Topic Models

Topic modeling has become an integral part of any text-analyst's toolbox, and MeTA will allow you to explore generating topics for any of your corpora in a general way. It supports several different flavors of inference for the LDA topic model: see the API documentation for `lda_model` for a list (doxygen/classmeta_1_1topics_1_1lda__model.html). At the time of writing, the following inference methods are present:

- `lda_cvb` , which uses collapsed variational bayes,
- `lda_gibbs` , which uses collapsed Gibbs sampling, and
- `parallel_lda_gibbs` , which uses an approximation of collapsed Gibbs sampling, but does so exploiting multi-processor systems

To run topic the topic modeling application bundled with MeTA, you need to configure the `lda_model` in your configuration file. Here is an example configuration:

```
[lda]
inference = "cvb" # or gibbs, or pargibbs
max-iters = 1000
alpha = 0.1
beta = 0.1
topics = 10
model-prefix = "lda-model"
```

(For more information on the parameters, please see the documentation for the corresponding inference method you’ve chosen; for instance, collapsed variational bayes (doxygen/classmeta_1_1topics_1_1lda__cvb.html), Gibbs sampling (doxygen/classmeta_1_1topics_1_1lda__gibbs.html), or approximate parallel Gibbs sampling (doxygen/classmeta_1_1topics_1_1parallel__lda__gibbs.html).)

Running `./lda config.toml` will now run your chosen inference method for either the maximum number of specified iterations, or until the model determines it has converged (whichever one is sooner). Because the number of iterations can’t be known ahead of time, progress output is given on a per-iteration basis only, and each different inference model may print out inference-specific summary information after each iteration (for example, for the Gibbs samplers or the maximum for collapsed variational bayes).

Viewing the Topics Found

MeTA bundles an executable `./lda-topics` that can report the top words in each topic found during inference. To use it, you need to specify the configuration file used to do inference, the path to the model’s `.phi` file (which stores for each), and the number of words per topic to output. Below is some sample output for a simple dataset where we found two topics and used the default filter chain for text analysis (which includes `porter2_stemmer`).

```
$ ./lda-topics config.toml lda-model.phi 15
Topic 0:
-----------------------
smoke (3274): 0.391293
smoker (3276): 0.0849028
restaur (3006): 0.0825975
ban (259): 0.0507666
cigarett (584): 0.0261952
non (2445): 0.0232204
health (1726): 0.0188438
seat (3134): 0.017816
smell (3267): 0.0168126
harm (1709): 0.0162674
peopl (2649): 0.0154918
complet (690): 0.0151501
japan (2010): 0.0137623
nonsmok (2448): 0.0128775
tobacco (3638): 0.0120414

Topic 1:
-----------------------
job (2019): 0.156795
part (2607): 0.127387
student (3443): 0.12414
colleg (652): 0.0905726
time (3629): 0.0708495
money (2341): 0.064298
work (4013): 0.0372687
studi (3444): 0.0274538
learn (2110): 0.0253973
import (1857): 0.0167224
earn (1076): 0.015446
experi (1257): 0.0141118
school (3119): 0.00852317
parent (2602): 0.00835476
societi (3296): 0.00793623
```

Documentation for MeTA: ModErn Text Analysis (https://github.com/meta-toolkit/meta)