Inference & Representation : Lab 5

As a group of ML consultants, you have decided to take on *one* of two different projects. Both involve modifying Latent Dirichlet Allocation to suit your client.

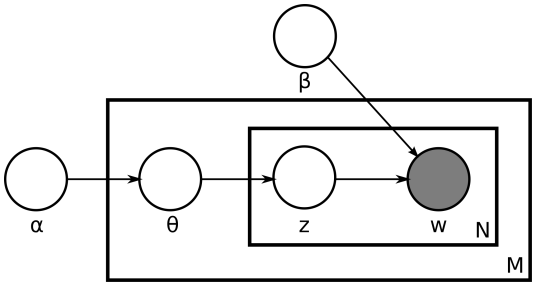
Instructions:

(1) Write down the names of people in your group at the top left.

(2) Draw your graphical model in plate notation or instantiated for specific values of N and M (c.f. Fig. 1)

(3) Write down how every random variable is distributed and how every conditional probability is parameterized.

Fig. 1: Plate model for Latent Dirichlet Allocation



1) A news organization has collected the entire transcript of the vice-presidential debate. There are thousands of documents; each one contains a single statement or response (eg. a one minute segment) by Mike Pence (Republican VP nominee), Tim Kaine (Democrat VP nominee) or the moderator. The documents do not contain the identity of the speaker but the organization is interested in knowing what broad ideas each of the speakers covered. Help them out! After you are finished you must answer the following questions:

a) Build a variant of LDA that models words in the documents
b) How would you tell the news organizations about the topics covered during the debate by each speaker? i.e Write down the probabilistic query. Why does this query represent the information you seek?
c) Given a new document, could you find out who spoke the words in it? Write down a query for it using the random variables in your model.

2) In its annual review, the Department of Justice has decided to hire your group and given you access to all court proceedings (anonymized of course) related to theft in the US over the last 20 years. Each document (court case) comprises three sections: arguments posed by the defense, arguments posed by the prosecution and the final verdict (who won the case). The DOJ wants to know about the topics used in arguments by the defence, topics used only by the prosecution and topics that both of them use.

a) Build a variant of LDA that models the arguments by either side as well as the outcome of the case.
b) Given a court proceedings with only the arguments by the defense and prosecution, write down the probabilistic query that would give you the verdict for the case.