11-711: Algorithms for NLP

Homework Assignment #0: Installing the Tools You'll Need

Out: August 26, 2014
Due: September 2, 2014
(No materials actually need to be turned-in)

1 Introduction

These instructions were developed on Linux and are also known to work on Mac OSX. During this first week, the TAs will be happy to assist you in installing these tools. After that, we'll assume that you were able to install them without any issues.

Windows users: If you intend to do NLP research, it's probably time to either get access to a Linux machine, or install a dual boot partition of Ubuntu Linux on your machine (see https://help.ubuntu.com/community/WindowsDualBoot). SCS also provides access to some general purpose Linux machines, which are *not* to be used for running CPU intensive tasks. You can find them at http://www.cs.cmu.edu/~help/accounts_passwords/ux_account.html and http://www.cs.cmu.edu/~help/windows/using.putty.html. However, we do not guarantee that the following software will install there, nor can we provide any form of support for these machines. Another option is running a virtual machine inside Windows using the free VirtualBox software at http://www.virtualbox.org; this will require significantly more system resources and, as some have reported that VirtualBox is not entirely stable, we will not be able to provide support for it within the scope of this class. We recommend an Ubuntu Linux dual boot for maximum compatibility with software you will need to use in this course and for NLP research in general.

Mac OS users: All of the tools required for this course will also run on Mac OS. However, many tools distributed by Apple can be much older or subtly different from their Linux counterparts, leading to additional installation steps to make software run properly. For maximum compatibility, we still recommend using Ubuntu Linux. Instructions for dual booting Ubuntu on an Apple machine can be found here: https://help.ubuntu.com/community/MactelSupportTeam/AppleIntelInstallation.

1.1 Typesetting Requirement

You will be required to typeset all assignments in this class. Handwritten assignments will not be accepted. Diagrams can be hand-drawn but must be clearly legible to receive credit. We highly recommend the LaTeX language for typesetting your assignments. LaTeX is also highly useful for typesetting your future research papers.

2 Installing LaTeX

To fulfill the typesetting requirement of this course, we recommend using LaTeX, a document typesetting language. While we don't require you to use LaTeX, most WYSIWYG word processors are not equipped to typeset mathematical text in a reasonable way.

Linux users: For Ubuntu and other Debian-based distributions, you can install LaTeX with the following command:

```
sudo apt-get install texlive-full
```

This can take awhile, but will install everything you need. Emacs, Vim, and many other common Linux text editors offer syntax highlighting out-of-the-box. For a a more graphical environment, we recommend gedit with the LaTeX plugin that includes code completion and live document preview:

```
sudo apt-get install gedit-latex-plugin
```

Mac OS users: To install LaTeX on Mac OS, visit the TeXShop page (http://pages.uoregon.edu/koch/texshop/obtaining.html) and follow the instructions to download and install the MacTeX package (MacTeX.mpkg.zip). As noted on the page, this is a large download.

For editing LaTeX code, we recommend Aquamacs (http://aquamacs.org/), a graphical version of Emacs with very good Mac OS integration. The Aquamacs distribution includes several tools for editing and compiling LaTeX, outlined here: http://aquamacs.org/latex.shtml.

Helpful LaTeX Resources

Getting to Grips with LaTeX: This is a series of short, well-written tutorials covering mathematics, tables, figures, and just about everything else you'll need to know to typeset your assignments: http://www.andy-roberts.net/writing/latex.

LaTeX Wikibook: This wikibook is more comprehensive than the tutorials but it also takes longer to sift through. If you want to use something that isn't covered in the above tutorials, you can check the corresponding section of the wikibook: http://en.wikibooks.org/wiki/LaTeX.

Detexify^2 - LaTeX symbol classifier: Find the LaTeX code for any symbol by drawing it on a nifty Web 2.0 canvas: http://detexify.kirelabs.org/classify.html. If a symbol doesn't show up on Detexify, you can check the comprehensive LaTeX symbol list (pdf): http://mirror.ctan.org/info/symbols/comprehensive/symbols-a4.pdf.

3 Installing OpenFST

First, download OpenFST, unpack it, compile it, and install it to /home/$USER/prefix. We'll assume you don't have root access and that your home directory is in the environment variable $HOME. If you have issues, see http://www.openfst.org/twiki/bin/view/FST/FstDownload and http://www.openfst.org/twiki/bin/view/FST/DistInstall. If you're running OSX 10.4 or 10.5, you have a buggy version of tr1/hashtable and will need to follow these instructions http://openfst.cs.nyu.edu/twiki/bin/view/FST/CompilingOnMacOSX.

```
# Change to a scratch directory, such as your home directory
mkdir $HOME/prefix
wget http://www.openfst.org/twiki/pub/FST/FstDownload/openfst-1.3.2.tar.gz
tar -xvzf openfst-1.3.2.tar.gz
cd openfst-1.3.2
./configure --prefix=$HOME/prefix
make -j4     # Expect 5-20 minutes of compile time
make install
```

Now, since we didn't install these into the root-only system directory, we'll put the libraries and binary on your search path. You'll have to either run this once per login or add this to your login script such as /home/$USER/.bashrc

```
export LD_LIBRARY_PATH=$HOME/prefix/lib:$LD_LIBRARY_PATH
export PATH=$PATH:$HOME/prefix/bin
```

Additional resources for OpenFST are available at http://www.openfst.org/twiki/bin/view/FST/FstQuickTour and http://www.openfst.org.

4 Installing Python

Some of the scripts provided in hands-on coding assignments depend on Python. You'll need at least version 2.5, and we recommend version 2.7. We also recommend Python (but of course do not require it) as a scripting language for aiding you in assignments as it is typically very easy to learn. Other scripting languages such as Perl, Scala, Ruby, etc. will also work just fine. There's of course nothing stopping you from using heavy weight languages such as C++ and Java, but since we will be giving you (relatively) small data sets, you shouldn't find it necessary in this class to spend the extra coding effort on these languages.

Newer Linux distributions and versions of OSX may already have Python 2.7. To check, run `python --version`.

- For Mac OSX 10.6 and up, you can use the GUI installer from https://www.python.org/download/releases/2.7.8/.
- For Mac OSX 10.3 or 10.5, you can use the GUI installer from https://www.python.org/ftp/python/2.7.8/python-2.7.8-macosx10.3.dmg or https://www.python.org/ftp/python/2.7.8/python-2.7.8-macosx10.5.dmg, respectively.

5 Installing virtualenv

There are several python modules that will be of use in exploring the course material. To ease the process of installing python modules, we recommend you install virtualenv, particularly if you are working on a machine where you can't `sudo`. To install virtualenv, if you have root access you can simply use: `sudo apt-get install python-virtualenv`. If you don't have root access, you can run:

```
wget https://pypi.python.org/packages/source/v/virtualenv/virtualenv-1.9.tar.gz
tar xzvf virtualenv-1.9.tar.gz
cd virtualenv-1.9
python virtualenv.py myVE
```

6 Installing GraphViz

GraphViz is used by OpenFST to visualize FST's.

For MacOSX, you can just use the GUI installer here: http://www.graphviz.org/pub/graphviz/stable/macos/leopard/graphviz-2.26.3.pkg (I don't recommend V2.28.0 – the installation failed on my machine).

On Ubuntu Linux (or other Debian-based distributions), you can simply use: `sudo apt-get install graphviz` if you have root access. If you run a different distribution or you don't have root access, then you'll need to follow the directions below.

```
wget http://www.graphviz.org/pub/graphviz/CURRENT/graphviz-working.tar.gz
tar -xvzf graphviz-working.tar.gz
cd graphviz-2.29.20110817.0445
./configure --prefix=$HOME/prefix
make
make install
```

Like OpenFST, since we installed this in the $HOME/prefix directory, you will need to keep the prefix/bin directory in your environment, as discussed above.

7 Getting to Know Shell Scripting

You will also likely find it helpful to write some basic shell scripts to automate and reproduce the steps you perform to complete this homework (and most NLP research tasks, for that matter). The bash shell is default on most Linux distributions and OSX. If you're new to shell scripting, consider reading http://tldp.org/LDP/Bash-Beginners-Guide/html/Bash-Beginners-Guide.html or for more advanced usage http://tldp.org/LDP/abs/html/abs-guide.html.