

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.

```

demouser@htrc-demo-guest:~/demo/htrc-data
File Edit View Search Terminal Help
demouser@htrc-demo-guest:~$ cd demo/htrc-data/
demouser@htrc-demo-guest:~/demo/htrc-data$ ./htrc-demo.sh
Downloading texts from HTRC Data API...
client_secret=2Zur_V2hQm7YcsKfnc3UG9r0okaagrnt_type=client_credentials&client_id=0gvnHUBw1VRXhYL89r6xrfFshsa
*** JSON: { 'u_token_type': 'bearer', 'u_access_token': 'u429bc8e7f32745db587caidcd35bc4', 'u_expires_in': 84733 }
*** parsed tokens: 429bc8e7f32745db587caidcd35bc4
obtained token: 429bc8e7f32745db587caidcd35bc4

Running topic modeling algorithms...
Processing /media/secure_volume/volumes/hvd.hxj4gg
/home/demouser/anaconda2/lib/python2.7/site-packages/unicode/___init___py:46: RuntimeWarning: Argument <type 'str'> is not an unicode object. Passing an encoded string will likely have unexpected results.
  warn_if_not_unicode(string)
Processing /media/secure_volume/volumes/hvd.hxjfm
Processing /media/secure_volume/volumes/uc2.ark+13960=t5w6bs1h

Corpus file found. Rebuild? [y/N] y
Building corpus from /media/secure_volume/volumes 1289 files, 3 dirs, 0 levels
with coll_corpus function
100%|#####| /media/secure_volume/volumes/..models/volumes-freq5.npz
Saving corpus as /media/secure_volume/volumes/..models/volumes-freq5.npz

Config file /media/secure_volume/volumes.ini exists. Overwrite? [Y/n] y
Writing configuration file /media/secure_volume/volumes.ini

TIP: Only initializing corpus object and config file.
Next prepare the corpus using:
topicexplorer prep /media/secure_volume/volumes.ini
Or skip directly to training LDA models using:
topicexplorer train /media/secure_volume/volumes.ini
Loading corpus...

TIP: This configuration can be automated as:
topicexplorer train /media/secure_volume/volumes.ini --lter 200 --context-type page -k 20

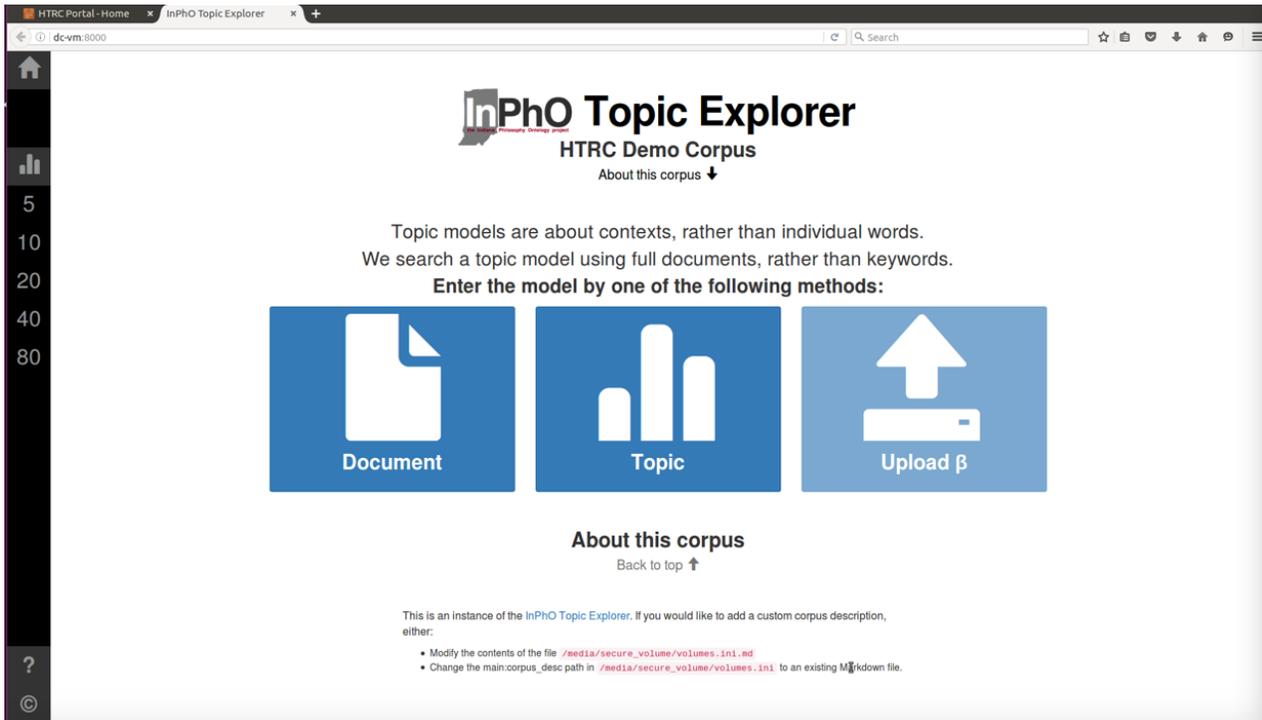
Training model for k=20 Topics with 1 Processes
Begin LDA training for 200 iterations
100%|#####| /media/secure_volume/volumes-freq5-LDA-K20-page-200.npz
Saving LDA model to /media/secure_volume/models/volumes-freq5-LDA-K20-page-200.npz

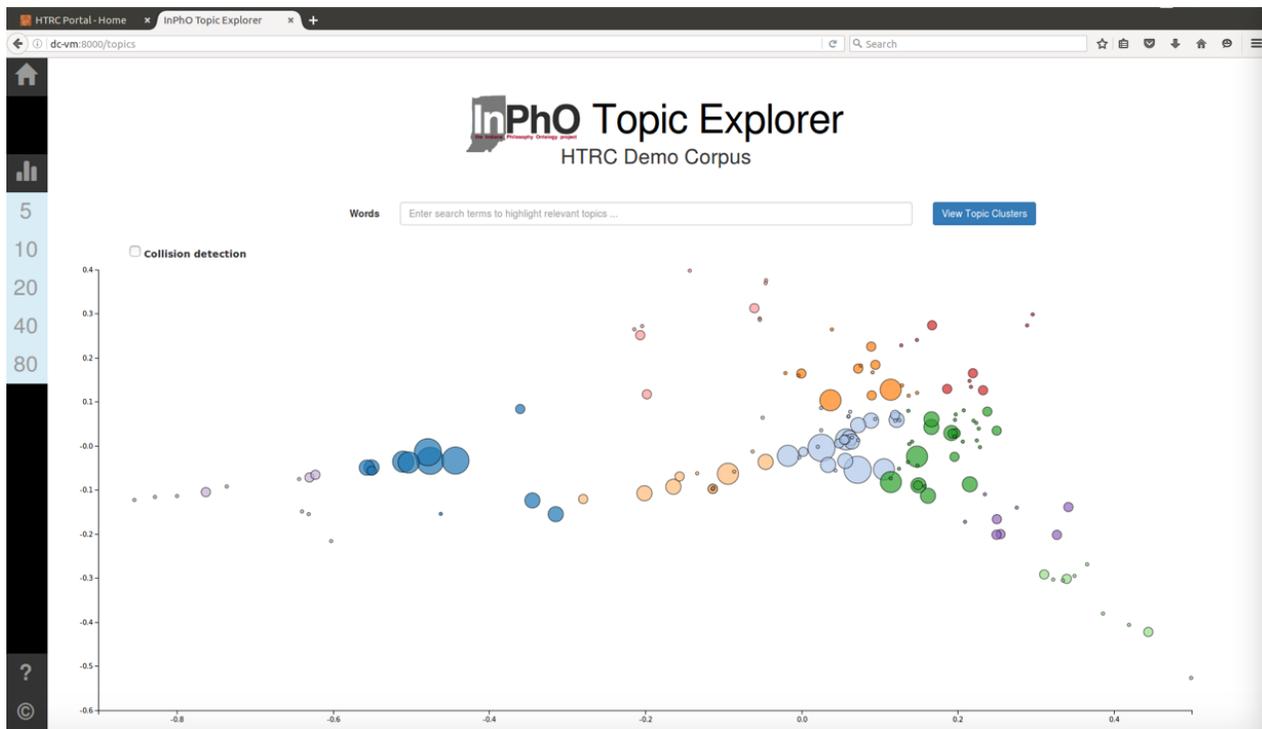
TIP: Launch the topic explorer with:
topicexplorer launch /media/secure_volume/volumes.ini
or the notebook server with:
topicexplorer notebook /media/secure_volume/volumes.ini

Getting metadata...
hvd.hxj4gg
hvd.hxjfm
uc2.ark+13960=t5w6bs1h
Launching topic explorer...
Imported label module
Loading HTRC metadata from /media/secure_volume/models/..metadata.json
Imported label function
using default ld function
Loading LDA data from /media/secure_volume/models/volumes-freq5-LDA-K20-page-200.npz
TIP: Browser launch can be disabled with the '--no-browser' argument:

```

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](#)