Установка и настройка Tesseract на Debian

Все команды делаются под пользователем «root»

# Загрузка исходных файлов Tesseract

Установка Git:

apt-get install git-core

Установка шрифтов Windows:

apt-get install ttf-mscorefonts-installer

Просмотр шрифтов:

fc-list

Предустанки:

$ apt-get install autoconf automake libtool

$ apt-get install libpng12-dev

$ apt-get install libjpeg62-dev

$ apt-get install libtiff4-dev

$ apt-get install zlib1g-dev

$ # (if you plan to make the training tools)

$ apt-get install libicu-dev

$ apt-get install libpango1.0-dev

$ apt-get install libcairo2-dev

Скачать Leptonica:

wget http://www.leptonica.org/source/leptonica-1.71.tar.gz

Установка Leptonica:

$ tar xzf leptonica-1.71.tar.gz

$ cd leptonica-1.71

$ ./configure

$ make

$ make install

Загрузка исходного кода Tesseract:

git clone https://code.google.com/p/tesseract-ocr/

Компиляция и настройка Tesseract:

./autogen.sh

./configure

make

make install

ldconfig

Установка инструмента обучения:

make training

sudo make training-install

Настройка пути:

export TESSDATA\_PREFIX=/usr/local/share/tessdata

Создания рабочей папки:

mkdir / opt/synergy/jboss/standalone/ocr-dir/work

Скопируйте туда скрипты из папки дистрибутива ocr-dir/work/convertToTiff.sh, execConcatFiles.sh, execTesseract.sh. Дайте им права 755 и владельцы папки на jboss:synergy

Добавьте право jboss на выполнение tesseract под sudo

visudo

Cmnd\_Alias TESSERACTEXEC = /usr/local/bin/tesseract

ZEBRAUSERS ALL = NOPASSWD: TESSERACTEXEC

Это просто… для растаривания:

tar xvzf file-1.0.tar.gz

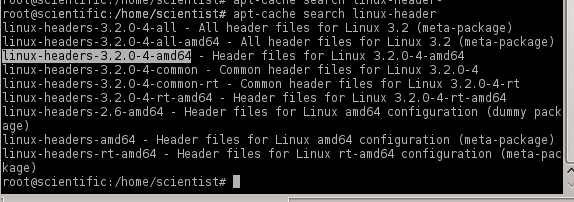
Обновление gcc:

#i have installed Debian wheezy CD1. And i don't see gcc and make tools already installed . So gonna install:

apt-get install gcc make

#now installing linux-headers-3.2.0-4-amd64

apt-cache search linux-header



#now installing linux-headers-3.2.0-4-amd64

apt-get install linux-headers-3.2.0-4-amd64

Установка poppler-utils:

apt-get install poppler-utils

Установка pdftk:

apt-get install pdftk

Скачать «tesseract-ocr-3.01.osd.tar.gz», «tesseract-ocr-3.02.eng.tar.gz», «tesseract-ocr-3.02.rus.tar.gz» по ссылке https://code.google.com/p/tesseract-ocr/downloads/list и разпоковать traindata в папку «tessdata»

Информация по установке шрифтов:

<http://www.cyberciti.biz/faq/howto-debian-install-use-ms-windows-truetype-fonts-under-xorg/>

Скачка Leptonica:

<http://www.leptonica.org/download.html>

Компиляция Tesseract:

<https://code.google.com/p/tesseract-ocr/wiki/Compiling>

<http://blog-en.openalfa.com/how-to-install-tesseract-ocr-in-debian>

Обучение Tesseract:

<https://code.google.com/p/tesseract-ocr/wiki/TrainingTesseract3>

Установка Java:

<http://stackoverflow.com/questions/15543603/installing-java-7-oracle-in-debian-via-apt-get>

Обновление gcc (установка VMTools)::

<http://blog.pusheax.com/2013/02/installing-vmware-tools-on-debian.html>

**Дополнительно:**

apt-get install alien

Скачать фонты: <http://avi.alkalay.net/software/webcore-fonts/>

Инструкция: <http://forum.lafox.net/index.php?showtopic=14313>

Преобразование: alien webcore-fonts-3.0-1.noarch.rpm

Установка: dpkg -i webcore-fonts\_3.0-2\_all.deb