

OCR Entwickler-Workshop

Aktuelle Entwicklungen in der Tesseract-Community

Stefan Weil, UB Mannheim

28./29. September 2017
BBAW Berlin

Tesseract Community News

- Tesseract-Community seit 2015 deutlich aktiver, insbesondere seit einem Jahr: 40 Contributors, 790 Commits in den letzten 12 Monaten, auch einige neue Entwickler (siehe [OpenHub](#))
- Auch der Hauptentwickler Ray Smith (Google) ist seitdem sichtbarer (Commits, Dokumentation, Diskussionen, siehe [GitHub](#))

Aktuelle Entwicklungen

- 1985: Start der Entwicklung
- 2007-03-06: Anfang des Git-Repositories
- 2010-10-01: Version 3.00.00
- 2017-06-01: Version 3.05.01
-
- 2016-11-07: LSTM Code wird öffentlich
- 2016-11-08: Version 4.00.00alpha
- 2016-11-28: LSTM Daten für 101 Sprachen
- 2017-07-31: verbesserte („best“) LSTM Daten
- 2017-09-14: „best“ und „fast“ veröffentlicht

Aktueller Stand Sprachmodelle

- 161 neue Sprachmodelle mit deutlich verbesserten Erkennungsraten
- Alle Modelle in 2 Ausprägungen: „best“ und „fast“
- Neue Modelle „Latin“ und „Fraktur“ mit vielen diakritischen Zeichen ergänzen
sprachspezifische Modelle wie „deu“ und „frk“
- Systematische **Fehler** in allen wichtigen Sprachmodellen schränken Nutzen ein

Aktueller Stand Tesseract

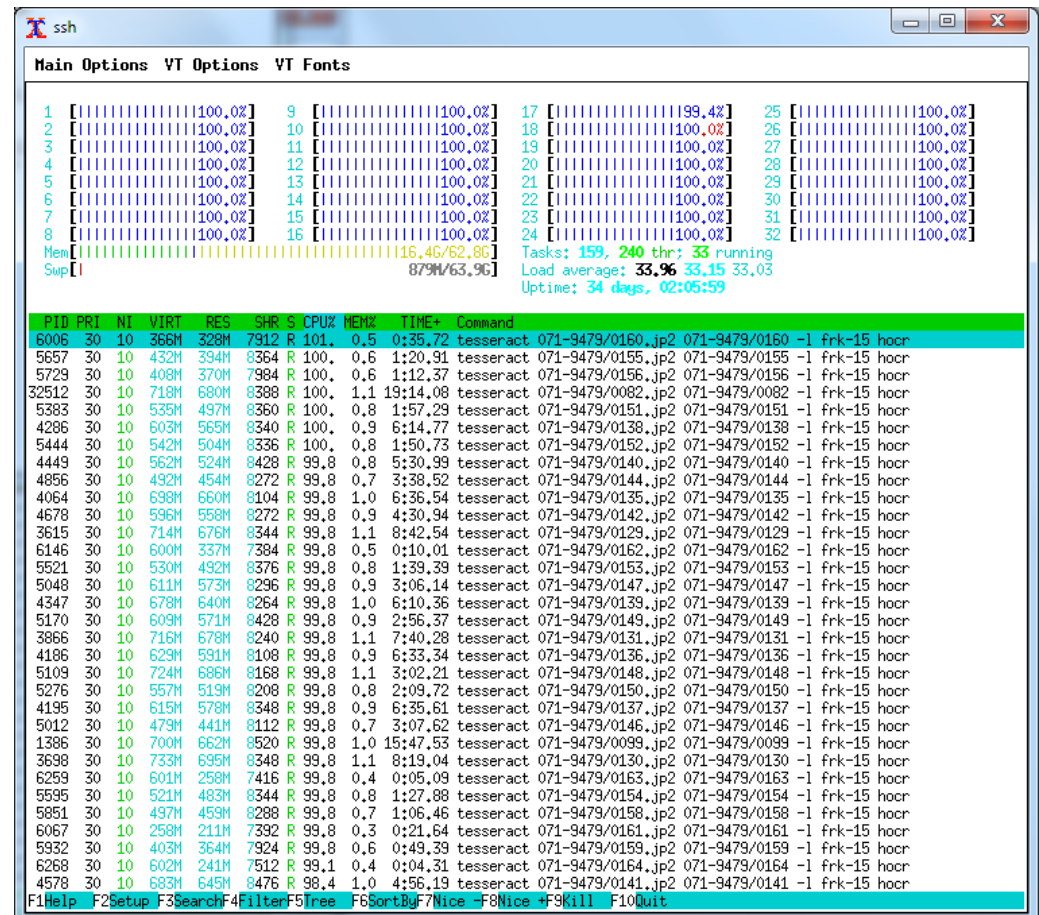
- tesseract : weitgehend stabil, unterstützt bis auf weiteres (Diskussion) auch alte 3.x Sprachmodelle (4.x erkennt keine Textattribute)
- OpenMP: Tesseract 4 nutzt teilweise bis zu 4 CPU Cores (abschaltbar, nicht geeignet für Massen-OCR wegen hohem Overhead)
- OpenCL: Rudimentärer Code für Nutzung von GPU (Grafikkarte), praktisch unbrauchbar, soll eventuell entfernt werden

Aktueller Stand Training

- Synthetische Trainingsdaten (Texte mit unterschiedlichen Fonts gerendert, eventuell künstlich verrauscht)
- Training grundsätzlich im [Wiki](#) dokumentiert
- Bisher keine Details zum Trainingsprozess für ausgelieferte Sprachmodelle bekannt, nur vage Aussagen (> 6000 Fonts, > 100000 Textzeilen)
- Trainingssoftware zeigt tw. noch Instabilitäten

Einsatz an UB Mannheim

- Texterkennung für Staats- und **Reichsanzeiger** (mehr als 700000 Zeitungsseiten) im Juli 2017
- Tesseract 4, aber mit 3.x Technologie und eigenem Sprachmodell **frk-015**



The screenshot shows a terminal window with the following content:

```
ssh
Main Options VT Options VT Fonts

1 [|||||100,0%] 9 [|||||100,0%] 17 [|||||199,4%] 25 [|||||100,0%]
2 [|||||100,0%] 10 [|||||100,0%] 18 [|||||100,0%] 26 [|||||100,0%]
3 [|||||100,0%] 11 [|||||100,0%] 19 [|||||100,0%] 27 [|||||100,0%]
4 [|||||100,0%] 12 [|||||100,0%] 20 [|||||100,0%] 28 [|||||100,0%]
5 [|||||100,0%] 13 [|||||100,0%] 21 [|||||100,0%] 29 [|||||100,0%]
6 [|||||100,0%] 14 [|||||100,0%] 22 [|||||100,0%] 30 [|||||100,0%]
7 [|||||100,0%] 15 [|||||100,0%] 23 [|||||100,0%] 31 [|||||100,0%]
8 [|||||100,0%] 16 [|||||100,0%] 24 [|||||100,0%] 32 [|||||100,0%]

Mem [|||||16,46/62,86] Tasks: 159, 240 thr: 33 running
Sup [|||||879M/63,96] Load average: 33,96 33,15 33,03
Uptime: 34 days, 02:05:59

PID PRI NI VIRT RES SHR S CPUZ MEM% TIME+ Command
6006 30 10 366M 328M 7912 R 101, 0,5 0:35,72 tesseract 071-9479/0160,,jp2 071-9479/0160 -1 frk-15 hocr
5657 30 10 432M 394M 8364 R 100, 0,6 1:20,91 tesseract 071-9479/0155,,jp2 071-9479/0155 -1 frk-15 hocr
5729 30 10 408M 370M 7984 R 100, 0,6 1:12,37 tesseract 071-9479/0156,,jp2 071-9479/0156 -1 frk-15 hocr
32512 30 10 718M 680M 8388 R 100, 1,1 19:14,08 tesseract 071-9479/0082,,jp2 071-9479/0082 -1 frk-15 hocr
5383 30 10 539M 497M 8360 R 100, 0,8 1:57,29 tesseract 071-9479/0151,,jp2 071-9479/0151 -1 frk-15 hocr
4286 30 10 603M 565M 8340 R 100, 0,9 6:14,77 tesseract 071-9479/0138,,jp2 071-9479/0138 -1 frk-15 hocr
5444 30 10 542M 504M 8336 R 100, 0,8 1:50,73 tesseract 071-9479/0152,,jp2 071-9479/0152 -1 frk-15 hocr
4449 30 10 562M 524M 8428 R 99,8 0,8 5:30,99 tesseract 071-9479/0140,,jp2 071-9479/0140 -1 frk-15 hocr
4856 30 10 492M 454M 8272 R 99,8 0,7 3:38,52 tesseract 071-9479/0144,,jp2 071-9479/0144 -1 frk-15 hocr
4064 30 10 698M 660M 8104 R 99,8 1,0 6:36,54 tesseract 071-9479/0135,,jp2 071-9479/0135 -1 frk-15 hocr
4678 30 10 596M 558M 8272 R 99,8 0,9 4:30,94 tesseract 071-9479/0142,,jp2 071-9479/0142 -1 frk-15 hocr
3615 30 10 714M 676M 8344 R 99,8 1,1 8:42,54 tesseract 071-9479/0129,,jp2 071-9479/0129 -1 frk-15 hocr
6146 30 10 600M 337M 7384 R 99,8 0,5 0:10,01 tesseract 071-9479/0162,,jp2 071-9479/0162 -1 frk-15 hocr
5521 30 10 530M 492M 8376 R 99,8 0,8 1:39,39 tesseract 071-9479/0153,,jp2 071-9479/0153 -1 frk-15 hocr
5048 30 10 611M 573M 8296 R 99,8 0,9 3:06,14 tesseract 071-9479/0147,,jp2 071-9479/0147 -1 frk-15 hocr
4347 30 10 678M 640M 8264 R 99,8 1,0 6:10,36 tesseract 071-9479/0139,,jp2 071-9479/0139 -1 frk-15 hocr
5170 30 10 609M 571M 8428 R 99,8 0,9 2:56,37 tesseract 071-9479/0149,,jp2 071-9479/0149 -1 frk-15 hocr
3866 30 10 716M 678M 8240 R 99,8 1,1 7:40,28 tesseract 071-9479/0131,,jp2 071-9479/0131 -1 frk-15 hocr
4186 30 10 629M 591M 8108 R 99,8 0,9 6:33,34 tesseract 071-9479/0136,,jp2 071-9479/0136 -1 frk-15 hocr
5109 30 10 724M 686M 8168 R 99,8 1,1 3:02,21 tesseract 071-9479/0148,,jp2 071-9479/0148 -1 frk-15 hocr
5276 30 10 557M 519M 8208 R 99,8 0,8 2:09,72 tesseract 071-9479/0150,,jp2 071-9479/0150 -1 frk-15 hocr
4195 30 10 615M 578M 8348 R 99,8 0,9 6:35,61 tesseract 071-9479/0137,,jp2 071-9479/0137 -1 frk-15 hocr
5012 30 10 479M 441M 8112 R 99,8 0,7 3:07,62 tesseract 071-9479/0146,,jp2 071-9479/0146 -1 frk-15 hocr
1386 30 10 700M 662M 8520 R 99,8 1,0 15:47,53 tesseract 071-9479/0099,,jp2 071-9479/0099 -1 frk-15 hocr
3698 30 10 733M 695M 8348 R 99,8 1,1 8:19,04 tesseract 071-9479/0130,,jp2 071-9479/0130 -1 frk-15 hocr
6259 30 10 601M 258M 7416 R 99,8 0,4 0:05,09 tesseract 071-9479/0163,,jp2 071-9479/0163 -1 frk-15 hocr
5595 30 10 521M 483M 8344 R 99,8 0,8 1:27,88 tesseract 071-9479/0154,,jp2 071-9479/0154 -1 frk-15 hocr
5851 30 10 497M 459M 8288 R 99,8 0,7 1:06,46 tesseract 071-9479/0158,,jp2 071-9479/0158 -1 frk-15 hocr
6067 30 10 258M 211M 7392 R 99,8 0,3 0:21,64 tesseract 071-9479/0161,,jp2 071-9479/0161 -1 frk-15 hocr
5932 30 10 403M 364M 7924 R 99,8 0,6 0:49,39 tesseract 071-9479/0159,,jp2 071-9479/0159 -1 frk-15 hocr
6268 30 10 602M 241M 7512 R 99,1 0,4 0:04,31 tesseract 071-9479/0164,,jp2 071-9479/0164 -1 frk-15 hocr
4578 30 10 683M 645M 8476 R 98,4 1,0 4:56,19 tesseract 071-9479/0141,,jp2 071-9479/0141 -1 frk-15 hocr

F1Help F2Setup F3Search F4Filter F5Tree F6SortBy F7Nice F8Nice F9Kill F10Quit
```