Newsreader virtual machines

Technical Report NWR-2014-4

Version FINAL

Aitor Soroa[1], Enrique Fernández[2]

[1]University of Basque Country
Donostia, Basque Country
a.soroa@ehu.es
[2]University of Basque Country
Donostia, Basque Country
kike.fernandez@ehu.es

Building structured event indexes of large volumes of financial and economic data for decision making

ICT 316404

Contents

List of Tables

1 Introduction

Linguistic processors are complex software packages which often require a large set of dependencies to be met in order to effectively perform their tasks. Deploying LPs often requires pre-installing a large set of common software modules on the same machine, which must be accessible to the LP. The capacity of replicate the results is very important within the project. One LP module applied to a particular input text has to produce the same output regardless the software framework (machine, operating system, etc.) where it is installed. Therefore, special care has to be taken on guaranteeing that the same version of the LP modules, along with the exact same dependencies, are deployed among machines.

The aforementioned reasons have lead us to use virtual machine (VM) technologies for deploying the LP modules. Virtualization is a widespread practice that increases the server utilization and addresses the variety of dependencies and installation requirements. Besides, virtualization is a 'de-facto' standard on cloud computing solutions, which offer the possibility of installing many copies of the virtual machines on commodity servers.

This document covers several aspects for running the VM defined within the Newsreader project, as well as for installing NLP modules inside the VM.

2 Files

You need two files for running the VM.

`https://siuc05.si.ehu.es/~sisfetek/datuak_deskargatzeko/centos64newsreader.img`
`https://siuc05.si.ehu.es/~sisfetek/datuak_deskargatzeko/newsreader-vm.xml`

The first is a big file with an "empty" VM with Centos 6.4 Linux operating system and one NLP module installed. The second is an short XML document needed for running the VM.

3 Prerequisites

The are some pre-requisites the host machine has to fulfil for running the VM inside it:

- Linux operating system (any recent flavor would do it)
- In-kernel KVM virtualization capabilities. You can also determine if your system processor supports KVM by running the following command:

```
% grep -E 'vmx|svm' /proc/cpuinfo
```

if this command returns output, then your system supports KVM. You also have to verify that the KVM-related feature is enabled in the machine's BIOS.

- 64 bit CPU ($x86_{64}$)

4 Running the VM

For running the VM, follow these steps:

4.1 Preliminary steps (do it once):

1. Install the necessary software into the host machine. On Deban/Ubuntu machines this includes the following packages:
   - `qemu-kvm`
   - `libvirt-bin`
   - `virt-manager`
2. Download `centos64newsreader.img` and `newsreader-vm.xml` into the host system.
3. Make sure that the path to the `centos64newsreader.img` file is accessible/readable to the user `qemu`
4. Make a copy of the XML doc and rename it to a proper name. For the sake of this document, the XML doc name will be `newsreader-EHU.xml`
5. Tweak the the XML file:
   (a) Put a proper name into the `<name>` element (line 3). For example, "newsreader-EHU".
   (b) Create a new UUID using 'uuidgen' program and paste it lo line 5 (into `<uuid>` element)
   (c) Put the absolute path to the IMG file in line 27 in the "file" attribute of `<source>` element.
   (d) Currently the VM is configured to use 8Gb RAM. You can change this value by editing around line 6 (`<memory>` and `<currentMemory>` elements).

6. If you experience problems running the VM, maybe you need to change line 23 and put the name of the kvm emulator executable in your system.

7. Alternatively, you can create a bare new VM using the IMG image. Use the `virt-manager` tool for this. The following link maybe useful:

`https://docs.google.com/document/d/1exv1X3zmtGT6lZihlKW9T-EXKLzj_ODOWmndMe4I-c8/edit?usp=sharing`

4.2 Accessing the VM

From the host machine, cd to where the IMG and XML documents are and run the following:

```
% virsh create newsreader-EHU.xml
Domain newsreader-EHU created from newsreader-EHU.xml
```

The machine should be running now. You can test this using the `virsh list` command:

```
% virsh list
 Id    Name                           State
----------------------------------------------------
 17    newsreader-EHU                 running
```

Note: The VM needs around 3/4 minutes to completely load all the modules and services, so please be patient until the login screen appears.

Now you can connect to the VM. There are several ways to run it:

4.2.1 Connecting from the console

Being on the host computer you can connect from the console using this command:

```
% virsh console newsreader-EHU
Connected to domain newsreader-EHU
Escape character is ^]
```

You can exit from the VM at any time by pressing the ^] key (Ctrl + ']').

4.2.2 Connecting from VNC

Start VNC in the host machine and connect to "localhost" using the "VNC" protocol. It will automatically show the console of the VM. Alternatively, run the `virt-manager` program, and double click into the running VM. It will open a console window.

4.2.3 Connecting via ssh

The VM is configured to get an IP address using DCHP. The VM receives a local IP address `192.168.122.X` from inside the host machine. To exactly know which local IP it has, you have to first access the VM via console or VNC. Once inside, you can get the IP address of the VM for instance running the following command:

```
$ ifconfig eth0
eth0      Link encap:Ethernet  HWaddr 52:54:00:8E:CD:B1
          inet addr:192.168.122.98  Bcast:192.168.123.255  Mask:255.255.252.0
          inet6 addr: fe80::5054:ff:fe8e:cdb1/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:568 errors:0 dropped:0 overruns:0 frame:0
          TX packets:253 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:49026 (47.8 KiB)  TX bytes:83040 (81.0 KiB)
```

In this case, the IP is `192.168.122.98` (look at the `inet addr` section).

Once we know which local IP the VM has, and being on the host computer, just ssh to the VM. In the example above, just type:

```
% ssh  newsreader@192.168.122.98
```

If you want to access the VM guest outside the host machine, perhaps the best way is to use a bridged network configuration (not explained here). Alternatively, you can use iptables for allowing external access through ssh to the VM. For example, you can access guest's ssh port using host's 2222:

```
iptables -t nat -A PREROUTING -p tcp --dport 2222 -j DNAT \\
--to-destination 192.168.122.98:22
iptables -t nat -A POSTROUTING -p tcp --dport 22 \\
-d 192.168.122.98 -j SNAT --to 192.168.122.1
iptables -D FORWARD 5 -t filter
iptables -D FORWARD 4 -t filter
```

In any case, consult with your IT staff to perform the above steps, as there are many alternatives.

4.2.4 Changing IP

You can set an static IP for the VM by editing `/etc/sysconfig/network-scripts/ifcfg-eth0` file inside the VM. For example, this lines would assign local IP `192.168.122.99`:

```
DEVICE=eth0
HWADDR=52:54:00:8e:cd:b1
TYPE=Ethernet
ONBOOT=yes
NM_CONTROLLED=no
BOOTPROTO=none
IPADDR=192.168.122.99
NETMASK=255.255.252.0
GATEWAY=192.168.122.1
DEFROUTE=yes
```

and then restarting network service:

```
$ /etc/init.d/network restart
```

5 Shut down the VM

Logout from VM user and then, from the host computer:

```
% virsh destroy newsreader-ixa
```

Alternatively, and being on the VM, run the following command:

```
$ sudo shutdown -h now
```

6 User and password

The VM has one user:

```
login: newsreader
pwd: readNEWS89
```

the `root` user has the same password as `newsreader`. You can run root commands within the newsreader user using the `sudo` command.

7 Directory structure

All the NLP modules and document directories are under the `/home/newsreader` directory. This directory structure is as follows:

`~/components`

Here lay the actual NLP modules.

`~/opt`

The dependencies of the modules should be installed under this directory (not system-wide). The idea is that we can synchronize the `~/components` and `~/opt` directories when a module is updated or a new module is deployed.

`~/docs`

The documents are stored here. Initially, input documents are placed in `~docs/input`. When the document is successfully processed, the compressed output NAF is placed into the `~docs/output` directory and removed from `~docs/input`. Alternatively, if the document can not be processed, it is moved to the `~docs/error` directory (and removed from the original place).

8 Using the NLP pipeline

This section describes how to actually use the pipeline. Documents are uploaded to the VM from outside, typically from the host machine. In th examples we use `IP` and `PORT` for specifying the VM IP address and the port of the service. See section on `Changing IP` above to know which IP the VM has. `PORT` will be usually 80.

8.1 Sending documents to VM

Use the `curl` command to send documents to the processing pipeline. For sending the document `doc.txt` to the virtual machine (with `IP` and `PORT`), use the following command:

```
% curl --form "file=@doc.txt" http://IP:PORT/cm_upload_text_file.php
```

8.2 Getting output documents

As said above, the documents uploaded to the VM are stored in the `~docs/input` directory. Once the document is processed, and if there is no error, the output NAF will be put in `~docs/output`. The name of the output NAF will be:

```
~docs/output/name.extension_MD5.naf.bz2
```

In the example above, the `doc.txt` document could get a name like:

```
~docs/output/doc.txt_8b45b51a553d702777bc627f262ea091.naf.bz2
```

Note that the output documents are compressed using the `bzip2` program.

8.3 Status of NLP processing

The VM has a service for knowing its internal status. Running this command:

```
% curl IP:PORT/cm_sysinfo.php
```

it gives information about the VM status:

```
VM: newsreader-EHU
Uptime:  15:25:58 up 7 days,  3:16,  7 users,  load average: 2.41, 1.58, 0.91
Free Memory:          1201708 kB
Free Disk: 12940.1640625 MB

Pipeline processing status:
Pending files-> 5
Finished files-> 3
Failed files-> 2
```

9 Deploying NLP modules

This Section explains how to deploy new NLP modules into the VM. All the modules should be installed under the `newsreader` user. The directory structure is as follows:

- Install the modules under the `~/components` directory, creating a subdirectory as appropriate (for instance, `~/components/EHU-ukb`).
- If the modules have dependencies, install the dependencies into the `~/opt` directory. If this is not an option, please let us know.

- There has to be a `run.sh` script inside each module which reads input from `STDIN`, runs the module, and write the output (NAF) to `STDOUT`. This script has no parameters.
- The `run.sh` script has to be callable (and will be called) from outside the directory where the module is. So make sure the `run.sh` script uses absolute paths or `cd`'s into the component's directory first.
- Please create an `INSTALL` document inside the module clearly specifying which steps are needed for deploying the module (how to install dependencies, etc).

You will find an example of a deployed module in `~components/EHU-ukb`.

10 Updating NLP modules

Modules are updated on the "master" VM at EHU. The address of the master VM is `u017940.si.ehu.es`, and the ssh port is 2223. Thus, the way to connect to the master VM is:

```
ssh -p 2223 newsreader@u017940.si.ehu.es
```

Once the modules are updated on the master VM, each VM copy can synchronize and update the modules by running the following steps:

1. Stop the running topology[1]:

   ```
   $ /home/newsreader/storm/cluster/storm-0.8.2/bin/storm kill -w 0 \\
   Newsreader-Pipeline
   ```

2. Run the update script:

   ```
   $ /home/newsreader/update_nlp_components.sh
   ```

   this script will connect to the master VM (asks for the usual password), and update all components.

3. Install the new topology (which excludes multiword and includes entity coreference). The topology jar is inside the "opt" directory. Run the following command (again a long line)

   ```
   $ /home/newsreader/storm/cluster/storm-0.8.2/bin/storm \\
   jar /home/newsreader/opt/topologies/Newsreader-Pipeline-Storm.jar \\
   TopologyMainProduction
   ```

[1] The folloeing command is one long line. The trailing \\ in the end of the line indicates that the command continues on the next line.

11 If something goes wrong

If the processing stalls or there is any other kind of problems, the easiest way to proceed is to just reboot the VM. From a shell command inside the VM, just run

```
$ sudo shutdown -r now
```

and the system will reboot. When the machines starts again, it will scan the input doc directory and start processing the documents present there. Remember that the machine needs 3/4 minutes to boot and launch all the services and daemons.

If you have any question, please do not hesitate to contact us:

Kike Fernandez <kike.fernandez@ehu.es>
Aitor Soroa <a.soroa@ehu.es>