1. **Introduction**
   1. Latent Dirichlet Allocation
   2. Why is it interesting?
2. **Background**
   1. What is it used for?
      1. What problem does it address?
         1. “generative probabilistic model for collections of discrete data such as text corpora” (LDA, 994)
         2. “The goal is to find short descriptions of the members of a collection that enable efficient processing of large collections while preserving the essential statistical relationships that are useful for basic tasks such as classification, novelty detection, summarization, and similarity and relevance judgments.” (LDA, 994)
   2. Description of the algorithm
      1. Generative process assumed for each document in a corpus (LDA, 996)
      2. Intractable posterior, therefore it is necessary to use an approximating method such as variational inference together with EM
   3. Applications of the algorithm
      1. Most well known in order to analyze text data
         * Determining and assigning documents in a corpora to topics
      2. Can also be used in other problems that have a structure similar to the term-document-topic structure
         1. Movie example in paper – data set in which users indicate the movies that they prefer, thus the users can be thought of as a document and the movies are the words, topics would classify the different movie choices
         2. Upon seeing some of the movie choices, be able to predict the held out set, this could potentially be used as part of a recommendation program
      3. Use in our research….lol
         1. There are many situations that LDA can be applied to
         2. Analyzing large amounts of text data, such as customer reviews to determine where people might be unhappy
   4. Other Methods/Algorithms
      1. Tf-idf matrix, latent semantic indexing, probabilistic LSI model
         1. “need to consider mixture models that capture the exchangeability of both words and documents” (994)
      2. Advantages
         1. Performs better in perplexity measurement
         2. Other techniques prone to overfitting
         3. Performs better in perplexity measurement
      3. Disadvantages
         1. “bag-of-words assumption allows words that should be generated by the same topic to be allocated to several different topics” (LDA 1008)
         2. “Exact inference is intractable for LDA” (LDA, 1014)
3. **Implementation**
   1. Testing – “Base cases”
      1. Number of topics is an integer greater than zero
      2. Check that corpus is not empty and that each element of the corpus was a string
      3. Tolerance is tolerance and NeedToSplit argument is 0 or 1
   2. Compare with Python Implementation
      1. Potential issue with the convergence stuff
4. **Speed –Up**
   1. Cython
   2. MCMC
   3. Gibbs Sampling
5. **Examples**
   1. State of the Unions
      1. How do Presidents’ topics change during their years in office?
      2. How are they similar to one another?
   2. Predictions with Movie Data
      1. Measure accuracy of predictions
6. **Conclusion**
   1. Findings
   2. Areas of Improvement

**References**

1. David M. Blei, Andrew Y. Ng, and Michael I. Jordan, *Latent Dirichlet Allocation*, Journal of Machine Learning Research 3, 2003, pg. 993-1022.
2. Max Sklar, *Fast MLE Computation for the Dirichlet Multinomial*, May 2014, <http://arxiv.org/pdf/1405.0099.pdf>. (accessed April 14, 2016)