INTA 6450: Enron Project

Natural Language Processing

Natural Language Processing (NLP) definition

Topic Modeling

Extracting various concepts or topics from a large corpus such as the Enron emails can be performed using topic modeling.

Topic modeling extracts features to generate clusters or groups of terms that are distinguishable from one another. Clusters of words form topics, which can be used to understand the main themes of a corpus.

- Latent semantic indexing (LSI)
- Latent Dirichlet allocation (LDA)
- Non-negative matrix factorization

First 2 methods are popular and have been around a long time. Non-negative matrix factorization is a newer method that is extremely effective and provides good results.

Text Analytics with Python Ch. 5 LDA

Background

LDA is a three-level generative model for collections of discrete data such as text corpora. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. Each email in the database is a document, and the corpus is a collection of documents. A document or email is a sequence of N words. LDA assumes that each email (document) exhibits K topics. It is important to note that the only observed variables are the words of the documents while topic related data are latent variables [1].

A graphical representation of the LDA model is shown in Figure 1 [1]. The model considers observations are the words, organized in the documents. The nth word in the dth document is $w_{d,n}$. Each word is an element in a fixed vocabulary of V terms. A topic β_k is a distribution over the vocabulary. In LDA there are K topics, and each topic k is a word distribution over the V terms. Each doument in the database is associated with a vector of topic proportions θ_d. Topic proportions are a distribution over topics, drawn from a Dirichlet distribution. Each word in each document is assumed to have been drawn from a single topic. The topic assignment for the nth word in the dth document is $z_{d,n}$.

Approach

Load and Parse Data

The Enron email dataset is loaded and parsed to extract the text content of the emails.

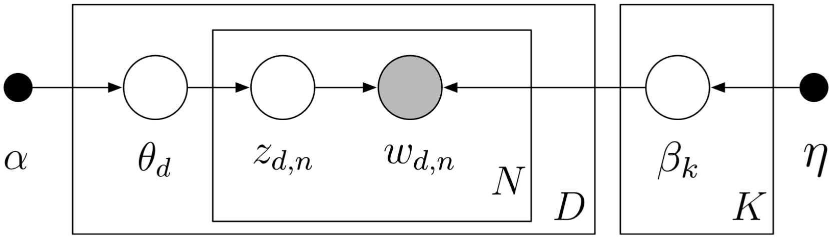

Figure 1: Graphical Model of LDA

Pre-Process Data

The text content of the emails is pre-processed to remove stop words, punctuation, and other non-essential information.

Develop Topic Model

To develop an LDA model with Python, the scikit-learn library can be utilized [2]. The scikit-learn package includes unsupervised learning with matrix decomposition algorithms. Under this umbrella is the `LatentDirichletAllocation` class, which is used to fit the LDA model to the Enron email dataset. LDA modeling can be performed by using the application programming interface (API) provided by scikit-learn [3] to access the `sklearn.decomposition` module that includes the `LatentDirichletAllocation` class.

`LatentDirichletAllocation` implements the online variational Bayes algorithm and is used with the batch update method. The batch method updates variational variables after each full pass through the data and is implemented by `learning_method='batch'` [4].

[1] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, “Stochastic variational inference,” *Journal of Machine Learning Research*, 2013.

[2] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[3] L. Buitinck *et al.*, “API design for machine learning software: Experiences from the scikit-learn project,” in *ECML PKDD workshop: Languages for data mining and machine learning*, 2013, pp. 108–122.

[4] F. Pedregosa *et al.*, *LatentDirichletAllocation*. scikit-learn developers, 2024.