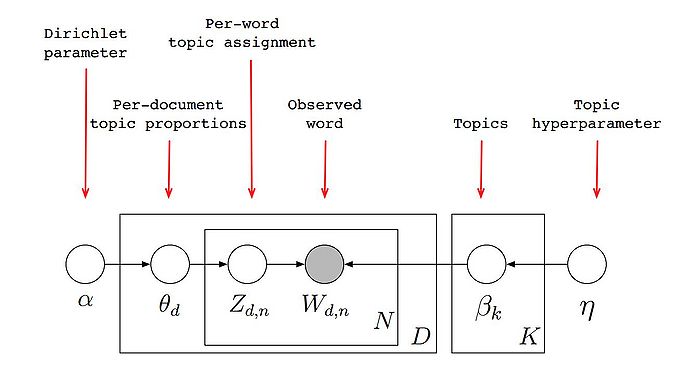
**Topic modelling using LDA.**

Topic Modeling is an unsupervised machine learning based technique to represent a document using a few topics which are composed of words from the document. These topics are formed such that they give the combination of words which best explain the document. This is a type of clustering but a little different in from the normal clustering because here words are grouped into topics such that the group represents a topic instead of numerical features in normal clustering. The topic modeling technique used in this paper is Latent Dirichlet Allocation which is a probabilistic approach. In the project, it is directly compared to Clustering i.e., a mathematical approach.

In topic modeling the goal is to find shared topics that show up in the documents. So, topic modelling is used to model a whole set of documents. Then it is figured out which topics show up in which documents, and it reveals which documents are similar as similar topics show up. The term document can be many different things, so for this project different users represent document and different products represent words. So, as topics containing different products are found, these topics are also used to cluster similar users together as they are affiliated to similar products. This is similar to clustering, but here it is a probabilistic approach.

The form of topic modelling where Dirichlet priors are used is known as LDA (Latent Dirichlet allocation). These priors are input to the model. The below model is then used to obtain optimized values for topic proportions and topics.



D denotes the number of documents,

N is number of words in a given document (document d has Nd words),

α is the parameter of the Dirichlet prior on the per-document topic distributions,

ƞ is the parameter of the Dirichlet prior on the per-topic word distribution,

θd is the topic distribution for document i,

βk is the word distribution for topic k,

zdn is the topic for the j -th word in document i,

wdn is the specific word.

So here the documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over all the words. So, to implement LDA we follow the below generative process.

1. Choose both the priors,

Θd ~ Dirichlet(α) where d is in {1,…,M}

Βk ~ Dirichlet(β) where k is in {1,….K}

1. For each of the word positions d, n where d is in {1,…,M}, and n is in {1,…,Ni}

Choose a topic zdn Multinomial (Θd)

Choose a word wdn Multinomial (Βkn)

The above generative process implemented using techniques such as MAP estimation using EM algorithm and Variational Expectation Maximization. Both techniques make use of expectation maximization step to obtain optimized values for topic proportions and topics.