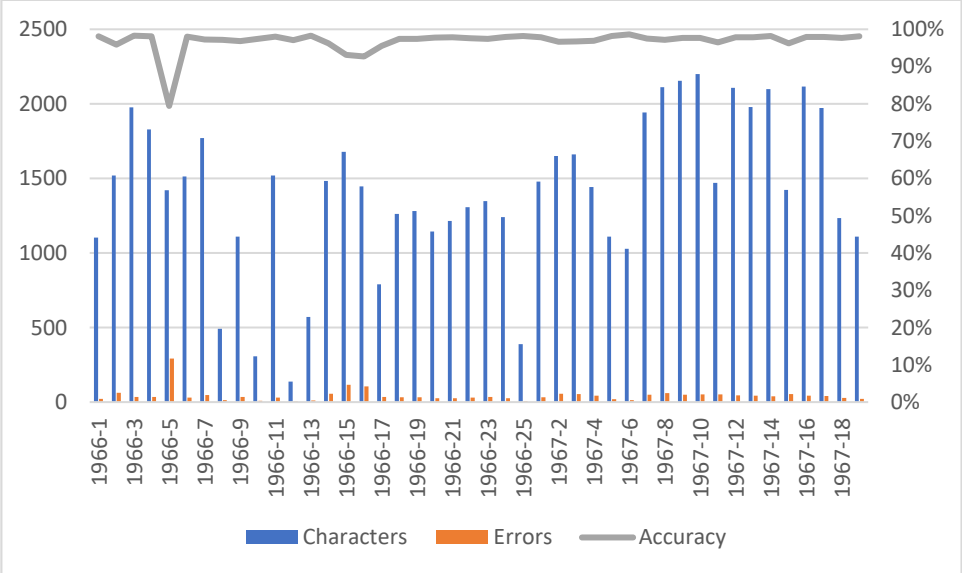
U rezultatima evaluacije uočljivo je da je Abbyy FineReader postigao lošije rezultate u odnosu na OCR sustave Zaklade Hathi jer manje od polovice rezultata svih triju skupina zadovoljava pravilo o zadovoljavajućoj stopi točnosti koja iznosi minimalno 98%.

Prosječna stopa točnosti za skupinu uzoraka koje je optički prepoznao i binarizirao Google (Grafikon 3) iznosi 96,76%.



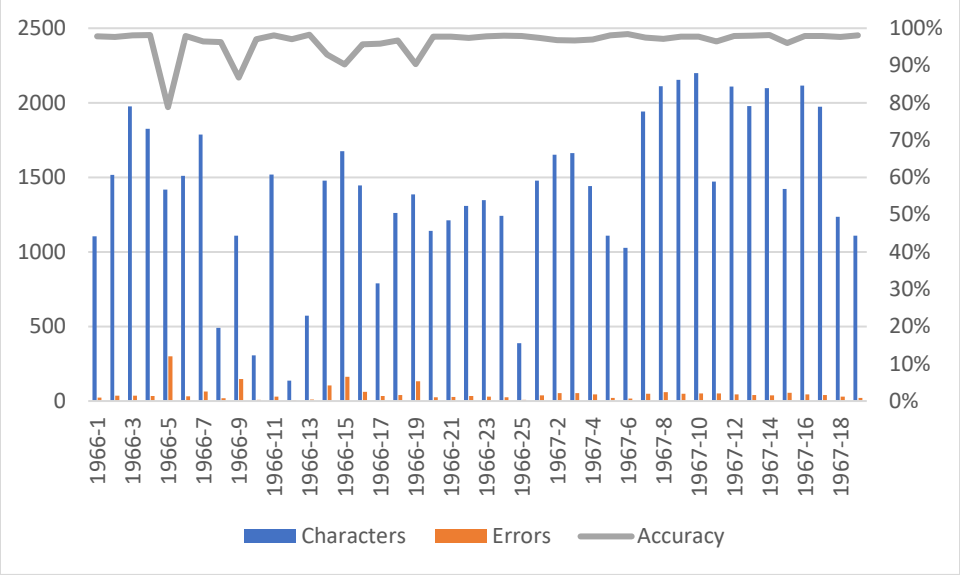
Grafikon 3. Google Abbyy FineReader 1966.-1967.

Prosječna stopa točnosti za skupinu uzoraka koju je optički prepoznao Internet Archive (Grafikon 4) je nešto veća i iznosi 96,87%.



Grafikon 4. IA Abbyy FineReader (nebinarizirano) 1966.-1967.

Prosječna stopa točnosti za skupinu uzoraka IA koja je binarizirana u sklopu istraživanja programom ImageMagick (Grafikon 5) nešto je lošija u odnosu na stope točnosti prijašnjih dviju skupina te iznosi 96,36%.



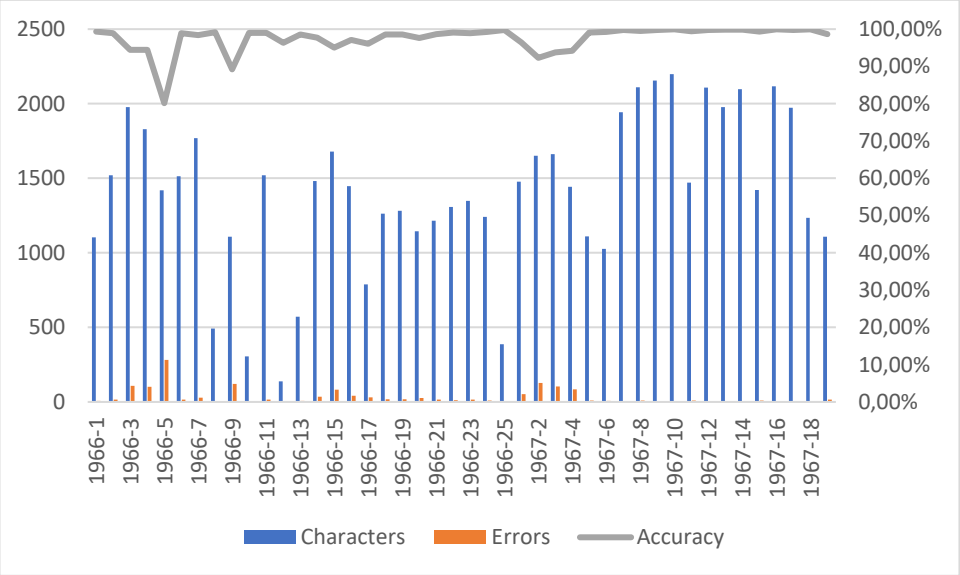
Grafikon 5. Grafikon 5. IA Abbyy FineReader 1966.-1967.

U usporedbi s rezultatima koje je postigao OCR sustav Zaklade Hathi, uočljivo je da je OCR sustav Abbyy FineReader postigao nešto lošije rezultate. U istraživanju kojeg su proveli autori H. Stančić i Ž. Trbušić i čiji su rezultati objavljeni u članku *Optimisation of archival processes involving digitisation of typewritten documents* navedeno je kako je stopa točnosti nakon provedene binarizacije, također, nešto lošija u odnosu na rezultate nad kojima nije provedena binarizacija.

11.3. Evaluacija OCR sustava Tesseract

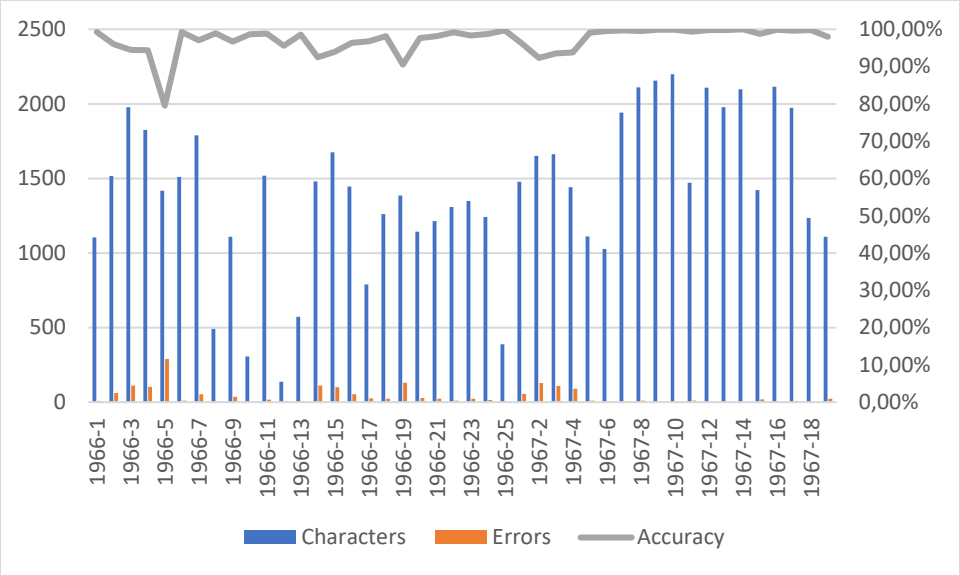
Rezultati evaluacije OCR sustava su iznenađujući. Više od polovice skupine uzoraka koje je optički prepoznao i binarizirao Google te skupine koju je optički prepoznao Internet Archive i čiji su uzorci binarizirani programom ImageMagick zadovoljavaju pravilo o zadovoljavajućoj stopi točnosti od 98%. Međutim, rezultati evaluacije skupine nebinariziranih uzoraka koju je optički prepoznao Internet Archive sadržavala je negativne vrijednosti što predstavlja anomaliju u testnim podacima te zahtijeva dodatno istraživanje. Iz tog razloga, navedena skupina izuzeta je iz analize podataka.

Prosječna stopa točnosti za binariziranu skupinu uzoraka koje je optički prepoznao Tesseract (Grafikon 6) iznosi 97,57%. U usporedbi s rezultatima koje je ostvario HathiTrust 97,55% te Abbyy FineReader 96,36%, stopa točnosti je veća.



Grafikon 6. Google Tesseract (binarizirano) 1966.-1967.

Prosječna stopa točnosti za skupinu uzoraka koje je optički prepoznao Internet Archive i koja je binarizirana programom ImageMagick (Grafikon 7) iznosi 97,19%. U odnosu na prosječnu stopu točnosti od 96,36% koju je postigao OCR sustav Abbyy FineReader, stopa točnosti je veća. No, u odnosu na prosječnu stopu točnosti od 97,55% koju je postigao HathiTrust, nešto je manja.



Grafikon 7. IA Tesseract (binarizirano) 1966-1967

11.4. Interpretacija rezultata

OCR sustavi Zaklade HathiTrust postigli su najbolje rezultate. Veća stopa točnosti Zaklade HathiTrust mogla bi se pripisati i tome da HathiTrust nudi mogućnost naknadnog ispravljanja optički prepoznatog teksta koji je mrežno dostupan korisnicima za slobodno korištenje.

Postoji mala razlika u stopi točnosti programa Tesseract i programa Abbyy FineReader. Stopa točnosti za oba OCR programa je i dalje visoka. Tesseract je pokazao veću stopu točnosti, no samo nad binariziranim uzorcima, dok je program Abbyy FineReader gotovo podjednako pouzdan kada se optičko prepoznavanje znakova provodi i nad binariziranim i nad nebinariziranim uzorcima. Primjena metode binarizacije kod OCR sustava Abbyy FineReader pokazala se suvišnom jer su rezultati bili lošiji u odnosu na rezultate nebinariziranih uzoraka. Takav rezultat je bio i očekivan, s obzirom na ranije spomenuto istraživanje koje su proveli autori H. Stančić i Ž. Trbušić.

No, rezultati ovog istraživanju donekle su oprečni s rezultatima istraživanja autora H. Stančića i Ž. Trbušića. U njihovom je istraživanju zaključeno kako Abbyy FineReader daje rezultate s većom točnošću, što nije slučaj i u istraživanju provedenom u sklopu ovog rada. Postoji mogućnost da anomalije u testnom uzorku nebinariziranih datoteka ukazuju na važnost provođenja metode binarizacije prije provođenja procesa optičkog prepoznavanja znakova korištenjem programa Tesseract, no tu tezu bi trebalo dalje istražiti.

12. Zaključak

Digitalizacija je donijela brojne promjene u načinu rada arhiva, ali i drugih kulturnih institucija koje imaju ulogu zaštite i očuvanja kulturne baštine. Posljedica brzog i značajnog tehnološkog napretka je i korištenje tehnologija u svakodnevnom životu. Stoga, ne iznenađuje da su arhivi usvojili taj trend. Korištenje tehnologija u arhivima donijelo je brojne prednosti, a najznačajnije prednosti su jednostavni pristup informacijama, ušteda vremena i zaštita gradiva

U digitalizacijskom procesu, korištenje OCR tehnologija neophodan je korak, a uspješna implementacija OCR-a tehnologija jedna je od zadaća arhivista. Uspješna implementacija OCR tehnologija u arhivima prema Smjernicama i OAIS referentnom modelu te u konačnici dobivanje kvalitetnih rezultata predstavljaju izazov u suvremenoj arhivistici te iziskuje nova znanja i vještine od arhivista, ali i cjeloživotno učenje kako bi arhivisti ostali u toku s tehnološkim promjenama i inovacijama.

Rezultat procesa optičkog prepoznavanja znakova je računalno kodirani tekst koji je ključan element digitalnih repozitorija zbog mogućnosti brzog i jednostavnog pretraživanja samog sadržaja gradiva. Gradivo pohranjeno u digitalnim repozitorijima od velike je važnosti jer služi kao izvor za istraživanja u raznim znanstvenim granama. Iz tog razloga razvitak OCR tehnologija i poboljšanja u vidu povećanja stope točnosti rezultata optički prepoznatog teksta predstavljaju veliki doprinos u daljnjem razvoju društvenih i humanističkih znanosti.

13. Literatura

1) Abbyy FineReader PDF. Dostupno na: https://pdf.abbyy.com/finereader-pdf/
2) Anderson, N., Muhlberger, G. Optical Character Recognition. Dostupno na: https://www.digitisation.eu/download/website-files/BPG/OpticalCharacterRecognition-IBPG_01.pdf
3) Awel, M. A., Abidi, A. I. Review on Optical Character Recognition. International Research Journal of Engineering and Technology 6, br. 6 (2019.)
4) Digitisation EU. https://www.digitisation.eu/download/website-files/BPG/OpticalCharacterRecognition-IBPG_01.pdf
5) Eikvil, Line. Optical Character Recognition. (1993.) Dostupno na: https://www.academia.edu/6214026/OCR_Optical_Character_Recognition_OCR_-Optical_Character_Recognition
6) Frutiger, A., Delamarre, N., Gurtler, A. OCR-B: A standardized character for optical recognition. Journal of Typographic Research, 1(2) (1967.)
7) Github: Tesseract. Dostupno na: https://github.com/tesseract-ocr/tesseract
8) HathiTrust: About. Dostupno na: https://www.hathitrust.org/about
9) ImageMagick. Dostupno na: https://imagemagick.org/index.php
10) Kleiner, A., Kurzweil, R. C. A Description of the Kurzweil Reading Machine and a Status Report on its Testing and Dissemination. Dostupno na: https://www.rehab.research.va.gov/jour/77/14/1/kleiner.pdf
11) Lee, Christopher A. Open Archival Information System (OAIS) Reference Model. Encyclopedia of Library and Information Sciences, Third Edition. (2010.). Dostupno na: https://ils.unc.edu/callee/p4020-lee.pdf
12) Memon, J., Sami, M., Khan, R. A., Uddin, M. Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR). IEEE Access, Vol. 8 (2020.). Dostupno na: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9151144
13) Patel, C. I., Patel, D. Optical Character Recognition by Open source OCR Tool Tesseract: A Case Study. International Journal of Computer Applications. Vol. 55, br. 10 (2012.), str. 50. Dostupno na:

https://www.researchgate.net/profile/Chirag-Patel-12/publication/235956427_Optical_Character_Recognition_by_Open_source_OCR_Tool_Tesseract_A_Case_Study/links/00463516fa43a64739000000/Optical-Character-Recognition-by-Open-source-OCR-Tool-Tesseract-A-Case-Study.pdf

14) PDF Association. Dostupno na: https://www.pdfa.org/resource/iso-19005-pdfa/
15) Rice, S. V., Nartker, T. A. The ISRI analytic tools for OCR evaluation. Information Science Research Institute (1996.). Dostupno na: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.216.9427&rep=rep1&type=pdf
16) Romein C. A., Kemman, M., Birkholz. J. M., Baker, J., Gruijter, M., Merono-Penuela, A., Ries, R., Ros, R., Scagliola, S. State of the Field: Digital History. The Journal of the Historical Association History 365, br. 105: 197-376 (2020.). Dostupno na: https://onlinelibrary.wiley.com/doi/10.1111/1468-229X.12969
17) Santos, E.A. OCR evaluation tools for the 21st century. National Research Council Canada. Dostupno na: https://aclanthology.org/W19-6004.pdf
18) Smjernice za digitalizaciju kulturne baštine. Dostupno na: https://min-kulture.gov.hr/vijesti-8/objavljene-smjernice-za-digitalizaciju-kulturne-bastine/21484
19) Stančić, H., Trbušić, Ž.. Optimisation.of archival processes involving digitisation of typewritten documents. Aslib Journal of Information Management. Vol. 72, br. 4 (2020.)
20) Suitter, J. Acuracy of Optical Character Recognition sofware Google Tesseract. Dostupno na: https://digitalcommons.usm.maine.edu/cgi/viewcontent.cgi?article=1042&context=thinking_matters
21) Špirenec, S., Ivanjko, T. Korisničko označavanje tekstualnih i vizualnih informacija: što mogu očekivati AKM ustanove? U: 16. seminar Arhivi , knjižnice, muzeji: mogućnosti suradnje u okruženju globalne informacijske strukture: zbornik radova. Hrvatsko knjižničarsko društvo, 2013.
22) Trbušić, Ž. Mogućnosti implementacije sustava za optičko prepoznavanje znakova tijekom prihvata gradiva u arhivske informacijske sustave. U: Radovi 52.

savjetovanja hrvatskih arhivista Zagreb/Šibenik. Hrvatsko arhivističko društvo, 2020., dostupno na: https://www.researchgate.net/publication/349723490_Mogucnosti_implementacije_sustava_za_opticko_prepoznavanje_znakova_tijekom_prihvata_gradiva_u_arhivske_informacijske_sustave

23) Wiggins, R. H., Christian Davidson, H., Ric Harnsberger, H., Lauman, J. R., Goede, P. A. Image File Formats: Past, Present, Future. RadioGraphics 21, br. 3 (2001.)

Popis slika

Popis grafikona

Prilozi

Prilog 1 – Programski kod za lakšu obradu podataka

```
with open("evaluacija.csv", "w", encoding="utf8") as dat:
    for i in range(1,10):
        with
open(r"C:\Users\Jopa\Desktop\Sumpor_OCR_2\OCR_Hathi\1967_Collected_reprints_I
nstitute_for_Oceanography\1967_Penn_State_Uni_Internet_Archive\IA_TXT\1967_IA_
0"+str(i)+r"_acc.txt", "r", encoding="utf8") as f:
            lista_linija=f.readlines()
        f.closed
        char=lista_linija[2].split()[0]
        error=lista_linija[3].split()[0]
        accuracy=lista_linija[4].split()[0]
##        print(char,error,accuracy)
        dat.write(";"+char+";"+error+";"+accuracy+";\n")
    for i in range(10,20):
        with
open(r"C:\Users\Jopa\Desktop\Sumpor_OCR_2\OCR_Hathi\1967_Collected_reprints_I
nstitute_for_Oceanography\1967_Penn_State_Uni_Internet_Archive\IA_TXT\1967_IA_"
+str(i)+r"_acc.txt", "r", encoding="utf8") as f:
            lista_linija=f.readlines()
        f.closed
        char=lista_linija[2].split()[0]
        error=lista_linija[3].split()[0]
        accuracy=lista_linija[4].split()[0]
##        print(char,error,accuracy)
        dat.write(";"+char+";"+error+";"+accuracy+";\n")
dat.closed
```

Prilog 2 – Primjer izlazne evaluacijske datoteke

```
ocreval Accuracy Report Version 7.0
-----------------------------------
    1478   Characters
      32   Errors
   97.83%  Accuracy

       0   Reject Characters
       0   Suspect Markers
       0   False Marks
    0.00%  Characters Marked
   97.83%  Accuracy After Correction

     Ins    Subst      Del   Errors
       0        0        0        0   Marked
       3       23        6       32   Unmarked
       3       23        6       32   Total

   Count   Missed   %Right
     219       18    91.78   ASCII Spacing Characters
     119        4    96.64   ASCII Special Symbols
      87        1    98.85   ASCII Digits
     170        1    99.41   ASCII Uppercase Letters
     882        1    99.89   ASCII Lowercase Letters
       1        1     0.00   General Punctuation
    1478       26    98.24   Total

  Errors   Marked   Correct-Generated
      15        0   {<\n>}-{ }
       5        0   {}-{ }
       3        0   {,<\n>T}-{ T }
       2        0   {-<\n>}-{}
       2        0   {.}-{,}
       1        0   {5}-{1}
       1        0   {<\n>}-{}
       1        0   {l}-{1}
       1        0   {}-{<FEFF>}
       1        0   {-}-{-}

   Count   Missed   %Right
      32       18    43.75   {<\n>}
     187        0   100.00   { }
       3        0   100.00   {(}
       3        0   100.00   {)}
      50        1    98.00   {,}
      13        1    92.31   {-}
      50        2    96.00   {.}
       4        0   100.00   {0}
      12        0   100.00   {1}
       9        0   100.00   {2}
       6        0   100.00   {3}
       8        0   100.00   {4}
      18        1    94.44   {5}
       8        0   100.00   {6}
       6        0   100.00   {7}
       3        0   100.00   {8}
      13        0   100.00   {9}
      16        0   100.00   {A}
       3        0   100.00   {B}
      10        0   100.00   {C}
       7        0   100.00   {D}
      10        0   100.00   {E}
       5        0   100.00   {F}
       8        0   100.00   {G}
       5        0   100.00   {H}
       5        0   100.00   {I}
       8        0   100.00   {J}
       1        0   100.00   {K}
       8        0   100.00   {L}
      13        0   100.00   {M}
      13        0   100.00   {N}
      11        0   100.00   {O}
       7        0   100.00   {P}
      13        0   100.00   {R}
```

Prilog 3 – Rezultati evaluacije za HathiTrust

	A	B	C	D	E	F	G	H	I	J	K
1	1966-1967_Google_Hathi						1966-1967_IA_Hathi				
2	Page	Characters	Errors	Accuracy			Page	Characters	Errors	Accuracy	
3	1966-1	1104	0	100,00%			1966-1	1104	7	99,37%	
4	1966-2	1517	66	95,65%			1966-2	1517	25	98,35%	
5	1966-3	1977	120	93,93%			1966-3	1977	33	98,33%	
6	1966-4	1826	97	94,69%			1966-4	1826	25	98,63%	
7	1966-5	1417	302	78,69%			1966-5	1417	297	79,04%	
8	1966-6	1511	5	99,67%			1966-6	1511	6	99,60%	
9	1966-7	1788	32	98,21%			1966-7	1788	33	98,15%	
10	1966-8	490	8	98,37%			1966-8	490	8	98,37%	
11	1966-9	1109	32	97,11%			1966-9	1109	39	96,48%	
12	1966-10	306	7	97,71%			1966-10	306	7	97,71%	
13	1966-11	1519	4	99,74%			1966-11	1519	29	98,09%	
14	1966-12	137	3	97,81%			1966-12	137	5	96,35%	
15	1966-13	572	6	98,95%			1966-13	572	8	98,60%	
16	1966-14	1479	103	93,04%			1966-14	1479	80	94,59%	
17	1966-15	1676	96	94,27%			1966-15	1676	45	97,32%	
18	1966-16	1446	93	93,57%			1966-16	1446	227	84,30%	
19	1966-17	789	11	98,61%			1966-17	789	60	92,40%	
20	1966-18	1261	0	100,00%			1966-18	1261	22	98,26%	
21	1966-19	1385	108	92,20%			1966-19	1385	123	91,12%	
22	1966-20	1142	8	99,30%			1966-20	1142	14	98,77%	
23	1966-21	1213	2	99,84%			1966-21	1213	25	97,94%	
24	1966-22	1308	6	99,54%			1966-22	1308	19	98,55%	
25	1966-23	1348	3	99,78%			1966-23	1348	17	98,74%	
26	1966-24	1241	4	99,68%			1966-24	1241	10	99,19%	
27	1966-25	388	3	99,23%			1966-25	388	3	99,23%	
28	1967-1	1478	34	97,70%			1967-1	1478	44	97,02%	
29	1967-2	1651	113	93,16%			1967-2	1651	31	98,12%	
30	1967-3	1662	91	94,52%			1967-3	1662	33	98,01%	
31	1967-4	1442	69	95,21%			1967-4	1442	25	98,27%	
32	1967-5	1110	4	99,64%			1967-5	1110	1	99,91%	
33	1967-6	1028	4	99,61%			1967-6	1028	4	99,61%	
34	1967-7	1942	2	99,90%			1967-7	1942	2	99,90%	
35	1967-8	2111	2	99,91%			1967-8	2111	5	99,76%	
36	1967-9	2155	3	99,86%			1967-9	2155	2	99,91%	
37	1967-10	2199	9	99,59%			1967-10	2199	2	99,91%	
38	1967-11	1471	20	98,64%			1967-11	1471	12	99,18%	
39	1967-12	2108	1	99,95%			1967-12	2108	2	99,91%	
40	1967-13	1978	4	99,80%			1967-13	1978	0	100,00%	
41	1967-14	2098	4	99,81%			1967-14	2098	3	99,86%	
42	1967-15	1422	18	98,73%			1967-15	1422	31	97,82%	
43	1967-16	2116	11	99,48%			1967-16	2116	2	99,91%	
44	1967-17	1973	13	99,34%			1967-17	1973	0	100,00%	
45	1967-18	1235	3	99,76%			1967-18	1235	5	99,60%	
46	1967-19	1109	22	98,02%			1967-19	1109	6	99,46%	
47											
48											
49											

Hathi | Abbyy | Tesseract

Prilog 4 – Rezultati evaluacije za Abbyy Fine Reader 15

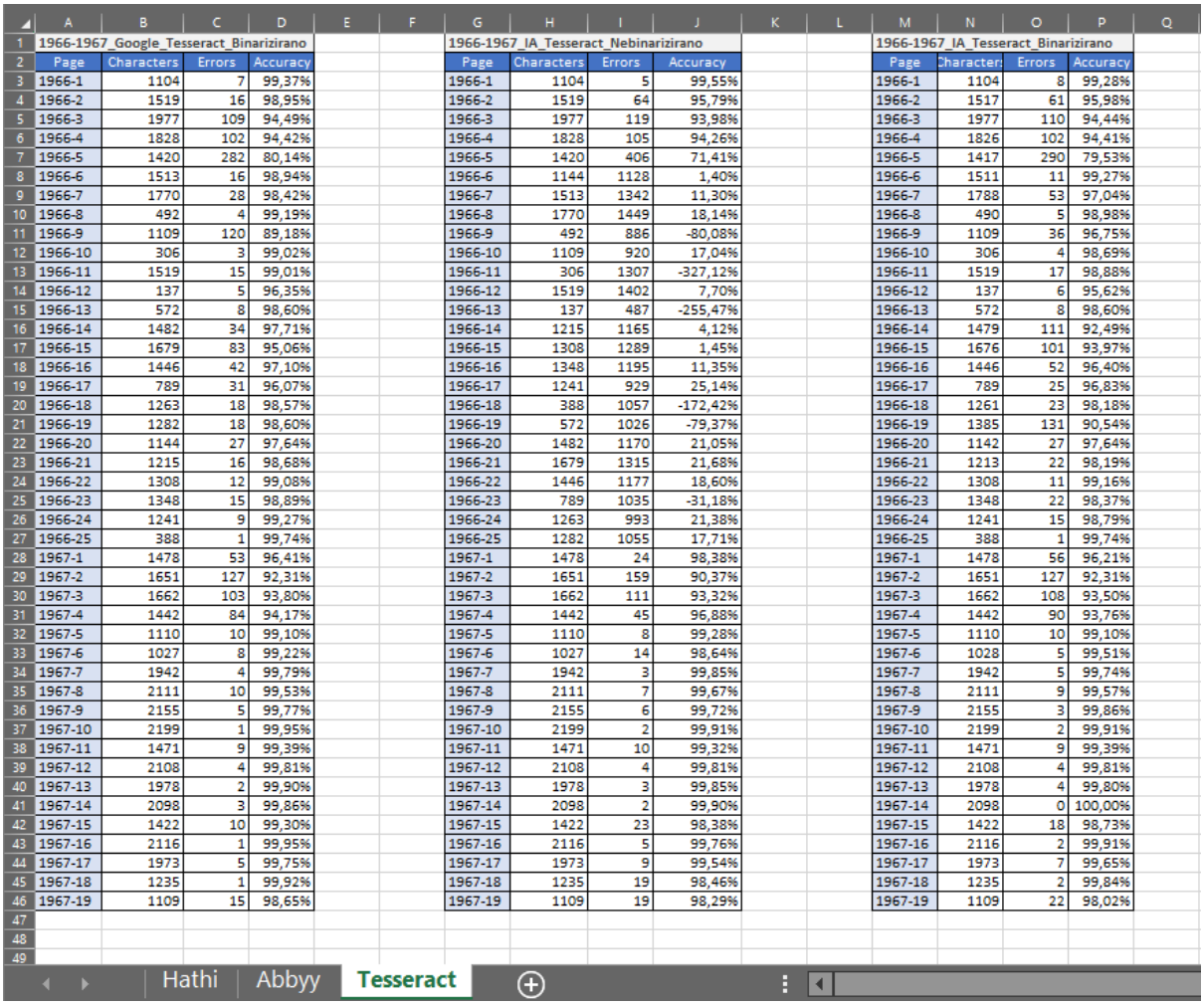
1966-1967_Google_Abbyy_Binarizirano			
Page	Characters	Errors	Accuracy
1966-1	1104	21	98,10%
1966-2	1519	33	97,83%
1966-3	1977	33	98,33%
1966-4	1828	33	98,19%
1966-5	1420	294	79,30%
1966-6	1513	30	98,02%
1966-7	1770	48	97,29%
1966-8	492	15	96,95%
1966-9	1109	31	97,20%
1966-10	306	10	96,73%
1966-11	1519	36	97,63%
1966-12	137	3	97,81%
1966-13	572	7	98,78%
1966-14	1482	61	95,88%
1966-15	1679	126	92,50%
1966-16	1446	72	95,02%
1966-17	789	97	87,71%
1966-18	1263	32	97,47%
1966-19	1282	31	97,58%
1966-20	1144	27	97,64%
1966-21	1215	28	97,70%
1966-22	1308	31	97,63%
1966-23	1348	33	97,55%
1966-24	1241	24	98,07%
1966-25	388	7	98,20%
1967-1	1478	29	98,04%
1967-2	1651	54	96,73%
1967-3	1662	47	97,17%
1967-4	1442	40	97,23%
1967-5	1110	25	97,75%
1967-6	1027	17	98,34%
1967-7	1942	50	97,43%
1967-8	2111	56	97,35%
1967-9	2155	51	97,63%
1967-10	2199	50	97,73%
1967-11	1471	54	96,33%
1967-12	2108	45	97,87%
1967-13	1978	40	97,98%
1967-14	2098	40	98,09%
1967-15	1422	66	95,36%
1967-16	2116	47	97,78%
1967-17	1973	42	97,87%
1967-18	1235	29	97,65%
1967-19	1109	23	97,93%

1966-1967_IA_Abbyy_Nebinarizirano			
Page	Characters	Errors	Accuracy
1966-1	1104	21	98,10%
1966-2	1519	62	95,92%
1966-3	1977	34	98,28%
1966-4	1828	34	98,14%
1966-5	1420	292	79,44%
1966-6	1513	30	98,02%
1966-7	1770	48	97,29%
1966-8	492	14	97,15%
1966-9	1109	35	96,84%
1966-10	306	8	97,39%
1966-11	1519	30	98,03%
1966-12	137	4	97,08%
1966-13	572	10	98,25%
1966-14	1482	56	96,22%
1966-15	1679	115	93,15%
1966-16	1446	106	92,67%
1966-17	789	35	95,56%
1966-18	1263	33	97,39%
1966-19	1282	33	97,43%
1966-20	1144	25	97,81%
1966-21	1215	26	97,86%
1966-22	1308	31	97,63%
1966-23	1348	35	97,40%
1966-24	1241	25	97,99%
1966-25	388	7	98,20%
1967-1	1478	32	97,83%
1967-2	1651	55	96,67%
1967-3	1662	54	96,75%
1967-4	1442	44	96,95%
1967-5	1110	20	98,20%
1967-6	1027	14	98,64%
1967-7	1942	49	97,48%
1967-8	2111	60	97,16%
1967-9	2155	49	97,73%
1967-10	2199	51	97,68%
1967-11	1471	52	96,46%
1967-12	2108	46	97,82%
1967-13	1978	42	97,88%
1967-14	2098	38	98,19%
1967-15	1422	54	96,20%
1967-16	2116	44	97,92%
1967-17	1973	41	97,92%
1967-18	1235	29	97,65%
1967-19	1109	21	98,11%

1966-1967_IA_Abbyy_Binarizirano			
Page	Characters	Errors	Accuracy
1966-1	1104	24	97,83%
1966-2	1517	36	97,63%
1966-3	1977	37	98,13%
1966-4	1826	34	98,14%
1966-5	1417	300	78,83%
1966-6	1511	32	97,88%
1966-7	1788	64	96,42%
1966-8	490	18	96,33%
1966-9	1109	147	86,74%
1966-10	306	9	97,06%
1966-11	1519	29	98,09%
1966-12	137	4	97,08%
1966-13	572	10	98,25%
1966-14	1479	104	92,97%
1966-15	1676	163	90,27%
1966-16	1446	62	95,71%
1966-17	789	33	95,82%
1966-18	1261	41	96,75%
1966-19	1385	133	90,40%
1966-20	1142	26	97,72%
1966-21	1213	27	97,77%
1966-22	1308	34	97,40%
1966-23	1348	29	97,85%
1966-24	1241	25	97,99%
1966-25	388	8	97,94%
1967-1	1478	38	97,43%
1967-2	1651	53	96,79%
1967-3	1662	54	96,75%
1967-4	1442	44	96,95%
1967-5	1110	21	98,11%
1967-6	1028	16	98,44%
1967-7	1942	49	97,48%
1967-8	2111	60	97,16%
1967-9	2155	49	97,73%
1967-10	2199	50	97,73%
1967-11	1471	52	96,46%
1967-12	2108	44	97,91%
1967-13	1978	40	97,98%
1967-14	2098	38	98,19%
1967-15	1422	56	96,06%
1967-16	2116	44	97,92%
1967-17	1973	41	97,92%
1967-18	1235	29	97,65%
1967-19	1109	21	98,11%

Hathi | Abbyy | Tesseract

Prilog 5 – Rezultati evaluacije za Tesseract



1966-1967_Google_Tesseract_Binarizirano			
Page	Characters	Errors	Accuracy
1966-1	1104	7	99,37%
1966-2	1519	16	98,95%
1966-3	1977	109	94,49%
1966-4	1828	102	94,42%
1966-5	1420	282	80,14%
1966-6	1513	16	98,94%
1966-7	1770	28	98,42%
1966-8	492	4	99,19%
1966-9	1109	120	89,18%
1966-10	306	3	99,02%
1966-11	1519	15	99,01%
1966-12	137	5	96,35%
1966-13	572	8	98,60%
1966-14	1482	34	97,71%
1966-15	1679	83	95,06%
1966-16	1446	42	97,10%
1966-17	789	31	96,07%
1966-18	1263	18	98,57%
1966-19	1282	18	98,60%
1966-20	1144	27	97,64%
1966-21	1215	16	98,68%
1966-22	1308	12	99,08%
1966-23	1348	15	98,89%
1966-24	1241	9	99,27%
1966-25	388	1	99,74%
1967-1	1478	53	96,41%
1967-2	1651	127	92,31%
1967-3	1662	103	93,80%
1967-4	1442	84	94,17%
1967-5	1110	10	99,10%
1967-6	1027	8	99,22%
1967-7	1942	4	99,79%
1967-8	2111	10	99,53%
1967-9	2155	5	99,77%
1967-10	2199	1	99,95%
1967-11	1471	9	99,39%
1967-12	2108	4	99,81%
1967-13	1978	2	99,90%
1967-14	2098	3	99,86%
1967-15	1422	10	99,30%
1967-16	2116	1	99,95%
1967-17	1973	5	99,75%
1967-18	1235	1	99,92%
1967-19	1109	15	98,65%

1966-1967_IA_Tesseract_Nebinarizirano			
Page	Characters	Errors	Accuracy
1966-1	1104	5	99,55%
1966-2	1519	64	95,79%
1966-3	1977	119	93,98%
1966-4	1828	105	94,26%
1966-5	1420	406	71,41%
1966-6	1144	1128	1,40%
1966-7	1513	1342	11,30%
1966-8	1770	1449	18,14%
1966-9	492	886	-80,08%
1966-10	1109	920	17,04%
1966-11	306	1307	-327,12%
1966-12	1519	1402	7,70%
1966-13	137	487	-255,47%
1966-14	1215	1165	4,12%
1966-15	1308	1289	1,45%
1966-16	1348	1195	11,35%
1966-17	1241	929	25,14%
1966-18	388	1057	-172,42%
1966-19	572	1026	-79,37%
1966-20	1482	1170	21,05%
1966-21	1679	1315	21,68%
1966-22	1446	1177	18,60%
1966-23	789	1035	-31,18%
1966-24	1263	993	21,38%
1966-25	1282	1055	17,71%
1967-1	1478	24	98,38%
1967-2	1651	159	90,37%
1967-3	1662	111	93,32%
1967-4	1442	45	96,88%
1967-5	1110	8	99,28%
1967-6	1027	14	98,64%
1967-7	1942	3	99,85%
1967-8	2111	7	99,67%
1967-9	2155	6	99,72%
1967-10	2199	2	99,91%
1967-11	1471	10	99,32%
1967-12	2108	4	99,81%
1967-13	1978	3	99,85%
1967-14	2098	2	99,90%
1967-15	1422	23	98,38%
1967-16	2116	5	99,76%
1967-17	1973	9	99,54%
1967-18	1235	19	98,46%
1967-19	1109	19	98,29%

1966-1967_IA_Tesseract_Binarizirano			
Page	Characters	Errors	Accuracy
1966-1	1104	8	99,28%
1966-2	1517	61	95,98%
1966-3	1977	110	94,44%
1966-4	1826	102	94,41%
1966-5	1417	290	79,53%
1966-6	1511	11	99,27%
1966-7	1788	53	97,04%
1966-8	490	5	98,98%
1966-9	1109	36	96,75%
1966-10	306	4	98,69%
1966-11	1519	17	98,88%
1966-12	137	6	95,62%
1966-13	572	8	98,60%
1966-14	1479	111	92,49%
1966-15	1676	101	93,97%
1966-16	1446	52	96,40%
1966-17	789	25	96,83%
1966-18	1261	23	98,18%
1966-19	1385	131	90,54%
1966-20	1142	27	97,64%
1966-21	1213	22	98,19%
1966-22	1308	11	99,16%
1966-23	1348	22	98,37%
1966-24	1241	15	98,79%
1966-25	388	1	99,74%
1967-1	1478	56	96,21%
1967-2	1651	127	92,31%
1967-3	1662	108	93,50%
1967-4	1442	90	93,76%
1967-5	1110	10	99,10%
1967-6	1028	5	99,51%
1967-7	1942	5	99,74%
1967-8	2111	9	99,57%
1967-9	2155	3	99,86%
1967-10	2199	2	99,91%
1967-11	1471	9	99,39%
1967-12	2108	4	99,81%
1967-13	1978	4	99,80%
1967-14	2098	0	100,00%
1967-15	1422	18	98,73%
1967-16	2116	2	99,91%
1967-17	1973	7	99,65%
1967-18	1235	2	99,84%
1967-19	1109	22	98,02%

Sažetak

Evaluacija OCR sustava

U završnom radu analiziran je optički prepoznat tekst pohranjen u sklopu repozitorija Zaklade Hathi. Izdvojeni su odabrani primjeri iz repozitorija i ekstrahirani kao optički prepoznati tekstovi i digitalizirani tekstovi u JPEG formatu. Sukladno ekstrahiranim optički prepoznatim tekstovima, izrađeni su tekstovi koji su istovjetni izvorniku kako bi se mogla provesti evaluacija. Evaluacija je provedena ISRI alatima za evaluaciju optički prepoznatog teksta. U radu je provedeno i prepoznavanje OCR sustavima Abbyy FineReader i Tesseract na istim primjerima. Prepoznavanje je provedeno na prethodno binariziranim i nebinariziranim JPEG formatima. U radu se uspoređuje i analizira točnost optičkog prepoznavanja OCR sustava Zaklade Hathi i OCR sustava Abbyy FineReader i Tesseract te se predstavljaju rezultati.

Ključne riječi: arhivsko gradivo, digitalizacija, Abbyy FineReader, Tesseract, repozitorij zaklade HathiTrust, optičko prepoznavanje znakova (OCR), binarizacija

Summary

Evaluation of OCR systems

In this thesis, the repository of the Hathi Foundation is analyzed. Selected examples are identified from the repository and extracted as optically recognized texts and digitized texts in JPEG format. In accordance with extracted optically recognized texts, texts identical to the originals were created so that the evaluation could be carried out. The evaluation was carried out using ISRI tools for evaluating optically recognized text. Recognition by OCR systems Abbyy FineReader and Tesseract are carried out using the same examples. Recognition is performed on the pre-binarized and non-binarized JPEG formats. The optical recognition accuracy of the Hathi Foundation OCR system and the Abbyy FineReader and Tesseract OCR systems are comparatively analyzed and the results are presented.

Key words: archival materials, digitization, Abbyy FineReader, Tesseract, HathiTrust repository, optical character recognition (OCR), binarization