Use Case: Perform Text Analytics Using Topic Explorer

Use the IPython interactive interface to fetch volume content, and then run vector space model and topic modeling on volumes' OCR content. It uses the in pho/vsm python package, a textual semantics package developed by Dr. Colin Allen and his team locally at IU.

This use case obtains some HTRC volume content, builds topic models based on the content, and then visualizes the topic models in a web browser.

VM Mode

This use case can be run in only ***secure*** mode in the VM. To export experiment results out of the VM, you need to release the result files in secure mode, and then receive results via email.

Example Use

First, switch the VM mode to ***secure*** mode (done in the HTRC portal).

In the VM, start a Terminal, and change directory to the htrc-data folder

```
cd /home/dcuser/HTRC-Demos/Python/topicexplorer-demo
```

List the files of this folder

```
ls
```

Following are the files related to this analysis.

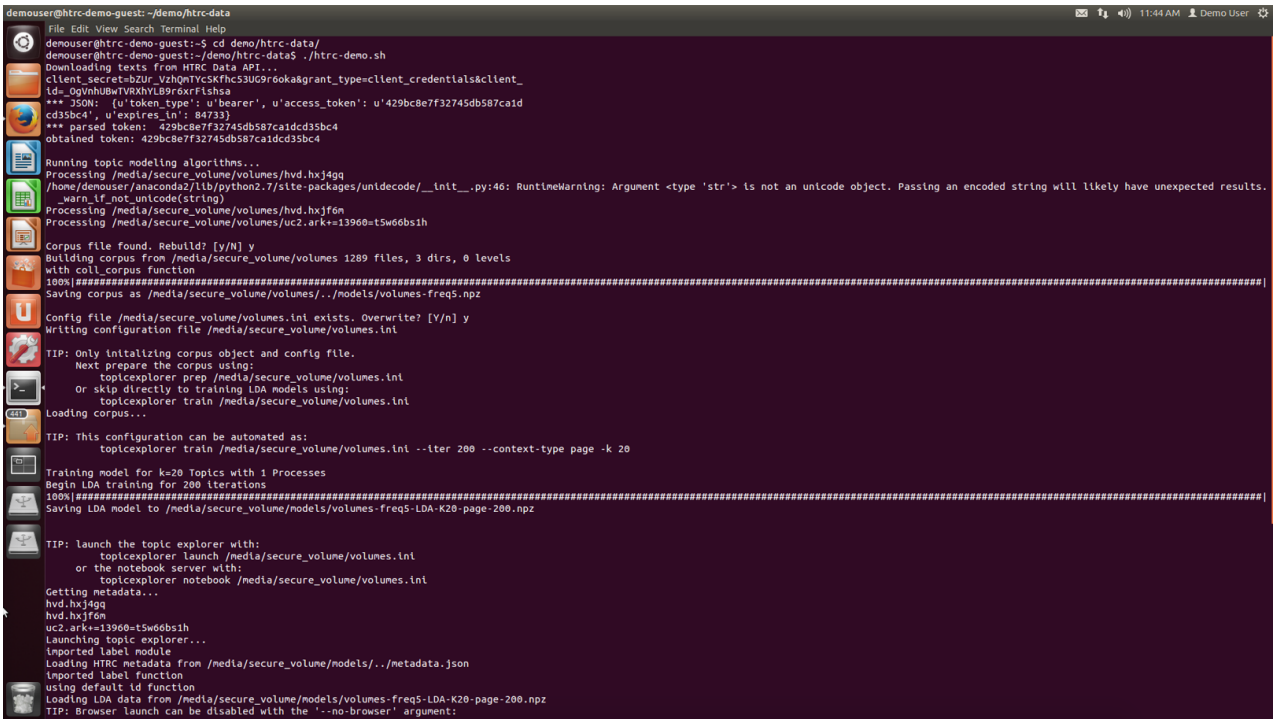
- htrc-demo.sh - This is the script for topic modeling analysis.
- htrc-id - This file contains the list of volume ids.

Run the topic modeling analysis

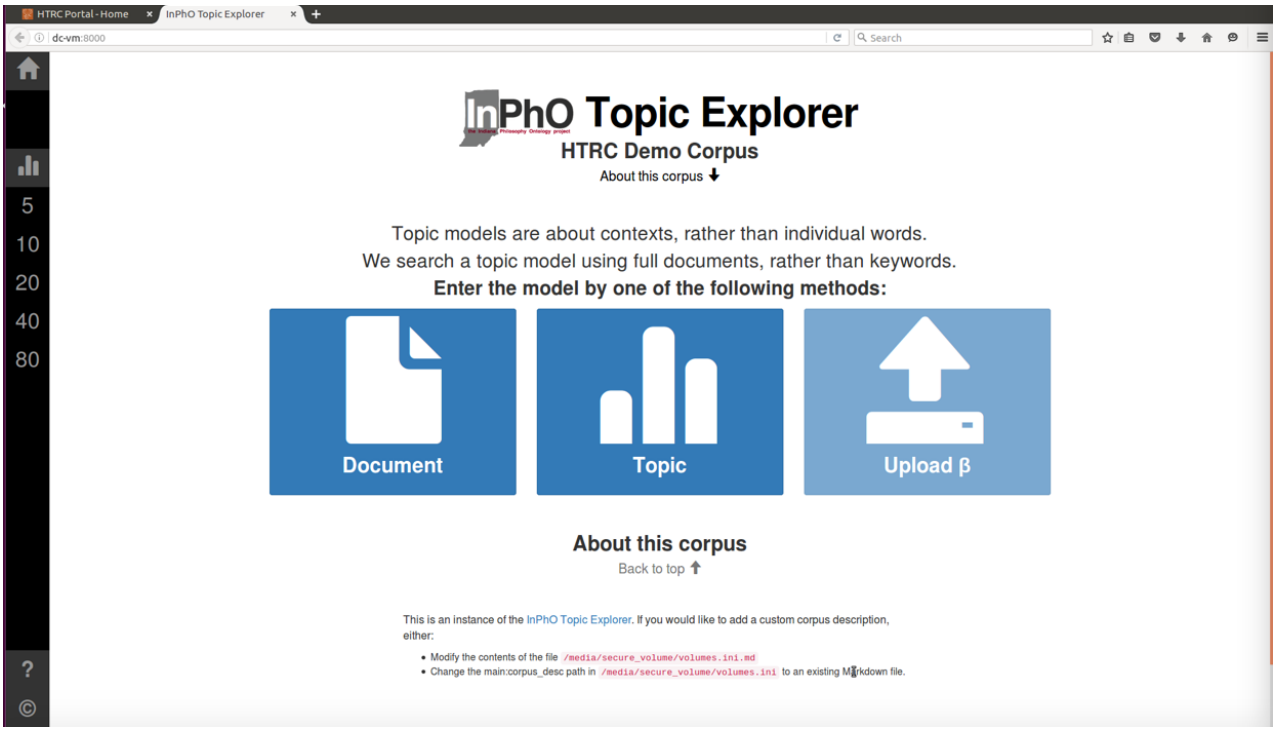
> ⚠ Before running the topic modeling analysis, please check the script whether the 'secure_volume' path is mentioned correctly. Correct path should be '/media/secure_volume'

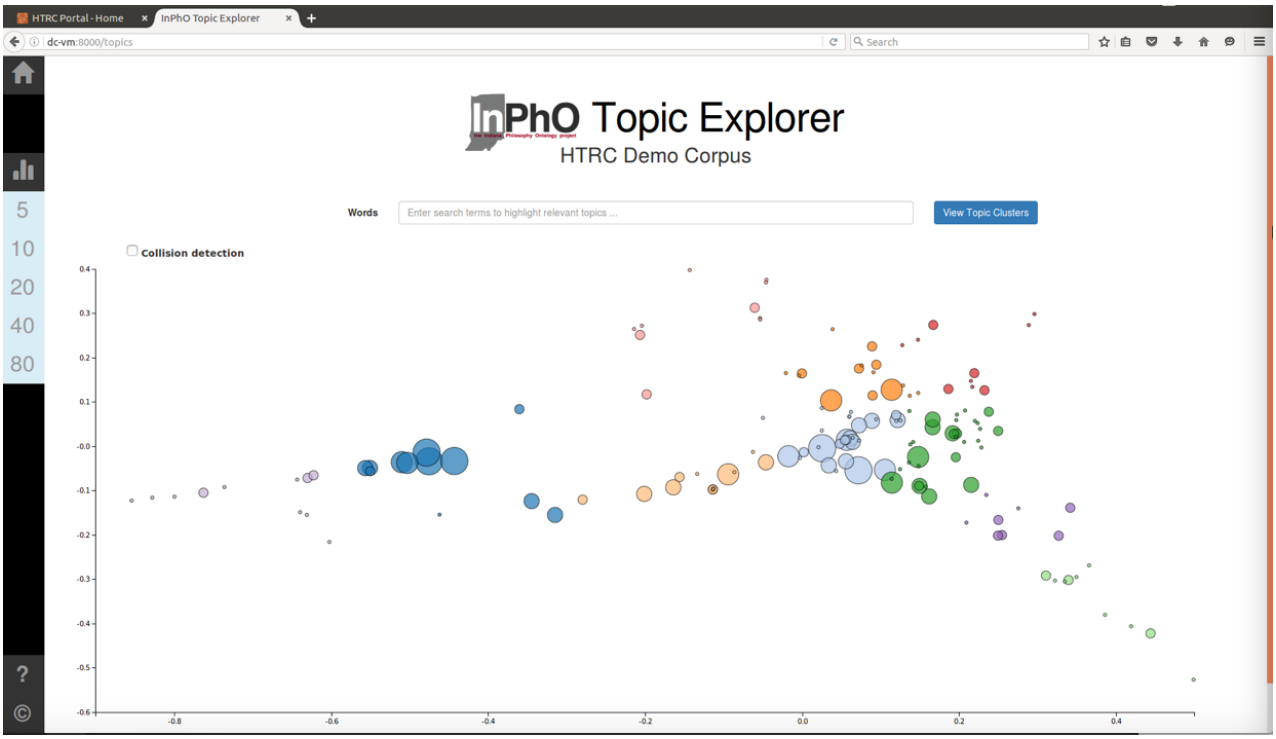
```
./htrc-demo.sh
```

You will see something like this in the console. This means the program is building topic models on the volume content.



It will take quite a while to finish the topic modeling due to the nature of this kind of computation. After the topic modeling process is done, you can view the result through the browser. (The browser will be automatically opened for you). Click on the "Topic" button.





You will find the scripts run into errors if the VM is in *maintenance* mode. It is because this use case fetches HTRC content by using the Data API, which is only accessible in the *secure* mode.

This demo code:

- loads data from 3 volumes in HathiTrust using the HTRC Data API
- builds an LDA topic model from the corpus
- save the LDA trained model
- view topics in a web browser in an interactive way

Here are the scripts used in this example: topicexplorer-demo.zip