Mozilla gibt Sprachdatensammlung frei

Originalartikel

Backup

<html> <p class=„printversionback-to-article printversion–hide“>zurück zum Artikel <figure class=„printversionlogo“></figure><figure class=„aufmacherbild“></figure> Die bisher durch das Common-Voice-Projekt zusammengetragenen Sprach-Samples lassen sich ab sofort frei verwenden. Es ist die größte Sammlung dieser Art. Mozilla hat seine Sprachdatensammlung Common Voice [1] öffentlich freigegeben. Mit 1361 Stunden – das entspricht knapp zwei Monaten – transkribierten Audiodaten ist es nach Angaben von Mozilla die größte frei zugängliche Sammlung der Welt. Ebenso wichtig wie die Größe ist Mozilla die Vielseitigkeit der Samples: 42.000 Sprecher wirkten daran mit und sprachen kurze Texte in 18 verschiedenen Sprachen ein. Die Veröffentlichung steht unter CC0-Lizenz – der freizügigsten Variante von Creative Commons („No rights reserved“) – der Öffentlichkeit zur Verfügung. Hauptziel der Sammlung ist es, hochwertige und frei verfügbare Sprachdatensätze zum Training für Spracherkennungssysteme zu schaffen – ein Gebiet, das bisher Cloud-Anwendungen großer Konzerne mit riesigen Sprachdatensammlungen dominieren. Mit DeepSpeech [2] entwickelt Mozilla eine eigene Open-Source-Spracherkennung, die bereits in Produkten wie Mycroft [3] oder Leon [4] eingesetzt oder getestet wird. Das Projekt startete Mitte 2017 [5] mit einer englischsprachigen Textsammlung; ein Jahr später öffnete sich Common Voice für andere Sprachen. Für Englisch hat Mozilla 685 Stunden von fast 36.000 Sprechern aufgezeichnet; Deutsch folgt auf Platz zwei mit 254 Stunden, an denen knapp 4000 Freiwillige mitwirkten. <h3 class=„subheading“ id=„nav_kabylisch0“>Kabylisch, Tatarisch, Walisisch</h3> Während sich kommerzielle Anbieter auf die Sprachen der wichtigsten Märkte konzentrieren, finden sich bei Common Voice auch viele, die sonst kaum im Internet repräsentiert sind, etwa Kabylisch (eine algerische Berbersprache), Tatarisch oder Walisisch.

Hier treiben oft wenige Enthusiasten das Projekt voran. Neuerdings kooperiert Mozilla mit der Deutschen Gesellschaft für Internationale Zusammenarbeit [6], um zum Beispiel Sprecher in dem afrikanischen Land Ruanda zu erreichen. Einige der großen Weltsprachen hinken dagegen noch hinterher, etwa Spanisch, Arabisch oder Russisch. Seit der Release finalisiert wurde, wuchs die Zahl der Sprachen in der Aufnahmephase auf 22 an; fast 200 Stunden Aufzeichnungen kamen hinzu. Bei 70 weiteren Sprachen läuft die Vorbereitungsphase, in der die Freiwilligen Sätze sammeln und die Website übersetzen. Auch wenn Deutsch in Common Voice gut vertreten ist, sucht das Projekt weiterhin Sprecher – das erklärte Ziel ist, für jede Sprache 1200 Stunden Material zu sammeln. Die Mitwirkung erfordert keine besonderen Kenntnisse und dauert nur wenige Minuten. Siehe dazu auch c't 18/2018: <ul class=„rtelist rtelist–unordered“>
Spracherkennung für alle: Mozillas Projekte Common Voice und DeepSpeech [7] ()<br class=„clear“/> <hr/> URL dieses Artikels:
<small><code>http://www.heise.de/-4323042</code></small> Links in diesem Artikel:
<small><code>[1] https://voice.mozilla.org/</code></small>
<small><code>[2] https://www.heise.de/meldung/Mozilla-Common-Voice-Sprachsteuerung-fuer-alle-und-ohne-Rueckgriff-auf-die-Cloud-3904454.html</code></small>
<small><code>[3] https://mycroft.ai/blog/deepspeech-update/</code></small>
<small><code>[4] https://getleon.ai/</code></small>
<small><code>[5] https://www.heise.de/meldung/Mozilla-sammelt-Stimmaufzeichnungen-fuer-offene-Spracherkennungs-Software-3780795.html</code></small>
<small><code>[6] https://www.giz.de/en/html/</code></small>
<small><code>[7] https://www.heise.de/select/ct/2018/18/1535420071631846</code></small>
<small><code>[8] mailto:dbe@ct.de</code></small>
 Copyright © 2019 Heise Medien </html>