How to Use

Programming Language and Packages

- Python 3.x
- Packages: pandas, numpy, scipy, sklearn, math, re, random, time

Executing Code

Fork or download Github repo.

Open in IDE and run file(s) or use command prompt (e.g., `python lda_var_inf_without_smoothing.py`). Start with `lda_var_inf_without_smoothing.py`. If unexpected results are encountered, try `lda_var_inf_without_smoothing_v2.py`. Optionally, you can also try the other variations with `lda_gibbs_sampling.py` and `lda_var_inf.py`.

To use a different input dataset, your file will need text and classification columns. Modify the source file (`input_path`) and column settings (`text_column`, `label_column`) in the `load_csv` function call.

```
(vocabulary_size,
     training_term_doc_matrix,
     training_labels,
     testing_term_doc_matrix,
     testing_labels,
     vocabulary) = load_csv(input_path = 'FA-KES-Dataset.csv',
                            test_set_size=100,
                            training_set_size=200,
                            num_stop_words=50,
                            min_word_freq=5,
                            text_column='article_content',
                            label_column='labels',
                            label_dict = {'1': 1, '0': 0})
```

`lda_var_inf_without_smoothing_v2.py` has both datasets (fake news and spam) coded. Comment/uncomment to switch between datasets.

Setting Parameters

Set the following parameters to tune the model:

- `num_topics`: number of topics to model

```
lda.train(num_topics=10, term_doc_matrix=training_term_doc_matrix, iterations=20,
e_iterations=10, e_epsilon=0.1, initial_training_set_size=50, initial_training_iterations=20)
```

See video walk-thru for additional information.