Digitalizacija i obrada slike i teksta u sustavima za optičko prepoznavanje znakova u domeni bibliotekarstva

Kmetič, Stela

SVEUČILIŠTE U ZAGREBU

FILOZOFSKI FAKULTET

ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI

SMJER BIBLIOTEKARSTVO

Ak. god. 2022./2023.

Stela Kmetič

Digitalizacija i obrada slike i teksta u sustavima za optičko prepoznavanje znakova u domeni bibliotekarstva

Diplomski rad

Mentor: prof. dr. sc. Sanja Seljan

Zagreb, lipanj 2023.

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.

Veliko hvala mojoj obitelji na strpljenju i podršci.

Sadržaj

1. Uvod

Digitalizacija društva više nije novi fenomen, već stvarnost modernog života. Zadnjih desetljeća svjedoči se sve većim i uspješnijim naporima za migraciju onoga što je nekoć bilo analogno u digitalnu inačicu: knjige, računi, dokumenti (Seljan i sur., 2017[1]), ugovori, pa čak i 3D prostor i objekti (Reljić i Dunđer, 2019[2]; Reljić i sur., 2019[3]) neke su od stvari koje se često susreću u digitalnom obliku.

Ova migracija nije zaobišla niti knjižnice koje se danas nalaze pred financijskim, vremenskim, legalnim i strukovnim izazovom digitalizacije svoga inventara.

Što, zašto i kako digitalizirati, kako riješiti pitanje autorskoga prava, online posudba ili otvorena vrata, kada i zašto koristiti sustave za optičko prepoznavanje znakova, koju tehnologiju odabrati?

Mnogo je pitanja na koja knjižnica mora odgovoriti prije nego se upusti u avanturu digitalizacije, a ono na koje će se ovaj rad fokusirati jest uporaba tehnologija za optičko prepoznavanje znakova u kontekstu tekstualne građe koju najčešće pronalazimo u knjižnicama - knjige.

Tehnologije za optičko prepoznavanje znakova možemo pojednostavljeno opisati kao „kompjutersko čitanje slike“. Od tih tehnologija očekuje se da kada u njih unesemo sliku, npr. stranice knjige, da će je uspješno prevesti u tekst koji krajnji korisnik zatim može pretraživati kako bi što brže pronašao ono što je za njega relevantno. Omogućavanje tako brzog snalaženja u tekstu može bitno ubrzati nečije istraživanje, pisanje rada ili edukaciju.

Tekst generiran ovim tehnologijama također može povećati pristupačnost našeg inventara slijepima integracijom sa, sada već uistinu dobro razvijenim, „tekst u govor“ alatima., koji se mogu koristiti i u kombinaciji sa strojnim prevođenjem (Seljan i Dunđer, 2014)[4].

[1] Seljan, S., Dunđer, I., Stančić, H. (2017). Extracting terminology by language independent methods. Forum Translationswissenschaft: Translation Studies and Translation Practice 19, 141-147.
[2] Reljić, I., Dunđer, I. Application of Photogrammetry in 3D Scanning of Physical Objects. TEM Journal 8 (1), 94
[3] Reljić, I., Dunđer, I., Seljan, S. Photogrammetric 3D Scanning of Physical Objects: Tools and Workflow, TEM Journal 8 (2), 383-388
[4] Seljan, S., Dunđer, I. (2014). Combined Automatic Speech Recognition and Machine Translation in Business Correspondence Domain for English-Croatian. Journal of Computer, Information, Systems and Control Engineering. WASET 8 (11), 1069.

Optičko prepoznavanje znakova bi stoga trebalo pronaći svoje mjesto i u hrvatskim knjižnicama, pogotovo onima koje raspolažu zavidnim inventarom znanstvene građe i građe edukacijskog karaktera.

Cilj ovoga rada je opisati proces digitalizacije teksta i rad sustava za optičko prepoznavanje znakova, te predložiti načine na koje se može dobiti što bolje rezultate očitanja znakova prilikom rada sa knjigama.

Rad je podijeljen u dvije osnovne cjeline: teorijski i praktični dio. U teorijskome dijelu opisuje se proces digitalizacije, povijest i rad sustava za optičko prepoznavanje znakova, te se daje pregled trenutno najpoznatijih tehnologija za optičko prepoznavanje znakova.

U praktičnom dijelu provodi se istraživanje utjecaja pretprocesiranja na kvalitetu rezultata očitavanja znakova u sustavima za optičko prepoznavanje znakova *Tesseract* i *Google DocumentAI*. Istraživanje se provodi nad tekstovima preuzetima iz knjiga tiskanih u 19. i 20. stoljeću pošto su one najizgledniji kandidati za digitalizaciju uz primjenu tehnologija za optičko prepoznavanje znakova.

2. Digitalizacija

U najširem smislu, digitalizacija se može definirati kao prevođenje analognog signala u digitalni. Detaljnija definicija bila bi: proces pretvaranja „*teksta, slike, zvuka, pokretnih slika (filmova i videa) ili trodimenzijskog oblika nekog objekta u digitalni oblik, u pravilu binaran kôd zapisan kao računalna datoteka sa sažimanjem podataka ili bez sažimanja podataka, koji se može obrađivati, pohranjivati ili prenositi računalima i računalnim sustavima. Postupci digitalizacije, kao i uređaji kojima se ona obavlja (analogno-digitalni pretvornici), ovise o vrsti gradiva koje se digitalizira.* “ („Leksikografski zavod Miroslav Krleža“, 2021)[5].

Digitalizacija u kontekstu domene knjižničarstva odnosi se primarno na digitalizaciju građe.

U širem kontekstu rada, proces digitalizacije građe prvi je korak prema obradi teksta, odnosno prema korištenju sustava za optičko prepoznavanja znakova, te će stoga fokus ovoga poglavlja biti stavljen na digitalizaciju teksta. Kratko će se opisati potrebe i razlozi za digitalizaciju u knjižnicama, procesi i alati za digitalizaciju tekstualne građe, te digitalizacija specifično u svrhe rada sa sustavima za optičko prepoznavanje znakova.

2.1. Ciljevi i svrhe digitalizacije

U cilju pravilnog odabira građe za digitalizaciju, potrebno je prvo utvrditi koji su najčešći razlozi za digitalizaciju. Prema Stančiću (2009)[6] i Ministarstvu kulture i medija [MKM] (2020)[7] najčešći ciljevi i svrhe digitalizacije su:

Digitalizacija radi zaštite izvornika

U ovome slučaju digitalizira se kako bi se korisnicima moglo umjesto originala na uvid dati digitalnu verziju neke građe, tako smanjujući korištenje izvornika. Ova vrsta digitalizacije pomaže očuvati originalnu građu, a također se primjenjuje i radi izrade sigurnosne kopije izvornika.

Digitalizacija radi povećanja dostupnosti

Jednu knjigu može koristiti jedan korisnik, no jednu digitalnu knjigu može čitati neograničeni broj korisnika. Digitalizacija radi povećanja dostupnosti može doprinijeti povećanoj vidljivosti ustanove, čime ona ujedno proširuje i svoju korisničku bazu. Također treba spomenuti i

[5] Leksikografski zavod Miroslav Krleža. (n.d.). digitalizacija. U Hrvatska enciklopedija, mrežno izdanje. Preuzeto 31.3.2023. s http://www.enciklopedija.hr/Natuknica.aspx?ID=68025
[6] Stančić, H. (2009). Digitalizacija. Zagreb: Zavod za informacijske studije. str. 10-11.
[7] Ministarstvo kulture i medija Republike Hrvatske. (2020). Smjernice za digitalizaciju kulturne baštine.

pridonose znanosti, istraživanjima i promociji kulturne baštine ovisno o vrsti građe koja se digitalizira.

Digitalizacija radi stvaranja nove ponude i usluga

Digitalizacija gradiva otvara cijeli niz opcija za poboljšanje poslovanja: razmjena metapodataka među ustanovama, istraživanje i analiza svih vrsta gradiva računalnim alatima, virtualno spajanje sadržaja itd. Pod ovom svrhom digitalizacije može se smatrati i digitalizacija radi upotpunjavanja fonda i suradnje.

Digitalizacija na zahtjev

Iako digitalizacija na zahtjev ne bi smjela biti primarni pristup odabiru građe za digitalizaciju, ona može postojati kao dodatna knjižnična usluga i/ili smjernica za digitalizaciju prema tome koje sadržaje korisnici najviše traže.

2.2. Načini digitalizacije tekstualne građe

Stančić (2009)[8] kao načine digitalizacije tekstualne građe navodi:

2.2.1. Prepisivanje gradiva

Iako najjednostavniji oblik digitalizacije gradiva, prepisivanje je skup i dugotrajan proces. Primjena ove metode preporučuje se ukoliko su ciljevi digitalizacije apsolutna podudarnost originalnog teksta s digitaliziranim i mogućnost pretraživanja teksta. Također je preporučljivo ovu tehniku digitalizacije koristiti ukoliko digitaliziramo rukopise, građu s nedovoljno kontrasta, tekst s rukom pisanim bilješkama ili označeni (engl. mark-up) tekst. U tim slučajevima digitalizacija prepisivanjem može biti jeftinija, pa čak i brža od drugih metoda.

2.2.2. Skeniranje gradiva

Skeniranje gradiva odnosi se na digitalizaciju teksta skenerom. Time naravno ne dobivamo odmah obradivi tekst, već sliku s kojom dalje možemo raditi u sustavima za optičko prepoznavanje znakova.

2.2.2.1. Karakteristike skenera

A) **Brzina** – Brzina skeniranja, učestalost zaglavljivanja papira, brzina prijenosa datoteka između uređaja za skeniranje i računala.

[8] Stančić, H. (2009). Digitalizacija. Zagreb: Zavod za informacijske studije. str. 33-46, 55-57.

B) **Razlučivost** (rezolucija) – Broj piksela koje skener može očitati. Mjeri se u točkama po inču (engl. DPI – dots per inch). Viša rezolucija daje više detalja, no ujedno zahtjeva duže skeniranje. Razlikujemo optičku i interpoliranu razlučivost. Optička razlučivost je maksimalna razlučivost koju možemo postići korištenjem optike u uređaju, dok je interpolirana ona u kojoj postižemo višu razlučivost dodavanjem piksela koji nisu uistinu ondje uz pomoć izračuna susjednih piksela. Interpolirana razlučivost unosi šum u rezultirajući sken, te ju je stoga bolje izbjegavati.

C) **Dinamički raspon** – Odnosi se na vjernost boje i kontrast slike: što je viši, to su veći kontrast i dubina boje. Plošni skeneri obično imaju dinamički raspon od 2.4, kvalitetniji skeneri 3.0, a skeneri za film i mikrooblike 3.8.

D) **Polje skeniranja** – Najveća površina koju skener može skenirati odjednom.

E) **Vezni uređaji** – Spojka između skenera i računala: utječe na brzinu prijenosa podataka.

F) **Softver za skeniranje** – Softver za skeniranje često dolazi u paketu sa samim skenerom, no može biti i odvojen. On omogućava automatsko primjenjivanje nekih od tehnika pretprocesiranja spomenutih u poglavlju 4.2. i time olakšava proces obrade slike koji često slijedi nakon samog procesiranja, a gotovo uvijek ukoliko skeniramo u svrhu korištenja softvera za optičko prepoznavanje znakova.

G) **Opseg skeniranja** – Karakteristika skenera koja nam govori koliko stranica skener u danu može skenirati.

2.2.2.2. Tipovi skenera

Skenere se dijeli u 2 generalne skupine: koračne i protočne. Koračni skeneri zahtijevaju ljudsku intervenciju kako bi skenirali veće količine građe od onoga što skener trenutno „vidi“, dok protočni skeneri samostalno mogu uvlačiti nove stranice ili listati knjige.

2.2.2.2.1 Koračani skeneri

Ručni skeneri

Ručnim skenerima rukuje se manualno, prelaženjem preko građe koju želimo skenirati. Kako bi skeniranje bilo uspješno, potrebno je mirno i jednolično pomicanje skenera preko građe. Ručni skeneri najjeftiniji su na tržištu, no radi točnosti, ukoliko je to moguće, preporučljivo je koristiti plošni skener.

Plošni skeneri

Naziva se još i stolni ili refleksni skener i najčešći je na tržištu. Takav skener, bez potrebe za velikim znanjem o rukovanju, omogućava uspješno skeniranje svih vrsta dvodimenzionalnih predmeta (dokument, slika, fotografija, umjetničko djelo i sl.) ali i blago trodimenzionalnih kao što su nakit i botanički uzorci.

Moguće ih je nadograditi opremom za protočno skeniranje kao što je uvlakač papira (automatski uvlači nove stranice) i opremom za skeniranje prozirnog gradiva dodavanjem adaptera za prozirno gradivo. Pod prozirnim gradivom obično se misli na mikrofilm. Iako plošni skener mikrofilm može uspješno skenirati, rezolucija skena obično je niža nego kod skenera za filmove.

Rotacioni skeneri (engl. Drum Scanner)

Rotacioni skeneri veoma su skupi, pa se stoga najčešće koriste u studijima za digitalizaciju, a omogućuju visoku rezoluciju skena. Materijal koji skeniramo postavlja se na cilindar ili bubanj koji se zatim okreće oko središnjeg mehanizma. Radi mehanizma rada ovog tipa skenera, njima je moguće skenirati isključivo gradivo koje se nalazi na savitljivim listovima ili na filmu. Rotacioni skeneri omogućuju direktnu konverziju u CMYK, autoizoštravanje, veći dinamički raspon i veću površinu slike.

Reprografski skeneri

Ovaj tip skenera namijenjen je digitalizaciji gradiva velikog formata ili gradiva koje je nezgodno drugačije skenirati (npr. uokvirene umjetničke slike). Sastoji se od prostrane podloge na koju smještamo gradivo, dobrog osvjetljenja i glave za snimanje postavljene visoko iznad podloge.

Skeneri za knjige

Skeneri za knjige mogu pripadati i koračnoj (čovjek je potreban za okretanje stranica) i protočnoj (uređaj sam okreće stranice) kategoriji skenera. Služe za digitalizaciju uvezenih tekstova koje nije moguće skenirati plošnim skenerom (radi npr. osjetljivosti). Radi njihove kompleksnosti cijena im je često visoka.

2.2.2.2.2 Protočni skeneri

Jedna od najbitnijih karakteristika ove skupine skenera jest njihova brzina skeniranja zahvaljujući automatskom uvlakaču stranica – mogu skenirati od nekoliko desetaka do nekoliko stotina dokumenata u minuti bez ljudske intervencije.

2.2.3. Digitalizacija fotoaparatom

Fotoaparatom ili reprografskim skenerom digitaliziramo onu građu koju ne možemo digitalizirati skenerom radi veličine, osjetljivosti ili drugih razloga. Digitalizacija fotoaparatom često se koristi u sklopu reprografskih skenera.

2.3. Digitalizacija za korištenje sa sustavima za optičko prepoznavanje znakova

Prema MKM (2020)[9], ukoliko se digitalizira u svrhe optičkog prepoznavanja znakova, treba se pridržavati niza pravila kako bi rezultati digitalizacije bili podobni za korištenje u sustavima za optičko prepoznavanje znakova:

1.) Datoteke trebamo pohraniti u TIFF (engl. Tag Image File Format) formatu.

2.) Rezolucija slike treba biti 300 ppi (engl. pixels per inch) do 600 ppi za sitniji tekst

3.) Slike trebaju biti pohranjene u 8-bitnoj sivoj skali.

Stančić (2009)[10] pak predlaže pohranu slike u binarnoj, odnosno crno-bijeloj inačici, dok Booth i Gelb (2006)[11] tvrde da bilo koji oblik pretprocesiranja šteti finalnim rezultatima optičkog prepoznavanja znakova. Tesseract (n.d.)[12] predlaže dodatne operacije poput morfološkog procesiranja, a slične dodatne operacije spominju se i u MKM (2020)[13].

Pretprocesiranje, odnosno grafička obrada slike za konzumaciju u sustavu za optičko prepoznavanje znakova, proces je o kojemu se više može pročitati u poglavlju 4.2., a biti će i fokus istraživanja (poglavlje 5).

[9] Ministarstvo kulture i medija Republike Hrvatske. (2020). Smjernice za digitalizaciju kulturne baštine.
[10] Stančić, H. (2009). Digitalizacija. Zagreb: Zavod za informacijske studije. str. 56.
[11] Booth, J. M., i Gelb, J. (2006). Optimizing OCR Accuracy on Older Documents: A Study of Scan Mode, File Enhancement, and Software Products. Office of Innovation and New Technology. Preuzeto 31.3.2023. s https://www.govinfo.gov/media/WhitePaper-OptimizingOCRAccuracy.pdf
[12] Tesseract. (n.d.). Improving the quality of the output. U GitHub. Preuzeto 31.3.2023. s https://tesseract-ocr.github.io/tessdoc/ImproveQuality.html
[13] Ministarstvo kulture i medija Republike Hrvatske. (2020). Smjernice za digitalizaciju kulturne baštine.

2.4. Studio za digitalizaciju

Archive.org diljem svijeta je poznat po svojoj zavidnoj zbirci digitalizirane građe, te posluje i s privatnim korisnicima u sektoru usluga digitalizacije (Internet Archive, n.d.)[14].

Ovdje će se ukratko opisati njihov proces digitalizacije kao primjer profesionalnog procesa digitalizacije u svrhu korištenja građe u sustavu za optičko prepoznavanje znakova.

Prema Internet Archive (2010)[15], Archive.org koristi vlastiti sustav za digitalizaciju koji omogućava efikasnu digitalizaciju bez oštećenja korištene građe. Njihov sustav za digitalizaciju zove se Scribe, a sastoji se od aluminijske podloge s 2 kamere postavljene visoko iznad građe, te držača za knjige u obliku slova V koji dozvoljava prirodni položaj knjige prilikom skeniranja.

Inicijalne fotografije uzete su u boji u formatu .jpg, te archive.org vrši dio operacija pretprocesiranja nad dobivenim materijalom. Ovdje spomenute operacije pretprocesiranja podrobnije su opisane i objašnjene u poglavlju 4.2. Prvi korak je širenje kontrasta iz raspona 30 - 240 u raspon 0 – 255. Zatim se slika izrezuje (engl. crop), prema potrebi rotira i radi se kompenzacija svjetlosti. Daljnje procesiranje sastoji se od kompresije i pripreme građe za krajnjeg korisnika.

[14] Internet Archive. (n.d.). Internet Archive Digitization Services - Partner Documents. Preuzeto 31.3.2023. s https://archive.org/details/partnerdocs
[15] Internet Archive. (2010). Step 3 - Process Document. Preuzeto 31.3.2023. s https://archive.org/details/ProcessDocument/page/n1/mode/2up?view=theater

3. Optičko prepoznavanje znakova

Optičko prepoznavanje znakova jest proces klasifikacije optičkih uzoraka unutar digitalne slike koji odgovaraju alfanumeričkim ili drugim znakovima (Chaudhuri, Mandaviya, Badelia i Ghosh, 2016)[16].

Neke primjene sustava za optičko prepoznavanje znakova, osim onih spomenutih u poglavlju 2.2, su: automatsko prepoznavanje registarskih tablica i prometnih znakova, te razne primjene u automatskoj obradi teksta i unosu informacija. (IBM, 2022)[17], zatim u daljnjoj analizi teksta koja slijedi nakon postupka optičkog prepoznavanja znakova kao što je izrada terminoloških baza (Seljan i sur., 2013[18]), strojno prevođenje (Seljan i Dunđer, 2015[19]) ili klasifikacija dokumenata (Dunđer i sur., 2015[20]).

Sustavi za optičko prepoznavanje znakova danas imaju točnost od 99% u prepoznavanju strojno ispisanih znakova, a 90% u prepoznavanju rukopisa (Alginahi, 2010)[21]. Prema Stančiću (2009)[22], donja granica za isplativost korištenja sustava za optičko prepoznavanje jest 99,95%.

U ovome poglavlju ponuditi će se pregled povijesnog razvoja ove tehnologije, te pregled najpoznatijih sustava za optičko prepoznavanje znakova na tržištu.

3.1. Povijest Sustava za optičko prepoznavanje znakova

Prvi pokušaji u stvaranju tehnologija za optičko prepoznavanje znakova započeli su već 1870. kada je C.R. Carey izumio sustav prijenosa slika mozaikom fotoćelija. 20 godina kasnije, 1890., P. Nipkow u Poljskoj izumio je Nipkow disk, uređaj za prepoznavanje uzoraka (Trbušić, 2019)[23].

[16] Chaudhuri, A., Mandaviya, K., Badelia, P., i Ghosh, S. K. (2016). Optical Character Recognition Systems for Different Languages with Soft Computing. Švicarska: Springer. str. 9.
[17] IBM. (18.2.2022.). What Is Optical Character Recognition (OCR)? Preuzeto 31.3.2023. s https://www.ibm.com/cloud/blog/optical-character-recognition
[18] Seljan, S., Dunđer, I., Gašpar, A. (2013). From digitisation process to terminological digital resources. MIPRO 2013, 1053-1058.
[19] Seljan, S., Dunđer, I. (2015). Automatic quality evaluation of machine-translated output in sociological-philosophical-spiritual domain. Information Systems and Technologies (CISTI 2015), 1-4.
[20] Dunđer, I., Seljan, S., Stančić, H. (2015). Koncept automatske klasifikacije registraturnoga i arhivskoga gradiva. 48. savjetovanje Zaštita arhivskoga gradiva u nastajanju.
[21] Alginahi, Y. (2010). Preprocessing Techniques in Character Recognition. U Mori, M. (ur.), Character Recognition. Rijeka: Sciyo.
[22] Stančić, H. (2009). Digitalizacija. Zagreb: Zavod za informacijske studije. str.72.
[23] Trbušić, Ž. (2019). Zašto je arhivima potreban sustav za optičko prepoznavanje znakova?. @rhivi, 6, 6-7.

Njih su naslijedili Emmanuel Goldberg 1912. i Edmund Fournier D'Albe 1914. Iako nisu radili zajedno, Goldberg i Fournier D'Albe gotovo simultano su razvili uređaje koji su bili sposobni prebaciti tiskane znakove u drugi format.

Goldberg je svoj rad patentirao 1912. u Chicagu, a radilo se o uređaju koji je čitao znakove i prebacivao ih u telegrafski kod. Njegov uređaj pročitao bi poruku, prebacio je u papirnatu traku u obliku Morseovog koda i zatim s trake direktno poslao telegrafski kod, bez ikakve ljudske intervencije.

Fournier D'Albu je optičko prepoznavanje znakova zainteresiralo iz želje da pomogne slijepima, pa je tako 1914. izumio uređaj koji je nazvao „Optophone". „Optophone" je bio skener koji je emitirao zvuk kada bi ga se pomicalo nad tiskanim tekstom. Svaki zvuk bio je ekvivalent jednome znaku (slovu, broju, itd.), te kada bi slijepa osoba naučila koji zvuk odgovara kojemu znaku, mogla je „čitati"(Schantz, 1983).[24]

Prvi patent za ono što danas smatramo sustavom za optičko prepoznavanje znakova podnio je Gustav Tauschek 1928. godine u Austriji pod nazivom „stroj za čitanje". Tauschekov stroj za čitanje bio je mehanički stroj s fotodetektorom. Fotografija s tekstom postavila bi se ispred prozora za čitanje na stroju, zatim bi se s njome poravnali uzorci slova. Ukoliko bi došlo do podudaranja, do fotodetektora ne bi došla izravna svjetlost (Diem i Sablatnig, 2010).[25]

1931. Emmanuel Goldberg i 1932. Paul H. Handel patentirali su tzv. „statističke uređaje", odnosno uređaje koji su mogli iščitavati statističke podatke kroz optičko prepoznavanje uzoraka po uzoru na Tauscheka.

Goldbergov uređaj oslanjao se na fotografije dokumenata koje su zatim u uređaju bile uspoređene s uzorkom koji se tražio. Ukoliko bi se uzorci podudarali, svijetlo fotoćelije bilo bi u potpunosti blokirano, te bi se tako pronašli podudarni dokumenti.

Handelov uređaj također je radio na principu fotoćelija, no umjesto fotografija koristio je šablone, tako tražeći podudarnost između uzorka i drugih šablona.

[24] Schantz, H.F. (1982). The history of OCR, optical character recognition. Manchaster Center, Vermont, SAD: Recognition Technologies Users Association. str.3.
[25] Diem, M., i Sablatnig. R. (2010). Recognizing Degraded Handwritten Characters. Institute of Computer Aided Automation. Preuzeto 31.3.2023. s https://www.researchgate.net/publication/236130411_Recognizing_Degraded_Handwritten_Characters

Glavni nedostatak ovih patenata bila je njihova nepouzdanost – radi potrebe za iznimno preciznim poravnavanjem uzorka i potencijalnih rezultata, često se znalo dogoditi da svijetlost prođe iako je pronađeno preklapanje (Schantz, 1983).[26]

1950-ih tehnološka revolucija već je bila u snažnom zamahu i količina podataka koju je trebalo procesirati svakim danom sve se više povećavala, a mogućnosti kako te podatke procesirati često su uključivale i tehnologije optičkog prepoznavanja znakova.

David Shepard osnovao je Tvrtku za istraživanje inteligentnih strojeva (eng. Intelligent Machines Research Corporation – IMR) kojoj je primarni cilj bio riješiti probleme nastale oko procesiranja podataka. Tako je 1953. patentirao aparat za čitanje koji je pretvarao tiskani tekst u bušene kartice koje su tadašnji kompjuteri procesirali. U suradnji s IMR-om, Reader's Digest postao je prvi povijesni komercijalni korisnik tehnologija za optičko prepoznavanje znakova – koristili su ovaj sustav kako bi u kompjutere, u početku unosili prodajne izvještaje, a kasnije i procesirali pošiljke od 15 000 000 – 20 000 000 knjiga godišnje. Korištenje tehnologija za optičko prepoznavanje znakova skratilo im je vrijeme za procesiranje narudžbe s mjesec dana na malo više od jednog dana. (Schantz, 1983)[27]

Prema Berchmans i Kumar (2014)[28], daljnji razvitak tehnologija za optičko prepoznavanje znakova može se generalno klasificirati u 4 generacije prema njihovoj efikasnosti, robusnosti i prilagodljivosti.

U prvoj generaciji, ranih 1960-ih, sustavi za optičko prepoznavanje znakova mogli su čitati samo određene fontove i oblike znakova. Najrasprostranjeniji sustav za optičko prepoznavanje znakova ove generacije bio je IBM 1418 koji je radio na principu logičnog podudaranja predložaka (Berchmans i Kumar, 2014)[29].

Drugom generacijom smatraju se sustavi od sredine 1960-ih do ranih 1970-ih. U ovome periodu osnovana je i Grupa za automatsko procesiranje podataka Američkog instituta za standarde (eng. American Standards Institute – ANSI) kojoj je cilj bio uspostavljanje standardiziranog fonta za korištenje sa sustavima za optičko prepoznavanje znakova. Korak prema ostvarenju toga cilja ANSI je postigao već 1966. kada je Committee X3A odredio font

[26] Schantz, H.F. (1982). The history of OCR, optical character recognition. Manchaster Center, Vermont, SAD: Recognition Technologies Users Association. str.4-5.
[27] Schantz, H.F. (1982). The history of OCR, optical character recognition. Manchaster Center, Vermont, SAD: Recognition Technologies Users Association. str.9-11.
[28] Berchmans, D., i Kumar, S. S. (2014). Optical character recognition: An overview and an insight. 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT). doi:10.1109/iccicct.2014.6993174
[29] Ibid.

USASI-A (poznat i kao OCR-A) kao standardan font za rad sa sustavima za optičko prepoznavanje znakova (Schantz, 1983)[30]. Ovu generaciju obilježava korištenje mješavine digitalnih i analognih tehnologija za optičko prepoznavanje znakova (Berchmans i Kumar, 2014)[31].

Trećoj generaciji pripadaju sustavi u periodu od 1975.-1985. Ovu generaciju obilježava rad s rukopisima i tekstovima sa šumovima (Berchmans i Kumar, 2014)[32].

U četvrtoj generaciji, kojoj pripada i moderno doba, razvijaju se sustavi sposobni čitati kompleksne dokumente (sa slikama, grafovima, tablicama i sl.), simbole i dokumente u boji (Berchmans i Kumar, 2014)[33].

3.2. Sustavi za optičko prepoznavanje znakova

U ovome poglavlju biti će predstavljeno nekoliko najpoznatijih sustava za optičko prepoznavanje znakova.

3.2.1. Google DocumentAI

Google DocumentAI alat je koji spada u domene računalnog vida (engl. computer vision), prirodne obrade jezika (engl. NLP - natural language processing), tehnika optičkog prepoznavanja znakova i strojno učenje (engl. machine learning). On koristi svoj vlastiti sustav za optičko prepoznavanje znakova koji nije dostupan javnosti, ali poznato je da je treniran na velikim količinama podataka i sposoban raditi s raznim vrstama datoteka i jezika (Chang, 2021)[34].

Računalni vid koristi kako bi prepoznao i izvadio informacije iz raznih sekcija dokumenta poput tablica, fotografija, dijagrama i sl. Zatim koristi prirodno procesiranje jezika da bi dobivene informacije dalje segmentirao u imena, adrese, datume i sl. (Chang, 2021)[35]

[30] Schantz, H.F. (1982). The history of OCR, optical character recognition. Manchaster Center, Vermont, SAD: Recognition Technologies Users Association. str.37.
[31] Berchmans, D., i Kumar, S. S. (2014). Optical character recognition: An overview and an insight. 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT). doi:10.1109/iccicct.2014.6993174
[32] Ibid.
[33] Ibid.
[34] Chang, T. (2021). All Things Google Document AI. Preuzeto 31.3.2023. s https://nanonets.com/blog/document-ai/
[35] Ibid.

Ove značajke veoma su korisne za poslovnu primjenu ovoga alata za procesiranje računa, narudžbenica, uputnica itd., no mogu pomoći i u domeni bibliotekarstva davanjem konteksta tekstu u knjizi čime ćemo dobiti bolje krajnje rezultate.

Google DocumentAI također daje mogućnost treniranja i stvaranja osobnih modela na setovima podataka po želji. (Google, 2023b)[36]

Usluge koje *Google DocumentAI* nudi podijeljene su u 2 kategorije: generalni procesori i specijalizirani procesori. Generalnih procesora ima 4:

1.)Optičko prepoznavanje znakova u dokumentima

2.)Raščlanjivač obrazaca (izvlači unesene podatke prema mjestu u obrascu)

3.)Inteligentni procesor kvalitete dokumenata (daje ocjenu dokumentu prema čitljivosti teksta)

4.)Raščlanjivač dokumenata (ako učitamo više dokumenata u istoj PDF datoteci, ovaj procesor će zaključiti na kojemu mjestu raščlaniti dokumente)

Specijaliziranih procesora ima cijelo mnoštvo, a ovdje će biti spomenuto tek nekoliko zanimljivih kako bi čitatelj dobio osjećaj za mogućnosti alata: procesori za čitanje osobnih iskaznica, vozačkih dozvola i putovnica, procesori za raščlambu računa, raščlambu troškova itd. (Google, 2023c)[37]

Google DocumentAI ne temelji se na kodu otvorenog izvora, već se njegovo korištenje plaća prema korištenom procesoru i količini procesiranih dokumenata. Procesor za optičko prepoznavanje teksta za količine do 5 000 000 mjesečno stoji $1.50 za 1000 stranica, a za količine iznad 5 000 000 $0.60 za 1000 stranica. (Google, 2023)[38]

3.2.2. Tesseract

Prema Tesseract (2023)[39], *Tesseract* je sustav za optičko prepoznavanje znakova koji je originalno razvio Helwett-Packard (mnogima poznatiji pod kraticom HP) između 1985. i 1994.

[36] Google. (2023). Document AI. Preuzeto 31.3.2023. s https://cloud.google.com/document-ai
[37] Google. (2023). Full processor and detail list. Preuzeto 31.3.2023. s https://cloud.google.com/document-ai/docs/processors-list
[38] Google. (2023). Document AI pricing. Preuzeto 31.3.2023. s https://cloud.google.com/document-ai/pricing
[39] Tesseract. (2023). Tesseract OCR. U GitHub. Preuzeto 3.4.2023. s https://github.com/tesseract-ocr/tesseract/blob/main/README.md

2005. HP daje kod alata *Tesseract* javnosti na korištenje, a potom ga preuzima Google koji ga dalje razvija do 2018.

Danas je *Tesseract* u potpunosti besplatan i dostupan za korištenje i modifikaciju javnosti. *Tesseract* je pisan u C++-u, a njegova zadnja stabilna verzija objavljena je 30.11.2021.

Starije verzije Tesseacta oslanjale su se na algoritme za prepoznavanje uzoraka u tekstu, a verzijom 4 *Tesseract* uvodi LSTM neuronske mreže o kojima je moguće više pročitati u poglavlju 4.5.

Tesseract podržava preko 100 jezika i 35 pisama, a moguće je i trenirati vlastite modele za jezike i pisma koji nisu podržani (Tesseract, n.d.c)[40]

3.2.3. Amazon Textract

Amazon Textract usluga je koju od 2018. nudi Amazon u sklopu Amazon Web Servisa (Amazon Web Services [AWS], 2018)[41]. Prema AWS (n.d.)[42], *Textract*, kao i ostali do sada navedeni sustavi, oslanja se na strojno učenje za ekstrakciju teksta, rukopisa i podataka iz skeniranih ili fotografiranih dokumenata, no više o detaljima nije poznato (Kurama, 2022)[43].

Amazon Textract nudi i niz usluga povrh samog sustava za optičko prepoznavanje znakova (Amazon, n.d.b)[44]. Neke od tih usluga su:

a) Analiza zajmova - specijalizirani alat koji izvlači i klasificira podatke iz raznih dokumenata koji se koriste za procesiranje zajmova,

b) Ekstrakcija tablica - alat za izvlačenje podataka iz tablica, uz opciju strukturirane ponovne pohrane ili prebačaja u bazu podataka,

c) Detekcija potpisa - automatsko prepoznavanje potpisa, alat također daje procjenu koliko je siguran da je to uistinu potpis,

---

[40] Tesseract. (n.d.). Tesseract User Manual. U GitHub. Preuzeto 3.4.2023. s https://tesseract-ocr.github.io/tessdoc/

[41] Amazon Web Services. (2018). Introducing Amazon Textract: Now in Preview—easily extract text and data from virtually any document. Preuzeto 3.4.2023. s https://aws.amazon.com/about-aws/whats-new/2018/11/introducing-amazon-textract-now-in-preview-easily-extract-text-and-data-from-virtually-any-document/

[42] Amazon Web Services. (n.d.). Amazon Textract. Preuzeto 3.4.2023. s https://aws.amazon.com/textract/

[43] Kurama, V. (2022). AWS Textract Teardown - Pros and Cons of using Amazon's Textract in 2023. Preuzeto 3.4.2023. s https://nanonets.com/blog/aws-textract-teardown-pros-cons-review/

[44] Amazon Web Services. (n.d.). Amazon Textract Features. Preuzeto 15.5.2023. s https://aws.amazon.com/textract/features/?pg=ln&sec=hs

d) Ekstrakcija bazirana na upitu - Textract dozvoljava davanje upita na temelju kojih može dati odgovor iz dokumenta. Primjer pitanja na koje Textract može dati odgovor je: „Koje je ime mušterije s identifikacijskim brojem 12?“,

e) Prepoznavanje rukopisa,

f) Prepoznavanje računa,

g) Podešavanje praga procjene točnosti - Textract za izvlačenje informacija daje postotak koliko je siguran da je točno označio pojedine dijelove dokumenta. Korisnik može podesiti prag za ocjenu točnosti, a ukoliko je rezultat ispod toga praga, dokument će biti poslan ljudskom djelatniku na ocjenjivanje.

4. Faze rada sustava za optičko prepoznavanje znakova

Svi sustavi za optičko prepoznavanje znakova prolaze kroz određene faze: dobivanje slike, pretprocesiranje, segmentacija, izdvajanje značajki, klasifikacija i postprocesiranje (Mittal i Garg, 2020)[45]. U svim dijelovima ovoga procesa moguća je, a ponekad i potrebna, ljudska intervencija kako bi rezultat očitavanja znakova bio što bolji. Kao što je spomenuto u poglavlju 2.4., MKM (2020)[46], Stančić (2009)[47] i Tesseract (n.d.)[48] predlažu obavljanje određenih koraka pretprocesiranja prije procesiranja slika u sustavu za optičko prepoznavanje znakova.

Ovo poglavlje pokriti će sve faze rada sustava za optičko prepoznavanje znakova, s posebnim naglaskom na pretprocesiranje pošto ono najčešće zahtjeva ljudsku intervenciju.

4.1. Dobivanje slike

Prije korištenja samog sustava za optičko prepoznavanje znakova, prvo je potrebno građu digitalizirati kao što je opisano u poglavlju 2. Prihvatljivi formati datoteka za korištenje u sustavima za optičko prepoznavanje znakova zavise od sustava do sustava, no prema Vijayarani i Sakila (2015)[49] najčešće se koriste TIFF, PNG, JPG, BMP, GIF i PDF. Dobivena slika može imati cijeli niz problema radi samog dokumenta koji obrađujemo. Poteškoće se mogu javiti radi kvalitete papira (npr. vide se slova s poleđine stranice), korištenih fontova, osvjetljenja, oštećenja i samog aranžmana slika i teksta na stranici (Alginahi, 2010)[50].

4.2. Pretprocesiranje

Pretprocesiranje proces je u kojemu dobivenu slikovnu datoteku dalje obrađujemo kako bi je sustav za optičko prepoznavanje znakova što bolje prepoznao (Thorat, Bhat, Sawant, Bartakke i Shirsath, 2022)[51], a pri tome se često mogu ispraviti problemi spomenuti u 4.1.

[45] Mittal, R., i Garg, A. (2020). Text extraction using OCR: A Systematic Review. U 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), (str. 357-362). Indija, Coimbatore.
[46] Ministarstvo kulture i medija Republike Hrvatske. (2020). Smjernice za digitalizaciju kulturne baštine.
[47] Stančić, H. (2009). Digitalizacija. Zagreb: Zavod za informacijske studije. str. 56.
[48] Tesseract. (n.d.). Improving the quality of the output. U GitHub. Preuzeto 31.3.2023. s https://tesseract-ocr.github.io/tessdoc/ImproveQuality.html
[49] Vijayarani, S., i Sakila, A. (2015). Performance Comparison of OCR Tools. International Journal of UbiComp, 6(3), str. 19–30. https://doi.org/10.5121/iju.2015.6303
[50] Alginahi, Y. (2010). Preprocessing Techniques in Character Recognition. U Mori, M. (ur.), Character Recognition. Rijeka: Sciyo.
[51] Thorat, C., Bhat, A., Sawant, P., Bartakke, I., i Shirsath, S. (2022). A Detailed Review on Text Extraction Using Optical Character Recognition. U ICT Analysis and Applications (str. 719–728).

Alginahi (2010)[52] iz tog razloga smatra pretprocesiranje fazom od iznimne važnosti za što bolje rezultate u optičkom prepoznavanju znakova.

Tehnika za pretprocesiranje ima uistinu mnogo, tako da će ovo poglavlje izostaviti tehnike koje se koriste specifično za prepoznavanje rukopisa i druge domene obrade slike poput čitanja rendgenskih snimaka. Također će samo ukratko biti pokriveno frekvencijsko filtriranje slike pošto je ono iznimno resursno zahtjevno (Alginahi, 2010).

4.2.1. Tehnike pretprocesiranja

Poboljšanje slike (engl. Image enhancement) je tehnika koja se koristi u pretprocesiranju kako bi se uklonile smetnje (engl. noise) i zamućenja (engl. blur), povećao kontrast i generalno poboljšala vidljivost detalja na slici (Alginahi, 2010)[53]. U operacijama za poboljšanje slike razlikujemo 2 domene: spacijalnu i onu frekvencija. Kada se radi u spacijalnoj domeni, direktno se manipulira piksele u slici, a kada se radi u frekvencijskoj domeni, slika se procesira konverzijom iste u frekvencije kojima se dalje manipulira (Fisher, Perkins, Walker i Wolfart, 2003)[54].

4.2.1.1. Frekvencijsko filtriranje slike (engl. Frequency domain image filtering)

Kako bi se manipuliralo frekvencijskom domenom slike, prvo je potrebno sliku prevesti u frekvencije. Za to se koristi Fourierova transformacija čiji je rezultat slika na kojoj možemo vidjeti sve frekvencije sadržane u određenim točkama na originalnoj slici (Fisher i sur. 2003)[55].

Zatim se na dobivenom frekvencijskom prikazu koriste filteri. Filteri se najčešće koriste kako bi se supresirale visoke frekvencije (tzv. zaglađivanje, eng. smoothing) ili niske frekvencije (pomaže u jačanju vidljivosti rubova) (Alginahi, 2010)[56].

Prema Alginahi (2010)[57], frekvencijsko filtriranje slike veoma je resursno intenzivno, te se stoga predlaže korištenje 3x3 konvolucijske maske (eng. convolution mask/kernel/matrix) kao zamjenu za frekvencijsko filtriranje, uz tehnike spacijalnog filtriranja slike.

---

[52] Alginahi, Y. (2010). Preprocessing Techniques in Character Recognition. U Mori, M. (ur.), Character Recognition. Rijeka: Sciyo.
[53] Ibid.
[54] Fisher, R., Perkins, S., Walker, A., i Wolfart, E. (2003). Fourier Transform. Preuzeto 3.4.2023. s https://homepages.inf.ed.ac.uk/rbf/HIPR2/fourier.htm
[55] Ibid.
[56] Alginahi, Y. (2010). Preprocessing Techniques in Character Recognition. U Mori, M. (ur.), Character Recognition. Rijeka: Sciyo.
[57] Ibid.

Konvolucijska maska kod procesiranja slika sadrži vrijednosti koje se zatim primjenjuju na originalnu sliku piksel po piksel. Svaki piksel u originalnoj slici ima svoju vrijednost koja odgovara njegovoj boji u rasponu od 0 do 255. (Ludwig, n.d.)[58]

Tablica 1 - primjer konvolucijske maske

0	0	0
0	1	0
0	0	0

Tablica 2 - primjer rasporeda piksela unutar slike

125	118	92	12	0
16	179	0	11	25
49	58	63	25	14

U Tablici 1 prikazana je maska koja će, nakon primjene na sliku (Tablica 2), rezultirati bez promjena. Konvolucija funkcionira tako da množimo oko centra maske:

Tablica 3 - primjena maske na sliku

125	118	92
16	179	0
49	58	63

[58] Ludwig, J. (n.d.). Image Convolution [prezentacija]. Portland State University, Portland, SAD. Preuzeto 3.4.2023. s https://web.pdx.edu/~jduh/courses/Archive/geog481w07/Students/Ludwig_ImageConvolution.pdf

U Tablici 3 kao centralni piksel maske odabrali smo 179, pa primjenom konvolucijske maske množimo sve okolne vrijednosti onima u masci:

Tablica 4 - primjena maske na sliku

125*0	118*0	92*0
16*0	179*1	0*0
49*0	58*0	63*0

Rezultat konvolucije za naš odabrani piksel je suma svih vrijednosti u masci odnosno 0*8+179 = 179, dakle originalna vrijednost.

Mijenjanjem vrijednosti u masci postižemo razne efekte poput zamućivanja, izoštravanja, pronalaska rubova i slično.

Konvoluciju koristi i jedan od danas najpoznatijih programa za obradu slika, *Adobe Photoshop* (Photoshop, 2022)[59], a moguće ju je, kao i sve ostale tehnike koje će se spominjati u ovome poglavlju, i samostalno primijeniti korištenjem raznih programskih jezika .

4.2.1.2. Spacijalno filtriranje slike (eng. Spatial image filtering)

Kao što je već rečeno, operacije spacijalnog filtriranja provode se direktno nad pikselima unutar slike. U spacijalnom filtriranju razlikujemo:

-operacije nad točkom, odnosno operacije u kojima je rezultat na koordinatama (x,y) isključivo zavisan od vrijednosti na istim tim koordinatama originala,

-lokalne operacije, odnosno operacije u kojima je rezultat na koordinatama (x,y) zavisan o vrijednosti susjednih piksela onome na koordinatama (x,y)

-i globalne operacije u kojima je rezultat na koordinatama (x,y) zavisan od svih vrijednosti unutar slike koju obrađujemo (Javed, n.d.)[60].

[59] Photshop. (2022). List of filters supporting 16-bit/channel and 32-bit/channel documents. Preuzeto 3.4.2023. s https://helpx.adobe.com/photoshop/using/filter-effects-reference.html

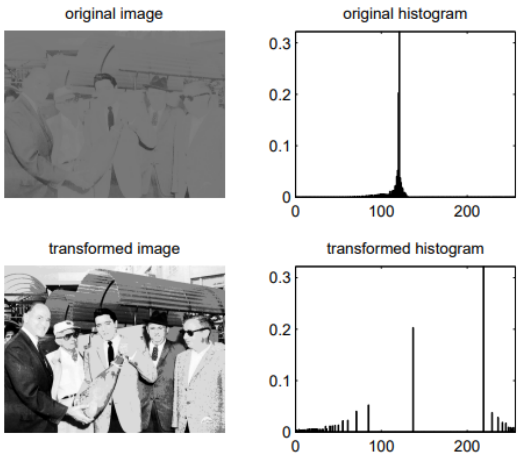
[60] Javed, A. (n.d.). Digital Image Processing: Lecture #5, Image Enhancement in Spatial Domain- I [prezentacija]. University of Engineering and Technology, Taxila, Pakistan. Preuzeto 3.4.2023. s https://web.uettaxila.edu.pk/CMS/AUT2010/seDIPbs/notes%5CLecture_05%20Image%20Enhancement%20in%20spatial%20domain.pdf

U ovome potpoglavlju bit će navedene neke od operacija pretprocesiranja redoslijedom kojim bi se trebale obnašati. Ove operacije pretprocesiranja predominantno spadaju pod procesiranje točke, no neke spadaju i u druge skupine, te će za te biti posebno istaknuto kojoj skupini pripadaju. Procesiranje točke, odnosno konverzija starog piksela u novi, koristi se primarno za pojačanje kontrasta i može se obnašati i na više piksela odjednom (Bebis, 2004)[61].

4.2.1.2.1 Procesiranje histograma

Histogram je grafička reprezentacija intenziteta sive boje u slici izražena u „binovima“, a dobiva se dodavanjem piksela „binu“ prema vrijednosti sive boje od 0 za crnu do 255 za bijelu za sve piksele unutar slike (Alginahi, 2010)[62]. Histogram je moguće napraviti i za slike u boji. Procesiranjem histograma možemo u daljnjim operacijama pretprocesiranja dobiti bolje rezultate, te se s toga procesiranje histograma preporuča kao prvi korak u pretprocesiranju (Alginahi, 2010)[63]. U procesiranju histograma razlikujemo 2 operacije: izjednačavanje histograma i specifikaciju histograma.

Izjednačavanje histograma spada pod globalne operacije i omogućava nam ravnomjerno raspoređivanje vrijednosti sive boje u slici, a time i jačanje kontrasta (Alginahi, 2010)[64]. Primjer histograma prije izjednačavanja i nakon izjednačavanja prikazan je na Slici 1:



Slika 1 – Preuzeto iz https://www.math.uci.edu/icamp/courses/math77c/demos/hist_eq.pdf

[61] Bebis, G. (2004). Image Operations. University of Nevada, Reno, Department of Computer Science and Engineering, Nevada, SAD. Preuzeto 3.4.2023. s https://www.cse.unr.edu/~bebis/CS791E/Notes/PointProcess.pdf
[62] Alginahi, Y. (2010). Preprocessing Techniques in Character Recognition. U Mori, M. (ur.), Character Recognition. Rijeka: Sciyo.
[63] Ibid.
[64] Ibid.

Cilj specifikacije histograma engl. (histogram specification/matching) je uskladiti histogram slike koju obrađujemo s već postojećim histogramom.(Sears-Collins, 2021)[65]. Ova operacija također se koristi za poboljšanje kontrasta s minimalnim zahtjevima za resursima (Thomas, Flores-Tapia i Pistorius, 2011)[66].

4.2.1.2.2 Širenje kontrasta (eng. Contrast Stretching)

Slika koju obrađujemo sastavljena je od piksela koje možemo brojčano prikazati kao vrijednost, odnosno razinu, sive boje koju piksel sadrži. Operacijama širenja kontrasta možemo proširiti raspon vrijednosti koje piksel može sadržavati i tako povećati ili smanjiti kontrast. U tim operacijama svaki piksel originalne slike mapira se direktno na obrađenu sliku (Sahidan, Mashor, Wahab, Salleh i Ja'afar, 2008)[67].

Jedna od metoda širenja kontrasta je lokalno širenje kontrasta (eng. Local contrast stretching). Lokalno širenje kontrasta omogućava ujednačenje kontrasta u slici, odnosno ono uzima presvijetla i pretamna područja slike i ujednačuje ih s ostatkom slike (Sahidan i sur., 2008)[68].

Druga metoda je globalno širenje kontrasta. Globalni kontrast je prosjek kontrasta manjih dijelova slike. Visoki globalni kontrast vizualno prikazuje više detalja, dok niski, iako s manje detalja, izgleda uniformnije. Globalno širenje kontrasta određuje maksimalnu i minimalnu vrijednost sive (ili RGB, za slike u boji) boje za cijelu sliku i prilagođuje vrijednosti piksela tome rasponu (Radha, 2012)[69].

4.2.1.2.3 Određivanje graničnog praga (eng. Thresholding)

Određivanje graničnog praga je metoda odvajanja objekata na slici od pozadine binarizacijom, odnosno pretvaranjem slike u potpuno crno-bijelu (Alginahi, 2010)[70]. Razlikujemo globalno, lokalno i adaptivno određivanje graničnog praga. **Globalno određivanje graničnog praga**

[65] Sears-Collins, A. (2021). Difference Between Histogram Equalization and Histogram Matching. Preuzeto 3.4.2023. s https://automaticaddison.com/difference-between-histogram-equalization-and-histogram-matching/
[66] Thomas, G., Flores-Tapia, D., i Pistorius, S. (2011) Histogram Specification: A Fast and Flexible Method to Process Digital Images. IEEE Transactions on Instrumentation and Measurement, 60(5), str. 1565-1578. doi: 10.1109/TIM.2010.2089110.
[67] Sahidan, S. I., Mashor, M. Y., Wahab, A. A., Salleh, Z., i Ja'afar, H. (2008). Local and Global Contrast Stretching For Color Contrast Enhancement on Ziehl-Neelsen Tissue Section Slide Images. 4th Kuala Lumpur International Conference on Biomedical Engineering (str.583–586).
[68] Ibid.
[69] Radha, N. (2012). Comparison of Contrast Stretching methods of Image Enhancement Techniques for Acute Leukemia Images. International Journal of Engineering Research & Technology (IJERT), 1(6).
[70] Alginahi, Y. (2010). Preprocessing Techniques in Character Recognition. U Mori, M. (ur.), Character Recognition. Rijeka: Sciyo.

uzima vrijednost T kao granicu između pretvorbe u crni ili bijeli piksel prema vrijednosti sive boje sadržane u tome pikselu. Za određenje vrijednosti T postoje razne metode (Alginahi, 2010)[71]:

-Uzimanje medijana vrijednosti sive boje svih piksela.

-Stvaranjem histograma za sliku. Pomoću histograma, T se određuje uzimanjem vrijednosti dola na prikazanom grafu.

-Iterativnom metodom. Iterativna metoda slijedi 5 koraka:

1.) Na bilo koji način, čak i nasumce, određuje se inicijalna vrijednost T.

2.) Sliku se segmentira koristeći T u R1 (pozadina) i R2 (objekt).

3.) Izračunava se prosjek vrijednosti sadržanih u R1 (u1) i R2 (u2) na originalnoj slici.

4.)Novi T dobiva se formulom T=1/2(u1+u2).

5.)Ponavljamo korake 2-5 dok dva puta za redom ne dobijemo isti T ili prethodno utvrđenu razliku između 2 T (Kumar Chaubey, 2016)[72].

Za globalno određivanje praga postoji par desetaka, ako ne i više algoritama (Alginahi, 2010)[73], a inovacije su česte u ovome polju kao npr. Huang, Gao i Cai (2005)[74] koji su i citirani u ovome radu.

Prema pregledu literature u Alginahi (2010)[75], najbolji algoritam za globalno određivanje praga koji možemo koristiti je Otsu metoda. Otsu metoda zapisuje se formulom: $\sigma_W^2 = W_b\,\sigma_b^2 + W_f\,\sigma_f^2$.

Otsu metoda je iterativna, odnosno podrazumijeva prolaženje kroz niz mogućih vrijednosti praga T kako bi pronašla najmanju sumu ponderiranih varijanci. Finalni rezultat ove jednadžbe nije vrijednost praga, već suma ponderiranih varijanci (engl. weighted variance) w1 i w2, odnosno crne i bijele boje.

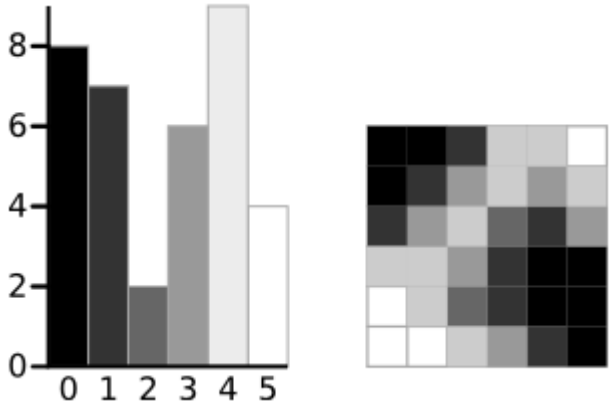
---

[71] Ibid.
[72] Kumar Chaubey, A. (2016). Comparison of The Local and Global Thresholding Methods in Image Segmentation. World Journal of Research and Review (WJRR), 2(1), str.1-4.
[73] Alginahi, Y. (2010). Preprocessing Techniques in Character Recognition. U Mori, M. (ur.), Character Recognition. Rijeka: Sciyo.
[74] Huang, Q., Gao, W. i Cai, W. (2005). Thresholding technique with adaptive window selection for uneven lighting image. Pattern Recognition Letters 26, str. 801-808.
[75] Alginahi, Y. (2010). Preprocessing Techniques in Character Recognition. U Mori, M. (ur.), Character Recognition. Rijeka: Sciyo.

Varijanca je mjera udaljenosti broja u zadanom setu od prosjeka brojeva u istome setu (Hayes, 2023)[76], a ponderirana varijanca je vrsta varijance u kojoj određene vrijednosti (ponderi ili engl. weights) imaju veće značenje od drugih (Toshkov, 2012)[77].

Vrijednosti koje se koriste u Otsu metodi nastaju stvaranjem histograma slike.

Ovdje je prikazan izračun Otsu algoritma na primjeru i slikama preuzetima s http://www.labbookpages.co.uk/software/imgProc/otsuThreshold.html uz potvrdu točnosti kalkulacija iz Bangare, Dubal, Patil i Bangare (2015)[78]:



Slika 1 – *Lijevo je prikazan histogram slike: na osi y nalazi se broj piksela, a na osi x imaginarna vrijednost sive boje u pikselima. Desno se nalazi originalna slika 6 x 6.*

Prema histogramu sliku ćemo razdvojiti u crni i bijeli dio, a kao prag za razdvajanje uzet ćemo sredinu vrijednosti piksela, odnosno 3. U algoritmu se vrijednosti označuju kao:

- N = ukupan broj piksela

- n = broj piksela u skupini

- w1, w2… wn = broj piksela iste histogramske vrijednosti

- h1, h2… hn = vrijednost razine sive boje u histogramu

Dalje ćemo izračunati:

[76] Hayes, A. (2023). What Is Variance in Statistics? Definition, Formula, and Example. Investopedia. Preuzeto 3.4.2023. s https://www.investopedia.com/terms/v/variance.asp
[77] Toshkov, D.D. (2012). Weighted variance and weighted coefficient of variation. RE-DESIGN. Preuzeto 3.4.2023. s http://re-design.dimiter.eu/?p=290
[78] Bangare, L.S., Dubal, A., Bangare, P.S., i Patil, S.T. (2015). Reviewing Otsu's Method For Image Thresholding. International Journal of Applied Engineering Research, 10(9), str. 21777-21783. https://dx.doi.org/10.37622/IJAER/10.9.2015.21777-21783

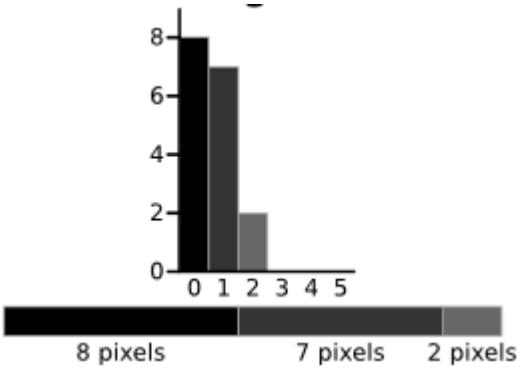
-težinski faktor: W=n/N,

-srednju vrijednost: M = ((h1 x w1)+(h2 x w2)+…(hn x wn))/n

-i varijancu: σ^2 = (((h1-M)^2 *w1) + ((h2- M)^2 *w2)+…. ((hn-M)^2 *wn))/n

za svaku skupinu.

1.) Izračun za skupinu ispod 3 (crni pikseli):



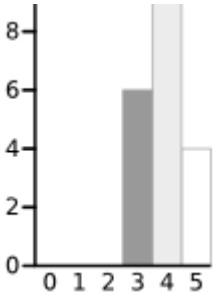
Slika 2 – *Prikaz histograma samo za crnu skupinu*

W_b = (8+7+2)/36=0.4722

M_b = ((0x8) + (1x7) + (2x2))/17=0.6471

σ^2_b = (((0-0.6471)^2 x 8) + ((1-0.6471)^2 x 7) + ((2-0.6471)^2 x 2))/17 = 0.4637

2.) Izračun za skupinu iznad 3 (bijeli pikseli):



Slika 3 – *Prikaz histograma samo za bijelu skupinu*

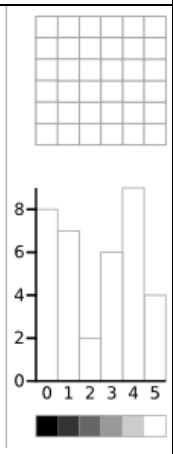
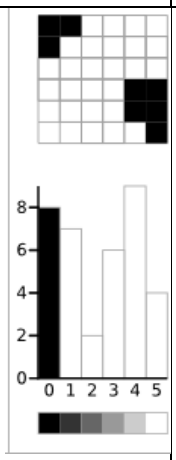
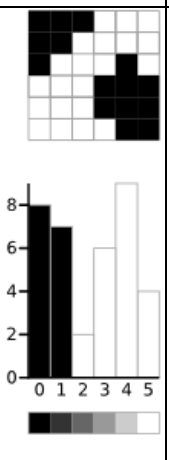
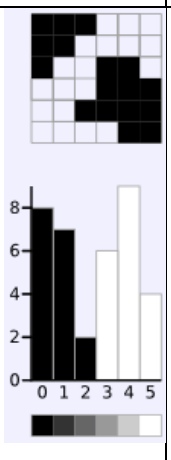
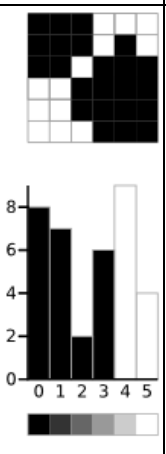
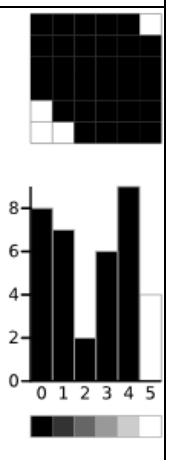
W_f = (6+9+4)/36=0.5278

M_f = ((3x6) + (4x9) + (5x4))/19 = 3.8947

σ^2_f = (((3-3.8947)^2 x 6) + ((4-3.8947)^2 x 9) + ((5-3.8947)^2 x 4))/19 = 0.5152

Rezultate ćemo uvrstiti u Otsu jednadžbu:

σ^2_W = 0.4722 x 0.4637 + 0.5278 x 0.5152 = 0.4909

Prag	T = 0	T = 1	T = 2	**T = 3**	T = 4	T = 5
Binarizirana slika i njen histogram						
W_b	$W_b = 0$	$W_b = 0.222$	$W_b = 0.4167$	**$W_b = 0.4722$**	$W_b = 0.6389$	$W_b = 0.8889$
M_b	$M_b = 0$	$M_b = 0$	$M_b = 0.4667$	**$M_b = 0.6471$**	$M_b = 1.2609$	$M_b = 2.0313$
σ^2_b	$\sigma^2_b = 0$	$\sigma^2_b = 0$	$\sigma^2_b = 0.2489$	**$\sigma^2_b = 0.4637$**	$\sigma^2_b = 1.4102$	$\sigma^2_b = 2.5303$
W_f	$W_f = 1$	$W_f = 0.7778$	$W_f = 0.5833$	**$W_f = 0.5278$**	$W_f = 0.3611$	$W_f = 0.1111$
M_f	$M_f = 2.3611$	$M_f = 3.0357$	$M_f = 3.7143$	**$M_f = 3.8947$**	$M_f = 4.3077$	$M_f = 5.000$
σ^2_f	$\sigma^2_f = 3.1196$	$\sigma^2_f = 1.9639$	$\sigma^2_f = 0.7755$	**$\sigma^2_f = 0.5152$**	$\sigma^2_f = 0.2130$	$\sigma^2_f = 0$
Rezultat	$\sigma^2_W = 3.1196$	$\sigma^2_W = 1.5268$	$\sigma^2_W = 0.5561$	**$\sigma^2_W = 0.4909$**	$\sigma^2_W = 0.9779$	$\sigma^2_W = 2.2491$

Tablica 5 - *Izračun Otsu jednadžbe*

Iz Tablice 5 može se iščitati da je najniži rezultat ujedno i najbolji.

Globalno određivanje praga nije moguće koristiti u svim situacijama. Ukoliko je osvjetljenje loše i slika koju pokušavamo obraditi ima sjene, varirajuće razine sive u pozadini ili vodeni

žig, preporučljivo je koristi **lokalno određivanje praga** (Huang i sur., 2005)[79]. Prema Firdousi i Parveen (2014)[80], u procesu lokalnog određivanja praga također se određuje prag T, ali za svaki piksel individualno. Prema Ni (2023)[81] i Firdousi i Parveen (2014)[82], za lokalno određivanje praga najčešće se koriste sljedeći algoritmi:

-Niblackov algoritam – Najčešće se koristi u optičkom prepoznavanju znakova i prema Alginahi (2010)[83] najefektivniji je algoritam za lokalno određivanje praga. Niblackov algoritam primjenjuje se nad prethodno određenom području veličine w x w, a slijedi formulu $T(x, y) = m(x, y) + k* \delta(x, y)$. U ovoj formuli T je prag koji tražimo za koordinate (x,y), m je srednja vrijednost sive boje svih piksela u prozoru w x w, δ je standardna devijacija vrijednosti sive boje svih piksela u prozoru w x w, a k je statistička pristranost kojoj Firdousi (2014)[84] dodjeljuje vrijednost -0.2.

-Sauvolin algoritam – Rezultira u smanjenju šumova i pomaže u očuvanju oblika čestica.

-Algoritam za ispravljanje pozadine – smanjuje šumove u velikim, praznim područjima

-T.R. Singh algoritam – Računarno manje zahtjevan od Niblackova i Sauvolina algoritma sa sličnim rezultatima.

Adaptivno određivanje graničnog praga varijacija je globalnog određivanja graničnog praga. Umjesto korištenja jednog praga za određivanje vrijednosti piksela unutar cijele slike, adaptivno određivanje praga određuje zasebne pragove za određene segmente slike, koristeći algoritme za globalno određivanje graničnog praga (Roy, Dutta, Dey, Chakraborty i Ray, 2014)[85].

[79] Huang, Q., Gao, W. i Cai, W. (2005). Thresholding technique with adaptive window selection for uneven lighting image. Pattern Recognition Letters 26, str. 801-808.
[80] Firdousi, R., i Parveen, S. (2014). Local Thresholding Techniques in Image Binarization. International Journal Of Engineering And Computer Science, 3(3), str. 4062-4065.
[81] Ni. (2023). Thresholding. Preuzeto 3.4.2023. s https://www.ni.com/docs/en-US/bundle/ni-vision-concepts-help/page/thresholding.html
[82] Firdousi, R., i Parveen, S. (2014). Local Thresholding Techniques in Image Binarization. International Journal Of Engineering And Computer Science, 3(3), str. 4062-4065.
[83] Alginahi, Y. (2010). Preprocessing Techniques in Character Recognition. U Mori, M. (ur.), Character Recognition. Rijeka: Sciyo.
[84] Firdousi, R., i Parveen, S. (2014). Local Thresholding Techniques in Image Binarization. International Journal Of Engineering And Computer Science, 3(3), str. 4062-4065.
[85] Roy, P., Dutta, S., Dey, N., Dey, G., Chakraborty, S., i Ray, R. (2014). Adaptive thresholding: A comparative study. 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT). doi:10.1109/iccicct.2014.6993140

4.2.1.2.4 Log transformacije

Log transformacija može pomoći bolje prikazati detalje na slici jer ona uzima mali raspon sivih boja i proširuje ih (Manikpuri i Yadav, 2014)[86].

4.2.1.2.5 Gamma korekcija (engl. Power Law transformation/ Gamma correction)

Gama korekcija, kao i log transformacija, koristi se za poboljšanje detalja na slici. Ona također uzima mali raspon sivih boja i proširuje ih, ali i kompresira dinamički raspon slika s velikim varijacijama u vrijednostima piksela (Alginahi, 2010)[87]. Gama korekcija važna je kada je potrebno imati vjeran prikaz slike (npr. prikaz slika medicinske radiologije ili priprema materijala za tisak) (Lončarić, n.d.)[88].

4.2.1.3. Procesiranje maske ili ispravljanje šumova

Procesiranje maske spada pod lokalne operacije, jer piksel poprima novu vrijednost izračunom zavisnim od njegovih susjeda. Operacije nad maskom više su resursno intenzivne, no zato daju i bolje rezultate (Alginahi, 2010)[89]. Procesiranje maske bilo je spomenuto i u potpoglavlju o frekvencijskom filtriranju maske u kontekstu konvolucijske matrice, a često se koristi za popravljanje šumova unutar slike.

4.2.1.3.1 Filteri za zaglađivanje (engl. Smoothing/Low-pass filters)

Postoje razni filteri za zaglađivanje, a ovdje će biti spomenuta 2: filter srednje vrijednosti i filter medijana. Prema Alginahi (2010)[90], filter srednje vrijednosti najčešće se koristi za smanjenje šumova unutar slike i zaglađivanje. On je podvrsta konvolucijske matrice koju smo opisali u 4.2.1.1.

Filter medijana uzima vrijednosti piksela na određenom području slike (npr. 3x3, 5x5, 7x7,…). Slično kao i u konvoluciji, on računa medijan vrijednosti piksela koji okružuju piksel u sredini i zamjenjuje piksel u sredini s tom vrijednosti. Filter medijana efektivan je u uklanjanju impulsnih šumova kao što je npr. nasumično pojavljivanje crnih i bijelih piksela.

---

[86] Manikpuri, U., i Yadav, Y. (2014). Image Enhancement Through Logarithmic Transformation. International Journal of Innovative Research in Advanced Engineering (IJIRAE), 1(8), str. 357-362.
[87] Alginahi, Y. (2010). Preprocessing Techniques in Character Recognition. U Mori, M. (ur.), Character Recognition. Rijeka: Sciyo.
[88] Lončarić, S. (n.d.). Poboljšanje slika u prostornoj domeni [prezentacija]. Fakultet elektrotehnike i računarstva, Sveučilište u Zagrebu. Preuzeto 3.4.2023. s https://www.fer.unizg.hr/_download/repository/opdos05a.pdf
[89] Alginahi, Y. (2010). Preprocessing Techniques in Character Recognition. U Mori, M. (ur.), Character Recognition. Rijeka: Sciyo.
[90] Ibid.

4.2.1.3.2 Filteri za izoštravanje (engl. Sharpening/High-pass filters)

Filteri za izoštravanje koriste se za naglašavanje detalja slike (Alginahi, 2010)[91]. Kao i filteri za zaglađivanje, tip su konvolucijske matrice, no za razliku od matrica korištenih za zaglađivanje, matrice za izoštravanje koriste i negativne vrijednosti unutar maske.

4.2.1.3.3 Minimalni, maksimalni i filter raspona

Prema Alginahi (2010)[92], maksimalni filter dobiva se na istome principu kao i filter medijana: u određenom području slike gledamo piksele koji okružuju centralni piksel i dodjeljujemo tom pikselu najvišu vrijednost koju jedan od okolnih piksela ima. Ovaj filter efektivan je u eliminiranju piksela koji nepotrebno odskaču od svoga okruženja.

Minimalni filter pojačava tamne boje unutar slike, a implementira se isto kao maksimalni filter, samo što se uzima minimalna vrijednost od okolnih piksela.

Filter raspona uzima razliku maksimalne i minimalne vrijednosti piksela u određenome području i zamjenjuje centralni piksel sa vrijednošću te razlike.

4.2.1.3.4 Ispravljanje kosine (engl. skew correction)

Većina sustava za optičko prepoznavanje znakova osjetljiva je na rotaciju slike koju pokušavamo „pročitati“, te je zato bitno sliku ispravno postaviti prije učitavanje iste u sustav. Alginahi (2010)[93] razlikuje više grupa sustava za otkrivanje kosine (engl. skew detection):

1.)Houghova transformacija - Tehnika koja omogućuje kompjutersku detekciju oblika (linije, krugovi, elipse) (Mukhopadhyay i Chaudhuri, 2015)[94],

2.)Analiza profila projekcije - Metoda mjerenja kosine prema histogramu crnih piksela na horizontalnim ili vertikalnim linijama (Jain i Borah, 2014)[95]

3.)Korelacije između linija - Metoda ispravljanja kosine prema korelacijama između 2 linije unutar slike koje imaju zadanu udaljenost (Yan, 1993)[96]

---

[91] Ibid.
[92] Ibid.
[93] Ibid.
[94] Manikpuri, U., i Yadav, Y. (2014). Image Enhancement Through Logarithmic Transformation. International Journal of Innovative Research in Advanced Engineering (IJIRAE), 1(8), str. 357-362.
[95] Jain, B., i Borah, M. (2014). A Comparison Paper on Skew Detection of Scanned Document Images Based on Horizontal and Vertical Projection Profile Analysis. International Journal of Scientific and Research Publications, 4(6).
[96] Yan, H. (1993). Skew Correction of Document Images Using Interline Cross-Correlation. CVGIP: Graphical Models and Image Processing, 55(6), str. 538-543. https://doi.org/10.1006/cgip.1993.1041

4.)Grupiranje (engl. clustering) - Metoda koje se koristi kao nadogradnja na jednu od metoda koja detektira linije. Linije se u ovoj metodi dalje grupiraju, kako bi se otkrio kut kosine koja se treba ispraviti.(Ahmad, Naz i Razzak, 2021)[97]

4.3. Segmentacija

Segmentacija je proces u kojemu izdvajamo značajke slike: raspored stranice (slike, grafovi, tekst, paragrafi, rečenice i sl.) i same znakove. Postoje i sustavi u koji ne koriste segmentaciju kao dio procesa prepoznavanja znakova, već se oslanjaju na druge tehnike (Breuel, Ul-Hasan, Al-Azawi i Shafait, 2013)[98].

4.3.1. Segmentacija stranice

Segmentacija stranice provodi se nad dokumentima koji imaju miješani sadržaj - tekst, slike, grafovi i sl. Prema Alginahi (2010)[99] segmentacije stranice može se klasificirati u 3 glavne kategorije:

1.) Odozgora prema dolje - segmentira se veće regije dokumenta u manje

2.) Odozdo prema gore -grupira se piksele u veće blokove ili povezane komponente poput slova, koja se zatim grupiraju u riječi, pa u rečenice itd.

3.) Hibridne tehnike

4.3.2. Segmentacija znakova odnosno slova

Segmentacija znakova posebno je kompleksna i nužna u jezicima u kojima se riječi pišu povezano kao npr. arapski (Alginahi, 2010)[100], no nužan je korak i za latinična pisma.

4.3.3. Morfološko procesiranje

Morfološko filtriranje ili procesiranje skup je tehnika kojima se rezultate segmentacije dalje procesira kako bi se osiguralo da se radi šumova određene linije nisu slučajno povezale u slova, ili određena slova razlomila. Morfološko procesiranje ponekad se radi prije prepuštanja slike sustavu za optičko procesiranje znakova, pogotovo u slučaju rukopisa i starijeg tiska

[97] Ahmad, R., Naz, S., i Razzak, I. (2021). Efficient skew detection and correction in scanned document images through clustering of probabilistic hough transforms. Pattern Recognition Letters, 152, str. 93-99. https://doi.org/10.1016/j.patrec.2021.09.014

[98] Breuel, T. M., Ul-Hasan, A., Al-Azawi, M. A., i Shafait, F. (2013). High-Performance OCR for Printed English and Fraktur Using LSTM Networks. 12th International Conference on Document Analysis and Recognition (str. 683-687). doi:10.1109/icdar.2013.140

[99] Alginahi, Y. (2010). Preprocessing Techniques in Character Recognition. U Mori, M. (ur.), Character Recognition. Rijeka: Sciyo.

[100] Ibid.

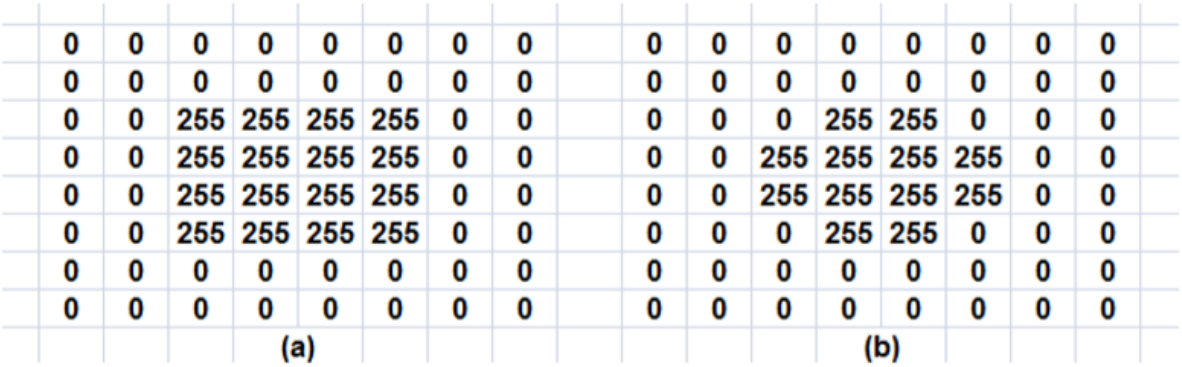
(*Tesseract*, n.d.)[101]. Morfološko procesiranje možemo obavljati isključivo nad binariziranim slikama (Alginahi, 2010)[102].

4.3.3.1. Erozija i proširenje

Erozija omogućuje umanjivanje objekta uklanjanjem piksela koji ga okružuju, dok proširenje omogućuje uvećavanje dodavanjem piksela oko objekta. (Alginahi, 2010)[103]

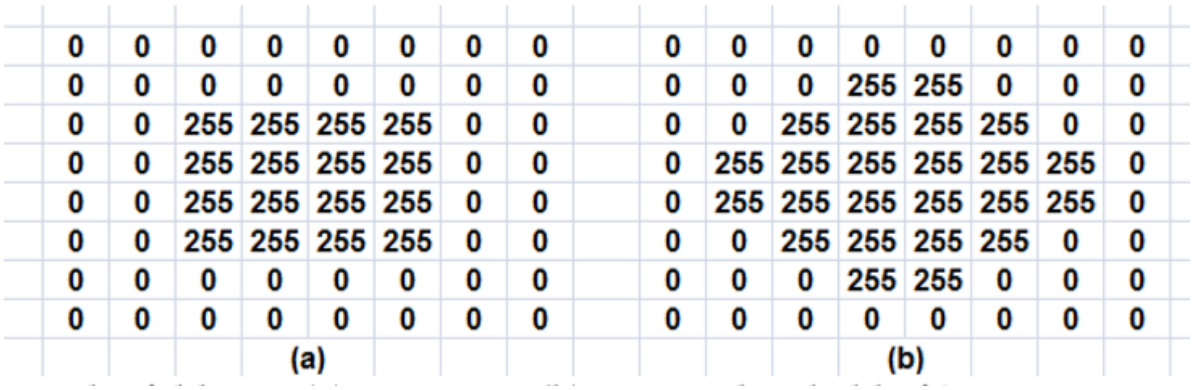
Eroziju i proširenje može se implementirati s dvije tehnike: određivanjem graničnog praga i procesiranjem maske.

Određivanje graničnog praga gleda susjedne piksele pikselu x i mijenja njihovo stanje (iz crnog u bijelo i bijelog u crno) prema tome da li prelaze prethodno određeni granični prag dozvoljenih piksela različitih od piksela x. U slučaju primjene za eroziju, ti pikseli koji prelaze granični prag biti će pretvoreni u pozadinske piksele (Slika 5), dok će u slučaju primjene za proširenje (Slika 6) pikseli koji prelaze granični prag biti „asimilirani“ u objekt. (Alginahi, 2010)[104]



Slika 4 - *Erozija graničnim pragom* (preuzeto iz Mori, 2009)

[101] Tesseract. (n.d.). Improving the quality of the output. U GitHub. Preuzeto 31.3.2023. s https://tesseract-ocr.github.io/tessdoc/ImproveQuality.html
[102] Alginahi, Y. (2010). Preprocessing Techniques in Character Recognition. U Mori, M. (ur.), Character Recognition. Rijeka: Sciyo.
[103] Ibid.
[104] Ibid.



Slika 5 - Proširenje graničnim pragom (preuzeto iz Mori, 2009)

Druga tehnika koristi masku (3x3, 5x5, i sl.) u kojoj se nalaze jedinice i nule i s njome prelazi preko svih piksela unutar slike. Maska može prolaziti vertikalno ili horizontalno preko slike i koristiti se i za eroziju i za proširenje.

4.3.3.2. Zatvaranje i otvaranje

Zatvaranje i otvaranje imaju sličan efekt na sliku kao i erozija i proširenje. Otvaranje razdvaja objekte koji su preblizu jedan drugome, dotiču se a ne trebaju i povećava rupe u objektima. Otvaranje možemo primijeniti malenim brojem erozija i/ili proširenja.

Zatvaranje spaja rastrgane elemente i ispunjava neželjene rupe u objektima, a primjenjuje se na istome principu kao i otvaranje (Alginahi, 2010)[105].

4.3.3.3. Skeletizacija

Skeletizacija je proces stanjivanja slova do najmanjeg mogućeg broja piksela. Postoje dvije osnovne tehnike skeletizacije: osnovno stanjivanje i transformacija medijalne osi (engl. medial axis transform). Stanjivanje je proces u kojemu se slovo erodira do te mjere da je slovo sastavljeno samo od pojedinih piksela ali zadržava originalni oblik. Transformacija medijalne osi mjeri udaljenost svakog piksela koji sastavlja slovo od ruba slova i prema onim pikselima koji su najmanje udaljeni od jednog ili više piksela ruba proizvodi kostur slova također konstruiran samo od pojedinih piksela (Alginahi, 2010)[106].

[105] Ibid.
[106] Ibid.

4.4. Izvlačenje značajki (engl. feature extraction)

Izvlačenje značajki postupak je u kojemu sustav izvlači najrelevantnije informacije iz slike kako bi omogućio prepoznavanje znakova (Singh i Budhiraja, 2011)[107]. Detaljnije, ovo je proces u kojemu se svakome znaku zadaje vektor koji će ga reprezentirati, a prema tome vektoru prepoznat će se druga pojavljivanja toga znaka (Mittal i Garg, 2020)[108]. Prema Mittal I Garg (2020)[109], neke od poznatih tehnika izvlačenja značajki u sustavima za optičko prepoznavanje znakova su:

A) **Zoniranje** (engl. zoning) - U procesu zoniranja znak je podijeljen u zone, najčešće 2x2, 4x4 i sl. Zatim se u zonama smanjuje gustoća piksela kako bi se dobila reprezentacija toga znaka. (Mittal i Garg, 2020)[110]

B) **Projekcijski histogram** (engl. projection histogram) - Računa broj piksela s vrijednosti 1 u određenom smjeru. Postoje 3 tipa projekcijskog histograma: horizontalni, vertikalni i dijagonalni. (Gharde, Baviskar i Adhiya, 2013)[111]

C) **Značajke profila udaljenosti** (engl. distance profile features) - Mjeri se udaljenost od ruba okvira u kojemu je znak do rubova znaka sa sve 4 strane (Singh i Budhiraja, 2011)[112].

D) **Pozadinska distribucija smjera** (engl. Background directional distribution) - Obavlja se s maskom u čijem je središtu piksel znaka. Maskom se prelazi preko cijeloga znaka, te se bilježi distribucija piksela pozadine od piksela znaka do ruba maske u svim smjerovima (Singh i Budhiraja, 2011)[113].

E) **Neuralne mreže** - neuralne mreže mogu se koristi za izvlačenje značajki i za klasifikaciju ili kombinirati oba procesa (Subasi, 2020)[114]. Detaljnije o njima biti će rečeno u sljedećem poglavlju.

---

[107] Singh, P., i Budhiraja, S. (2011). Feature Extraction and Classification Techniques in O.C.R. Systems for Handwritten Gurmukhi Script – A Survey. International Journal of Engineering Research and Applications (IJERA), 1(4), str. 1736-1739.
[108] Mittal, R., i Garg, A. (2020). Text extraction using OCR: A Systematic Review. U 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), (str. 357-362). Indija, Coimbatore.
[109] Ibid.
[110] Ibid.
[111] Gharde, S.S., Baviskar, P.V., i Adhiya, K.P. (2013). Identification of Handwritten Simple Mathematical Equation Based on SVM and Projection Histogram. International Journal of Soft Computing and Engineering (IJSCE), 3(2), str. 425-429.
[112] Singh, P., i Budhiraja, S. (2011). Feature Extraction and Classification Techniques in O.C.R. Systems for Handwritten Gurmukhi Script – A Survey. International Journal of Engineering Research and Applications (IJERA), 1(4), str. 1736-1739.
[113] Ibid.
[114] Subasi, A. (2020). Practical Machine Learning for Data Analysis Using Python.

4.5. Klasifikacija

Klasifikacija koristi vektore prepoznavanja dobivene u fazi izvlačenja značajki kako bi stvorila uzorke prema kojima će prepoznati tekst (Suganya, 2015[115]; Kovač i sur., 2022[116]; Seljan i sur., 2023[117]). Neke od metoda klasifikacije su:

A)**Klasifikator probabilističke neuronske mreže** - Klasifikator koji mapira ulazni obrazac nizu klasifikacija i određuje kojoj klasi pripada prema funkciji gustoće vjerojatnosti (engl. probability density function) (Mittal i Garg, 2020)[118].

B) **Klasifikator potpornog vektorskog stroja** (engl. Support Vector Machine) - Klasifikator baziran na strojnom učenju kojemu se daje uzorak za treniranje u kojemu isti jasno pripada jednoj od dvije moguće klase. Prema tome, algoritam pokušava izgraditi model odlučivanja prema kojemu će uvrstiti buduće uzorke u klase. (Gharde i sur., 2013)[119]

C) **Klasifikator K- najbliži susjed** - Najjednostavniji klasifikator koji odlučuje o klasi kojoj uzorak pripada prema najbližim susjedima uzorku. (Singh i Budhiraja, 2011)[120]

D)**Long Short-Term Memory (LSTM) neuronske mreže** - Neuronske mreže napravljene su da simuliraju proces ljudskog razmišljanja i zaključivanja (Elements of AI, n.d.)[121], a LSTM je jedna od podvrsta neuronskih mreža, točnije ponavljajućih neuronskih mreža (engl. Recurrent Neural Networks). Ponavljajuće neuronske mreže razlikuju se od svojih prethodnika po tome što mogu zadržati informacije o prethodnim ulaznim podacima prema tome kolika je bitnost tih podataka korištenjem pondera, odnosno imaju mogućnost korištenja konteksta u obradi

[115] Suganya, S. (2015). Analysis of Feature Extraction of Optical Character detection in Image Processing Systems, National Conference on Recent Trends in Engineering and Technology, 3(4).
[116] Kovač, A., Dunđer, I., Seljan, S. (2022). An overview of machine learning algorithms for detecting phishing attacks on electronic messaging services. MIPRO 2022 - International Convention on Information, Communication and Electronic Technology. 954-961
[117] Seljan, S., Tolj, N., Dunđer, I. (2023). 595-Information Extraction from Security-Related Datasets. MIPRO 2022 - International Convention on Information, Communication and Electronic Technology.
[118] Mittal, R., i Garg, A. (2020). Text extraction using OCR: A Systematic Review. U 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), (str. 357-362). Indija, Coimbatore.
[119] Gharde, S.S., Baviskar, P.V., i Adhiya, K.P. (2013). Identification of Handwritten Simple Mathematical Equation Based on SVM and Projection Histogram. International Journal of Soft Computing and Engineering (IJSCE), 3(2), str. 425-429.
[120] Singh, P., i Budhiraja, S. (2011). Feature Extraction and Classification Techniques in O.C.R. Systems for Handwritten Gurmukhi Script – A Survey. International Journal of Engineering Research and Applications (IJERA), 1(4), str. 1736-1739.
[121] Elements of AI. (n.d.). Neural network basics. Preuzeto 3.4.2023. s https://course.elementsofai.com/5/1

podataka (Brownlee, 2021)[122]. Problem osnovnih ponavljajućih mreža je nestajući gradijent (vrijednost kojom se osvježuju ponderi mreže). Kako se procesiraju novi podaci, tako se potiskuje utjecaj starih podataka iz prethodnih koraka, odnosno gradijent se kada dođe do starih podataka toliko smanji da stariji podaci prestanu imati utjecaj na učenje unutar mreže (Phi, 2020)[123]. LSTM mreže rješavaju taj problem preradom modula za ponavljanje, o čemu se detaljnije može pročitati na https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

Uloga LSTM-ova u optičkom prepoznavanju znakova ona je klasifikatora. Kao i prethodno navedene metode, LSTM uzima značajke iz prethodnog procesa i prema podacima na kojima je treniran predviđa o kojim se znakovima ili riječima radi (Kutvonen, 2022)[124].

4.6. Postprocesiranje

Postprocesiranje je zadnja faza rada sustava za optičko prepoznavanje znakova. U ovoj fazi, sustav pokušava pronaći i označiti greške, generirati listu potencijalnih zamjena za grešku i odabrati najbolju zamjenu (Nguyen, Jatowt, Coustaty i Doucet, 2021)[125]. Prema Nguyen i sur. (2021)[126] neke od metoda postprocesiranja su:

A)Spajanje rezultata više sustava - Ista slika provlači se kroz više sustava za optičko prepoznavanje znakova te se njihovi rezultati spajaju pomoću algoritama za usklađivanje.

B)Ispravljanje grešaka leksikonom - u ovoj metodi kontekst u kojemu se riječ nalazi ne igra ulogu, već se riječ traži u leksikonu koji je dostupan sustavu i prema tome leksikonu se generiraju greške i potencijalne zamjene korištenjem mjera udaljenosti.

C)Modeli grešaka - Modeli grešaka probabilistički su modeli koji se koriste sami ili u kombinaciji s drugim jezičnim modelima.

D)Modeli jezika bazirani na temi - Ovi modeli fokusiraju se na temu dokumenta kako bi što brže pronašli moguće greške i njihove zamjene

[122] Brownlee, J. (2021). A Gentle Introduction to Long Short-Term Memory Networks by the Experts. Preuzeto 3.4.2023. s https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/
[123] Phi, M. (2020). Illustrated Guide to LSTM's and GRU's: A step by step explanation. Preuzeto 3.4.2023. s https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21
[124] Kutvonen, A. (2022). Get started with deep learning OCR - Towards Data Science. Preuzeto 3.4.2023. s https://towardsdatascience.com/get-started-with-deep-learning-ocr-136ac645db1d
[125] Nguyen, T.T.H., Jatowt A., Coustaty, M., i Doucet, A. (2021). Survey of Post-OCR Processing Approaches. ACM Computing Surveys, 54(6). https://doi.org/10.1145/3453476
[126] Ibid.

5. Istraživanje

Cilj istraživanja je procijeniti utjecaj pretprocesiranja na rezultate očitanja znakova u sustavima za optičko prepoznavanje znakova *Tesseract* i *Google DocumentAI*.

Istraživanje se provodi nad 40 stranica teksta preuzetih iz 5 knjiga objavljenih u 19. i 20. stoljeću. Iz svake knjige preu O samoj korištenoj građi može se više pročitati u sljedećem poglavlju.

Korištene metode pretprocesiranja u istraživanju su binarizacija Otsu metodom, te pretvorba u sivu skalu.

Svaka od 40 slika biti će očitana u sustavima *Tesseract* i *Google DocumentAI* u svojoj originalnoj inačici, a zatim nakon pretvorbe u sivu skalu i finalno nakon binarizacije.

Ta očitanja u sustavima *Tesseract* i *Google DocumentAI* zatim će biti uspoređena sa unaprijed pripremljenim ručno prepisanim tekstom koji se smatra 100% točnim.

Alat koji će se koristiti za usporedbu je *OcrevalUAtion*, a bilježiti će se broj grešaka u znakovima (engl. Character Error Rate - CER) i broj grešaka u riječima (engl. Word Error Rate - WER).

Finalno, rezultati će se prikazati za svaku pojedinu skupinu građe, odnosno pojedinih 8 stranica iz svake pripadajuće knjige, prema korištenome sustavu za optičko prepoznavanje znakova.

5.1. Odabrana građa

Za potrebe istraživanja odabrano je 5 knjiga, a iz svake je preuzeto po 8 stranica. Dakle ukupan broj stranica, ili građe, nad kojima se provodi istraživanje je 40. Sva građa preuzeta je s archive.org[127] web stranice. Archive.org ispod svake knjige navodi opis skenera ili fotoaparata koji je korišten za digitalizaciju, te daje pristup originalnim fotografijama/skenovima dovoljno velikog formata za daljnje korištenje u sustavima za optičko prepoznavanje znakova. Sve knjige nad kojima se provodi istraživanje otisnute su strojem.

Knjige iz kojih su preuzete fotografije su:

[127] Internet Archive. (n.d.). Internet Archive. Preuzeto 15.6.2023. s archive.org

1.) **Plebiscit u koruškoj Sloveniji -** Ova knjiga izdana je 1920. godine te okvirno predstavlja razdoblje tekstova kojima je već isteklo autorsko pravo i stoga su podobni za digitalizaciju bez dodatnog zadatka reguliranja pristupa dokumentu (zavisno o slučaju). Knjiga je preuzeta s https://archive.org/details/plebiscitukorusk00rozi/page/8/mode/2up.

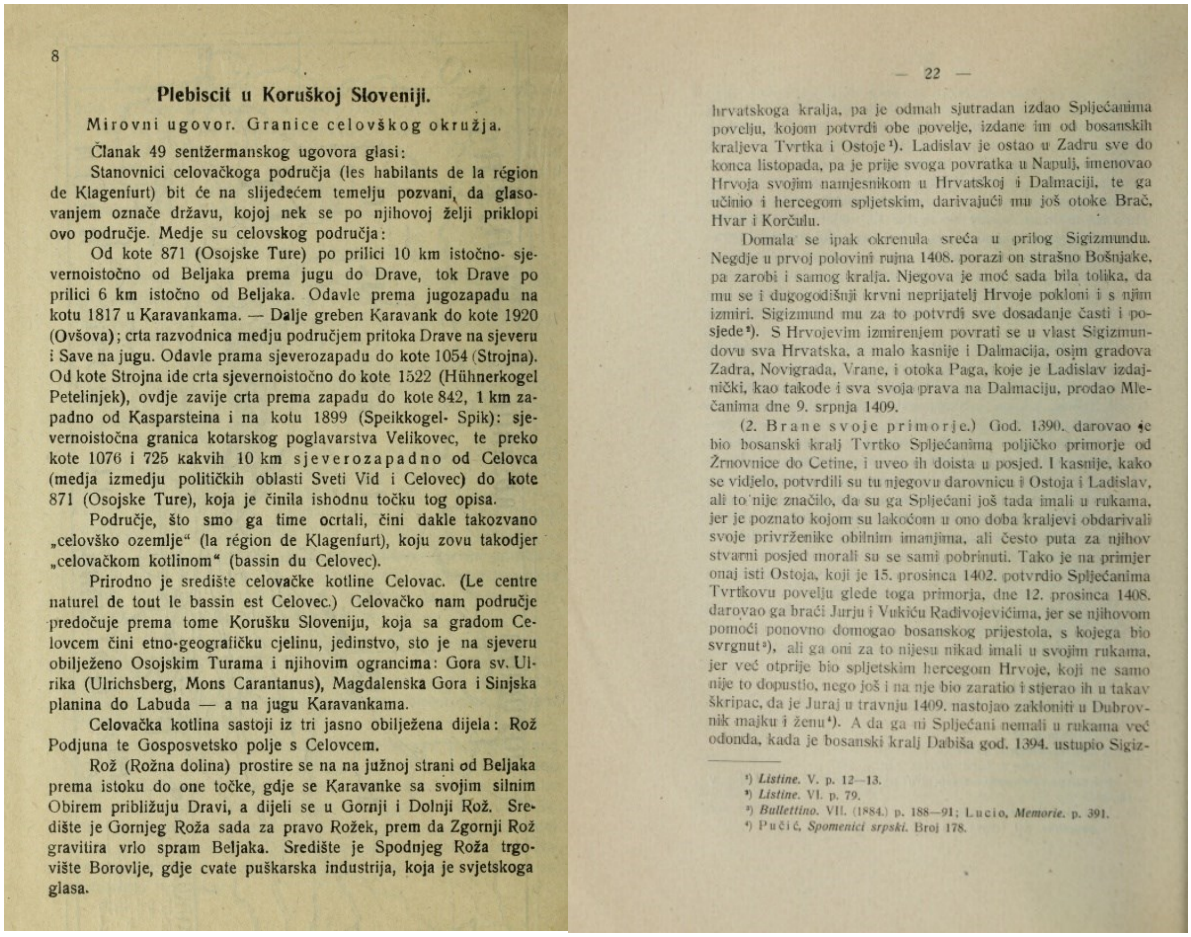
2.) **Povijest Poljica** - Ova knjiga je izdana 1921. godine, te je odabrana iz istog razloga kao i Plebiscit o koruškoj Sloveniji. Knjiga je preuzeta s https://archive.org/details/povijestpoljica00pive/page/6/mode/2up.

3.) **Runje i pahuljice; pesni porugljive i pastirske ponajveć Dubrovačke -** Knjiga je izdana 1866. godine, a pisana je arhaičnom inačicom hrvatskoga jezika. Uz samu arhaičnost jezika, ona sadrži i elemente čakavskoga narječja, što se pretpostavlja izazovnim sustavu za optičko prepoznavanje znakova. Knjiga je preuzeta s https://archive.org/details/runjeipahuljicep00kure/page/n9/mode/2up.

4.) **Djačko društvo** - Knjiga je izdana 1874. godine, a stranice su joj dodatno ukrašene - tekst je uokviren i dodani su razni vizualni elementi koji potencijalno mogu biti prepoznati kao slova u sustavima za optičko prepoznavanje znakova. Knjiga je preuzeta s https://archive.org/details/djackodruztvovel00unse/page/10/mode/2up.

5.) **Optimizam** - Ova knjiga izdana je 1931. godine i pisana je mješavinom hrvatskog i srpskog jezika. Originalne fotografije imaju neravnomjerno osvjetljenje, što se pretpostavlja izazovnim za uspješno pretprocesiranje i očitanje u sustavima za optičko prepoznavanje znakova. Knjiga je preuzeta s https://archive.org/details/optimizam00hele/page/22/mode/2up.

Na slikama 6-10 mogu se vidjeti reprezentativni uzorci iz svake od prethodno navedenih knjiga.



8

Plebiscit u Koruškoj Sloveniji.

Mirovni ugovor. Granice celovškog okružja.

Članak 49 sentžermanskog ugovora glasi:

Stanovnici celovačkoga područja (les habilants de la région de Klagenfurt) bit će na slijedećem temelju pozvani, da glasovanjem označe državu, kojoj nek se po njihovoj želji priklopi ovo područje. Medje su celovskog područja:

Od kote 871 (Osojske Ture) po prilici 10 km istočno- sjevernoistočno od Beljaka prema jugu do Drave, tok Drave po prilici 6 km istočno od Beljaka. Odavle prema jugozapadu na kotu 1817 u Karavankama. — Dalje greben Karavank do kote 1920 (Ovšova); crta razvodnica medju područjem pritoka Drave na sjeveru i Save na jugu. Odavle prama sjeverozapadu do kote 1054 (Strojna). Od kote Strojna ide crta sjevernoistočno do kote 1522 (Hühnerkogel Petelinjek), ovdje zavije crta prema zapadu do kote 842, 1 km zapadno od Kasparsteina i na kotu 1899 (Speikkogel- Spik): sjevernoistočna granica kotarskog poglavarstva Velikovec, te preko kote 1076 i 725 kakvih 10 km sjeverozapadno od Celovca (medja izmedju političkih oblasti Sveti Vid i Celovec) do kote 871 (Osojske Ture), koja je činila ishodnu točku tog opisa.

Područje, što smo ga time ocrtali, čini dakle takozvano „celovško ozemlje“ (la région de Klagenfurt), koju zovu također „celovačkom kotlinom“ (bassin du Celovec).

Prirodno je središte celovačke kotline Celovac. (Le centre naturel de tout le bassin est Celovec.) Celovačko nam područje predočuje prema tome Koruško Sloveniju, koja sa gradom Celovcem čini etno-geografičku cjelinu, jedinstvo, sto je na sjeveru obilježeno Osojskim Turama i njihovim ograncima: Gora sv. Ulrika (Ulrichsberg, Mons Carantanus), Magdalenska Gora i Sinjska planina do Labuda — a na jugu Karavankama.

Celovačka kotlina sastoji iz tri jasno obilježena dijela: Rož Podjuna te Gosposvetsko polje s Celovcem.

Rož (Rožna dolina) prostire se na na južnoj strani od Beljaka prema istoku do one točke, gdje se Karavanke sa svojim silnim Obirem približuju Dravi, a dijeli se u Gornji i Dolnji Rož. Središte je Gornjeg Roža sada za pravo Rožek, prem da Zgornji Rož gravitira vrlo spram Beljaka. Središte je Spodnjeg Roža trgovište Borovlje, gdje cvate puškarska industrija, koja je svjetskoga glasa.

— 22 —

hrvatskoga kralja, pa je odmah sjutradan izdao Spljećanima povelju, kojom potvrdi obe povelje, izdane im od bosanskih kraljeva Tvrtka i Ostoje[1]). Ladislav je ostao u Zadru sve do konca listopada, pa je prije svoga povratka u Napulj, imenovao Hrvoja svojim namjesnikom u Hrvatskoj i Dalmaciji, te ga učinio i hercegom spljetskim, darivajući mu još otoke Brač, Hvar i Korčulu.

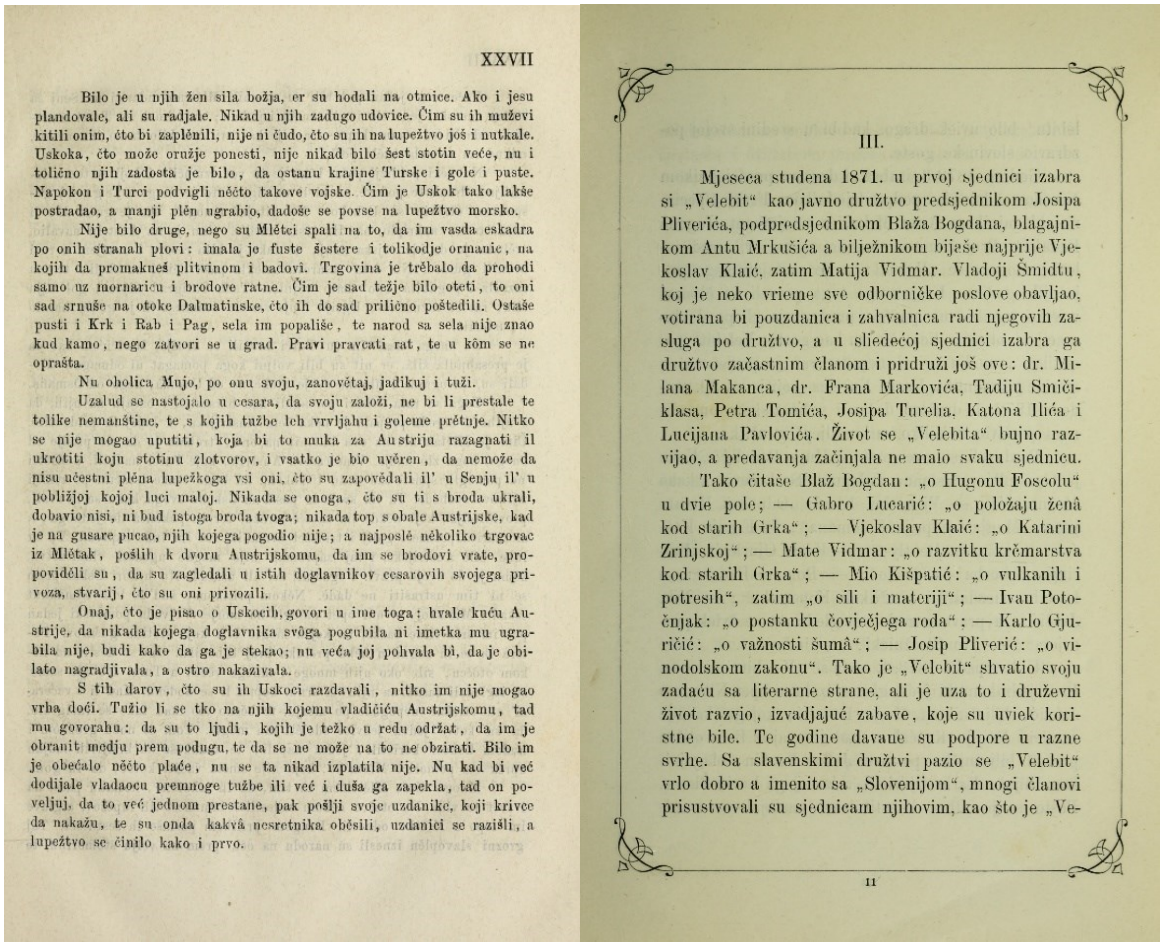
Domala se ipak okrenula sreća u prilog Sigizmundu. Negdje u prvoj polovini rujna 1408. porazi on strašno Bošnjake, pa zarobi i samog kralja. Njegova je moć sada bila tolika, da mu se i dugogodišnji krvni neprijatelj Hrvoje pokloni i s njim izmiri. Sigizmund mu za to potvrdi sve dosadanje časti i posjede[2]). S Hrvojevim izmirenjem povrati se u vlast Sigizmundovu sva Hrvatska, a malo kasnije i Dalmacija, osim gradova Zadra, Novigrada, Vrane, i otoka Paga, koje je Ladislav izdajnički, kao također i sva svoja prava na Dalmaciju, prodao Mlečanima dne 9. srpnja 1409.

(2. Brane svoje primorje.) God. 1390. darovao je bio bosanski kralj Tvrtko Spljećanima poljičko primorje od Žrnovnice do Cetine, i uveo ih doista u posjed. I kasnije, kako se vidjelo, potvrdili su tu njegovu darovnicu i Ostoja i Ladislav, ali to nije značilo, da su ga Spljećani još tada imali u rukama, jer je poznato kojom su lakoćom u ono doba kraljevi obdarivali svoje privrženike obilnim imanjima, ali često puta za njihov stvarni posjed morali su se sami pobrinuti. Tako je na primjer onaj isti Ostoja, koji je 15. prosinca 1402. potvrdio Spljećanima Tvrtkovu povelju glede toga primorja, dne 12. prosinca 1408. darovao ga braći Jurju i Vukiću Radivojevićima, jer se njihovom pomoći ponovno domogao bosanskog prijestola, s kojega bio svrgnut[3]), ali ga oni za to nijesu nikad imali u svojim rukama, jer već otprije bio spljetskim hercegom Hrvoje, koji ne samo nije to dopustio, nego još i na nje bio zaratio i stjerao ih u takav škripac, da je Juraj u travnju 1409. nastojao zakloniti u Dubrovnik majku i ženu[4]). A da ga ni Spljećani nemali u rukama već odonda, kada je bosanski kralj Dabiša god. 1394. ustupio Sigiz-

[1]) *Listine.* V. p. 12—13.
[2]) *Listine.* VI. p. 79.
[3]) *Bullettino.* VII. (1884.) p. 188—91; Lucio, *Memorie.* p. 391.
[4]) Pučić, *Spomenici srpski.* Broj 178.

Slika 6 - *Lijevo: Stranica knjige Plebiscit u Koruškoj Sloveniji*

Slika 7 - *Desno: Stranica knjige Povijest Poljica*



XXVII

Bilo je u njih žen sila božja, er su hodali na otmice. Ako i jesu plandovale, ali su radjale. Nikad u njih zadugo udovice. Čim su ih muževi kitili onim, čto bi zaplěnili, nije ni čudo, čto su ih na lupeštvo još i nutkale. Uskoka, čto može oružje ponesti, nije nikad bilo šest stotin veće, nu i tolično njih zadosta je bilo, da ostanu krajine Turske i gole i puste. Napokon i Turci podvigli něčto takove vojske. Čim je Uskok tako lakše postradao, a manji plěn ugrabio, dadoše se povse na lupeštvo morsko.

Nije bilo druge, nego su Mlětci spali na to, da im vasda eskadra po onih stranah plovi: imala je fuste šestere i tolikodje ormanic, na kojih da promakneš plitvinom i badovi. Trgovina je trěbalo da prohodi samo uz mornaricu i brodove ratne. Čim je sad težje bilo oteti, to oni sad srnuše na otoke Dalmatinske, čto ih do sad prilično poštedili. Ostaše pusti i Krk i Rab i Pag, sela im popališe, te narod sa sela nije znao kud kamo, nego zatvori se u grad. Pravi pravcati rat, te u kôm se ne oprašta.

Nu oholica Mujo, po onu svoju, zanovětaj, jadikuj i tuži.

Uzalud se nastojalo u cesara, da svoju založi, ne bi li prestale te tolike nemanštine, te s kojih tužbe Ich vrvljahu i goleme prětnje. Nitko se nije mogao uputiti, koja bi to muka za Austriju razagnati il ukrotiti koju stotinu zlotvorov, i vsatko je bio uvěren, da nemože da nisu učestni plěna lupežkoga vsi oni, čto su zapovědali il' u Senju il' u pobližjoj kojoj luci maloj. Nikada se onoga, čto su ti s broda ukrali, dobavio nisi, ni bud istoga broda tvoga; nikada top s obale Austrijske, kad je na gusare pucao, njih kojega pogodio nije; a najposlé několiko trgovac iz Mlětak, pošlih k dvoru Austrijskomu, da im se brodovi vrate, propovidĕli su, da su zagledali u istih doglavnikov cesarovih svojega privoza, stvarij, čto su oni privozili.

Onaj, čto je pisao o Uskocih, govori u ime toga: hvale kuću Austrije, da nikada kojega doglavnika svôga pogubila ni imetka mu ugrabila nije, budi kako da ga je stekao; nu veća joj pohvala bi, da je obilato nagradjivala, a ostro nakazivala.

S tih darov, čto su ih Uskoci razdavali, nitko im nije mogao vrha doći. Tužio li se tko na njih kojemu vladičiću Austrijskomu, tad mu govorahu: da su to ljudi, kojih je težko u redu održat, da im je obranit medju prem podugu, te da se ne može na to ne obzirati. Bilo im je obećalo něčto plaće, nu se ta nikad izplatila nije. Nu kad bi već dodijale vladaocu premnoge tužbe ili već i duša ga zapekla, tad on poveljuj, da to već jednom prestane, pak pošlji svoje uzdanike, koji krivce da nakažu, te su onda kakvá nesretnika obĕsili, uzdanici se razišli, a lupeštvo se činilo kako i prvo.

III.

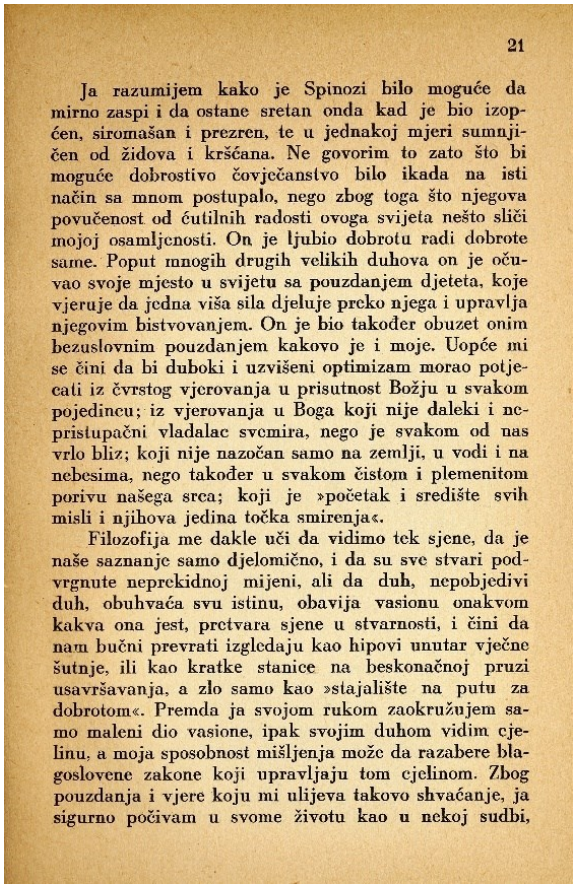
Mjeseca studena 1871. u prvoj sjednici izabra si „Velebit“ kao javno družtvo predsjednikom Josipa Pliverića, podpredsjednikom Blaža Bogdana, blagajnikom Antu Mrkušića a bilježnikom bijaše najprije Vjekoslav Klaić, zatim Matija Vidmar. Vladoji Šmidtu, koj je neko vrieme sve odborničke poslove obavljao, votirana bi pouzdanica i zahvalnica radi njegovih zasluga po družtvo, a u sliedećoj sjednici izabra ga družtvo začastnim članom i pridruži još ove: dr. Milana Makanca, dr. Frana Markovića, Tadiju Smičiklasa, Petra Tomića, Josipa Turelia. Katona Ilića i Lucijana Pavlovića. Život se „Velebita“ bujno razvijao, a predavanja začinjala ne malo svaku sjednicu.

Tako čitaše Blaž Bogdan: „o Hugonu Foscolu“ u dvie pole; — Gabro Lucarić: „o položaju ženâ kod starih Grka“; — Vjekoslav Klaić: „o Katarini Zrinjskoj“; — Mate Vidmar: „o razvitku krčmarstva kod starih Grka“; — Mio Kišpatić: „o vulkanih i potresih“, zatim „o sili i materiji“; — Ivan Potočnjak: „o postanku čovječjega roda“: — Karlo Gjuričić: „o važnosti šumâ“; — Josip Pliverić: „o vinodolskom zakonu“. Tako je „Velebit“ shvatio svoju zadaću sa literarne strane, ali je uza to i druževni život razvio, izvadjajuć zabave, koje su uviek koristne bile. Te godine davane su podpore u razne svrhe. Sa slavenskimi družtvi pazio se „Velebit“ vrlo dobro a imenito sa „Slovenijom“, mnogi članovi prisustvovali su sjednicam njihovim, kao što je „Ve-

11

Slika 8 - *Lijevo: Stranica knjige Runje i pahuljice; pesni porugljive i pastirske ponajveć Dubrovačke*

Slika 9 - *Desno*: *Stranica knjige Djačko društvo*



21

Ja razumijem kako je Spinozi bilo moguće da mirno zaspi i da ostane sretan onda kad je bio izopćen, siromašan i prezren, te u jednakoj mjeri sumnjičen od židova i kršćana. Ne govorim to zato što bi moguće dobrostivo čovječanstvo bilo ikada na isti način sa mnom postupalo, nego zbog toga što njegova povučenost od ćutilnih radosti ovoga svijeta nešto sliči mojoj osamljenosti. On je ljubio dobrotu radi dobrote same. Poput mnogih drugih velikih duhova on je očuvao svoje mjesto u svijetu sa pouzdanjem djeteta, koje vjeruje da jedna viša sila djeluje preko njega i upravlja njegovim bistvovanjem. On je bio također obuzet onim bezuslovnim pouzdanjem kakovo je i moje. Uopće mi se čini da bi duboki i uzvišeni optimizam morao potjecati iz čvrstog vjerovanja u prisutnost Božju u svakom pojedincu; iz vjerovanja u Boga koji nije daleki i nepristupačni vladalac svemira, nego je svakom od nas vrlo bliz; koji nije nazočan samo na zemlji, u vodi i na nebesima, nego također u svakom čistom i plemenitom porivu našega srca; koji je »početak i središte svih misli i njihova jedina točka smirenja«.

Filozofija me dakle uči da vidimo tek sjene, da je naše saznanje samo djelomično, i da su sve stvari podvrgnute neprekidnoj mijeni, ali da duh, nepobjedivi duh, obuhvaća svu istinu, obavija vasionu onakvom kakva ona jest, pretvara sjene u stvarnosti, i čini da nam bučni prevrati izgledaju kao hipovi unutar vječne šutnje, ili kao kratke stanice na beskonačnoj pruzi usavršavanja, a zlo samo kao »stajalište na putu za dobrotom«. Premda ja svojom rukom zaokružujem samo maleni dio vasione, ipak svojim duhom vidim cjelinu, a moja sposobnost mišljenja može da razabere blagoslovene zakone koji upravljaju tom cjelinom. Zbog pouzdanja i vjere koju mi ulijeva takovo shvaćanje, ja sigurno počivam u svome životu kao u nekoj sudbi,

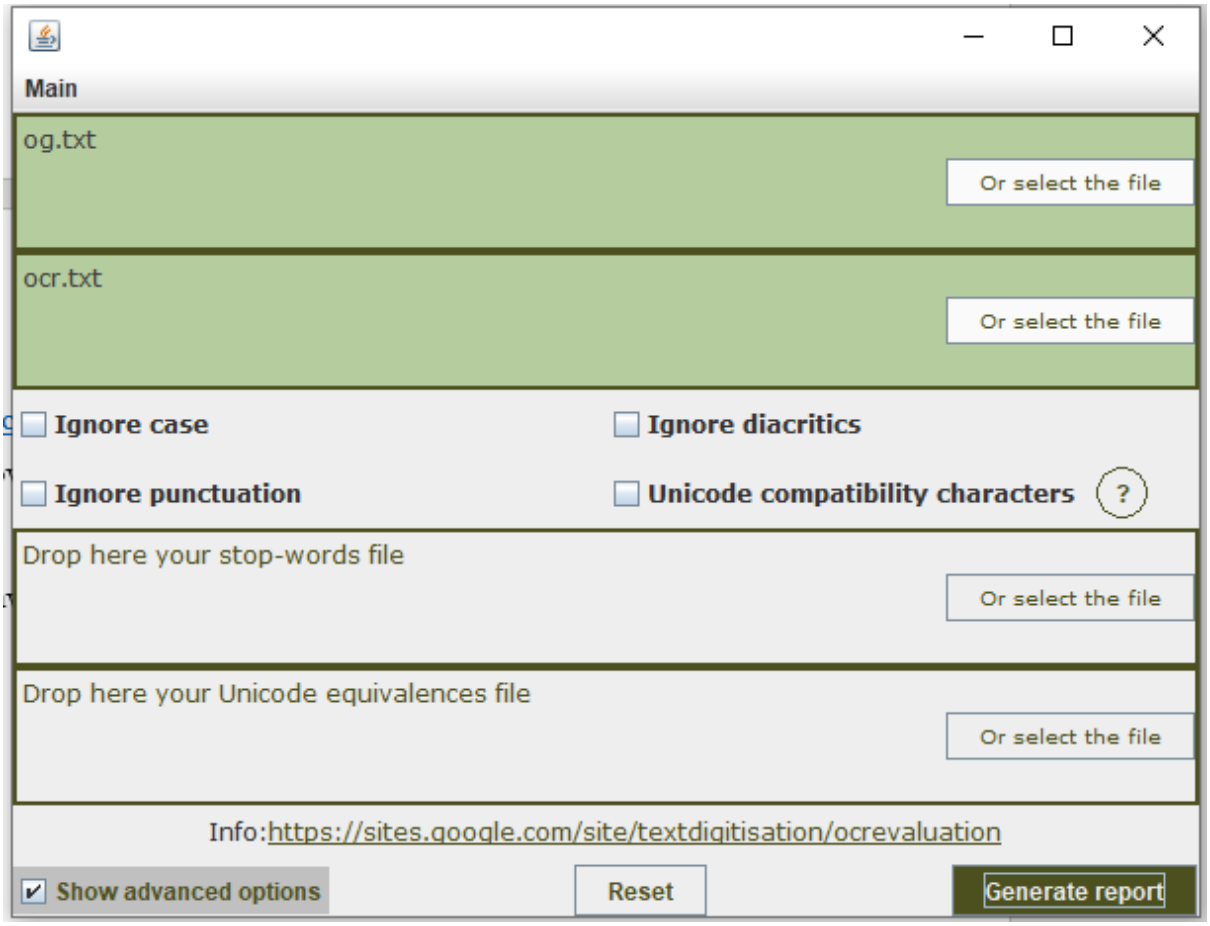
Slika 10 - *Stranica knjige Optimizam*

5.2. Evaluacija

Za određivanje točnosti prepoznavanja znakova koristiti će se alat *OcrevalUAtion*[128]. *OcrevalUAtion* alat je otvorenog koda baziran na Javi. Njegove značajke podosta su osnovne u usporedbi s alatom *ISRI*[129] i njegovom modernijom inačicom, alatom *OCReval*[130], no za potrebe ovoga rada dostatne.

Alat *OcrevalUAtion* može se koristiti putem CMD-a ili jednostavnog sučelja koje se može vidjeti na Slici 11.

[128] OcrevaluUAtion. (n.d.). OcrevalUAtion. U GitHub. Preuzeto 13.6.2023. s https://github.com/impactcentre/ocrevalUAtion
[129] ISRI OCR evaluation tools. (n.d.). U Google Code. Preuzeto 13.6.2023. s https://code.google.com/archive/p/isri-ocr-evaluation-tools/
[130] Stamos, E.A. (n.d.). OCReval. U GitHub. Preuzeto 13.6.2023. s https://github.com/eddieantonio/ocreval



Slika 11 - *OcrevalUAtion sučelje*

Potrebno je učitati dvije datoteke - potpuno točan prijepis teksta koji se pokušava „pročitati“ u sustavu za optičko prepoznavanje znakova, te rezultat procesiranja dobiven iz sustava za optičko prepoznavanje znakova. Alat prihvaća .txt, .hOCR, .html te više vrsta XML datoteka (Carasco, n.d.b).

Dalje se mogu podesiti dodatne opcije kao što je ignoriranje određenih značajki teksta, te dodavanje dodatnih znakova za prepoznavanje u formatu Unicode.

Nakon podešavanja značajki, generiraju se rezultati usporedbe koji se pohranjuju u .html datoteci.

Slika 12 prikazuje primjer dobivenih rezultata u alatu *OcrevalUAtion*. Pod naslovom „*Difference spotting*“ nalaze se 2 teksta od kojih svaki sadrži istu rečenicu s jednom promijenjenom riječi. Na slici se može vidjeti da alat crvenom bojom označava mjesto gdje se riječi ne podudaraju, a zelenom znakove koji su višak.

General results

CER	41.67
WER	25.00
WER (order independent)	25.00

Difference spotting

og.txt	ocr.txt
I am a šljam	I am a džip

Error rate per character and type

Character	Hex code	Total	Spurious	Confused	Lost	Error rate
	20	3	0	0	0	0.00
I	49	1	0	0	0	0.00
a	61	3	0	1	0	33.33
j	6a	1	0	1	0	100.00
l	6c	1	0	1	0	100.00
m	6d	2	0	0	1	50.00
š	161	1	0	1	0	100.00

Slika 12 - *Prikaz rezultata alata OcrevalUAtion*

CER stoji za „Character Error Rate“ i u postotcima izražava broj „krivih“ znakova unutar teksta. No, CER se ne izračunava doslovno. U primjeru gdje se uspoređuje „Ernest“ u točnome tekstu s „rnst“ u očitanome tekstu, ako bismo gledali samo pozicije u tekstu, ova usporedba bi generirala 100% CER (E>r, r>n, n>s, e>t, s>_, t>_), zato *OcrevalUAtion* u CER kalkulacijama primjenjuje Levenshteinovu udaljenost (Carasco, n.d.).

Levenshteinova udaljenost je mjera najmanjeg broja promjena znakova kako bi se dobio točan rezultat (Leung, 2022).

Formula za CER primjenom Levenshteinove udaljenosti bila bi:

CER = (S+D+I)/N

gdje je:

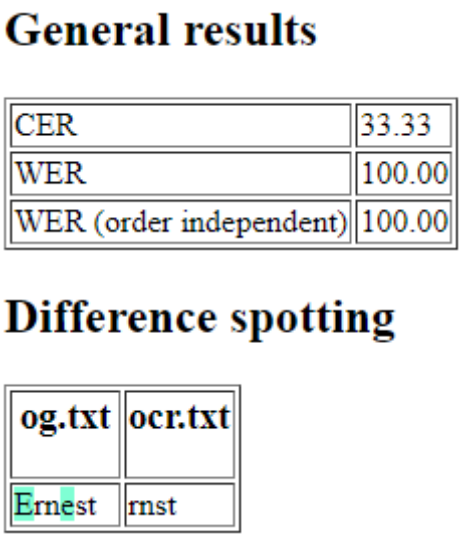
S = najmanji broj potrebnih zamjena slova

D = najmanji broj potrebnih brisanja slova

I = najmanji broj znakova koje je potrebno ubaciti

N = broj znakova u točnoj riječi

Izračun ranije navedenog primjera „Ernest“ i „rnst“, prema Levenshteinovoj udaljenosti, bio bi (0+0+2)/6, čime bi se dobilo CER od 33%. Na slici 13 mogu se vidjeti rezultati mjere CER dobiveni primjenom alata *OcrevalUAtion*.



General results

CER	33.33
WER	100.00
WER (order independent)	100.00

Difference spotting

og.txt	ocr.txt
Ernest	rnst

Slika 13 - *OcrevalUAtion rezultati za primjer „Ernest“ i „rnst“*

Druga kategorija rezultata vidljiva na slikama 12 i 13 je WER ili Word Error Rate, odnosno broj pogrešnih riječi. Ovdje se također primjenjuje Levenshteinova udaljenost, samo na razini riječi (Leung, 2022):

WER = (S+D+I)/N

gdje je:

S = broj supstitucija riječi u rezultatu (npr. pore u bore)

I = broj umetaka riječi u rezultatu (npr. mravojed u mravo jed)

D = broj izostavljenih riječi u rezultatu (npr. Išao medu u dućan -> Išao medo dućan)

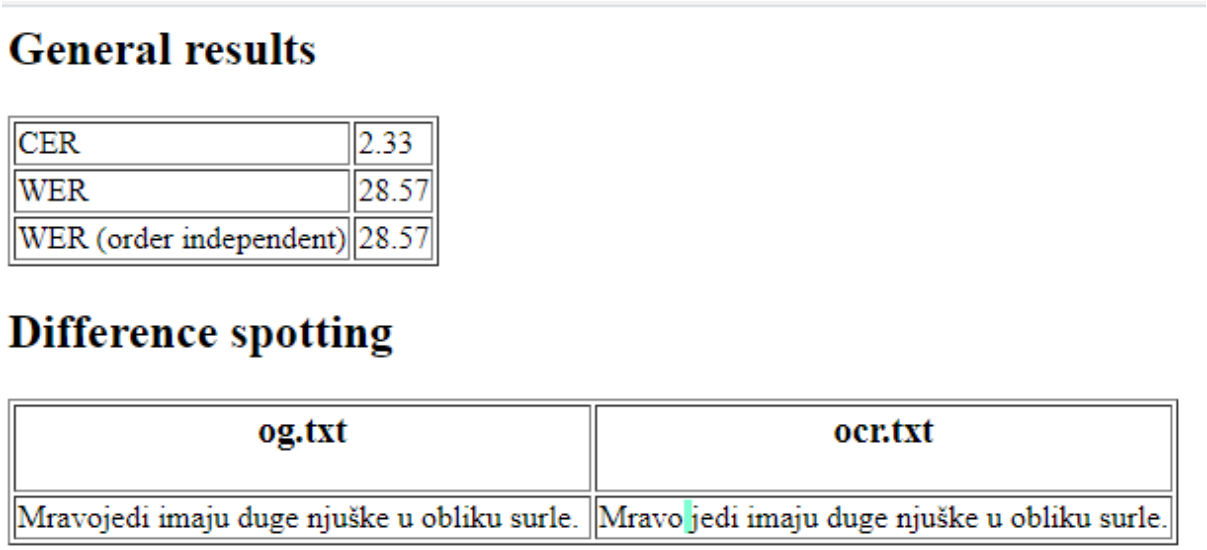
N = ukupan broj riječi

Npr. ako u izvornom tekstu stoji rečenica „Mravojedi imaju duge njuške u obliku surle.“, a rezultati očitanja u sustavu za optičko prepoznavanje znakova su „Mravo jedi imaju duge njuške u obliku surle.“, WER formula izračunala bi se na sljedeći način:

Riječ mravojedi prvo je zamijenjena („Mravojedi“ u „Mravo“), a zatim je umetnuta „jedi“.

Uvrštavanjem u formulu dobivamo 2/7=28.57%.

Na Slici 14 može se vidjeti rezultate u alatu *OcrevalUAtion* za isti ovaj primjer.



General results

CER	2.33
WER	28.57
WER (order independent)	28.57

Difference spotting

og.txt	ocr.txt
Mravojedi imaju duge njuške u obliku surle.	Mravo jedi imaju duge njuške u obliku surle.

Slika 14 - *OcrevalUAtion rezultati za primjer „Mravojedi“ u „Mravo jedi“.*

5.3. Pretprocesiranje

Najčešće faze pretprocesiranja koje se predlažu za procesiranje slika prije korištenja u sustavu za optičko prepoznavanje znakova su konverzija u sivu ljestvicu (MKM, 2020)[131] i

binarizacija (Stančić, 2009[132]; Tesseract, n.d.[133]), te će se upravo te operacije i koristi u istraživanju.

Bilježiti će se rezultati za slike prije obrade, te nakon primjene svake od navedenih operacija pretprocesiranja.

Operacije pretprocesiranja rađene su u Pythonu s knjižnicom OpenCV. U nastavku se može pronaći dio koda korištenog za operacije pretvorbe u sivu skalu i Otsu binarizaciju.

[131] Ministarstvo kulture i medija Republike Hrvatske. (2020). Smjernice za digitalizaciju kulturne baštine.
[132] Stančić, H. (2009). Digitalizacija. Zagreb: Zavod za informacijske studije. str.56.
[133] Tesseract. (n.d.). Improving the quality of the output. U GitHub. Preuzeto 31.3.2023. s https://tesseract-ocr.github.io/tessdoc/ImproveQuality.html

```
import cv2
import os


#pretvorba u sivu skalu i binarizacija
def sivo_i_bin(folder, t_folder, t_folder_b):
    for slika in os.listdir(folder):
        lokacija = os.path.join(folder, slika)
        slika_og = cv2.imread(lokacija)
        #siva skala
        grey = cv2.cvtColor(slika_og, cv2.COLOR_BGR2GRAY)
        naziv = str.replace(slika, '.tiff', '')
        cv2.imwrite(os.path.join(t_folder, 'grey'+naziv+'.tiff'), grey)
        #otsu binarizacija, prima sivu skalu
        tVal,otsu = cv2.threshold(grey,0,255,cv2.THRESH_BINARY+cv2.THRESH_OTSU)
        cv2.imwrite(os.path.join(t_folder_b, 'otsu'+naziv+'.tiff'), otsu)
```

5.4. Rad s alatom Tesseract

Tesseract nema grafičko sučelje za rad, već se u njemu slike učitavaju putem terminala. Prema Tesseract (n.d.), *Tesseract* kao argumente prima:

-oem 1 kako bi koristio LSTM mrežu za čitanje ili **--oem 0** kako bi koristio starije algoritme za prepoznavanje nizova.

-l lang se koristi kako bi se definirao tekst jezika koji pokušavamo „pročitati“. „lang“ se zamjenjuje s kraticom od tri znaka koja predstavlja jezik. Ovaj argument podržava korištenje više jezika istovremeno: -l eng+hrv pročitat će tekst koristeći engleski kao primarni jezik.

hocr / **pdf** / **tsv** / **txt** ... određuje u kojemu formatu ćemo dobiti finalan rezultat prepoznavanja znakova. Za potrebe ovoga rada koristiti će se .txt pošto je s njime najjednostavnije raditi.

-psm 3 je postavka koja će se primijeniti ako se ne specificira suprotno, a označava automatsku segmentaciju stranice bez orijentacije i detekcije korištenog pisma (engl. Orientation and Script

Detection). Za potrebe ovoga rada koristiti će se -psm 6 koji alatu *Tesseract* daje do znanja da se koriste uniformni blokovi teksta, odnosno da tekst nije u više stupaca kao npr. kazalo i da se radi o velikom bloku teksta.

Tesseract će se koristiti sa sličnim postavkama za sve dokumente, no za Plebiscit i Runje biti će određeni dodatni dopunski jezici koji će olakšati prepoznavanje dijakritičkih znakova netipičnih za hrvatski jezik. Osnovna komanda kojom će se pokretati *Tesseract* biti će:

Tesseract slike\ime.tiff slike\ime_rezultat -l hrv --psm 6 txt

Iteracija kroz slike implementirana je .bat datotekom čiji se dio može pronaći u nastavku.

```
@Echo off
SetLocal DisableDelayedExpansion

Set "_SourcePath=C:\Users\pc\Desktop\primjer\tiff\simple"
Set "_SourceMask=*.tiff"
Set "_OutputPath=C:\Users\pc\Desktop\primjer\tess-rezultat"
Set "_TesserFile=C:\Program Files\Tesseract-OCR\Tesseract.exe"

For /F "Delims=" %%A In (
    '""%__AppDir__%where.exe" /R "%_SourcePath%" "%_SourceMask%" 2>Nul"'
) Do Echo %%A...& "%_TesserFile%" "%%A" "%_OutputPath%\%%~nA" -l hrv --psm 6
txt
```

5.4.1. Određivanje kombinacije jezika primjerene za građu

Prije pretprocesiranja slika, na originalima s „problematičnim“ znakovima koji se ne koriste u modernome hrvatskome jeziku napravljeno je testiranje jezika kako bi se odredio jezik koji će dati najbolje rezultate. Testiranje s jezikom implementirano je prije pretprocesiranja kako bi se u što ranijoj fazi razaznala idealna kombinacija jezika za daljnje korištenje.

Važno je napomenuti da bi se kao alternativu korištenju drugih jezika moglo trenirati novi model za hrvatski jezik, usmjeren na specifične vremenske periode ili dijakritičke znakove.

Prvi dokument na kojemu se radilo je **Plebiscit u koruškoj Sloveniji**.

Dijakritički znakovi sadržani u tekstu, a netipični za hrvatski jezik su: é, ü

Isprobane su sljedeće kombinacije sa sljedećim rezultatima:

1.) Samo hrvatski jezik

raw_plebiscit_rezultat.txt

8 di Plebiscit u Koruškoj Sloveniji. ' > Mirovni ugovor. Granice celovškog okružja. Članak 49 sentžermanskog ugovora glasi: Stanovnici celovačkoga područja (les habilants de la rćgion de Klageniurt) bit će na slijedećem temelju pozvani, da glasovanjem označe državu, kojoj nek se po njihovoj želji priklopi ovo područje. Medje su celovskog područja: | Od kote 871 (Osojske Ture) po prilici 10 km istočno-|sje- vernoistočno od Beljaka prema jugu do Drave, tok Drave po prilici 6 km istočno od Beljaka. Odavle prema jugozapadu na kotu 1817 u Karavankama. — Dalje greben Karavank do kote 1920 | (Ovšova)|; crta razvodnica medju područjem pritoka Drave na sjeveru i Save na jugu. Odavle prama sjeverozapadu do kote 1054 (Štrojna). Od kote Strojna ide crta sjevernoistočno do kote 1522 (Hiihnerkogel Petelinjek), ovdje zavije crta prema zapadu do kole 842, 1 km za- padno od Kasparsteina i na kotu 1899 (Speikkogel+- Spik): sje- vernoistočna granica kotarskog poglavarstva Velikovec, te preko kote 1076 i 725 kakvih 10 km sjeverozapadno od Celovca (medja izmedju političkih oblasti Sveti Vid i Celovec) do kote | 871 (Osojske Ture), koja je činila ishodnu točku tog opisa. Područje, što smo ga time ocrtali, čini dakle takozvano »celovško ozemlje“ (la rćgion de Klageniuri), koju zovu takodjer ' »celovačkom kotlinom“ (bassin du Celovec). Prirodno je središte celovačke kotline Celovac. (Le centre naturel de tout le bassin est Celovec.) Celovačko nam područje predočuje prema tome Korušku Sloveniju, koja sa gradom Ce- lovcem čini etno-geografičku cjelinu, jedinstvo, sto je na sjeveru obilježeno Osojskim Turama i njihovim ograncima: Gora sv. Ul- rika (Ulrichsberg, Mons Carantanus), Magdalenška Gora i Sinjska planina do Labuda — a na jugu Karavankama. Celovačka kotlina sastoji iz tri jasno obilježena dijela: Rož Podjuna te Gosposvetsko polje s Celovcem, Rož (Rožna dolina) prostire se na na južnoj strani od Beljaka prema istoku do one točke, gdje se Karavanke sa svojim silnim_| Obirem približuju Dravi, a dijeli se u Gornji i Dolnji Rož. Sre- dište je Gornjeg Roža sada za pravo Rožek, prem da Zgornji Rož gravitira vrlo spram Beljaka. Središte je Spodnjeg Roža trgo- vište Borovlje, gdje cvate puškarska industrija, koja je svjetskoga glasa.

Slika 15 - *Plebiscit, prepoznavanje samo s hrvatskim jezikom*

Kao što se može vidjeti na slici, prepoznavanje samo s hrvatskim jezikom nije uspjelo prepoznati problematične dijakritičke znakove, s rezultatom:

General results

CER	1.57
WER	3.52
WER (order independent)	3.52

Slika 16 - *Plebiscit, CER i WER samo s hrvatskim jezikom*

2.) Hrvatski i njemački

raw_plebiscit_rezultat_deu.txt

8 bh Plebiscit u Koruškoj Sloveniji. \ > Mirovni ugovor. Granice celovškog okru2ja. _ Članak 49 sentžermanskog ugovora glasi: Stanovnici celovačkoga područja (les habilants de la region de Klagenfurt) bit ce na slijedecem temelju pozvani, da glaso- vanjem označe državu, kojoj nek se po njihovoj želji priklopi ovo područje. Medje su celovskog područja: | Od kote 871 (Osojske Ture) po prilici 10 km istočno- | sje- vernoistočno od Beljaka prema jugu do Drave, tok Drave po prilici 6 km istočno od Beljaka. Odavle prema jugozapadu na kotu 1817 u Karavankama. — Dalje greben Karavank do kote 1920 | (Ovšova) |; crta razvodnica medju područjem pritoka Drave na sjeveru i Save na jugu. Odavle prama sjeverozapadu do kote 1054 (Strojna). Od kote Strojna ide crta sjevernoistočno do kote 1522 (Hiihnerkogel Petelinjek), ovdje zavije crta prema zapadu do kole 842, 1 km za- padno od Kasparsteina i na kotu 1899 (Speikkogel- Spik): sje- vernoistočna granica kotarskog poglavarstva Velikovec, te preko kote 1076 i 725 kakvih 10 km sjeverozapadno od Celovca (medja izmedju političkih oblasti Sveti Vid i Celovec) do kote | 871 (Osojske Ture), koja je činila ishodnu točku tog opisa. Područje, što smo ga time ocrtali, čini dakle takozvano »celovško ozemlje“ (la region de Klagenfurt), koju zovu takodjer ' „celovackom kotlinom“ (bassin du Celovec). Prirodno je središte celovačke kotline Celovac. (Le centre naturel de tout le bassin est Celovec.) Celovačko nam područje predočuje prema tome Korušku Sloveniju, koja sa gradom Ce- lovcem čini etno-geografic̈ku cjelinu, jedinstvo, sto je na sjeveru obilježeno Osojskim Turama i njihovim ograncima: Gora sv. Ul- rika (Ulrichsberg, Mons Carantanus), Magdalenška Gora i Sinjska planina do Labuda — a na jugu Karavankama. Celovačka kotlina sastoji iz tri jasno obilježena dijela: Rož Podjuna te Gosposvetsko polje s Celovcem, Rož (Rožna dolina) prostire se na na južnoj strani od Beljaka prema istoku do one točke, gdje se Karavanke sa svojim silnim Obirem približuju Dravi, a dijeli se u Gornji i Dolnji Rož. Sre- dište je Gornjeg Roža sada za pravo Rožek, prem da Zgornji Rož gravitira vrlo spram Beljaka. Središte je Spodnjeg Roža trgo- vište Borovlje, gdje cvate puškarska industrija, koja je svjetskoga glasa.

Slika 17 - *Plebiscit, rezultati s njemačkim i hrvatskim*

Kombinacija s njemačkim pomaže prepoznati hrvatski stil citiranja „“, no neki znakovi točno prepoznati u hrvatskome (č,ć,ž), pretvoreni su u c i 2, a problematični znakovi nisu prepoznati. Rezultat:

General results

CER	1.48
WER	4.40
WER (order independent)	4.40

Slika 18 - *Plebiscit, CER i WER s hrvatskim i njemačkim jezikom*

3.) Hrvatski i islandski

raw_plebiscit_rezultat_isl.txt

8 di Plebiscit u Koruškoj Sloveniji. ' > Mirovni ugovor. Granice celovškog okružja. Članak 49 sentžermanskog ugovora glasi: Stanovnici celovačkoga područja (les habilants de la région de Klagenfurt) bit će na slijedećem temelju pozvani, da glasovanjem označe državu, kojoj nek se po njihovoj želji priklopi ovo područje. Medje su celovskog područja: | Od kote 871 (Osojske Ture) po prilici 10 km istočno- sje- vernoistočno od Beljaka prema jugu do Drave, tok Drave po prilici 6 km istočno od Beljaka. Odavle prema jugozapadu na kotu 1817 u Karavankama. — Dalje greben Karavank do kote 1920 | (Ovšova) ; crta razvodnica medju područjem pritoka Drave na sjeveru i Save na jugu. Odavle prama sjeverozapadu do kote 1054 (Strojna). Od kote Strojna ide crta sjevernoistočno do kote 1522 (Hühnerkogel Petelinjek), ovdje zavije crta prema zapadu do kole 842, 1 km za- padno od Kasparsteina i na kotu 1899 (Speikkogel- Spik): sje- vernoistočna granica kotarskog poglavarstva Velikovec, te preko kote 1076 i 725 kakvih 10 km sjeverozapadno od Celovca (medja izmedju političkih oblasti Sveti Vid i Celovec) do kote | 871 (Osojske Ture), koja je činila ishodnu točku tog opisa. Područje, što smo ga time ocrtali, čini dakle takozvano »celovško ozemlje“ (la région de Klagenfurt), koju zovu takodjer ' »celovačkom kotlinom“ (bassin du Celovec). Prirodno je središte celovačke kotline Celovac. (Le centre naturel de tout le bassin est Celovec.) Celovačko nam područje predočuje prema tome Korušku Sloveniju, koja sa gradom Ce- lovcem čini etno-geografičku cjelinu, jedinstvo, sto je na sjeveru obilježeno Osojskim Turama i njihovim ograncima: Gora sv. Ul- rika (Ulrichsberg, Mons Carantanus), Magdalenška Gora i Sinjska planina do Labuda — a na jugu Karavankama. Celovačka kotlina sastoji iz tri jasno obilježena dijela: Rož Podjuna te Gosposvetsko polje s Celovcem, Rož (Rožna dolina) prostire se na na južnoj strani od Beljaka prema istoku do one točke, gdje se Karavanke sa svojim silnim Obirem približuju Dravi, a dijeli se u Gornji i Dolnji Rož. Sre- dište je Gornjeg Roža sada za pravo Rožek, prem da Zgornji Rož gravitira vrlo spram Beljaka. Središte je Spodnjeg Roža trgo- vište Borovlje, gdje cvate puškarska industrija, koja je svjetskoga glasa.

Slika 19 - *Plebiscit, rezultati kombinacije hrvatski i islandski*

Kombinacija s islandskim odabrana je radi dijakritičkih znakova sadržanih u islandskome jeziku. Prepoznati su svi potrebni znakovi izuzev znakova citiranja i slova ü s rezultatom:

General results

CER	1.17
WER	2.05
WER (order independent)	2.05

Slika 20 - *Plebiscit, CER i WER kombinacije hrvatski i islandski*

Idući, a ujedno i zadnji dokument s „problematičnim“ znakovima je **Runje i pahuljice; pesni porugljive i pastirske ponajveć Dubrovačke**

Znakovi unutar Runja netipični za hrvatski jezik: ě, ŏ, â

1.)Samo hrvatski jezik

raw_runje_rezultat.txt

XXVII Bilo je u njih žen sila božja, er su hodali na otmice, Ako i jesu plandovale, ali su radjale. Nikad u njih zadugo udovice. Čim su ih muževi kitili onim, čto bi zaplčnili, nije ni čudo, čto su ih na lupežtvo još i nutkale. Uskoka, čto može oružje ponesti, nije nikad bilo Šest stotin veće, nu i tolično njih zadosta je bilo, da ostanu krajine Turske i gole i puste. Napokon i Turci podvigli nččto takove vojske. Čim je Uskok tako lakše postradao, a manji pičn ugrabio, dadoše se povse na lupežtvo morsko. Nije bilo druge, nego su Mlčtci spali na to, da im vasda eskadra po onih stranah plovi: imala je fuste šestere i tolikodje ormanic, na kojih da promakneš plitvinom i badovi. Trgovina je trčbalo da prohodi samo uz mornaricu i brodove ratne. Čim je sad težje bilo oteti, to oni sad srnuše na otoke Dalmatinske, čto ih do sad prilično poštedili. Ostaše pusti i Krk i Rab i Pag, sela im popališe, te narod sa sela nije znao kud kamo, nego zatvori se u grad. Pravi pravcati rat, te u kom se ne oprašta. Nu oholica Mujo, po onu svoju, zanovčtaj, jadikuj i tuži. Uzalud se nastojalo u cesara, da svoju založi, ne bi li prestale te tolike nemanštine, te s kojih tužbe leh vrvljahu i goleme prčtnje. Nitko se nije mogao uputiti, koja bi to muka za Austriju razagnati il ukrotiti koju stotinu zlotvorov, i vsatko je bio uvčren, da nemože da nisu učestni pična lupežkoga vsi oni, čto su zapovčdali il' u Senju il' u pobližjoj kojoj luci maloj. Nikada se onoga, čto su ti s broda ukrali, dobavio nisi, ni bud istoga broda tvoga; nikada top sobale Austrijske, kad jena gusare pucao, njih kojega pogodio nije; a najposlč nčkoliko trgovac iz Mičtak, pošlih k dvoru Austrijskomu, da im se brodovi vrate, pro- povidčli su, da su zagledali u istih doglavnikov cesarovih svojega pri- voza, stvarij, čto su oni privozili. Onaj, čto je pisao o Uskocih, govori u ime toga: hvale kuću Au- strije, da nikada kojega doglavnika svoga pogubila ni imetka mu ugra- bila nije, budi kako da ga je stekao; nu veća joj pohvala bi, da je obi- lato nagradjivala, a ostro nakazivala. S tih darov, čto su ih Uskoci razdavali, nitko im nije mogao vrha doći. Tužio li se tko na njih kojemu vladičiću Austrijskomu, tad mu govorahu: da su to ljudi, kojih je težko u redu održat, da im je obranit medju prem podugu, te da se ne može na to ne obzirati. Bilo im je obećalo nččto plaće, nu se ta nikad izplatila nije. Nu kad bi već dodijale vladaocu premnoge tužbe ili već i duša ga zapekla, tad on po- veljuj, da to već jednom prestane, pak pošlji svoje uzdanike, koji krivce da nakažu, te su onda kakva nesretnika občsili, uzdanići se razišli, a lupežtvo se činilo kako i prvo.

Slika 21 - *Runje rezultati samo s hrvatskim jezikom*

Netipični znakovi nisu prepoznati s rezultatom:

General results

CER	1.59
WER	5.59
WER (order independent)	5.38

Slika 22 - Runje CER i WER samo s hrvatskim jezikom

2.)Hrvatski + slovački

raw_runje_rezultat_slk.txt
XXVII Bilo je u njih žen sila božja, er su hodali na otmice, Ako i jesu plandovale, ali su radjale. Nikad u njih zadugo udovice. Čim su ih muževi kitili onim, čto bi zaplčnili, nije ni čudo, čto su ih na lupežtvo još i nutkale. Uskoka, čto može oružje ponesti, nije nikad bilo šest stotin veće, nu i tolično njih zadosta je bilo, da ostanu krajine Turske i gole i puste. Napokon i Turci podvigli nččto takove vojske. Čim je Uskok tako lakše postradao, a manji pičn ugrabio, dadoše se povse na lupežtvo morsko. | Nije bilo druge, nego su Mlétci spali na to, da im vasda eskadra po onih stranah plovi: imala je fuste šestere i tolikodje ormanic, na kojih da promakneš plitvinom i badovi. Trgovina je trčbalo da prohodi samo uz mornaricu i brodove ratne. Čim je sad težje bilo oteti, to oni sad srnuše na otoke Dalmatinske, čto ih do sad prilično poštedili. Ostaše pusti i Krk i Rab i Pag, sela im popališe, te narod sa sela nije znao kud kamo, nego zatvori se u grad. Pravi pravcati rat, te u kom se ne oprašta. Nu oholica Mujo, po onu svoju, zanovčtaj, jadikuj i tuži. | Uzalud se nastojalo u cesara, da svoju založi, ne bi li prestale te tolike nemanštine, te s kojih tužbe leh vrvljahu i goleme prčtnje. Nitko se nije mogao uputiti, koja bi to muka za Austriju razagnati il ukrotiti koju stotinu zlotvorov, i vsatko je bio uvčren, da nemože da nisu učestni pléna lupežkoga vsi oni, čto su zapovédali il" u Senju il' u pobližjoj kojoj luci maloj. Nikada se onoga, čto su ti s broda ukrali, dobavio nisi, ni bud istoga broda tvoga; nikada top sobale Austrijske, kad jena gusare pucao, njih kojega pogodio nije; a najposlé nékoliko trgovac iz Mičtak, pošlih k dvoru Austrijskomu, da im se brodovi vrate, pro- povidčli su, da su zagledali u istih doglavnikov cesarovih svojega pri- voza, stvarij, čto su oni privozili. ||| Onaj, čto je pisao o Uskocih, govori u ime toga: hvale kuću Au- strije, da nikada kojega doglavnika svoga pogubila ni imetka mu ugra- bila nije, budi kako da ga je stekao; nu veća joj pohvala bi, da je obi- lato nagradjivala, a ostro nakazivala. S tih darov, čto su ih Uskoci razdavali, nitko im nije mogao vrha doći. Tužio li se tko na njih kojemu vladičiću Austrijskomu, tad mu govorahu: da su to ljudi, kojih je težko u redu održat, da im je obranit medju prem podugu, te da se ne može na to ne obzirati. Bilo im je obećalo nččto plaće, nu se ta nikad izplatila nije. Nu kad bi već dodijale vladaocu premnoge tužbe ili već i duša ga zapekla, tad on po- veljuj, da to već jednom prestane, pak pošlji svoje uzdanike, koji krivce da nakažu, te su onda kakva nesretnika občsili, uzdaniči se razišli, a lupežtvo se činilo kako i prvo.

Slika 23 - *Runje rezultati kombinacija slovački i hrvatski*

Vidimo da je dio slova ě promijenjen u é, što je za potrebe ovoga rada dostatno. Rezultati su:

General results

CER	1.55
WER	5.38
WER (order independent)	5.18

Slika 24 - *Runje CER i WER za kombinaciju slovačkog i hrvatskog*

CER je također bolji nego u slučaju korištenja samo s hrvatskim jezikom, tako da će se koristiti kombinacija slovačkoga i hrvatskoga za ovo djelo.

5.5. Rad s alatom Google DocumentAI

Za rad s alatom *Google DocumentAI*, potrebno je prvo napraviti korisnički račun na Google Cloud platformi, kreirati projekt i omogućiti *DocumentAI* API.

Unutar *Google DocumentAI* konzole zatim se kreira procesor za optičko prepoznavanje znakova.

Potom se na računalu instalira *Google Cloud Cli* koji služi za autentifikaciju s *Google DocumentAI* API-em. Nakon instalacije u terminalu se pokreće naredba *gcloud auth*

application-default login koja će kreirati podatke za autentifikaciju. Ti podaci će se automatski koristiti svaki puta kada ćemo kroz python kontaktirati *Google DocumentAI* API.

Kako bismo koristili *Google DocumentAI* s Pythonom, potrebno je instalirati dodatne knjižnice:

pip install google-cloud-core google-cloud-documentai

Slijedi dio Python datoteke korištene za obradu uzoraka s alatom *Google DocumentAI.*

```
import os
from google.api_core.client_options import ClientOptions
from google.cloud import documentai_v1 as documentai
#sve ove informacije treba prepisati sa Google Document AI platforme
project_id="*****************
location="eu"
proc_id="************"
#popis prihvatljivih formata na https://cloud.google.com/document-ai/docs/file-types
mime="image/tiff"

#pokreni instancu klijenta
docai_client = documentai.DocumentProcessorServiceClient(
    client_options=ClientOptions(api_endpoint=f"{location}-documentai.googleapis.com")
)

proc_name = docai_client.processor_path(project_id, location, proc_id)

#iteriraj kroz folder
for dokument in popis:
    with open(dokument, "rb") as slika:
        sadrzaj = slika.read()
    raw_sadrzaj = documentai.RawDocument(content=sadrzaj, mime_type=mime)
    zahtjev = documentai.ProcessRequest(name=proc_name, raw_document=raw_sadrzaj)
```

```
odgovor = docai_client.process_document(request=zahtjev)

rezultat = odgovor.document

doc_name=(dokument.rsplit("\\", 1)[1])

doc_name=str.replace(doc_name, '.tiff','.txt')
```

5.6. Rad s alatom OcrevalUAtion i obrada rezultata

Rad s alatom *OcrevalUAtion* automatiziran je kroz Python datoteku čiji se isječak može pronaći u nastavku.

```
os.system('cmd /c "java -cp OcrevalUAtion.jar eu.digitisation.Main -gt %s -ocr %s -o usporedba\\docai\\%s"'
%(item, rezultat, izlazna))
```

Automatizirana je i ekstrakcija rezultata iz datoteka, kao i njihov prikaz, također u Pythonu. Pošto se rezultati alata *OcrevalUAtion* pohranjuju u .html datoteci, za parsiranje iste korištena je knjižnica BeautifulSoup, a za prikaz rezultata knjižnica Matplotlib.

6. Rezultati

Rezultati će biti podijeljeni prema sustavu za optičko prepoznavanje znakova koji je korišten. Grafički će se prikazati svi rezultati CER i WER za pojedini sustav, a zatim tablično prosjek CER i WER prema skupini građe. Skupine građe su određene prema knjizi iz koje su stranice preuzete, te postoji 5 skupina: Djački, Optimizam, Plebiscit, Poljice i Runje. Finalno će se usporediti rezultati između oba sustava. Za potrebe grafičkog prikaza, vrijednosti iznad 20 ograničene su na 20.

6.1. Tesseract

Na osi X nalaze se skraćeni nazivi za pojedine stranice iz određene skupine. Dj za Djački, Op za Optimizam, Pl za Plebiscit, Po za Poljice i Ru za Runje. Svaka stranica ima 3 pripadajuća rezultata definirana legendom - za original, sivu skalu i binarizaciju.

Na osi Y nalazi se vrijednost CER u potpoglavlju Grafički prikaz rezultata - CER *Tesseract*, te WER u potpoglavlju Grafički prikaz rezultata - WER *Tesseract*.

Najmanje vrijednosti za CER ili WER označavaju najbolju metodu pretprocesiranja, odnosno onu koja će dati najmanje grešaka prilikom čitanja u sustavu za optičko prepoznavanje znakova.

Legenda:

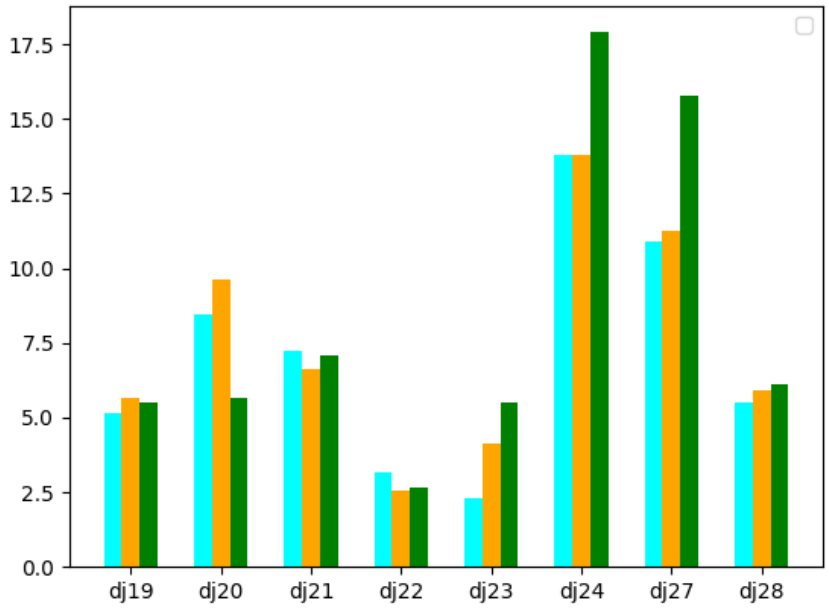
Plava boja = original bez pretprocesiranja,

narančasta boja = pretprocesiranje sa sivom skalom,

zelena boja = pretprocesiranje binarizacijom.

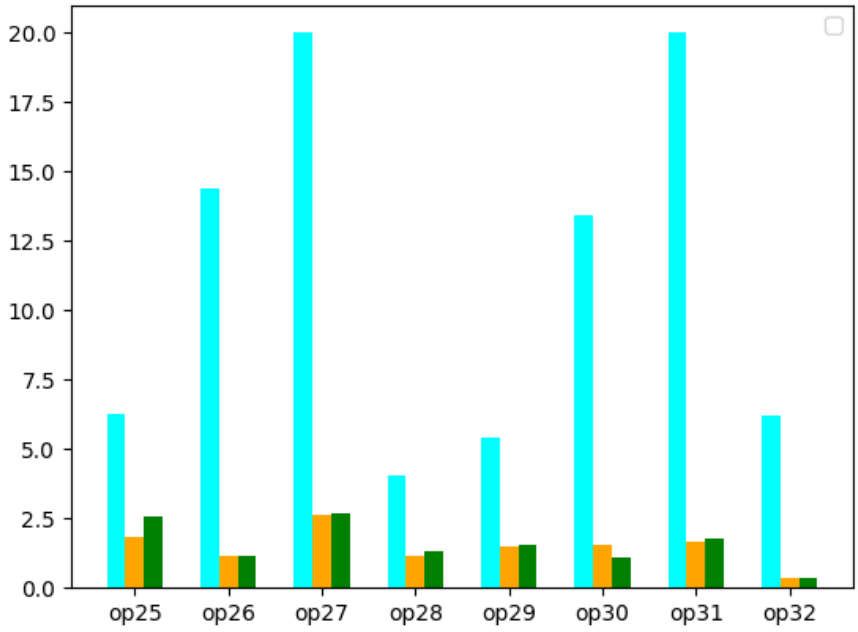
6.1.1. Grafički prikaz rezultata - CER Tesseract

Slika 25 prikazuje CER za skupinu Djački obradom u alatu *Tesseract*. Za stranice 19, 23, 27 i 28 optimalan je pristup bez pretprocesiranja. Za stranicu 20, optimalna je binarizacija. Za stranice 21 i 22 optimalna je pretvorba u sivu skalu, dok je za stranicu 24 podjednako optimalna pretvorba u sivu skalu i pristup bez preptprocesiranja.



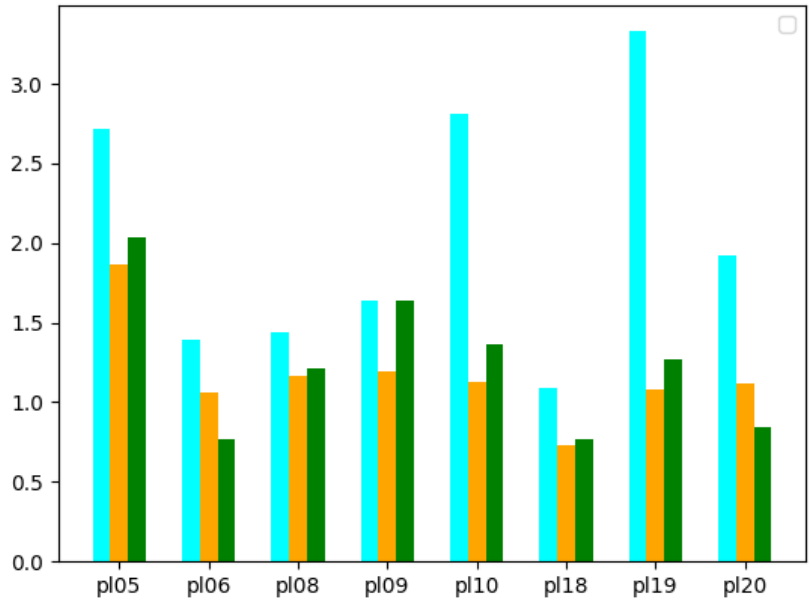
Slika 25 - *Tesseract: CER skupine Djački*

Slika 26 prikazuje CER za skupinu Optimizam obradom u alatu *Tesseract*. Za stranice 25, 27, 28, 29 i 31 optimalna je pretvorba u sivu skalu. Za stranicu 30, optimalna je binarizacija. Za stranice 26 i 32 podjednako je optimalna pretvorba u sivu skalu i binarizacija.



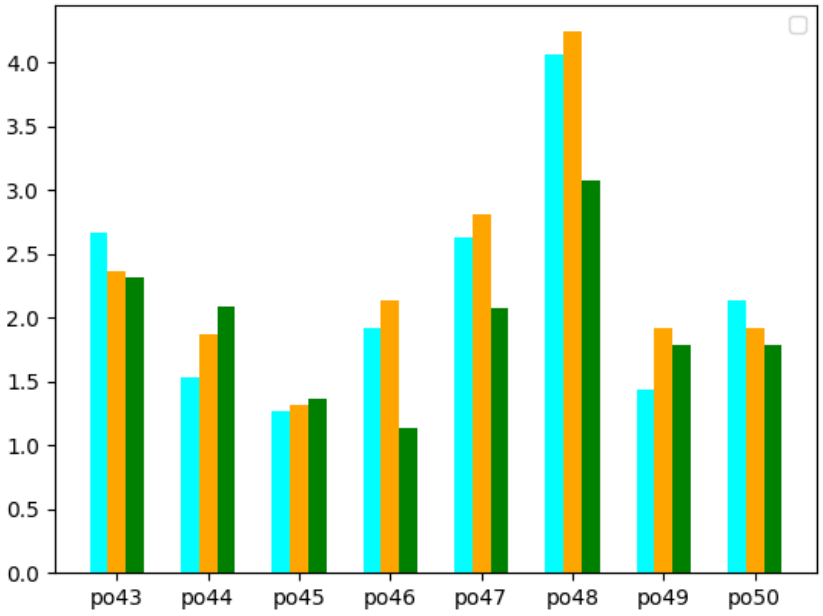
Slika 26 - *Tesseract: CER skupine Optimizam*

Slika 27 prikazuje CER za skupinu Plebiscit obradom u alatu *Tesseract*. Za stranice 5, 8, 9, 10, 18 i 19 optimalna je pretvorba u sivu skalu. Za stranice 6 i 20 optimalna je binarizacija.



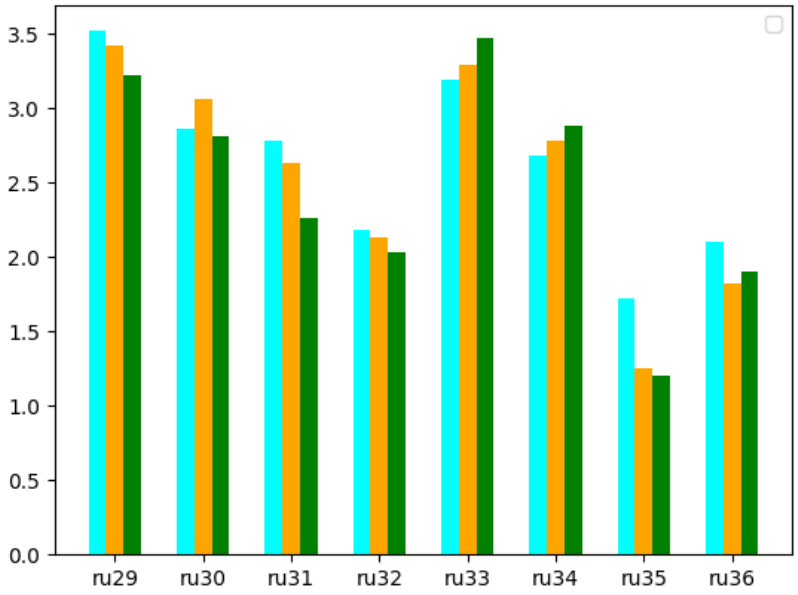
Slika 27 - *Tesseract: CER skupine Plebiscit*

Slika 28 prikazuje CER za skupinu Poljice obradom u alatu *Tesseract*. Za stranice 43, 46, 47, 48 i 50 optimalna je binarizacija. Za stranice 44, 45 i 49 optimalan je pristup bez pretprocesiranja.



Slika 28 - *Tesseract: CER skupine Poljice*

Slika 29 prikazuje CER za skupinu Runje obradom u alatu *Tesseract*. Za stranice 29, 30, 31, 32 i 35 optimalna je binarizacija. Za stranice 33 i 34, optimalan je pristup bez pretprocesiranja. Za stranicu 36 optimalna je pretvorba u sivu skalu.



Slika 29 - *Tesseract: CER skupine Runje*

Tablica 6 prikazuje prosjek CER vrijednosti prema skupini i prema načinu pretprocesiranja. Prosječno, za skupine Plebiscit, Poljice i Runje najbolji način pretprocesiranja je binarizacija. Prosječno, za skupinu Djački, najbolji je pristup bez pretprocesiranja.

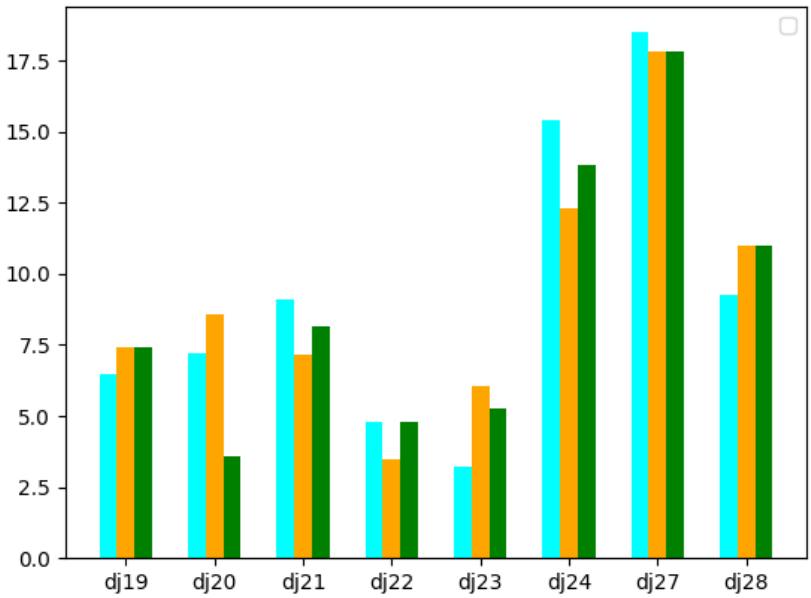
Prosječno, za skupinu Optimizam, najbolji način pretprocesiranja je pretvorba u sivu skalu.

Grupa	Original	Siva skala	Binarizacija
Plebiscit	2.0425	1.16875	1.2375
Poljice	2.20625	2.32	1.9525
Runje	2.62875	2.5475	2.47125
Djački	7.0625	7.4425	8.27125
Optimizam	37.615	1.475	1.56

Tablica 6 - *Tesseract*: *Prosjek vrijednosti CER prema skupini*

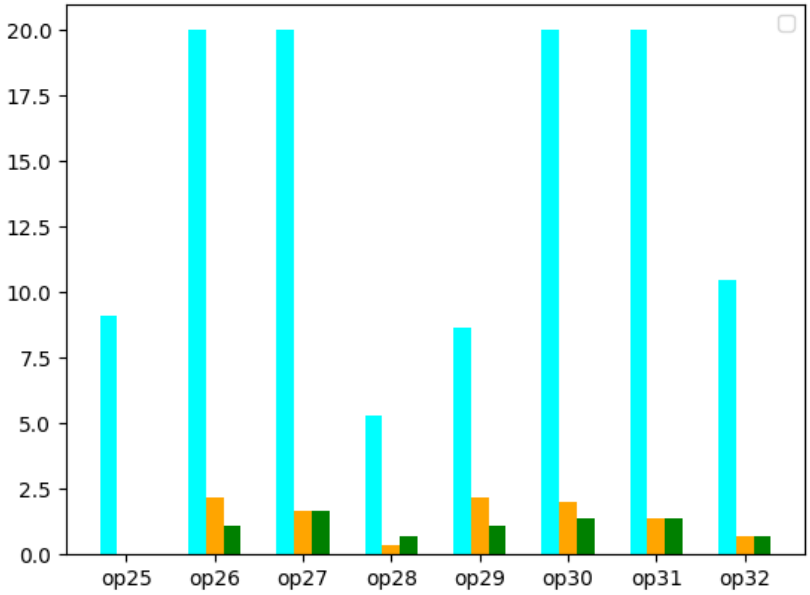
6.1.2. Grafički prikaz rezultata - WER Tesseract

Slika 30 prikazuje WER za skupinu Djački obradom u alatu *Tesseract*. Za stranicu 20 optimalna je binarizacija. Za stranice 19, 23 i 28 optimalan je pristup bez pretprocesiranja. Za stranice 21, 22 i 24 optimalna je pretvorba u sivu skalu. Za stranicu 27 podjednako je optimalna pretvorba u sivu skalu i binarizacija.



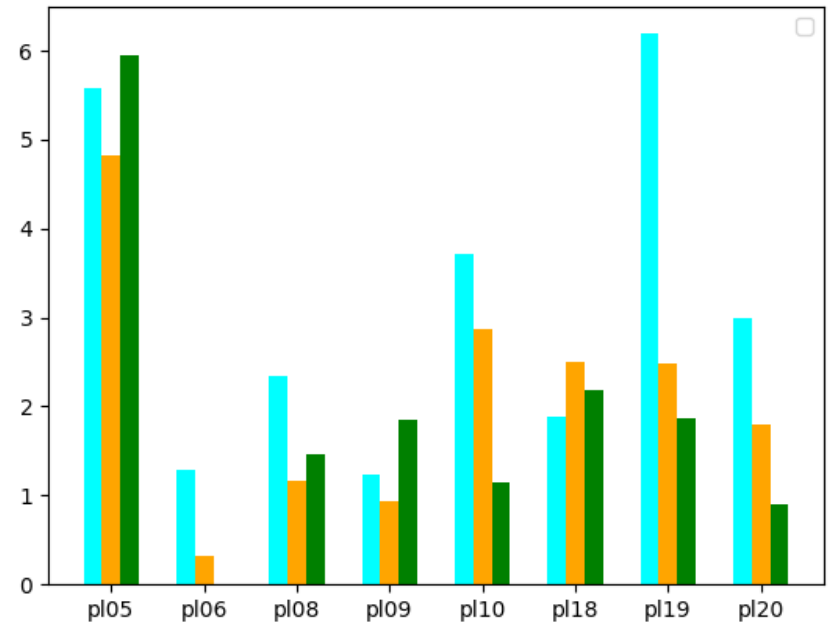
Slika 30 – *Tesseract: Prosjek WER za skupinu Djački*

Slika 31 prikazuje WER za skupinu Optimizam obradom u alatu *Tesseract*. Za stranice 26, 29 i 30 optimalna je binarizacija. Za stranicu 28 optimalna je pretvorba u sivu skalu. Za stranice 25, 27 31 i 32 podjednako je optimalna pretvorba u sivu skalu i binarizacija.



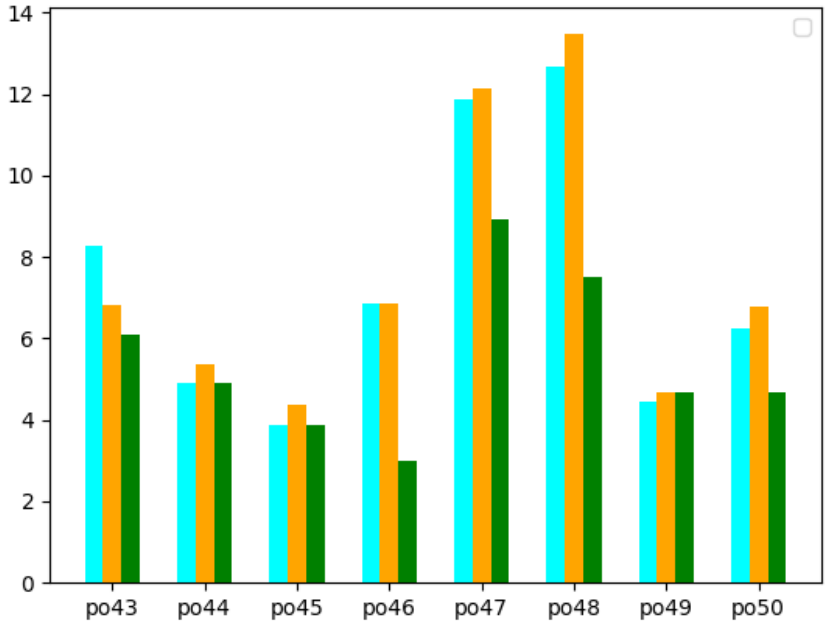
Slika 31 - *Tesseract: Prosjek WER za skupinu Optimizam*

Slika 32 prikazuje WER za skupinu Plebiscit obradom u alatu *Tesseract*. Za stranice 6, 10, 19 i 20 optimalna je binarizacija. Za stranicu 18 optimalan je pristup bez pretprocesiranja. Za stranice 5, 8 i 9 optimalna je pretvorba u sivu skalu.



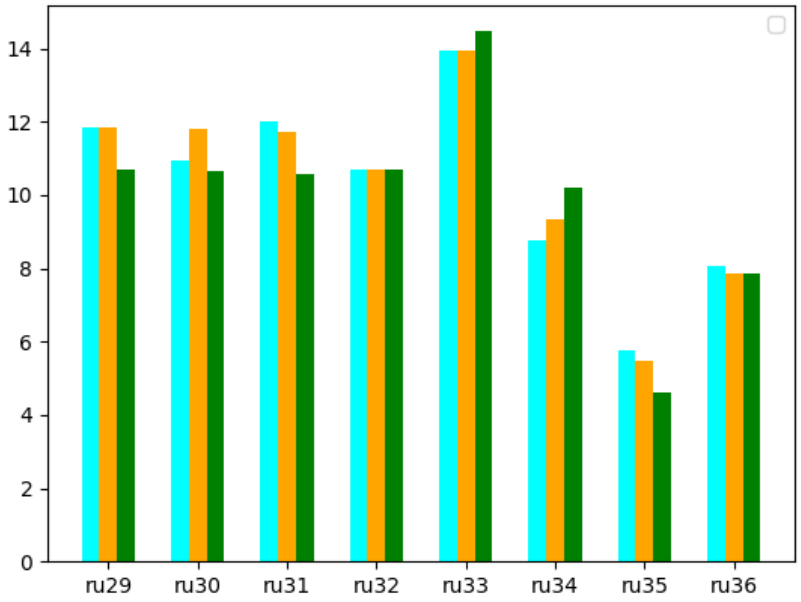
Slika 32 - *Tesseract: Prosjek WER za skupinu Plebiscit*

Slika 33 prikazuje WER za skupinu Poljice obradom u alatu *Tesseract*. Za stranice 43, 46, 47, 48 i 50 optimalna je binarizacija. Za stranicu 49 optimalan je pristup bez pretprocesiranja. Za stranice 44 i 45 podjednako je optimalna binarizacija i pristup bez pretprocesiranja.



Slika 33 - *Tesseract: Prosjek WER za skupinu Poljice*

Slika 34 prikazuje WER za skupinu Runje obradom u alatu *Tesseract*. Za stranice 29, 30, 31 i 35 optimalna je binarizacija. Za stranicu 32 optimalni su svi pristupi. Za stranicu 33 podjednako su optimalni pristup bez pretprocesiranja i pretvorba u sivu skalu. Za stranicu 34 optimalan je pristup bez pretprocesiranja. Za stranicu 36 podjednako su optimalni pretvorba u sivu skalu i binarizacija.



Slika 34 - *Tesseract: Prosjek WER za skupinu Runje*

Tablica 7 prikazuje prosjek WER vrijednosti prema skupini i prema načinu pretprocesiranja. Prosječno, za sve skupine, najbolji način pretprocesiranja je binarizacija.

Grupa	Original	Siva skala	Binarizacija
Plebiscit	3.1525	2.1125	1.92
Poljice	7.39875	7.56	5.45875
Runje	10.2575	10.34	9.96875
Djački	9.2425	9.22625	8.98
Optimizam	78.4375	1.30375	0.99125

Tablica 7 - *Tesseract: Prosjek vrijednosti WER prema skupini*

6.2. Google DocumentAI

Na osi X nalaze se skraćeni nazivi za pojedine stranice iz određene skupine. Dj za Djački, Op za Optimizam, Pl za Plebiscit, Po za Poljice i Ru za Runje. Svaka stranica ima 3 pripadajuća rezultata definirana legendom - za original, sivu skalu i binarizaciju.

Na osi Y nalazi se vrijednost CER u potpoglavlju Grafički prikaz rezultata - CER *Google DocumentAI*, te WER u potpoglavlju Grafički prikaz rezultata - WER *Google DocumentAI*.

Najmanje vrijednosti za CER ili WER označavaju najbolju metodu pretprocesiranja, odnosno onu koja će dati najmanje grešaka prilikom čitanja u sustavu za optičko prepoznavanje znakova.

Legenda:

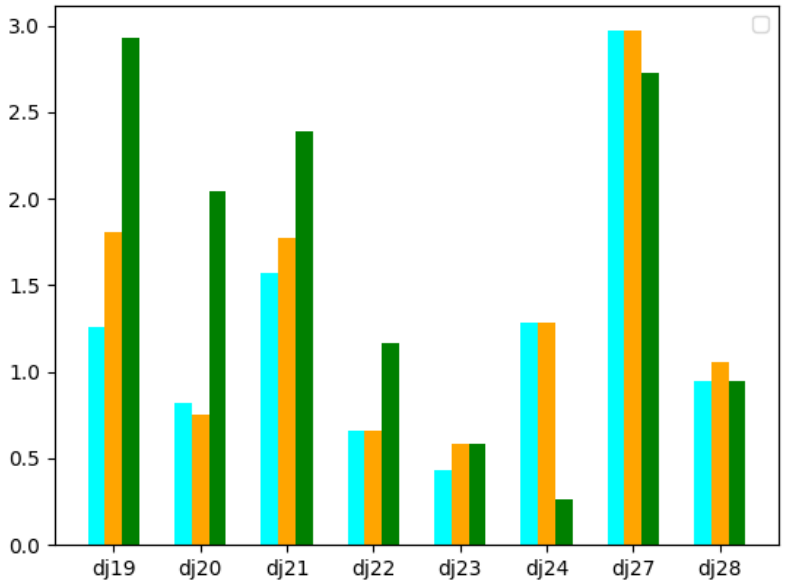
Plava boja = original bez pretprocesiranja,

narančasta boja = pretprocesiranje sa sivom skalom,

zelena boja = pretprocesiranje binarizacijom.

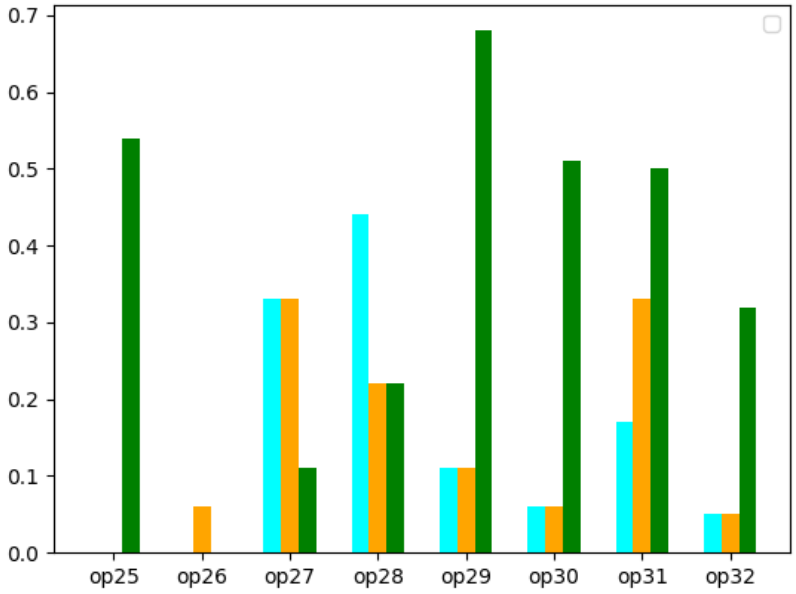
6.2.1. Grafički prikaz rezultata - CER Google Document AI

Slika 35 prikazuje CER za skupinu Djački obradom u alatu *Google DocumentAI*. Za stranice 24 i 27 optimalna je binarizacija. Za stranice 19, 21 i 23 optimalan je pristup bez pretprocesiranja. Za stranicu 20 optimalna je pretvorba u sivu skalu. Za stranicu 23 podjednako su optimalni pretvorba u sivu skalu i binarizacija. Za stranicu 22 podjednako su optimalni pristup bez pretprocesiranja i pretvorba u sivu skalu.



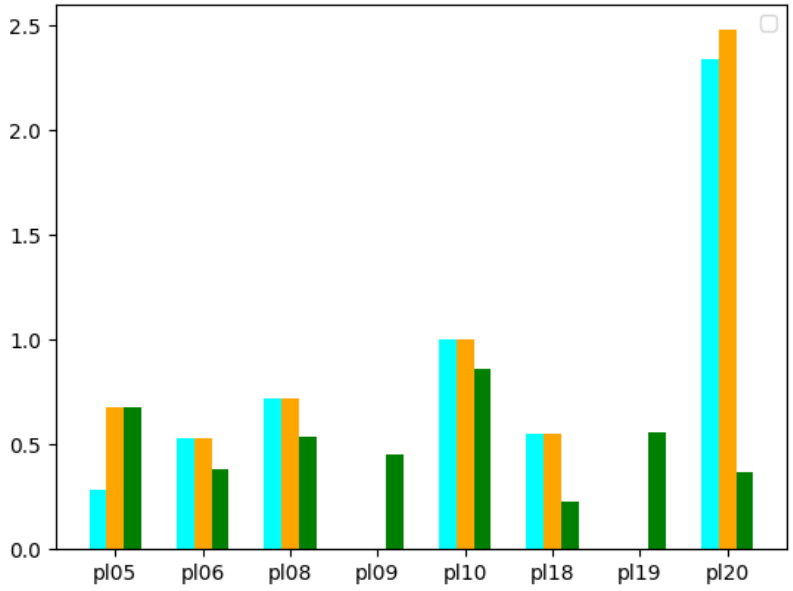
Slika 35 - *Google DocumentAI: Prosjek CER za skupinu Djački*

Slika 36 prikazuje CER za skupinu Optimizam obradom u alatu *Google DocumentAI*. Za stranicu 27 optimalna je binarizacija. Za stranicu 31 optimalan je pristup bez pretprocesiranja. Za stranice 25, 29, 30 i 32 podjednako su optimalni pristup bez pretprocesiranja i pretvorba u sivu skalu. Za stranicu 26 podjednako su optimalni pristup bez pretprocesiranja i binarizacija. Za stranicu 28 podjednako su optimalni pretvorba u sivu skalu i binarizacija.



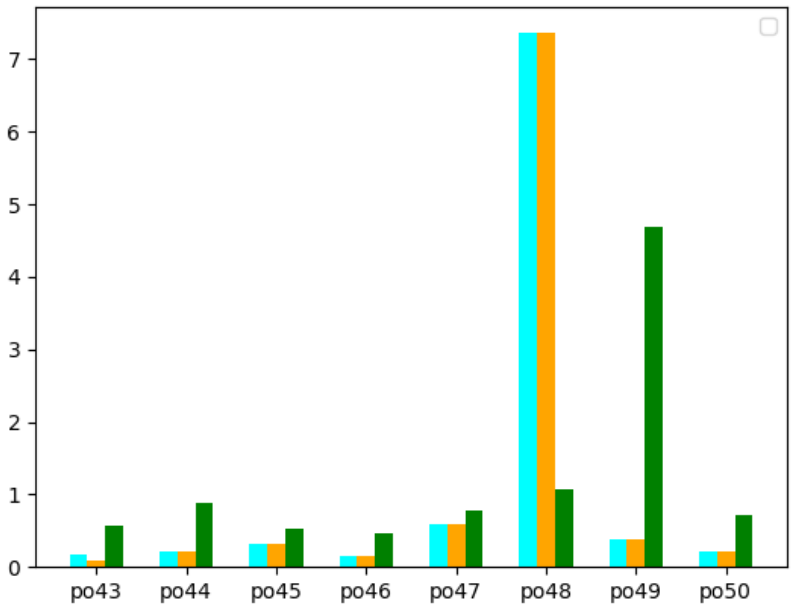
Slika 36 - *Google DocumentAI: Prosjek CER za skupinu Optimizam*

Slika 37 prikazuje CER za skupinu Plebiscit obradom u alatu *Google DocumentAI*. Za stranice 6, 8, 10, 18 i 20 optimalna je binarizacija. Za stranicu 5 optimalan je pristup bez pretprocesiranja. Za stranice 9 i 19 podjednako su optimalni pristup bez pretprocesiranja i pretvorba u sivu skalu.



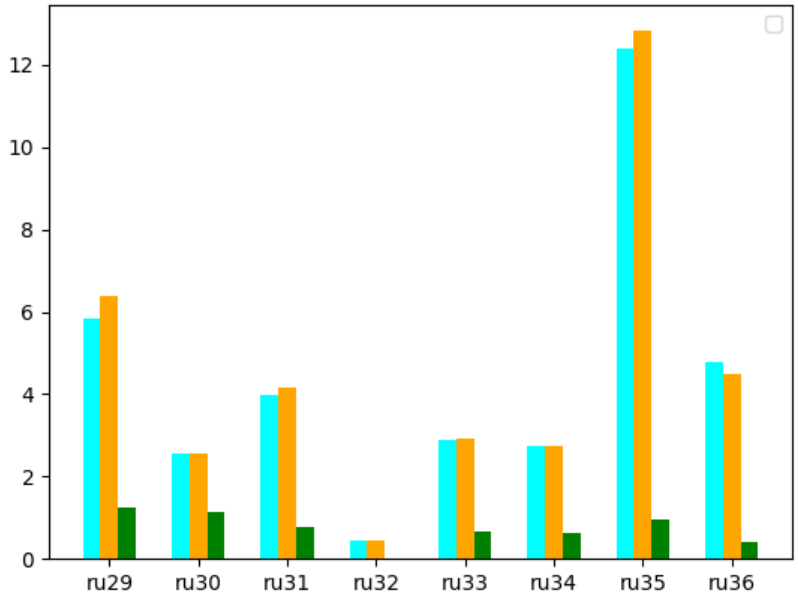
Slika 37 - *Google DocumentAI: Prosjek CER za skupinu Plebiscit*

Slika 38 prikazuje CER za skupinu Poljice obradom u alatu *Google DocumentAI*. Za stranicu 43 optimalna je pretvorba u sivu skalu. Za stranicu 48 optimalna je binarizacija. Za stranice 44, 45, 46, 47, 49 i 50 podjednako su optimalni pristup bez pretprocesiranja i pretvorba u sivu skalu.



Slika 38 - *Google DocumentAI: Prosjek CER za skupinu Poljice*

Slika 39 prikazuje CER za skupinu Runje obradom u alatu *Google DocumentAI*. Za sve stranice optimalna je binarizacija.



Slika 39 - *Google DocumentAI: Prosjek CER za skupinu Runje*

Tablica 8 prikazuje prosjek CER vrijednosti prema skupini i prema načinu pretprocesiranja. Prosječno, za skupine Plebiscit i Runje najbolji način pretprocesiranja je binarizacija. Prosječno, za skupinu Djački, najbolji je pristup bez pretprocesiranja.

Prosječno, za skupinu Poljice, najbolji način pretprocesiranja je pretvorba u sivu skalu.

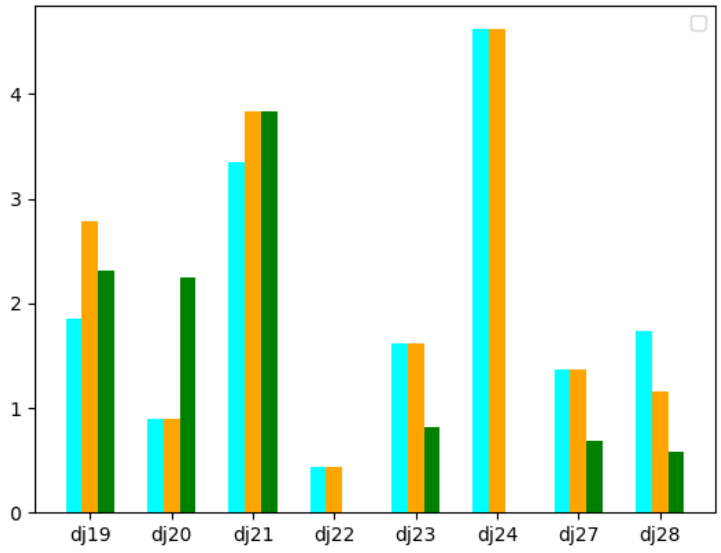
Prosječno, za skupinu Optimizam, podjednako su optimalni pristup bez pretprocesiranja i pretvorba u sivu skalu.

Grupa	Original	Siva skala	Binarizacija
Plebiscit	0.6775	0.745	0.50875
Poljice	1.1775	1.1675	1.2125
Runje	4.45625	4.56375	0.72375
Djački	1.2425	1.36	1.63125
Optimizam	0.145	0.145	0.36

Tablica 8 - *Google DocumentAI: Prosjek vrijednosti CER prema skupini*

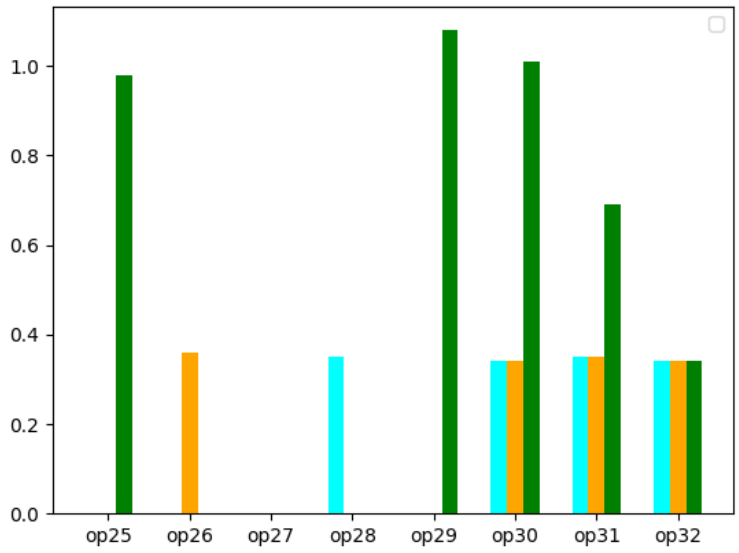
6.2.2. Grafički prikaz rezultata - WER Google Document AI

Slika 40 prikazuje WER za skupinu Djački obradom u alatu *Google DocumentAI*. Za stranice 22, 23, 24, 27 i 28 optimalna je binarizacija. Za stranice 19 i 21 optimalan je pristup bez pretprocesiranja. Za stranicu 20 podjednako su optimalni pristup bez pretprocesiranja i pretvorba u sivu skalu.



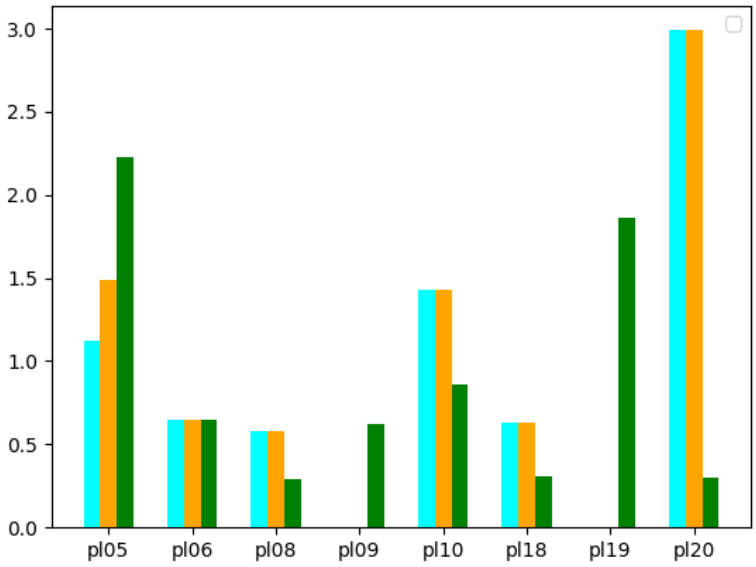
Slika 40 - *Google DocumentAI: Prosjek WER za skupinu Djački*

Slika 41 prikazuje WER za skupinu Optimizam obradom u alatu *Google DocumentAI*. Za stranice 25,29, 30 i 31 podjednako su optimalni pristup bez pretprocesiranja i pretvorba u sivu skalu. Za stranicu 26 podjednako su optimalni pristup bez pretprocesiranja i binarizacija. Za stranicu 28 podjednako su optimalni pretvorba u sivu skalu i binarizacija. Za stranice 27 i 32 podjednako su optimalni svi pristupi.



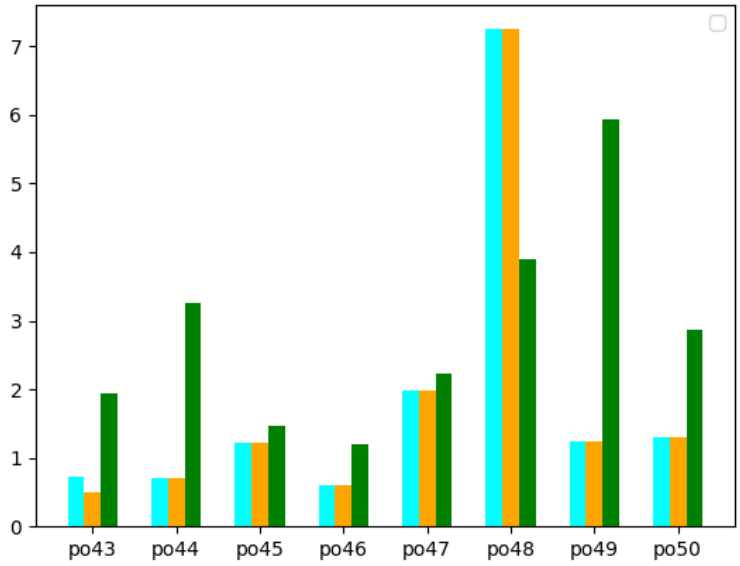
Slika 41 - *Google DocumentAI: Prosjek WER za skupinu Optimizam*

Slika 42 prikazuje WER za skupinu Plebiscit obradom u alatu *Google DocumentAI*. Za stranice 8, 10, 18 i 20 optimalna je binarizacija. Za stranicu 5 optimalan je pristup bez pretprocesiranja. Za stranice 9 i 19 podjednako su optimalni pristup bez pretprocesiranja i pretvorba u sivu skalu. Za stranicu 6 optimalni su svi pristupi.



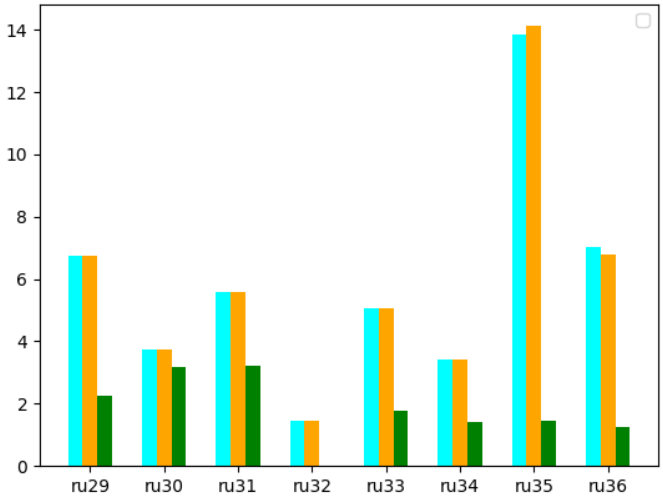
Slika 42 - *Google DocumentAI: Prosjek WER za skupinu Plebiscit*

Slika 43 prikazuje WER za skupinu Poljice obradom u alatu *Google DocumentAI*. Za stranicu 48 optimalna je binarizacija. Za stranicu 43 optimalna je pretvorba u sivu skalu. Za stranice 44, 45, 46, 47, 49 i 50 podjednako su optimalni pristup bez pretprocesiranja i pretvorba u sivu skalu.



Slika 43 - *Google DocumentAI: Prosjek WER za skupinu Poljice*

Slika 44 prikazuje WER za skupinu Runje obradom u alatu *Google DocumentAI*. Za sve stranice optimalna je binarizacija.



Slika 44 - *Google DocumentAI: Prosjek WER za skupinu Runje*

Tablica 9 prikazuje prosjek WER vrijednosti prema skupini i prema načinu pretprocesiranja. Prosječno, za skupine Plebiscit, Runje i Djački najbolji način pretprocesiranja je binarizacija. Prosječno, za skupinu Optimizam, najbolji je pristup bez pretprocesiranja.

Prosječno, za skupinu Poljice, najbolji način pretprocesiranja je pretvorba u sivu skalu.

Grupa	Original	Siva skala	Binarizacija
Plebiscit	0.925	0.97125	0.89
Poljice	1.87625	1.84625	2.84625
Runje	5.85625	5.865	1.82
Djački	1.985	2.09	1.3075
Optimizam	0.1725	0.17375	0.5125

Tablica 9 - *Google DocumentAI: Prosjek vrijednosti WER prema skupini*

6.3. Usporedba rezultata evaluacije

6.3.1. Tesseract - Google DocumentAI

Tablica 10 prikazuje usporedbu prosjeka vrijednosti CER u alatima *Tesseract* i *Google DocumentAI* prema skupini. Najbolji prosječni rezultati za sve skupine dobiveni su alatom *Google DocumentAI*. Najbolji rezultat za skupine Plebiscit i Runje dobiven je binarizacijom, za skupinu Poljice pretvorbom u sivu skalu, za skupinu Djački pristupom bez pretprocesiranja, a za skupinu Optimizam podjednako su optimalni pristup bez pretprocesiranja i pretvorba u sivu skalu.

CER - prosjek						
	Google Document Ai			Tesseract		
Grupa	Original	Siva skala	Binarizacija	Original	Siva skala	Binarizacija
Plebiscit	0.6775	0.745	0.50875	2.0425	1.16875	1.2375
Poljice	1.1775	1.1675	1.2125	2.20625	2.32	1.9525
Runje	4.45625	4.56375	0.72375	2.62875	2.5475	2.47125
Djački	1.2425	1.36	1.63125	7.0625	7.4425	8.27125

Optimizam	0.145	0.145	0.36	37.615	1.475	1.56

Tablica 10 - *Usporedba prosjeka vrijednosti CER u alatima Tesseract i Google DocumentAI prema skupini*

Tablica 11 prikazuje usporedbu prosjeka vrijednosti WER u alatima *Tesseract* i *Google DocumentAI* prema skupini. Najbolji prosječni rezultati za sve skupine dobiveni su alatom *Google DocumentAI*. Najbolji rezultat za skupine Plebiscit, Runje i Djački dobiven je binarizacijom, za skupinu Poljice pretvorbom u sivu skalu, a za skupinu Optimizam pristupom bez pretprocesiranja.

WER - prosjek						
	Google Document Ai			**Tesseract**		
Grupa	**Original**	**Siva skala**	**Binarizacija**	**Original**	**Siva skala**	**Binarizacija**
Plebiscit	0.925	0.97125	0.89	3.1525	2.1125	1.92
Poljice	1.87625	1.84625	2.84625	7.39875	7.56	5.45875
Runje	5.85625	5.865	1.82	10.2575	10.34	9.96875
Djački	1.985	2.09	1.3075	9.2425	9.22625	8.98
Optimizam	0.1725	0.17375	0.5125	78.4375	1.30375	0.99125

Tablica 11 - *Usporedba prosjeka vrijednosti WER u alatima Tesseract i Google DocumentAI prema skupini*

6.3.2. Najbolji rezultat prema pojedinačnom uzorku (WER)

Tablica 12 sadrži prikaz najboljih rezultata WER za svaku pojedinačnu stranicu korištenu u istraživanju. Prefiks dj označava skupinu Djački, op skupinu Optimizam, pl skupinu Plebiscit, po skupinu Poljice i ru skupinu Runje. Raspon vrijednosti WER među najboljim rezultatima kreće se od 0, što označava 100% točno očitane riječi, pa do 3.35, odnosno 96,65% točnosti u očitanim riječima. Predominantno je u rezultatima zastupljen *Google DocumentAI*, dok je *Tesseract* bolji samo u jednom slučaju (pl06), a za uzorak op25 dijeli najbolji rezultat s alatom *Google DocumentAI*. Izraženo u brojevima, u 38 od 40, odnosno 95% slučajeva bolji rezultati dobiveni su alatom *Google DocumentAI*. U 1 od 40, odnosno 2.5% slučajeva najbolji rezultati

dobiveni su alatom *Tesseract*, a u 1 od 40, odnosno 2.5% slučajeva oba alata su dala podjednako dobre rezultate.

Najbolji rezultati prema samoj metodi pretprocesiranja biti će razvrstani prema binarizaciji, pristupu bez pretprocesiranja i pretvorbi u sivu skalu. Pošto je za pojedine uzorke više metoda optimalno, prilikom izračuna postotka uzoraka za koje je optimalan pristup npr. binarizacija, u izračun će se uzeti i slučajevi u kojima je binarizacija samo jedna od optimalnih metoda za taj uzorak.

Prema metodi pretprocesiranja, pristup bez pretprocesiranja najbolje rezultate dao je u 19 od 40, odnosno 47.5% slučajeva. Binarizacija Otsu metodom najbolje rezultate dala je u 24 od 40, odnosno 60% slučajeva. Pretvorba u sivu skalu najbolje rezultate dala je u 17 od 40, odnosno 42.5% slučajeva.

ime	sustav	metoda	vrijednost
dj19	Document Ai	original	1.85
dj20	Document Ai/Document Ai	original/siva skala	0.9
dj21	Document Ai	original	3.35
dj22	Document Ai	otsu	0
dj23	Document Ai	otsu	0.81
dj24	Document Ai	otsu	0
dj27	Document Ai	otsu	0.68
dj28	Document Ai	otsu	0.58
op25	Document Ai/Document Ai/*Tesseract*/*Tesseract*	original/siva skala/siva skala/otsu	0
op26	Document Ai/Document Ai	original/otsu	0
op27	Document Ai/Document Ai/Document Ai	original/siva skala/otsu	0
op28	Document Ai/Document Ai	siva skala/otsu	0
op29	Document Ai/Document Ai	original/siva skala	0
op30	Document Ai/Document Ai	original/siva skala	0.34

op31	Document Ai/Document Ai	original/siva skala	0.35
op32	Document Ai/Document Ai/Document Ai	original/siva skala/otsu	0.34
pl05	Document Ai	original	1.12
pl06	*Tesseract*	otsu	0
pl08	Document Ai	otsu	0.29
pl09	Document Ai/Document Ai	original/siva skala	0
pl10	Document Ai	otsu	0.86
pl18	Document Ai	otsu	0.31
pl19	Document Ai/Document Ai	original/siva skala	0
pl20	Document Ai	otsu	0.3
po43	Document Ai	siva skala	0.49
po44	Document Ai/Document Ai	original/siva skala	0.7
po45	Document Ai/Document Ai	original/siva skala	1.22
po46	Document Ai/Document Ai	original/siva skala	0.6
po47	Document Ai/Document Ai	original/siva skala	1.98
po48	Document Ai	otsu	3.89
po49	Document Ai/Document Ai	original/siva skala	1.23
po50	Document Ai/Document Ai	original/siva skala	1.3
ru29	Document Ai	otsu	2.25
ru30	Document Ai	otsu	3.17
ru31	Document Ai	otsu	3.23
ru32	Document Ai	otsu	0
ru33	Document Ai	otsu	1.78
ru34	Document Ai	otsu	1.42
ru35	Document Ai	otsu	1.44
ru36	Document Ai	otsu	1.27

Tablica 12 - *Prikaz najboljih rezultata za vrijednost WER prema pojedinačnom uzorku*

7. Diskusija

Metrika CER se odnosi na jedan pogrešan znak, što može biti i nadodani razmak, dodatna točka i sl., dok se metrika WER koristi kako bi se izračunale pogrešno napisane riječi. Ukoliko se optičko prepoznavanje znakova primjenjuje za potrebe knjižnice, nekoliko dodanih znakova neće biti problem, no činjenica da korisnik pretraživanjem teksta neće moći naći ono što traži, predstavlja veći problem. Stoga bi za potrebe knjižnica fokus trebao biti stavljen na osiguranje što manjih vrijednosti WER metrike.

Rezultati istraživanja za *Tesseract*, prema Tablici 7, pokazali su da postotak WER metrike za rad u ovom alatu varira od 1.92% pa sve do 78.44%. Kao generalno najbolja metoda za rad s ovim alatom ističe se binarizacija, no prema Slikama 30-34 vidljivo je da za pojedinačne uzorke variraju optimalni pristupi pretprocesiranja.

Rezultati istraživanja za *Google DocumentAI* prema Tablici 9, pokazali su da postotak WER metrike varira između 0.17% i 5.86%, što je bitno bolje od WER metrike za alat *Tesseract*. Najbolje metode pretprocesiranja u ovome alatu također variraju, te se ne može odabrati jedna metoda kao najbolja: Za 3 skupine najbolja metoda je binarizacija, za 1 pristup bez pretprocesiranja, a za zadnju pretvorba u sivu skalu.

Prema Tablici 12, na kojoj su najbolji rezultati WER metrike za svaki pojedini uzorak, vidljivo je da je alat *Google DocumentAI* bolji u očitanju znakova od alata *Tesseract*. U čak 95% slučajeva očitanje mu je bilo najbolje, dok je u 2.5% slučajeva prvo mjesto podijelio s alatom *Tesseract*. *Tesseract* je bio bolji samo u jednom, odnosno 2.5% slučajeva.

Metode pretprocesiranja najviše zastupljene u najboljim rezultatima su: 60% Otsu binarizacija, 47.5% original i 42.5% siva skala. Prema tim rezultatima može se zaključiti da se generalno najbolje rezultate može dobiti korištenjem binarizacije i alata *Google DocumentAI*. No, pošto binarizacija u 40% slučajeva nije najbolji način pretprocesiranja, te alat *Google DocumentAI* u 2.5% slučajeva nije najbolji alat, drugačijim pristupom bilo bi moguće dobiti još točnije rezultate očitanja.

Na primjeru Slike 35, može se vidjeti da rezultati očitanja za stranice iste knjige bitno variraju prema korištenoj metodi pretprocesiranja, te da niti jedna metoda korištena samostalno ne bi dala optimalne rezultate. Na Slici 35 vidljivo je da je za 2 stranice optimalna binarizacija, za 3 korištenje originala, za 1 siva skala, za 1 su podjednako optimalne siva skala i binarizacija, a za zadnju podjednako dobre rezultate daje korištenje originala i pretvorba u sivu skalu. Ako

bismo se za ovu skupinu odlučili pretprocesirati samo binarizacijom, za stranicu npr. 19 dobili bismo duplo veći WER nego da smo procesirali u originalu.

Rezultati poput tih sa Slike 35 mogu se vidjeti na svim grafovima s izuzetkom Slike 39 i Slike 44 na kojima je skupina Runje obrađena alatom *Google DocumentAI*.

Jedan od pristupa odabiru najbolje metode pretprocesiranja potencijalno je statistička obrada uzoraka kao što je napravljeno u ovome radu. Dakle uzimanje nekoliko reprezentativnih stranica građe koju želimo obraditi u sustavu za optičko prepoznavanje znakova, digitalizacija istih, pretvorba u sivu skalu i binarizacija i zatim provođenje kroz sustav za optičko prepoznavanje znakova u sve 3 faze pretprocesiranja: original, siva skala i binarizacija. Finalno se rezultate očitanja uspoređuje za metrike CER i WER s prethodno pripremljenim i provjerenim točnim tekstovima kako bi se odredila najbolja metoda pretprocesiranja za upravo tu građu. No, kako bi se ovakva obrada napravila potrebno je odvojiti značajnu količinu vremena, pogotovo za pisanje (ili uređivanje) točnih uzoraka, a takva statistička obrada ne može osigurati da će taj pristup biti najbolji za svaku od stranica unutar iste knjige. Ono što takva obrada može osigurati je da će biti odabrana najbolja metoda za većinu uzoraka.

Drugi, manje vremenski zahtjevan pristup, bio bi onaj naveden u 4.6 prema Nguyen i sur. (2021)[134], koji kao potencijalnu metodu postprocesiranja navodi kombiniranje sustava za optičko prepoznavanje znakova.

Kao što je vidljivo u Tablici 12, *Tesseract* je jedan rezultat bolje očitao nego *Google DocumentAI*, iako je *Google DocumentAI* u ostalih 95% slučajeva bio bolji. Da se za potrebe ovoga rada koristio npr. *Textract* umjesto *Tesseract*a, rezultati bi vjerojatno bili više ravnomjerno raspoređeni, te bi imalo smisla kombinirati njihova očitanja kako bismo došli do preporučenih 99,5% točnosti za očitanje znakova.

Druga metoda kombiniranja odnosi se na faze pretprocesiranja. Prema Tablici 12, najbolji rezultati za uzorke iz knjige Djačko društvo dobiveni su Otsu binarizacijom u 62.5% slučajeva, dok je metodom bez pretprocesiranja najbolji rezultat dobiven u 37.5% slučajeva. Kombiniranjem svih rezultata prema fazama pretprocesiranja, po uzoru na gore navedeni primjer kombiniranja sustava, također bi se moglo postići slično poboljšanje rezultata.

[134] Nguyen, T.T.H., Jatowt A., Coustaty, M., i Doucet, A. (2021). Survey of Post-OCR Processing Approaches. ACM Computing Surveys, 54(6). https://doi.org/10.1145/3453476

8. Zaključak

U ovome radu obrađen je postupak digitalizacije teksta u domeni knjižničarstva i njegova daljnja obrada u sustavima za optičko prepoznavanje znakova kako bi se tekst iz slike pretvorio u pretraživi tekst.

Cilj ovoga rada bio je ukratko objasniti procese digitalizacije i rad sustava za optičko prepoznavanje znakova, te provesti istraživanje o utjecaju pretprocesiranja na rezultate očitanja znakova u alatima *Google DocumentAI* i *Tesseract*.

Istraživanje je provedeno nad 40 uzoraka preuzetih iz 5 knjiga izdanih u 19. i 20. stoljeću. Uzorci su zatim provedeni kroz 2 faze pretprocesiranja: pretvorbu u sivu skalu i binarizaciju Otsu metodom. Pretvorba u sivu skalu odnosi se na pretvaranje svih piksela unutar slike u tonove sive, dok binarizacijom sve piksele unutar slike pretvaramo u crnu ili bijelu boju. Unaprijed su pripremljeni točni tekstovi s kojima će se uspoređivati rezultati očitanja u prethodno navedenim sustavima za optičko prepoznavanje znakova. Usporedba je rađena alatom *OcrevalUAtion*, koji kao mjerilo koristi WER i CER metrike.

Rezultati su se bilježili zasebno za alate *Google DocumentAI* i *Tesseract* prema prethodno navedenim fazama pretprocesiranja i u originalu bez pretprocesiranja.

Rezultati istraživanja za *Tesseract* pokazali su da postotak WER metrike varira od 1.92% do 78.44%. Kao generalno najbolja metoda za rad s ovim alatom ističe se binarizacija, no za pojedinačne uzorke variraju optimalni pristupi pretprocesiranja.

Rezultati istraživanja za *Google DocumentAI* pokazali su da postotak WER metrike varira između 0.17% i 5.86%. Najbolje metode pretprocesiranja u ovome alatu također variraju, te se ne može odabrati jedna metoda kao najbolja: Za 3 skupine najbolja metoda je binarizacija, za 1 pristup bez pretprocesiranja, a za zadnju pretvorba u sivu skalu.

Finalni prikaz najboljih rezultata WER metrike za svaki pojedini uzorak u Tablici 12 ukazuje na višu točnost očitanja prilikom rada s *Google DocumentAI* alatom u usporedbi s alatom *Tesseract*. Također se iz Tablice 12 može iščitati da ne postoji jedna najbolja metoda pretprocesiranja koja će za sve uzorke dati najbolji rezultat, već da najbolje metode variraju.

Stoga se predlaže primjena sljedećih metoda kako bi se osiguralo dobivanje što boljih rezultata očitanja: statistička obrada uzoraka kako je napravljeno u ovome radu, spajanje rezultata više sustava za optičko prepoznavanje znakova, te korištenje više metoda pretprocesiranja i spajanje rezultata očitanja u svim fazama pretprocesiranja.

Sustavi za optičko prepoznavanje znakova danas prolaze svoje mjesto u mnogim industrijama, uključujući i knjižnice. U knjižnicama se uglavnom primjenjuju za omogućavanje brzog snalaženja u tekstu korisnicima, približavanje građe slijepima i automatiziranu ekstrakciju metapodataka.

Na tržištu danas postoji cijeli niz modernih sustava za optičko prepoznavanje znakova koji se s vremenom približavaju (a u nekim slučajevima i dostižu) brojci od 99,95% točnosti koja je u literaturi postavljena kao granica isplativosti korištenja sustava za optičko prepoznavanje znakova.

9. Literatura

1. Ahmad, R., Naz, S., i Razzak, I. (2021). Efficient skew detection and correction in scanned document images through clustering of probabilistic hough transforms. Pattern Recognition Letters, 152. https://doi.org/10.1016/j.patrec.2021.09.014

2. Alginahi, Y. (2010). Preprocessing Techniques in Character Recognition. U Mori, M. (ur.), Character Recognition. Rijeka: Sciyo.

3. Amazon Web Services. (2018). Introducing Amazon Textract: Now in Preview—easily extract text and data from virtually any document. Preuzeto 3.4.2023. s https://aws.amazon.com/about-aws/whats-new/2018/11/introducing-amazon-textract-now-in-preview-easily-extract-text-and-data-from-virtually-any-document/

4. Amazon Web Services. (n.d.). Amazon Textract Features. Preuzeto 15.5.2023. s https://aws.amazon.com/textract/features/?pg=ln&sec=hs

5. Amazon Web Services. (n.d.). Amazon Textract. Preuzeto 3.4.2023. s https://aws.amazon.com/textract/

6. Bangare, L.S., Dubal, A., Bangare, P.S., i Patil, S.T. (2015). Reviewing Otsu's Method For Image Thresholding. International Journal of Applied Engineering Research, 10(9). https://dx.doi.org/10.37622/IJAER/10.9.2015.21777-21783

7. Bebis, G. (2004). Image Operations. University of Nevada, Reno, Department of Computer Science and Engineering, Nevada, SAD. Preuzeto 3.4.2023. s https://www.cse.unr.edu/~bebis/CS791E/Notes/PointProcess.pdf

8. Berchmans, D., i Kumar, S. S. (2014). Optical character recognition: An overview and an insight. 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT). doi:10.1109/iccicct.2014.6993174

9. Booth, J. M., i Gelb, J. (2006). Optimizing OCR Accuracy on Older Documents: A Study of Scan Mode, File Enhancement, and Software Products. Office of Innovation and New Technology. Preuzeto 31.3.2023. s https://www.govinfo.gov/media/WhitePaper-OptimizingOCRAccuracy.pdf

10. Breuel, T. M., Ul-Hasan, A., Al-Azawi, M. A., i Shafait, F. (2013). High-Performance OCR for Printed English and Fraktur Using LSTM Networks. 12th International Conference on Document Analysis and Recognition. doi:10.1109/icdar.2013.140

11. Brownlee, J. (2021). A Gentle Introduction to Long Short-Term Memory Networks by the Experts. Preuzeto 3.4.2023. s https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/

12. Carasco, R.C. (n.d.). Text Digitisation. 2. Measuring OCR quality > 2.3 Computing error rates. Preuzeto 3.4.2023. s https://sites.google.com/site/textdigitisation/OcrevalUAtion/input

13. Carasco, R.C. (n.d.). Text Digitisation. 3 The OcrevalUAtion tool > 3.1 Input. Preuzeto 3.4.2023. s https://sites.google.com/site/textdigitisation/OcrevalUAtion/input

14. Chang, T. (2021). All Things Google Document AI. Preuzeto 31.3.2023. s https://nanonets.com/blog/document-ai/

15. Chaudhuri, A., Mandaviya, K., Badelia, P., i Ghosh, S. K. (2016). Optical Character Recognition Systems for Different Languages with Soft Computing. Švicarska: Springer.

16. Diem, M., i Sablatnig. R. (2010). Recognizing Degraded Handwritten Characters. Institute of Computer Aided Automation. Preuzeto 31.3.2023. s https://www.researchgate.net/publication/236130411_Recognizing_Degraded_Handwritten_Characters

17. Dunđer, I., Seljan, S., Stančić, H. (2015). Koncept automatske klasifikacije registraturnoga i arhivskoga gradiva. 48. savjetovanje Zaštita arhivskoga gradiva u nastajanju.

18. Elements of AI. (n.d.). Neural network basics. Preuzeto 3.4.2023. s https://course.elementsofai.com/5/1

19. Firdousi, R., i Parveen, S. (2014). Local Thresholding Techniques in Image Binarization. International Journal Of Engineering And Computer Science, 3(3).

20. Fisher, R., Perkins, S., Walker, A., i Wolfart, E. (2003). Fourier Transform. Preuzeto 3.4.2023. s https://homepages.inf.ed.ac.uk/rbf/HIPR2/fourier.htm

21. Gharde, S.S., Baviskar, P.V., i Adhiya, K.P. (2013). Identification of Handwritten Simple Mathematical Equation Based on SVM and Projection Histogram. International Journal of Soft Computing and Engineering (IJSCE), 3(2).

22. Google. (2023). Document AI pricing. Preuzeto 31.3.2023. s https://cloud.google.com/document-ai/pricing

23. Google. (2023). Document AI. Preuzeto 31.3.2023. s https://cloud.google.com/document-ai

24. Google. (2023). Full processor and detail list. Preuzeto 31.3.2023. s https://cloud.google.com/document-ai/docs/processors-list

25. Hayes, A. (2023). What Is Variance in Statistics? Definition, Formula, and Example. Investopedia. Preuzeto 3.4.2023. s https://www.investopedia.com/terms/v/variance.asp

26. Huang, Q., Gao, W. i Cai, W. (2005). Thresholding technique with adaptive window selection for uneven lighting image. Pattern Recognition Letters 26.

27. IBM. (18.2.2022.). What Is Optical Character Recognition (OCR)? Preuzeto 31.3.2023. s https://www.ibm.com/cloud/blog/optical-character-recognition

28. Internet Archive. (2010). Step 3 - Process Document. Preuzeto 31.3.2023. s https://archive.org/details/ProcessDocument/page/n1/mode/2up?view=theater

29. Internet Archive. (n.d.). Internet Archive Digitization Services - Partner Documents. Preuzeto 31.3.2023. s https://archive.org/details/partnerdocs

30. Internet Archive. (n.d.). Internet Archive. Preuzeto 15.6.2023. s https://archive.org

31. ISRI OCR evaluation tools. (n.d.). U Google Code. Preuzeto 13.6.2023. s https://code.google.com/archive/p/isri-ocr-evaluation-tools/

32. Jain, B., i Borah, M. (2014). A Comparison Paper on Skew Detection of Scanned Document Images Based on Horizontal and Vertical Projection Profile Analysis. International Journal of Scientific and Research Publications, 4(6).

33. Javed, A. (n.d.). Digital Image Processing: Lecture #5, Image Enhancement in Spatial Domain- I [prezentacija]. University of Engineering and Technology, Taxila, Pakistan. Preuzeto 3.4.2023. s https://web.uettaxila.edu.pk/CMS/AUT2010/seDIPbs/notes%5CLecture_05%20Image%20Enhancement%20in%20spatial%20domain.pdf

34. Kovač, A., Dunđer, I., Seljan, S. (2022). An overview of machine learning algorithms for detecting phishing attacks on electronic messaging services. MIPRO 2022 - International Convention on Information, Communication and Electronic Technology. 954-961

35. Kumar Chaubey, A. (2016). Comparison of The Local and Global Thresholding Methods in Image Segmentation. World Journal of Research and Review (WJRR), 2(1), str.1-4.

36. Kurama, V. (2022). AWS Textract Teardown - Pros and Cons of using Amazon's Textract in 2023. Preuzeto 3.4.2023. s https://nanonets.com/blog/aws-textract-teardown-pros-cons-review/

37. Kutvonen, A. (2022). Get started with deep learning OCR - Towards Data Science. Preuzeto 3.4.2023. s https://towardsdatascience.com/get-started-with-deep-learning-ocr-136ac645db1d

38. Leksikografski zavod Miroslav Krleža. (n.d.). digitalizacija. U Hrvatska enciklopedija, mrežno izdanje. Preuzeto 31.3.2023. s http://www.enciklopedija.hr/Natuknica.aspx?ID=68025

39. Leung, K. (2022). Evaluate OCR Output Quality with Character Error Rate (CER) and Word Error Rate (WER). Preuzeto 3.4.2023. s https://towardsdatascience.com/evaluating-ocr-output-quality-with-character-error-rate-cer-and-word-error-rate-wer-853175297510

40. Lončarić, S. (n.d.). Poboljšanje slika u prostornoj domeni [prezentacija]. Fakultet elektrotehnike i računarstva, Sveučilište u Zagrebu. Preuzeto 3.4.2023. s https://www.fer.unizg.hr/_download/repository/opdos05a.pdf

41. Ludwig, J. (n.d.). Image Convolution [prezentacija]. Portland State University, Portland, SAD. Preuzeto 3.4.2023. s https://web.pdx.edu/~jduh/courses/Archive/geog481w07/Students/Ludwig_ImageConvolution.pdf

42. Manikpuri, U., i Yadav, Y. (2014). Image Enhancement Through Logarithmic Transformation. International Journal of Innovative Research in Advanced Engineering (IJIRAE), 1(8).

43. Ministarstvo kulture i medija Republike Hrvatske. (2020). Smjernice za digitalizaciju kulturne baštine.

44. Mittal, R., i Garg, A. (2020). Text extraction using OCR: A Systematic Review. U 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA). Indija, Coimbatore.

45. Mukhopadhyay, P., i Chaudhuri, B.B. (2015). A survey of Hough Transform. Pattern Recognition, 48(3). https://doi.org/10.1016/j.patcog.2014.08.027

46. Nguyen, T.T.H., Jatowt A., Coustaty, M., i Doucet, A. (2021). Survey of Post-OCR Processing Approaches. ACM Computing Surveys, 54(6). https://doi.org/10.1145/3453476

47. Ni. (2023). Thresholding. Preuzeto 3.4.2023. s https://www.ni.com/docs/en-US/bundle/ni-vision-concepts-help/page/thresholding.html

48. OcrevaluUAtion. (n.d.). OcrevalUAtion. U GitHub. Preuzeto 13.6.2023. s https://github.com/impactcentre/ocrevalUAtion

49. Phi, M. (2020). Illustrated Guide to LSTM's and GRU's: A step by step explanation. Preuzeto 3.4.2023. s https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21

50. Photshop. (2022). List of filters supporting 16-bit/channel and 32-bit/channel documents. Preuzeto 3.4.2023. s https://helpx.adobe.com/photoshop/using/filter-effects-reference.html

51. Radha, N. (2012). Comparison of Contrast Stretching methods of Image Enhancement Techniques for Acute Leukemia Images. International Journal of Engineering Research & Technology (IJERT), 1(6).

52. Reljić, I., Dunđer, I. Application of Photogrammetry in 3D Scanning of Physical Objects. TEM Journal 8 (1), 94

53. Reljić, I., Dunđer, I., Seljan, S. Photogrammetric 3D Scanning of Physical Objects: Tools and Workflow, TEM Journal 8 (2), 383-388

54. Roy, P., Dutta, S., Dey, N., Dey, G., Chakraborty, S., i Ray, R. (2014). Adaptive thresholding: A comparative study. 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT). doi:10.1109/iccicct.2014.6993140

55. Sahidan, S. I., Mashor, M. Y., Wahab, A. A., Salleh, Z., i Ja'afar, H. (2008). Local and Global Contrast Stretching For Color Contrast Enhancement on Ziehl-Neelsen Tissue Section Slide Images. 4th Kuala Lumpur International Conference on Biomedical Engineering.

56. Schantz, H.F. (1982). The history of OCR, optical character recognition. Manchaster Center, Vermont, SAD: Recognition Technologies Users Association

57. Sears-Collins, A. (2021). Difference Between Histogram Equalization and Histogram Matching. Preuzeto 3.4.2023. s https://automaticaddison.com/difference-between-histogram-equalization-and-histogram-matching/

58. Seljan, S., Dunđer, I. (2014). Combined Automatic Speech Recognition and Machine Translation in Business Correspondence Domain for English-Croatian. Journal of Computer, Information, Systems and Control Engineering. WASET 8 (11), 1069.

59. Seljan, S., Dunđer, I. (2015). Automatic quality evaluation of machine-translated output in sociological-philosophical-spiritual domain. Information Systems and Technologies (CISTI 2015), 1-4.

60. Seljan, S., Dunđer, I., Gašpar, A. (2013). From digitisation process to terminological digital resources. MIPRO 2013, 1053-1058.

61. Seljan, S., Dunđer, I., Stančić, H. (2017). Extracting terminology by language independent methods. Forum Translationswissenschaft: Translation Studies and Translation Practice 19, 141-147.

62. Seljan, S., Tolj, N., Dunđer, I. (2023). 595-Information Extraction from Security-Related Datasets. MIPRO 2022 - International Convention on Information, Communication and Electronic Technology.

63. Singh, P., i Budhiraja, S. (2011). Feature Extraction and Classification Techniques in O.C.R. Systems for Handwritten Gurmukhi Script – A Survey. International Journal of Engineering Research and Applications (IJERA), 1(4).

64. Srinivasan, S. (2016). Fundamentals of digital image processing: A practical approach. CRC Press. Srinivassan

65. Stamos, E.A. (n.d.). OCReval. U GitHub. Preuzeto 13.6.2023. s https://github.com/eddieantonio/ocreval

66. Stančić, H. (2009). Digitalizacija. Zagreb: Zavod za informacijske studije.

67. Subasi, A. (2020). Practical Machine Learning for Data Analysis Using Python.

68. Suganya, S. (2015). Analysis of Feature Extraction of Optical Character detection in Image Processing Systems, National Conference on Recent Trends in Engineering and Technology, 3(4).

69. Tesseract. (2023). Tesseract OCR. U GitHub. Preuzeto 3.4.2023. s https://github.com/Tesseract-ocr/Tesseract/blob/main/README.md

70. Tesseract. (n.d.) Command Line Usage. U GitHub. Preuzeto 3.4.2023. s https://Tesseract-ocr.github.io/tessdoc/Command-Line-Usage.html

71. Tesseract. (n.d.). Improving the quality of the output. U GitHub. Preuzeto 31.3.2023. s https://Tesseract-ocr.github.io/tessdoc/ImproveQuality.html

72. Tesseract. (n.d.). Tesseract User Manual. U GitHub. Preuzeto 3.4.2023. s https://Tesseract-ocr.github.io/tessdoc/

73. Thomas, G., Flores-Tapia, D., i Pistorius, S. (2011) Histogram Specification: A Fast and Flexible Method to Process Digital Images. IEEE Transactions on Instrumentation and Measurement, 60(5). doi: 10.1109/TIM.2010.2089110.

74. Thorat, C., Bhat, A., Sawant, P., Bartakke, I., i Shirsath, S. (2022). A Detailed Review on Text Extraction Using Optical Character Recognition. U ICT Analysis and Applications.

75. Toshkov, D.D. (2012). Weighted variance and weighted coefficient of variation. RE-DESIGN. Preuzeto 3.4.2023. s http://re-design.dimiter.eu/?p=290

76. Trbušić, Ž. (2019). Zašto je arhivima potreban sustav za optičko prepoznavanje znakova?. @rhivi, 6, 6-7.

77. Vijayarani, S., i Sakila, A. (2015). Performance Comparison of OCR Tools. International Journal of UbiComp, 6(3). https://doi.org/10.5121/iju.2015.6303

78. Yan, H. (1993). Skew Correction of Document Images Using Interline Cross-Correlation. CVGIP: Graphical Models and Image Processing, 55(6). https://doi.org/10.1006/cgip.1993.1041

Digitalizacija i obrada slike i teksta u sustavima za optičko prepoznavanje znakova u domeni bibliotekarstva

Sažetak

U svijetu koji ide prema digitalizaciji svega, pa tako i tekstualnih podataka koji nisu izvorno digitalni, optičko prepoznavanje znakova ima centralnu ulogu. Takvi sustavi danas još nisu usavršeni, te ne postoji sustav koji ima 100%-tnu točnost prepoznavanja znakova. Donja granica za isplativost korištenja sustava za optičko prepoznavanje znakova često se citira kao 99,95%, a kako bismo se približili toj vrijednosti potrebna je i ljudska intervencija u raznim fazama rada sustava. Najčešće se takva intervencija predlaže u fazi pretprocesiranja slika prije korištenja istih u sustavima za optičko prepoznavanje znakova. Ovaj rad daje pregled procesa digitalizacije tekstualnog gradiva, pregled povijesti razvitka sustava za optičko prepoznavanje znakova, te kratak opis danas najpoznatijih i najefikasnijih takvih sustava na tržištu. Također istražuje i sve faze rada sustava za optičko prepoznavanje znakova, s naglaskom na razne metode pretprocesiranja, te njihovu implementaciju i efikasnost u poboljšavanju rezultata očitanja znakova u sustavima *Tesseract* i *Google Document AI*.

Ključne riječi: optičko prepoznavanje znakova, pretprocesiranje, poboljšanje slike, binarizacija, siva skala

Digitization, image and text processing in optical character recognition systems in the field of Librarianship

Summary

In a world that seems to be going towards digitising everything, including non-digitally native textual data, optical character recognition plays a central role. Optical character recognition systems haven't yet been perfected - a system with an accuracy rate of 100% in all cases is, as of the writing of this thesis, non-existent. 99.95% is often cited as the bottom barrier for accuracy necessary for the usage of optical character recognition systems to be cost-effective. To get near that number, human intervention is often necessary in many phases of an optical character recognition system's work cycle. In literature, interventions are most often recommended in the phase of preprocessing, e.g., before we even feed an image to an optical character recognition system.

This thesis gives a brief summary of the process of digitising textual materials, the history of optical character recognition and the most popular and effective optical character recognition systems available today. It also explores all phases of an optical character recognition system's work cycle, with emphasis on preprocessing implementation and efficiency in increasing the rate of recognition in *Tesseract* and *Google Document AI*.

Key words: optical character recognition, preprocessing, image enhancement, binarization, greyscale