NAKI-II-USTR-UKONCENE - Task #4062

OCR - zhlukovanie stranok podla vizualneho obsahu, odstranenie grafickych artefaktov z okraju stranky

20.10.2016 12:14 - Hrúz Marek

Status:	Closed	**Start date:**	20.10.2016
Priority:	Normal	**Due date:**	25.11.2016
Assignee:	Neduchal Petr	**% Done:**	100%
Category:		**Estimated time:**	40.00 hours
Target version:			

Description

- analyzovat vizualny obsah okraju stranky a navrhnut postup zhlukovania stranok
- konzultovat zistene skutocnosti s veducim pracovnikom (MHR)

- prejednat s Honzom Zelinkom moznosti pouzitia jeho algoritmu pre odstranovanie artefaktov z okraju stranky
- vyskusat dany algoritmus na nase data + analyzovat vysledky/uspesnost
- konzultovat zistene vysledky s veducim pracovnikom (MHR)

History

#1 - 30.01.2017 14:27 - Bureš Lukáš

- % Done changed from 0 to 80

- detekce radek - bude dolazeno
- detekce der - bude dolazeno
- konzultace byla provedena

#2 - 06.02.2017 18:12 - Zajíc Zbyněk

Ze schůzky [[https://wikky.zcu.cz/redmine/projects/naki-ii-ustr/wiki/Schuzka16-10-03]]:

- **PNe** detekce bloků v textu:
  - natočení dokumentu, získání řádků,
  - zahozena světlá místa = není text,
  - označení bloků v textu -> ty rozpoznány Tesserackem (dva možné vstupy - binary nebo šedotón obraz)
- **dodělat** -> spojit bloky na úrovni řádky (aby měl Tesserac větší kontext)

- **PNe** detekce děr po děrovačce
  - pro spojení dokumentů patící do jedné série
  - moc nefunguje
- **dodělat** -> přidat informaci o historii předchozích i následujících obrázků a v nich existenci děr
  -> pracovat s celou segvencí dokumentů v adresáři a v nich maximalizovat ppst. hypotézy jedné/více serií dokumentů

#3 - 20.04.2017 09:03 - Neduchal Petr

- % Done changed from 80 to 90

Zjištění z poslední fáze:

Detekce bloků:
- Hledání bloků samo o sobě nepřináší zlepšení výsledků. Dojde k tomu, že se některé znaky přečtou lépe za cenu toho, že se rozpoznání jiných zhorší. Stejně tak to dopadá s různými způsoby předzpracování. Dle vyhodnocení na anotovaném vzorku dokumentů se samotný tesseract dostal na 79%. Nejlepší nalezená úprava dosahla téměř 84%. Vzhledem k tomu, že těch cca 80% se u různých metod předzpracování skládá z části z jiných dobře rozpoznaných znaků, tak se jako logický krok zdá získání výsledků z více předzpracování a ty dále tanalyzovat. Z toho důvodu je momentálně dořešuje napojení na kód, který je schopen vrátit lattice (základní verze už nám funguje --> bude možné nagenerovat data pro zpracování textu.)

Detekce děr --> respektive detekce stejné části dokumentu v rámci jednoho svazku:
- Aktuálně v řešení.

#4 - 09.07.2018 09:14 - Neduchal Petr

Bylo vyzkoušeno shlukování postavené na SVM a s pomocí neuronové sítě. Obě metody dosahovaly úspěšnosti 70-74% oproti manuálně oanotovaným dokumentům pomocí SW.

Dalším krokem bylo vytvoření syntetizátoru dokumentů. Postup shrnut v článku na SPECOM 2018.

#5 - 19.11.2018 07:52 - Neduchal Petr

- Status changed from Assigned to Resolved

- % Done changed from 90 to 100

Práce na syntetizátoru dokumentů, z nich bude pravděpodobně možné natrénovat klasifikátor. Úspěšnost bez syntetizátoru viz předchozí aktualizace úkolu. Pro teď úkol nastavuji jako resolved jelikož za mě je práce hotová.

#6 - 18.11.2019 09:08 - Zajíc Zbyněk

- Status changed from Resolved to Closed