Příprava obrázků k OCR

Obhajoba maturitní práce

Jakub Ambroz

24. května 2020

Základní pojmy

- *AI* = *Artificial Intelligence* = umělá inteligence
- *OCR* = *Optical Character Recognition* = optické rozpoznávání znaků

Tesseract

- OCR program příkazové řádky
- napsaný v C++ a nyní má otevřený zdrojový kód na Githubu
- K dispozici jsou už naučené (natrénované) verze
- Pro zlepšení výsledků doporučuje dokumentace úpravu obrázků

Otázka maturitní práce

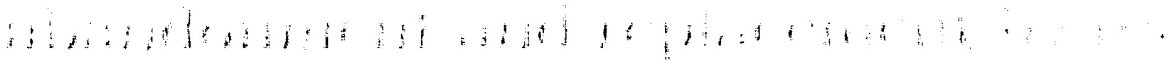
Jak (a jestli) je možné upravit obrázek, tak aby byla zvýšena kvalita rozpoznání znaků Tesseractem?

OpenCV

- *Open Source Computer Vision Library* = Knihovna s otevřeným zdrojovým kódem pro počítačové vidění
- Psaná v C++, ale má rozhraní i v Pythonu

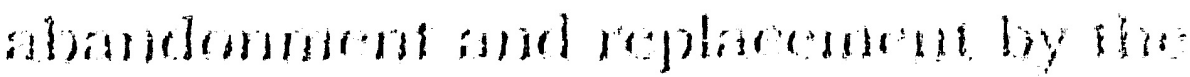
Binarizace

- Převede obrázek na černé a bílé pixely
- Práh, který je rozdělí je buď globální nebo adaptivní (bere v potaz okolí bodu)



Obrázek-1: Gaussovský adaptivní práh

Obrázek-2: Gaussovský adaptivní práh s jinými parametry



Obrázek-3: Adaptivní práh

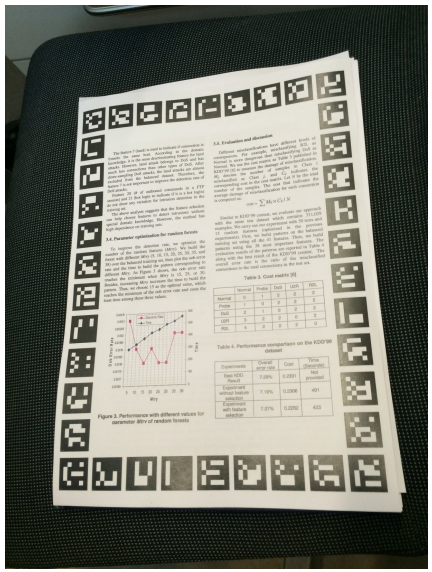
Obrázek-4: Otsuovo práh

Další úpravy

- Grayscale - převede obrázek do úrovní šedi
- Rozmazání - nějakým způsobem průměruje okolí pixelu
- Škálování - změna velikosti obrázku
- Odstranění šumu

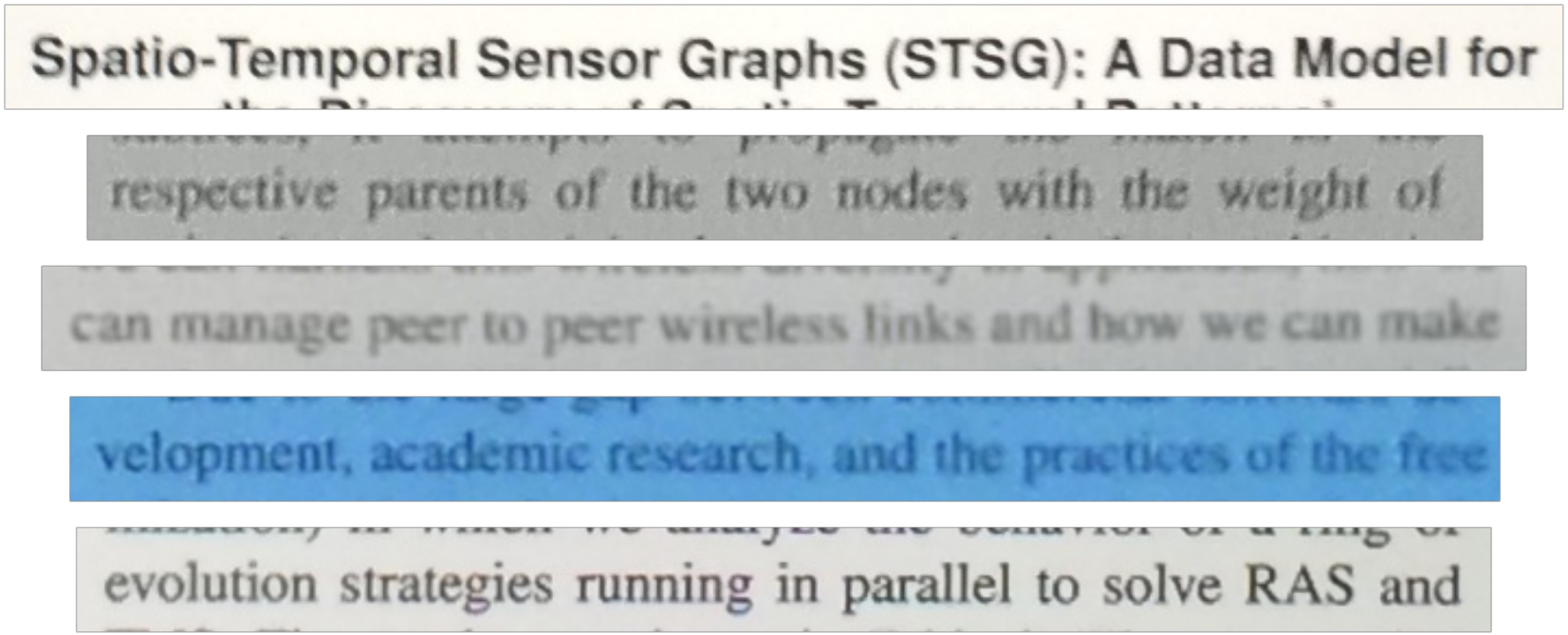
B-MOD

- *Brno Mobile OCR Dataset* - z VUTBR
- Rozmanitý: foceno různými zařízením za různých světelných vzdáleností z různých úhlů
- Rozstřihané fotky po jednotlivých řádcích



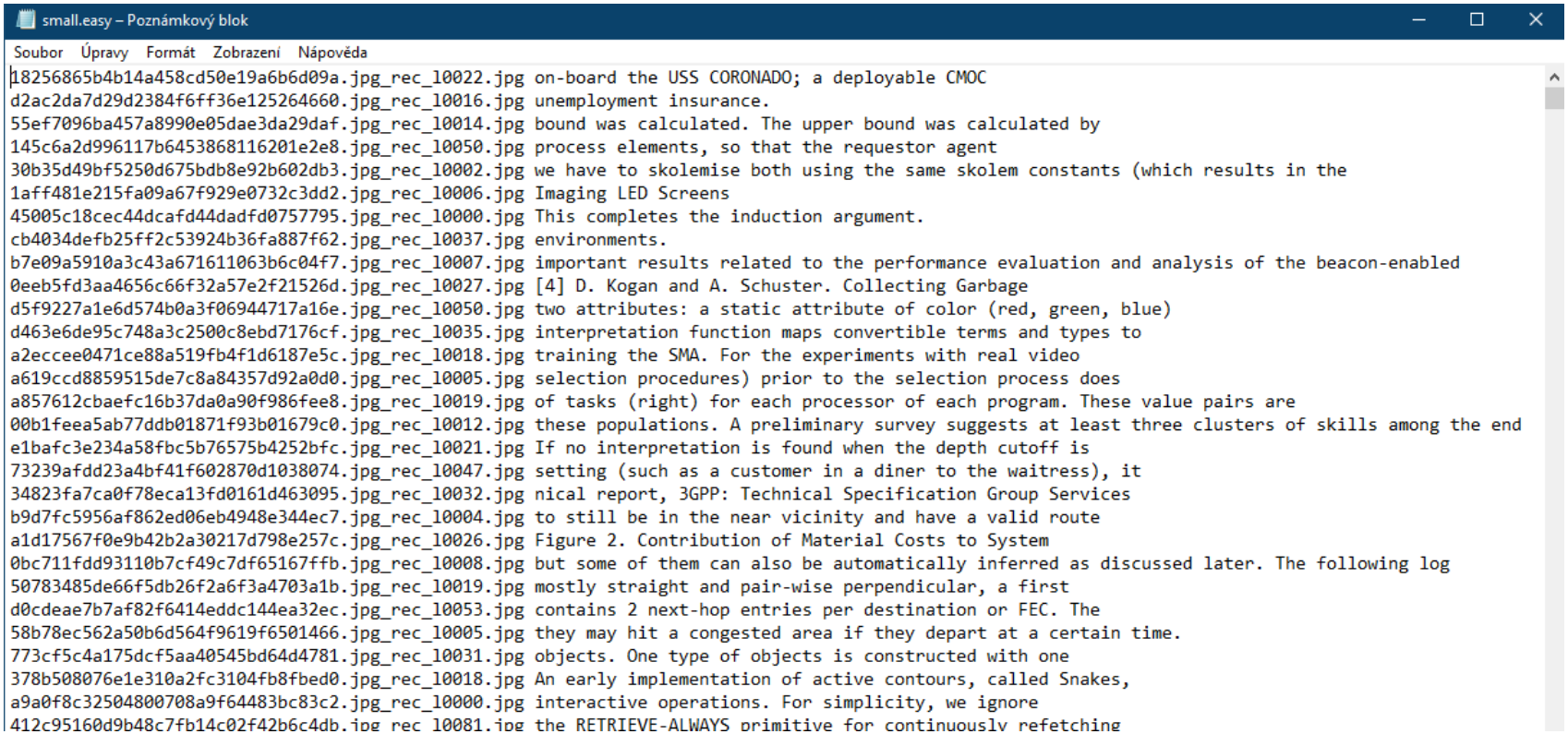
Obrázek-5: Ukázka z datasetu

Mnou použité



Obrázek-6: Po řádcích

Mnou použité 2



Obrázek-7: Textový soubor se správným řešením

Cíle a automatizace

- Zjistit, jak upravit obrázek, tak aby se zlepšili výsledky Tesseractu
- Zkusit velké množství úprav obrázků, zkusit rozpoznat Tesseractem a porovnat úspěšnosti

OpenCV

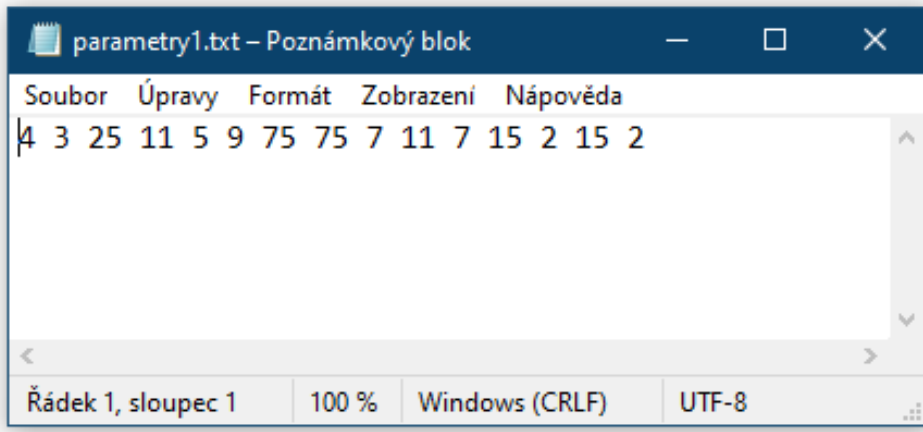
- Vytvořil jsem si kódy jednotlivým funkcím
- Program, který je vykoná v takovém pořadí v jakém jsou předloženy
- Potřebné informace získá program z jedné řádky v textovém dokumentu
- Program provede operaci pro všechny řádky v souboru

OpenCV parametry

prikazy-pro-opencv3.old – Poznámkový blok

Soubor Úpravy Formát Zobrazení Nápověda

```
lines-02(4)-11- small.easy parametry3.txt 43
lines-02(4)-11- small.easy parametry1.txt 31
lines-02(4)-11- small.easy parametry1.txt 32
lines-02(4)-11- small.easy parametry1.txt 33
lines-02(4)-11- small.easy parametry1.txt 34
lines-02(4)-11- small.easy parametry2.txt 31
lines-02(4)-11- small.easy parametry2.txt 32
lines-02(4)-11- small.easy parametry2.txt 33
lines-02(4)-11- small.easy parametry2.txt 34
lines-02(4)-11- small.easy parametry3.txt 31
lines-02(4)-11- small.easy parametry3.txt 32
lines-02(4)-11- small.easy parametry3.txt 33
lines-02(4)-11- small.easy parametry3.txt 34
lines-02(4)-11-21(3,25,11)- small.easy parametry1.txt 31 41
lines-02(4)-11-21(3,25,11)- small.easy parametry1.txt 31 42
lines-02(4)-11-21(3,25,11)- small.easy parametry1.txt 31 43
lines-02(4)-11-21(3,25,11)- small.easy parametry1.txt 32 41
lines-02(4)-11-21(3,25,11)- small.easy parametry1.txt 32 42
lines-02(4)-11-21(3,25,11)- small.easy parametry1.txt 32 43
lines-02(4)-11-21(3,25,11)- small.easy parametry1.txt 33 41
```

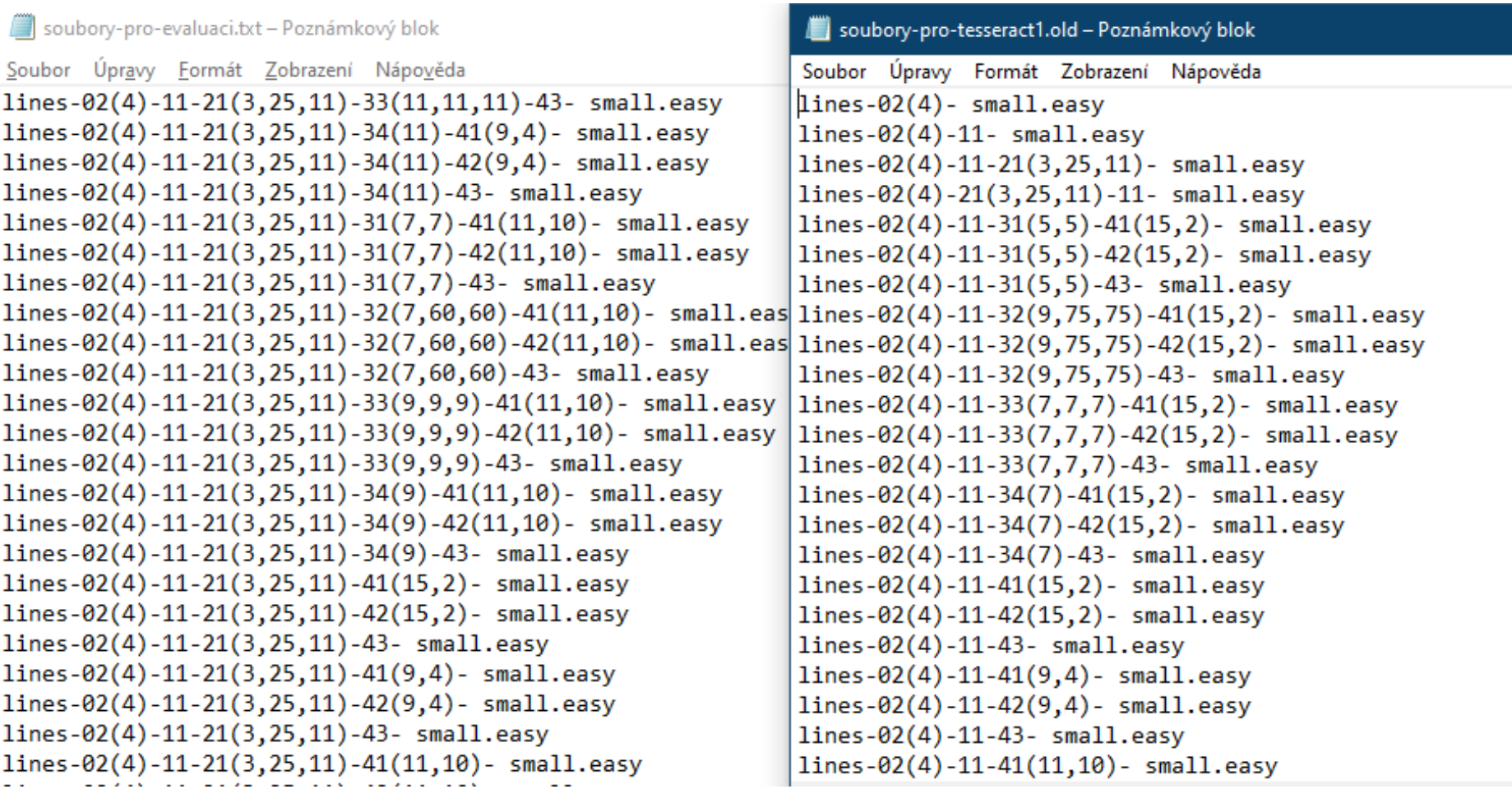


Obrázek-8: Textový soubor pro OpenCV

Tesseract a pytesseract

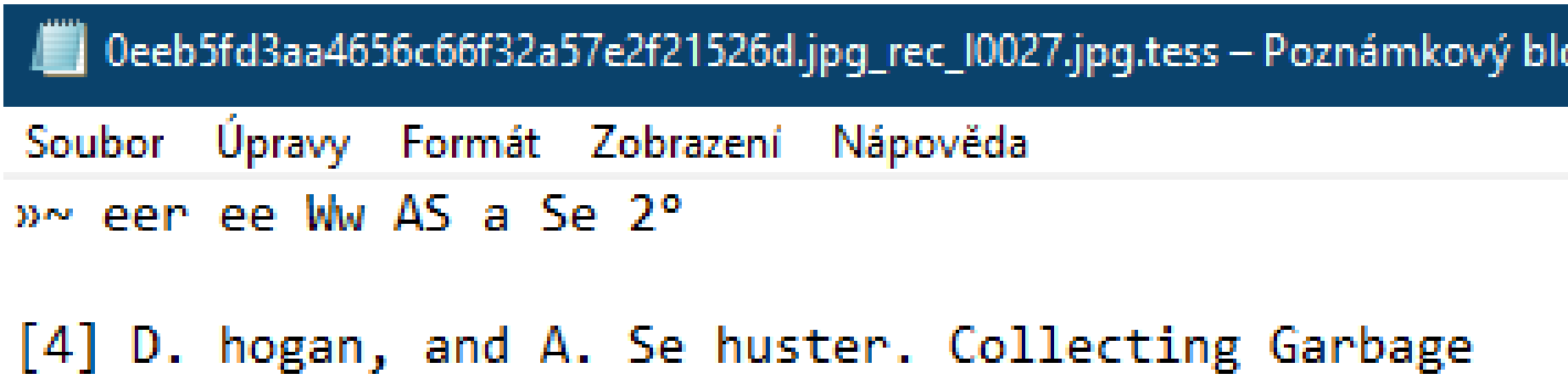
- Je nutné dobře nainstalovat Tesseract
- Je třeba ho nastavit jako příkaz v cmd
- A propojit s Pythonem pomocí *pytesseractu*
- Poté zavolat Tesseract a uložit jeho výstup do textového souboru

Příkazy pro Tesseract



Obrázek-9: Textový soubor pro Tesseract

Výstup Tesseractu

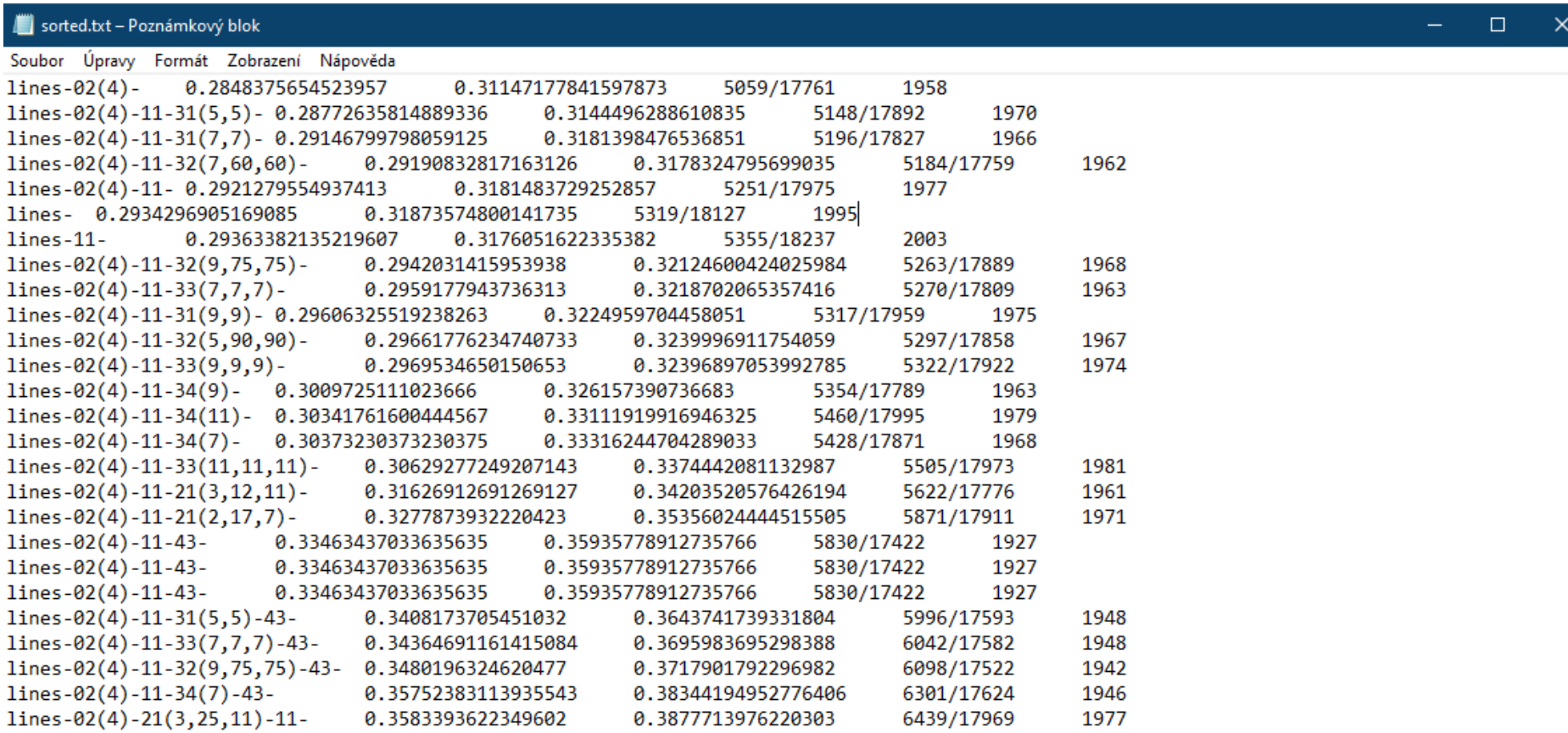


Obrázek-10: Textový soubor vytvořený Tesseractem

Úspěšnost

- Převede znak nové řádky na mezeru
- Rozdělí podle mezer
- Ve vzniklém poli hledá slova v řešení (to také rozdělíme podle mezer)
- Vypočítá jaké procento slov to nenašlo, tj. byly chybně identifikovány

Výstup programu



Obrázek-11: Textový soubor vytvořený evaluačním programem

Tabulka

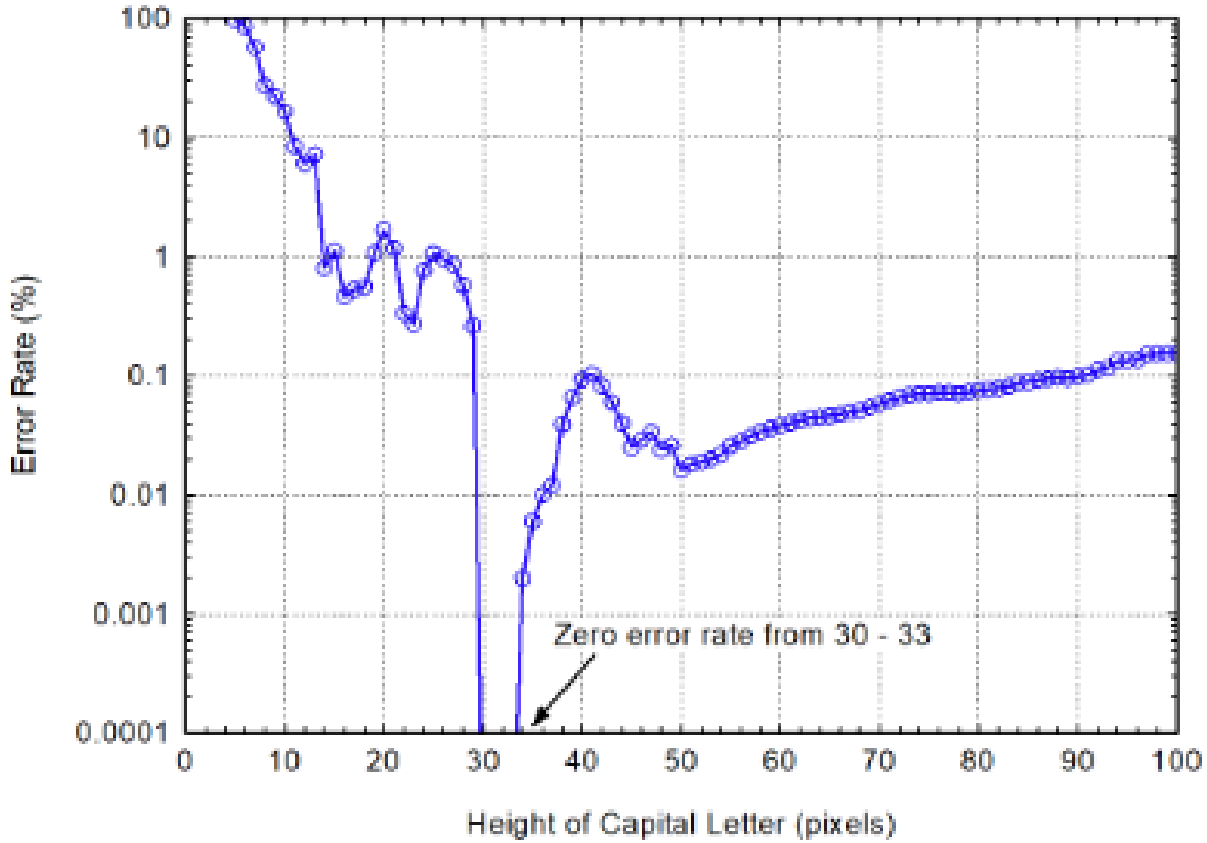
Název složky	Nerozpoznaných slov v %	Průměrný řádek v %
lines-02(4)-	28.48375654523957	31.147177841597875
lines-02(4)-11-31(5,5)-	28.772635814889334	31.44496288610835
lines-02(4)-11-31(7,7)-	29.146799798059124	31.813984765368506
lines-02(4)-11-32(7,60,60)-	29.190832817163127	31.78324795699035
lines-02(4)-11-	29.212795549374132	31.81483729252857
lines-	29.342969051690847	31.873574800141737
lines-11-	29.363382135219606	31.760516223353818
lines-02(4)-11-32(9,75,75)-	29.42031415953938	32.124600424025984
lines-02(4)-11-33(7,7,7)-	29.59177943736313	32.18702065357416

Obrázek-12: Tabulka nejlepších kombinací

Tabulka

Willus Dotkom

the fact that there is an optimum letter size (in pixels) is quite unexpected



Obrázek-13: Optimální velikost pro Tesseract

Nedostatky

- Nepřehledné kódování funkcí OpenCV
- Nesystematické procházení možností

Děkuji za pozornost!

Pokud máte zájem, položte dotazy!