Latent Dirichlet Allocation for Topic Modelling

Introduction

Topic models group words together in a document into different topics. For example in the sentence "This boy loves to play in the park and likes ice cream" can be said to consists of two important topics, 1) Food, because of "ice cream", and 2) Activities, because of "play" and "park". Latent Dirichlet Allocation (LDA) is a unsupervised learning algorithm used to discover different topics and their associated indicators (words relating to topic) in a collection of documents. LDA is based on the idea that words often have strong semantic relationships to certain topics, and so topics in a given document will consist of a group of similar words.

Like in K-means clustering unsupervised algorithm, LDA requires us to pick the number of topics for it to discover, and outputs the words in the text corpus (a set of documents) that frequently occur together within the topic. LDA assumes that a document is a mixture of a set of latent (unknown) topics, and each topic is another mixture of words (collection of words that help identify the topic).

Document generating process

LDA is known to be a generative model, in the context of text analysis this assumes the documents are generated through some statistical process. Given a document *d* is a text corpus *D* (a set of documents), then *d* is generated by,

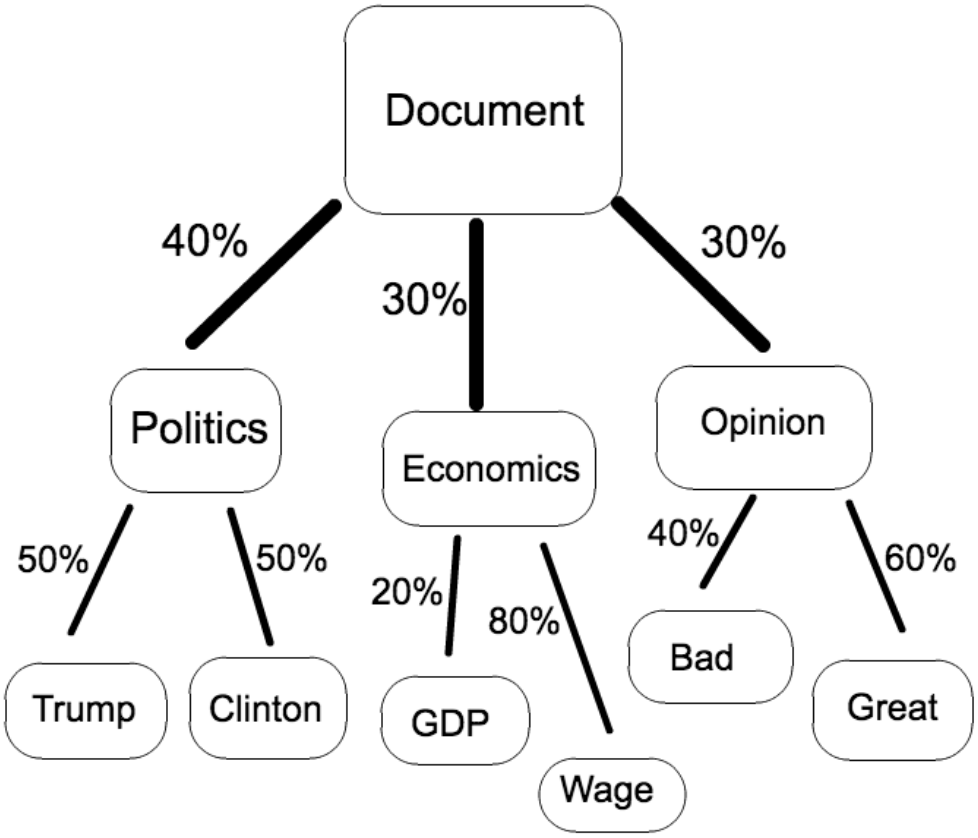
1) Number of words in document *d*, represented by N_d, is drawn from poisson distribution. That is $N_d \sim Poisson(\eta)$

2) The mixture of topics in document *d*, represented by $Theta_d$ is drawn from dirichlet distribution. That is $Theta(\alpha) \sim Dirichlet(\alpha)$ this is the topic-document distribution.

3) Assign each word w_i, i = 1… N_d a topic, z_i in a way so that it is consistent with the document-topic distribution in 2). That is $z_i \sim$ Multinomial ($Theta_d$)

4) Now that we know the topic z_i of each word w_i we can draw word w_i from the topic-word distribution $\emptyset(\beta)$ That is we choose w_i with probability $P_r(w_i|z_i, \beta)$

The figure below illustrates this idea of a document being a mixture of topics (politics, economics, and opinions of people) and each of those topics are represented by some set of words.



To summarize, LDA assumes a document is a mixture of topics, where the topics are drawn from the topic-document distribution, and topics consist of words, where the words are drawn from the topic-word distribution. In practice we already have a text corpus, a set of documents. So we are usually not interested in generating new documents, but rather doing inference on how the document is generated from varying topics and words.

Intuition behind topic analysis

Now that we have assumed a data generating process for each document in our text corpus, we can work backwards by doing inference on the unknown parameters in the document generating process. We are particularly interested in finding the distribution of topics for each document, and the distribution of words for each topic.

Ignoring some formalities with priors and posteriors that we will discuss later in this notebook, the pseudocode for the LDA algorithm looks like,

1) Iterate through each document d

2) Iterate through each word w in document

3) For given document d and word w, find the probability topic t generated w. Using Bayes rule, we can do this by

$$Pr(\text{word} = \text{w} \mid \text{topic} = \text{t}, \text{document} = \text{d}) = Pr(\text{topic} = \text{t} \mid \text{document} = \text{d}) Pr(\text{word} = \text{w} \mid \text{topic} = \text{t}),$$

where

$$Pr(\text{topic} = \text{t} \mid \text{document} = \text{d}) = \frac{\text{Number of words in topic} = \text{t and document} = \text{d}}{\text{Number of words in document} = \text{d}},$$

and

$$Pr(\text{word} = \text{w} \mid \text{topic} = \text{t}) = \frac{\text{Number of words in topic} = \text{t and word} = \text{w}}{\text{Number of words in topic} = \text{t}}.$$

4) Reassign the word w to topic = t with probability computed in step 3), $Pr(\text{word} = \text{w} \mid \text{topic} = \text{t}, \text{document} = \text{d})$.

Steps 1-4 above would be on iteration of the LDA algorithm, we need to run steps 1-4 several times, then we will have discovered the distribution of topics that make up any given document, and also the distribution of words for all topics.