Assignment on LDA

CPSC 503 - Pedagogical Project

April 2019

This is a self-assessed assignment intended to assess students' understanding of the Latent Dirichlet Allocation (LDA) topic modelling algorithm. Possible answers are provided here.

1 Warm-up (30 points)

Determine an application of LDA (does not have to be in academic research). Describe for what purpose and how LDA has been used in this applciation. What do you like and/or dislike about this application and its use of LDA?

2 Comprehension (30 points)

a. What is a generative model, i.e. what makes a topic model generative? What are the core assumptions?

b. How is the LDA different from the pLSI model?

c. Is perplexity a satisfactory measure for evaluating the results of the LDA model? What are/can be some alternative measures?

3 Application (40 points)

In this question, we will use the Usenet dataset consisting of 20,000 messages sent to 20 bulletin boards in 1993. The Usenet bulletin boards include newsgroups for topics like politics, religion, cars, sports, and cryptography, and offer a rich set of text written by many users. This dataset is quite popular for exercises in text analysis and machine learning, and it is publicly available.

For this question, you can use a programming language of your choice. Make sure to submit your code with sufficient documentation and comments.

The Data

- Go to http://qwone.com/~jason/20Newsgroups/
- Download the `20news-bydate.tar.gz` file

Analysis

Within the `20news-bydate-train` folder, you will see subfolders, each representing a different bulletin board. Using the messages within the bulleting boards that are related to science (folder names begin with `sci.`), apply an LDA model.

Report on the words with the highest probability in each topic visually.