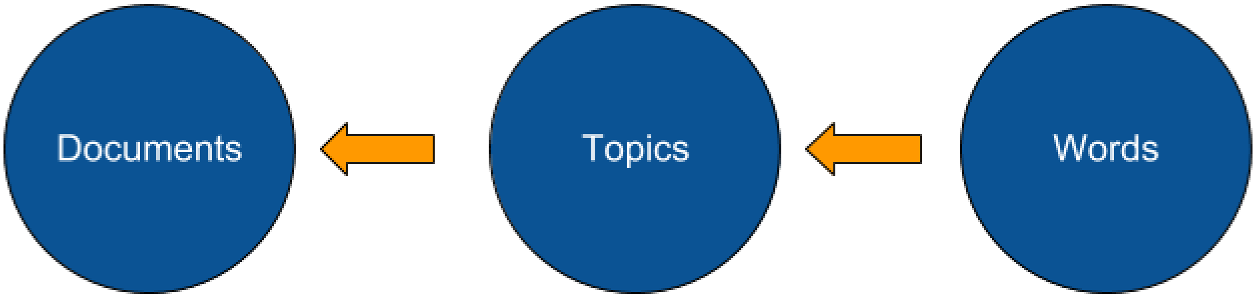
Topic Modelling:

I have used the NLTK and Gensim libraries in Python for topic modelling.
The technique used for modelling is the Latent Dirichlet Allocation (LDA). It is an unsupervised learning method that identifies labels representing the document.

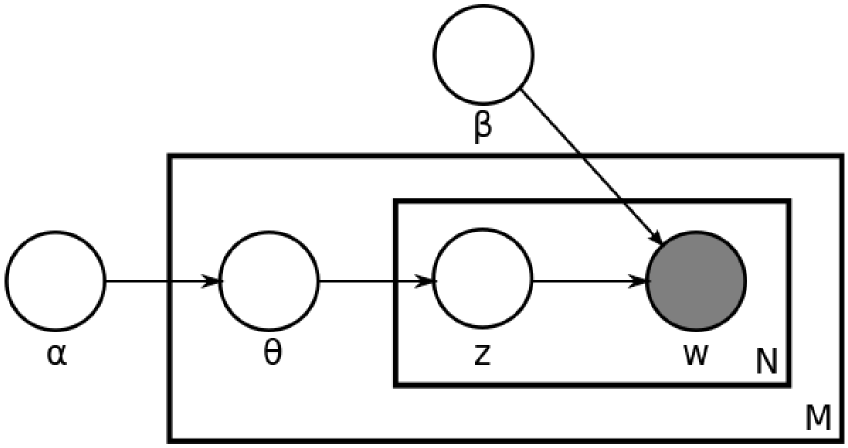
This technique models each document as a set of topics and each topic is a collection of a group of words.



Data format:
The code takes a .json file as an input which contains 1838 input documents. I have converted each document to a .csv file for ease of use.

Text cleaning and topic extraction:
Each document in the .csv format is parsed to tokenize terms, group different forms of the same words into one term (lemmatize) and stop words are eliminated. The Python Spacy library makes this very easy with in-built methods for tokens and lemma conversions.
The cleaned document is converted to a list of topics. This list is stored as a vector in a Bag of Words (BoW) representation.

LDA Model:
The LDA model is trained on the corpus of the document and the number of topics for each word are customized for the model. The number of words are set for each document.



The boxes are "plates" representing replicates, which are repeated entities. The outer plate represents documents, while the inner plate represents the repeated word positions in a given document, each of which position is associated with a choice of topic and word. M denotes the number of documents, N the number of words in a document.

Output:
The output is the document expressed as a set of 5 topics. Each topic is expressed as a probabilistic representation of the group of words.

Further Improvements:
The code can be made much more efficient and visually appealing by using some additional functionalities of the Python packages.
The corpus and the dictionary can be updated after each document to minimize the processing for each new document.
The PyLDAvis library can be used to visualize the corpus as a connection between topics and words.

References:

https://radimrehurek.com/gensim/models/ldamodel.html
https://radimrehurek.com/gensim/corpora/textcorpus.html
https://towardsdatascience.com/topic-modelling-in-python-with-nltk-and-gensim-4ef03213cd21