INTERLUDE

Creating a healthy Python Data Science Environment

INSTALL BASE PACKAGES

```
# Setup the Python environment
#
# Install matplotlib (pip install matplotlib will fail)
# and the Python dev package (the headers are required by numpy and friends)
# Make sure the python versions match the one installed
# libzmq is required by pyzmq (or its install will fail)
#
echo "[INFO] Installing Python 3, virtual environment support and SciPy packages"
read -p "Press a key when ready or Ctrl-C to abort"

sudo apt-get install -y python-setuptools python3-matplotlib python3.4-dev libzmq-dev
sudo easy_install pip
sudo pip install virtualenvwrapper
if [[ $? != 0 ]]; then
    echo "[ERROR] Could not install virtualenvwrapper, please check pip logs"
    exit 1
fi

mkdir -p "${VENVS}"

# Add the virtualenvwrapper env for .bashrc
#
echo "export WORKON_HOME=\"${VENVS}\"
source /usr/local/bin/virtualenvwrapper.sh" >> .bashrc
source .bashrc
```

source: https://gist.github.com/massenz/5d392763d21f980a4f2f

BUILD THE VIRTUAL ENVIRONMENT

```
# Create the virtual environment
# The use of --system-site-packages is necessary to make matplotlib work in Python 3
#
if [[ ! -d "${SCIPY_DIR}" ]]; then
    mkvirtualenv -p `which python3` --system-site-packages ${SCIPY_VENV}
    if [[ $? != 0 ]]; then
        echo "[ERROR] Could not create a virtualenv, please check the error message, if any"
        exit 1
    fi
    workon ${SCIPY_VENV}
fi

# Add all the necessary packages (see the SparkLab-requirements.txt file)
# NOTE - this will take a long time to run, with no output to stdout
if [[ ! -f SparkLab-requirements.txt ]]; then
    echo "[ERROR] Missing SparkLab-requirements.txt file, please copy it to the ubuntu user home dir"
    exit 1
fi
echo "[INFO] Installing SciPy packages, this will take forever: go grab a book..."
read -p "Press a key when ready or Ctrl-C to abort"
pip install -r SparkLab-requirements.txt
if [[ $? != 0 ]]; then
    echo "[ERROR] Your virtual environment may miss critical packages"
fi
read -p "Press a key when ready or Ctrl-C to abort"
```

yes, you do want to use virtual environments when messing around with Python…

source: https://gist.github.com/massenz/5d392763d21f980a4f2f

RECOMMENDED PACKAGES

```
# SparkLab Python dev requirements

Jinja2==2.7.3
cassandra-driver==2.1.3
futures==2.2.0
ipython==2.3.1
nose==1.3.4
numpy==1.9.1
pandas==0.15.2
pymongo==2.7.2
pyparsing==2.0.3
python-dateutil==2.4.0
pytz==2014.10
pyzmq==14.5.0
six==1.9.0
tornado==4.0.2
```

```
# Other stuff you may find useful…

bpython
mocks
unittests
```

source: https://gist.github.com/massenz/5d392763d21f980a4f2f

WRITE HEALTHY PYTHON

- Build for Python 3
  https://docs.python.org/3/howto/pyporting.html

- Use Virtual Environments (and `virtualenvwrapper`)

- Test! Test! Test!
  (`unittests nose mocks`)

- Write readable, consistent code
  PEP8 (https://www.python.org/dev/peps/pep-0008/) & `pylint`

SUGGESTED READING

- Beazley, Python Essential Reference, Addison Wesley, 4th ed.
- Python Weekly, http://www.pythonweekly.com
- RTFM :) https://docs.python.org/3/
- McKinney, Python for Data Analysis, O'Reilly
- boto reference, https://boto.readthedocs.org/en/latest/

None of this is required for this course, though!