Abstract

Latent Dirichlet Allocation (LDA) is a Bayes statistical model of text corpora proposed in a paper by David Blei, Andrew Ng, and Michael Jordan in 2003. The model is hierarchical in nature, in that each word in a document is generated from a topic, and the mixture of topics for each document is determined by the document’s distribution of topics. This distribution of mixtures and the probability of generating a particular word given a topic are parameters that help characterize a corpus. The aim of this project is to implement LDA in Python and perform classification. In particular, to test that the algorithm works, given training documents from 5 separate corpora (Enron emails, NIPS papers, KOS blog entries, NYTimes news articles, and PubMed abstracts), the goal will be to infer the corpora-level parameters of each and then identify the corpora of unknown test documents. This will be done by calculating the marginal probability of generating each test document for each of the inferred corpus parameters. The corpus with the highest probability will then be selected as the source. The accuracy of this test will be compared to other text classification algorithms, such as the simple Latent Semantic Analysis approach done in a previous homework and functions from the gensim package.

Preliminary section and subsection headings:

1. Background
2. Algorithm Description
3. Data Set

https://archive.ics.uci.edu/ml/datasets/Bag+of+Words

1. Implementation
2. Test
3. Application and Comparison
4. Conclusion