SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 1605

SUSTAV ZA OPTIČKO RASPOZNAVANJE TEKSTA STARIJIH KNJIGA NA HRVATSKOME JEZIKU

Dominik Agejev

Zagreb, rujan, 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 4. ožujka 2024.

ZAVRŠNI ZADATAK br. 1605

Pristupnik:	**Dominik Agejev (0036537505)**
Studij:	Elektrotehnika i informacijska tehnologija i Računarstvo
Modul:	Računarstvo
Mentor:	izv. prof. dr. sc. Tomislav Hrkać

Zadatak: **Sustav za optičko raspoznavanje teksta starijih knjiga na hrvatskome jeziku**

Opis zadatka:

Optičko raspoznavanje teksta izazovan je problem računalnog vida koji podrazumijeva pretvorbu skeniranog ili fotografiranog teksta u strojno čitljiv oblik prikazan u nekom od standardnih računalnih tekstovnih formata. Dodatni izazovi javljaju se pri takvoj pretvorbi starijih knjiga, primjerice zbog narušene kvalitete papira, slabije kvalitete tiska, arhaičnog jezika i slično. U okviru ovog završnog rada potrebno je: proučiti najznačajnije slobodno dostupne alate za optičko raspoznavanje teksta i opisati njihove osnovne značajke, ispitati primjenjivost odabranih sustava za raspoznavanje teksta starijih knjiga na hrvatskome jeziku uz primjenu prikladnih tehnika predobrade i naknadne obrade, pripremiti bazu slika za ispitivanje sustava, analizirati ponašanje ostvarenog sustava te prikazati i ocijeniti ostvarene rezultate. Radu priložiti izvorni i izvršni kod razvijenih postupaka, ispitne podatke i rezultate, uz potrebna objašnjenja i dokumentaciju te navesti korištenu literaturu.

Rok za predaju rada: 14. lipnja 2024.

Jakovu, Lovri i Roku, hvala na pomoći!

Sadržaj

1. Uvod

Cilj rada nadići je uspješnost gotovih sustava za optičko raspoznavanje teksta (eng. *Optical Character Recognition* ili *OCR*) na starijim knjigama hrvatskoga jezika koristeći se nenadziranim metodama učenja, odnosno bez označenih podataka za trening modela, uz predobradu i naknadnu obradu.

Iako suvremeni sustavi poput DTrOCR-a [1] postižu gotovo savršene rezultate u raznim primjenama, optičko raspoznavanje teksta nipošto nije riješen problem. Još uvijek i najbolji sustavi, poput gore navedenog, pogrešno prepoznaju više od 10% riječi na fotografijama teksta „u divljini“ i gotovo 20% riječi u rukopisima na kineskom jeziku.

Uz to, zbog ovisnosti o jezičnim modelima, novija rješenja općenito nisu primjenjiva bez dodatne prilagodbe na manje zastupljene jezike, poput hrvatskog, ili se pak oslanjaju na veliku količinu označenih podataka ili na sintetičke podatke, generirane modelima koji za rjeđe jezike nisu dostupni, te zahtijevaju značajne računalne resurse.

Nadalje, zbog suviše uske primjene, specifični problemi, poput predmeta ovog rada, raspoznavanja teksta antikvarnih knjiga i to na jeziku ograničene uporabe, redovito se zaobilaze u prilog doprinosima aktualnim primjenama. [2]

U okviru ovog rada najprije se uvodi u područje, metode i temeljne pojmove koji će se koristiti u radu. Zatim je detaljno izložen zadatak i njegove specifičnosti uz pregled dosadašnjih postignuća unutar područja.

Prelazeći na izvedbu rješenja, razmatraju se najznačajniji slobodno dostupni OCR alati prikladni zadatku, a to su Tesseract [3], OCR sustav opće namjene koji održava Google, te Ocular [**?**], razvijen specifično za primjenu na antikvarnim dokumentima.

Nakon treniranja i optimiziranja hiperparametara Oculara, uspoređen je s Tesserac-

tom na već predobrađenim ispitnim podacima gdje se pokazuje da usprkos starijoj arhitekturi u bitnome nadjačava Tesseract, ali uz određena ograničenja.

Konačno, izveden je sustav glasanja kojim se postiže veća uspješnost od one samostalnih modela.

2. Uvod u OCR

Optičko raspoznavanje teksta grana je računalnog vida koja se bavi izdvajanjem teksta iz slika, bilo dokumenata, rukopisa ili scenskih fotografija, radi lakog pretraživanja i uređivanja, jednostavnijeg arhiviranja ili pak dostupnosti sadržaja slabovidnima i slijepima.

U odnosu na sadržaj ulaznih slika najčešće govorimo o prepoznavanju teksta tiskanih dokumenata, rukopisa ili teksta „u divljini“, npr. natpisa na pročeljima trgovina, s tim da je potonje uže povezano s drugim granama računalnog vida poput detekcije i klasifikacije objekata.

OCR sustavi često se razvijaju i za još uže definirane zadatke, primjerice prepoznavanje teksta na računima ili antikvarnim dokumentima. Takvi sustavi, kakvima se bavi i ovaj rad, nazivaju se jednonamjenskima (eng. *task-specific*), dok se sustavi prilagođeni raznim uporabama zovu sustavima opće namjene (eng. *general purpose*). [4]

Optičko raspoznavanje teksta podrazumijeva u bitnome pet koraka: predobradu, segmentaciju, izdvajanje značajki, klasifikaciju te naknadnu obradu. [5] U ovom poglavlju izložit će se ugrubo najznačajnije metode i pojmovi koji će se koristiti u ostatku rada.

2.1. Predobrada

Predobrada se odnosi na postupak prilagodbe ulazne slike radi uspješnijeg raspoznavanja znakova. Načela po kojima se ravna predobrada uključuju pojednostavljenje ulaza izostavljanjem nebitnih informacija, što čine binarizacija i eliminacija šuma, te ispravljanje fizičkih nesavršenosti, uzrokovanih bilo tiskom bilo digitalizacijom ulaza, što rade metode poput ispravljanja nagnuća (eng. *skew correction*).

2.1.1. Binarizacija

Cilj binarizacije razlučivanje je između teksta i pozadine. Najjednostavniji način za to postavljanje je praga (eng. *thresholding*) za koji su svi pikseli s RGB ili sivotonskim (eng. *grayscale*) vrijednostima nižim od praga obojani crno, tj. označeni kao tekst, a pikseli iznad praga označeni kao pozadina. [6]

Razlikujemo globalne i lokalne metode binarizacije. Globalne, poput često korištene Otsuove metode [7], postavljaju jedan prag za cijeli dokument, što je vremenski učinkovito i uspješno u idealnom slučaju s jednoličnom pozadinom, no zakazuje pri nejednakom osvjetljenju ili sjeni uslijed loše skeniranog pregiba knjige. Lokalne, poput metode adaptivnog kontrasta [8], temeljenoj na prepoznavanju rubova pomoću kontrasta susjednih piksela, nešto su resursno zahtjevnije, ali zato daju bolje rezultate. [7, 8]

2.1.2. Ispravljanje nagnuća

Cilj ispravljanja nagnuća zaokrenuti je retke teksta tako da su vodoravni. Utvrđivanje kuta nagnuća redaka binarizirane slike može se svesti na pronalazak pravca koji najbolje aproksimira redak, a u tu svrhu najčešće se koristi Houghova transformacija. [**?**] Konceptualno, Houghova transformacija za svaku rubnu točku slike pronalazi parametre (m, c) pravaca koji se kroz nju mogu provući. Pronađeni skup parametara zapravo je pravac u m, c prostoru, a odredivši pripadajući pravac svakoj točci, ako neki pravci imaju zajedničko sjecište, kroz njima pripadne točke moguće je povući pravac koji u konkretnom slučaju određuje redak teksta.

2.2. Segmentacija

2.3. Izdvajanje značajki

Neuralne mreže se koriste i za izdvajanje značajki, što sad s organizacijom poglavlja? prof. Hrkać: Ništa, možete samo spomenuti da su starije metode koristile ručno definirane značajke (npr. tzv. Granlundove koeficijente) dok se kod današnjih metoda značajke uče, pa se izdvajanje značajki obavlja u sklopu iste neuronske mreže koja se koristi i za klasifikaciju (raniji slojevi izdvajaju (naučene) značajke, dok završni sloj provodi klasifikaciju). Dakle bit će vrlo kratko podpoglavlje :).

2.4. Klasifikacijske arhitekture

2.4.1. Usporedba predložaka

2.4.2. CNN

2.4.3. RNN

2.4.4. LSTM

long short-term memory (*LSTM*)

2.5. Naknadna obrada

Tesseract zamjena ’,’ (dva zareza) s " (ravni navodnici)

3. OCR starijih tekstova

Iz specifičnosti zadatka, tj. starosti knjiga, proizlaze određene poteškoće, naime:

- Čest višak ili manjak tinte pri tisku pojedinih znakova
- Zastarjela znakovlja (fontovi) s neuobičajenim znakovima
- Otežana predobrada zbog spremanja na mikrofilmu
- Arhaičan jezik
- Neravan tisak
- Istrošenost i oštećenja papira

3.1. Pregled literature

[**?**] [9] [10] [11] [12] [13]

4. Tesseract

Tesseract [3] je najpoznatiji i najprecizniji slobodno dostupan OCR sustav opće namjene koji podržava i hrvatski jezik, a od 4. inačice temeljen je na LSTM neuronskim mrežama. U ovom poglavlju objasnit će se ugrubo Tesseractov proces prepoznavanja teksta prema zadanim postavkama.

4.1. Predobrada

Tesseract ima ugrađena tri koraka predobrade: binarizaciju, eliminaciju šuma i analizu uređenja stranica.

4.1.1. Binarizacija

Tesseract se koristi Otsuovom metodom, no ne za postavljanje jednog globalnog praga za čitavu stranicu, već rabi implementaciju Leptonica biblioteke [14] koja dijeli stranicu u jednake blokove te na njima postavlja prag. Takvim pristupom nadvladavaju se varijacije u svjetlini na makro razini slike, ali se zadržava i veća resursna učinkovitost uslijed paralelizacije i izbjegavanja složenijih računa lokalnih metoda.

Za slike koje nisu više-manje dvobojne nego pate od većih nejednakosti u osvijetljenju Tesseract podržava i Sauvolinu [15] lokalnu metodu binarizacije koja utvrđuje prag za svaki piksel slike na temelju srednje vrijednosti i standardne devijacije okolnih piksela.

4.1.2. Eliminacija šuma

Razlučivši pozadinu od ostatka slijedi brisanje šuma, tj. smetnji poput prolivene tinte. To se postiže pronalaskom spojenih piksela te usporedbom karakteristika nakupine s tipičnim karakteristikama teksta.

Tijekom analize razmatra se: [16]

Širina poteza

Potezi jednolične širine vjerojatnije pripadaju znaku.

Veličina nakupina

Skupine piksela koje se protežu izvan uobičajene visine retka vjerojatno nisu znakovi.

Obujmljena površina

Gledajući površinu koju skupina piksela okružuje možemo procijeniti je li znak ili nije.

Broj nakupina po retku

Brojeći skupine piksela u retku provjerava se omjer malih nakupina naspram skupina veličine znaka.

Odnos među točkama

Ako se detektira velik broj susjednih točaka na istoj visini ne odbacuju se već su označene kao "vodeće točke" sadržaja.

4.1.3. Analiza uređenja stranice

Line Detection Tesseract uses Leptonica to find and remove rule/separator lines in the input image. This helps to separate text from graphical elements like tables or forms.

Photo Region Detection The `FindImages` function from Leptonica is used to detect photo regions in the input. This allows Tesseract to distinguish between text and image areas.

Connected Component Analysis This step identifies individual characters or character fragments. It is performed by the `find_components` function, which scans the binary image pixel by pixel, labels connected black pixels, and groups them into connected components representing potential characters or parts of characters.

Orientation and Script Detection If enabled, Tesseract analyzes the layout to detect text orientation and script. This is particularly useful for documents that may contain text in multiple orientations or scripts.

Column Detection If `PSM_COL_FIND_ENABLED` is true for the selected page segmentation mode, Tesseract attempts to divide the image into columns. This is crucial for correctly processing multi-column documents.

Text Line Formation Tesseract analyzes the spatial relationships between connected components to detect text lines and words. This step uses statistical approaches based on the spacing between components.

[17] o LSTM-u

5. Ocular

Ocular [18] je sustav za optičko raspoznavanje teksta razvijena specifično za rad s povijesnim dokumentima, i koji je, kada je izdan i svojevremeno unaprjeđen [19], bio vrhunac tehnologije za to područje (eng. *state-of-the-art*).

Njegove glavne značajke su: [**?**]

- Nenadzirano učenje nepoznatih znakovlja rabeći slike ulaznog dokumenta i korpus teksta na ciljnom jeziku.
- Prilagođenost radu sa šumovitim dokumentima.
- Podrška za višejezične dokumente.
- Nenadzirano učenje ortografskih varijacija uslijed arhaičnog pravopisa.
- Istovremen ispis doslovnog teksta i normaliziranog oblika (prilagođenog standardnom jezku).

6. Metodologija

Ključan dio razvoja boljeg rješenja evaluacija je preciznosti Tesseracta i Oculara. U tu svrhu potreban je ispitni skup podataka prilagođen ograničenjima sustava.

Zatim treba prikupiti podatke za treniranje Oculara i namjestiti njegove hiperparametre da daju zadovoljavajuće rezultate, što će zapravo biti najznačajniji dio rada.

6.1. Ispitni skup podataka

Budući da Tesseract ima ugrađenu predobradu slika, radi pravednije usporedbe same klasifikacije teksta izabrani su već obrađeni dokumenti.

Izvadci iz:

- Fra Jozo Garić, biskup – Korizmena okružnica (1932.)
- Prof. dr. Antun Heinz – Nekoliko misli o definiciji i klasifikaciji plodova (1897.)
- Sv. Petar Kanizije – Summa nauka christianskoga (1583.)

Oba sustava imaju određena ograničenja na ulaze:

Ocular radi jedino s PDF dokumentima zastarjele verzije 1.4 te ih je stoga bilo potrebno pretvoriti u taj format. Za to je korišten Ghostscript, [**?**] slobodno dostupan alat otvorenog koda.

Tesseract pak radi jedino na slikama te je zato bilo potrebno ekstrahirati ih iz PDF-a prije prepoznavanje teksta. Ovdje je zgodno napomenuti da treba paziti da se prilikom pretvorbe ne smanji DPI rezolucija jer to ima poguban utjecaj na preciznost.

6.2. Mjere uspješnosti

Za uspoređivanje znakovnih nizova najčešće korištena mjera uspješnosti je **Levenshteinova udaljenost** koja bilježi broj potrebnih zamjena, brisanja ili umetanja znakova da bi se iz jednog niza dobio drugi. [20]

Budući da ta mjera ovisi o duljini teksta, dijeljenjem Levenshteinove udaljenosti ukupnim brojem znakova dobivamo **stopu pogreške za znakove** (eng. *Character Error Rate*). Obično se dobrom vrijednošću smatra 1-2%.

Stopa pogreške za riječi (eng. *Word Error Rate*) dobiva se uzimanjem riječi za najmanju jedinicu zamjene pri računanju Levenshteinove udaljenosti, tj. ako su jedan ili više znakova u riječi pogrešni čitava riječ broji se kao pogrešna, te dijeljenjem te udaljenosti s ukupnim brojem riječi.

6.2.1. F1?

6.2.2. Recall?

7. Optimizacija Oculara

Za razliku od Tesseracta, Ocular nije univerzalno primjenjiv za različite jezike i znakovlja već je potrebno naučiti jezični model, za koji je potreban korpus teksta na ciljnom jeziku, te model fonta, koji izučava na temelju slika čiji će tekst kasnije prepoznavati.

7.1. Jezični model

Pri izgradnji skupa podataka za trening jezičnog modela najrelevantnije su dvije stavke: broj podataka i tematika teksta.

U izvornom radu pokazano je kako model malo precizniji (4 WER postotna boda) na dokumentima čija je tematika pokrivena u jezičnom modelu. Budući da je sustav namijenjen starijim knjigama, od kojih je dobar dio vjerske tematike, uključeno je Sveto Pismo i druge duhovne knjige pored novijeg i starijeg štiva koje doprinosi većoj raznolikost izričaja i opsežnijem rječniku. [18]

Izvorno trenirano na tekstovima javno dostupnih knjiga, poput djela Augusta Šenoe, Marije Jurić-Zagorke, Charlesa Dickensa i sl. (7.5 milijuna riječi) uz Šarićev prijevod Svetog Pisma (670k riječi) i još 11 knjiga vjerske tematike (500k riječi).

Ukupan broj riječi od 8.7 milijuna usporediv je sa skupom podataka korištenim u izvornom radu koji ih ima 10 milijuna.

Iteracije treninga	**Veličina snopa**		**CER**	**WER**
	Trening	**Transkripcija**		
3x3	10	10	2.03	3.58
3x3	50	50	**1.06**	**2.76**
3x3+2x2	50	50	1.19	2.9

Tablica 7.1. Trebala bi ići u drugi odjeljak za trening fonta

7.1.1. Proširenje podacima van domene

7.1.2. Proširenje nesavršenim podatcima

Veličina snopa		CER	WER
Trening	Transkripcija		
40	40	1.59	3.64
40	120	1.81	3.77
120	50	2.21	5.04
120	120	2.29	5.78

Tablica 7.2. dd

7.2. Ispitivanje ortografskih značajki

8. Sinteza rješenja

Budući da je izuzev ispuštanja određenih redaka Ocular točniji od Tesseracta, kako je utvrđeno u prethodnom poglavlju, ovdje predlažemo jednostavan sustav glasanja kojim je Ocularova manjkavost otklonjena bez gubitka preciznosti.

Algoritam glasanja čita redak po redak Tesseractov ispis i traži odgovarajući redak Ocularovog ispisa na temelju sličnosti izračunate pomoću Levenshteinove udaljenosti. Ako pronađe dovoljno sličan redak odabire ga kao izlaz, inače preferira Tesseractov ispis.

```
      for t_line in tesseract_output:
        for c_line in ocular_output:
          distance = Levenshtein.distance(t_line, c_line)
          if distance < threshold * len(t_line):
            output.append(c_line)
          break
        output.append(t_line)
```

Slika 8.1. Algoritam glasanja predstavljen Python kodom

OCR sustav	Pojedinosti	CER	WER
Ocular		8.83	10.94
Ocular	Zanemareni retci s CER>20	1.06	2.76
Tesseract		1.52	3.77
Predložen sustav		**0.96**	**2.48**

Tablica 8.1. Uspješnosti sustava

Ideje za nadogradnje: počeci i krajevi redaka su obično Ocularu kritični. Tesseract češće prepoznaje interpunkciju kada treba i kada ne treba.

9. Diskusija

Komentirati konvergenciju računalnog vida, neuralnih mreža, NLP-a i OCR-a.

10. Zaključak

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Literatura

[1] M. Fujitake, "Dtrocr: Decoder-only transformer for optical character recognition", 8 2023. [Mrežno]. Adresa: http://arxiv.org/abs/2308.15996

[2] K. Olejniczak i M. Šulc, "Text detection forgot about document ocr", 10 2022.

[3] R. Smith, "An overview of the tesseract ocr engine", u *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2*. IEEE, 9 2007., str. 629–633. https://doi.org/10.1109/ICDAR.2007.4376991

[4] E. Borovikov, "A survey of modern optical character recognition techniques", 2014.

[5] P. Dhande i R. Kharat, "Recognition of cursive english handwritten characters", *Proceedings - International Conference on Trends in Electronics and Informatics, ICEI 2017*, sv. 2018-January, str. 199–203, 7 2017. https://doi.org/10.1109/ICOEI.2017.8300915

[6] Jyotsna, S. Chauhan, E. Sharma, i A. Doegar, "Binarization techniques for degraded document images-a review", *2016 5th International Conference on Reliability, Infocom Technologies and Optimization, ICRITO 2016: Trends and Future Directions*, str. 163–166, 12 2016. https://doi.org/10.1109/ICRITO.2016.7784945

[7] N. Otsu, "Threshold selection method from gray-level histograms." *IEEE Trans Syst Man Cybern*, sv. SMC-9, str. 62–66, 1979. https://doi.org/10.1109/TSMC.1979.4310076

[8] B. Su, S. Lu, i C. L. Tan, "Robust document image binarization technique for degraded document images", *IEEE Transactions on Image Processing*, sv. 22, str. 1408–1417, 2013. https://doi.org/10.1109/TIP.2012.2231089

[9] U. Springmann, D. Najock, H. Morgenroth, H. Schmid, A. Gotscharek, i F. Fink, "Ocr of historical printings of latin texts: Problems, prospects, progress", u *ACM International Conference Proceeding Series*. Association for Computing Machinery, 2014., str. 71–75. https://doi.org/10.1145/2595188.2595205

[10] M. Christy, A. Gupta, E. Grumbach, L. Mandell, R. Furuta, i R. Gutierrez-Osuna, "Mass digitization of early modern textswith optical character recognition", u *Journal on Computing and Cultural Heritage*, sv. 11. Association for Computing Machinery, 12 2017. https://doi.org/10.1145/3075645

[11] C. Wick, C. Reul, i F. Puppe, "Improving ocr accuracy on early printed books using deep convolutional networks", 2 2018. [Mrežno]. Adresa: http://arxiv.org/abs/1802.10033

[12] D. Garrette, H. Alpert-Abrams, T. Berg-Kirkpatrick, i D. Klein, "Unsupervised code-switching for multilingual historical document transcription", str. 1036–1041, 2015.

[13] D. Garrette i H. Alpert-Abrams, "An unsupervised model of orthographic variation for historical document transcription", 2016.

[14] "Leptonica: Leptonica reference documentation". [Mrežno]. Adresa: https://tpgit.github.io/Leptonica/index.html

[15] J. Sauvola, T. Seppanen, S. Haapakoski, i M. Pietikainen, "Adaptive document binarization", *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, sv. 1, str. 147–152, 1997. https://doi.org/10.1109/ICDAR.1997.619831

[16] "Tesseract open source ocr engine (main repository)". [Mrežno]. Adresa: https://github.com/tesseract-ocr/tesseract/

[17] T. M. Breuel, A. Ul-Hasan, M. A. Al-Azawi, i F. Shafait, "High-performance ocr for printed english and fraktur using lstm networks", *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, str. 683–687, 2013. https://doi.org/10.1109/ICDAR.2013.140

[18] T. Berg-Kirkpatrick, G. Durrett, i D. Klein, “Unsupervised transcription of historical documents”, str. 207–217, 2013. [Mrežno]. Adresa: https://aclanthology.org/P13-1021/

[19] T. Berg-Kirkpatrick i D. Klein, “Improved typesetting models for historical ocr”, str. 118–123, 2014.

[20] V. I. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals”, *Doklady Akademii Nauk SSSR*, sv. 163, str. 845–848, 1965.

Sažetak

SUSTAV ZA OPTIČKO RASPOZNAVANJE TEKSTA STARIJIH KNJIGA NA HRVATSKOME JEZIKU

Dominik Agejev

Cilj rada nadići je uspješnost gotovih sustava za optičko raspoznavanje teksta na starijim knjigama hrvatskoga jezika koristeći se nenadziranim metodama učenja, uz predobradu i naknadnu obradu. Razmatraju se najznačajniji slobodno dostupni OCR alati prikladni zadatku, Tesseract, OCR sustav opće namjene koji održava Google, te Ocular, razvijen specifično za primjenu na antikvarnim dokumentima. Nakon treniranja i optimiziranja hiperparametara Oculara, uspoređen je s Tesseractom gdje se pokazuje da usprkos starijoj arhitekturi u bitnome nadjačava Tesseract, ali uz određena ograničenja. Konačno, izveden je sustav glasanja kojim se postiže veća uspješnost od one samostalnih modela.

Ključne riječi: OCR; optičko raspoznavanje teksta; računalni vid; Ocular; Tesseract;

Abstract

Optical character recognition system for older books in Croatian

Dominik Agejev

The aim of the paper is to surpass the accuracy of out-of-the-box systems at Optical Character Recognition of historical documents in the Croatian language relying on unsupervised learning methods, preprocessing and postprocessing. The most appropriate freely available OCR tools are evaluated, namely Tesseract, a general-purpose OCR system maintained by Google, and Ocular, developed specifically for use on historical documents. After training and optimizing Ocular's hyperparameters it is compared to Tesseract where it is shown that despite its older architecture Ocular in the main still bests Tesseract, with certain caveats. Finally, a voting-based system is implemented which achieves greater success than each model alone.

Keywords: OCR; Optical Character Recogniton; Computer Vision; Ocular; Tesseract;

Privitak A: The Code

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam

rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.