NAKI-II-USTR-UKONCENE - Task #3803

Task # 3633 (Closed): Etapa 01 - Příprava dat a datových struktur, testy existujících metod

OCR - jazykový model v Tesseractu

09.03.2016 13:15 - Zajíc Zbyněk

Status:	Closed	**Start date:**	18.04.2016
Priority:	High	**Due date:**	05.11.2018
Assignee:	Neduchal Petr	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:			

Description

aktualizovat LM v Tesseractu, tak aby jsme z něj mohli dostat eventuálně: slovní/znakový přepis, lattice

Subtasks:

Task # 3875: Zjistit jak dostat z Tesseractu lattices pro další zpracování/rescoring	**Closed**
Task # 3960: Natrénovat vlastní LM pro Tesseract dle jejich receptu	**Closed**
Task # 3961: Dekódování grapheme-lattice na word-lattice	**Closed**
Task # 3962: Porovnat výsledky na OCR pro náš výsledek z TesseractApi a z modelu trénov...	**Closed**
Task # 4482: Balík OCR	**Closed**

History

#1 - 29.04.2016 14:30 - Hrúz Marek

- Assignee changed from Neduchal Petr to Soutner Daniel

#2 - 23.06.2016 15:17 - Soutner Daniel

Před jazkyovým modelel bude třeba také asi natrénovat na font "psací stroj". Návod by mohl být zde:
http://www.joyofdata.de/blog/a-guide-on-ocr-with-tesseract-3-03/

#3 - 23.06.2016 16:03 - Soutner Daniel

Oficiální popis tréninku LM zde:
https://github.com/tesseract-ocr/tesseract/wiki/tesstrain.sh

Není mi z toho jasné jeslti je to i tzv. "cube" language model, který má být lepší. Dokumentace k němu asi není (?), info se dá najít na googlegroups tesseract-ocr.
Opředeno tajemstvím :)

#4 - 27.06.2016 14:56 - Zajíc Zbyněk

- zatím lze získat z OCR jen 1Best hypotézu, ale zle získat fonémový lattice (slovní lattice asi nedostupný) - lze pak zpracovat vlasním LM
- zapojení vlastního LM nahráním trénovcacích dat

DS- doplnit data do Tesseractu a vyzkouší zpracovat lattice vlastními metodami

#5 - 06.05.2017 18:57 - Zajíc Zbyněk

z lattice vygenerovat slovní přepis (doplnit např. konfuzní tabulku, ...)

#6 - 08.11.2017 15:54 - Zajíc Zbyněk

- WER cca 45% (s LM z novin) https://docs.google.com/spreadsheets/d/1d3UJSlz3XRccygMNnTyxDNwLrrHfCox9Ywnu9LdxATQ/edit#gid=0
- získat lepší LM
- přidat info o pozici slova na stránce

#7 - 05.04.2018 12:34 - Zajíc Zbyněk

vygenerovat a poslat JŠ mřížky

#8 - 11.04.2018 16:59 - Soutner Daniel

Mřížky poslány JŠ, jsou tady: /data-kky/public/dsoutner/ocr-lattice

#9 - 25.06.2018 11:33 - Soutner Daniel

- Assignee changed from Soutner Daniel to Neduchal Petr

#10 - 18.11.2019 09:08 - Zajíc Zbyněk

- Status changed from Assigned to Closed