# Week 1 Report 24/25 S2

## 1. Code

A diagram of a diagram

Description automatically generated

Implemented hLDA model in terms of a tree. The implementation is organized into two primary classes within the hlda\_final.py file. The implementations follows this paper:  
<https://www.cs.columbia.edu/~blei/papers/BleiGriffithsJordanTenenbaum2003.pdf>

### HLDA\_Node

* Represents the tree node (i.e. a certain topic).
* Contains information such as parent, children, number of documents passing through the current node, etc.

### HLDA\_Tree (HierarchicalLDA)

* Represents the hierarchical structure.
* Includes hyperparameters
  1. Gamma: related to CRP process. Larger gamma makes the tree more branched out.
  2. Alpha: related to how many topics a single document is represented
  3. Eta: smoothing parameter. Smaller eta encourages fewer words in each topic. Hence, more topics are required to explain the documents.
  4. Level: Maximum number of levels
* Implements the Gibbs sampling, there are 2 steps to the Gibbs sampling:
  1. Level sampling
  2. Path sampling

The following code are utility functions which is contained in the file hlda\_utils.py

### Preprocessing steps

1. Lowercasing
2. Tokenizing
3. Removing non-alphabetic tokens and stop words
4. Applying stemming and lemmatization
5. Filtering out very short tokens

### Synthetic Corpus generation

1. Fit a hLDA model using some corpus. The result is taken to be the ground truth.
2. Sample a document and retrieve its path.
3. Partition length of the document among the levels using a Dirichlet(alpha) where alpha is a hyperparameter.
4. Sample words alone the topics alone the path.

## 2. Test

### 2.1 Data Description

The 20 News Groups dataset was selected for testing the HLDA model due to the availability of a ground truth hierarchical topic structure. This dataset comprises approximately 20,000 newsgroup posts across 20 distinct categories which can be arranged into a hierarchical structure.

A screenshot of a computer

Description automatically generated

Only the first 2000 documents are used for testing. After the pre-processing step, the average number of tokens per document is 90.

It takes around 4 hours to run the model on the 2000 documents for 2000 iterations.

With the following parameter,

gamma = 0.05, eta = 0.01, alpha = 10, levels= 3

**Initial Results**:

* Topic Hierarchy: The resulting tree contained 16 topics directly beneath the root node.
* Comparison with Ground Truth: This number significantly exceeded the number of subtopics defined in the ground truth hierarchy.
* Level 3 Topics: Similarly, the number of topics at the third level surpassed those in the ground truth, indicating over-segmentation.

### 2.2 Hyperparameter testing

The above result may be due to the wrong choice of hyperparameters. Hence, I have tried different combinations of hyperparameters such as:

1. gamma = 0.01, eta = 0.01, alpha = 10, levels=3

2. gamma = 0.05, eta = 0.01, alpha = 5, levels=3

3. gamma = 1, eta = 0.01, alpha = 5, levels= 5

4. gamma = 0.001, eta = 0.01, alpha = 10, levels= 3

5. gamma = 0.05, eta = 0.01, alpha = 20, levels= 3

6. gamma = 10, eta = 0.01, alpha = 5, levels= 3

**Findings:**

* Gamma Sensitivity: The model exhibited low sensitivity to changes in gamma. Significant alterations in gamma values were required to observe notable effects on the topic hierarchy, suggesting that gamma may not be a critical parameter within the tested range.
* Eta Sensitivity: Due to time constraints, the impact of eta was not thoroughly examined during this period.
* Alpha Adjustments: Variations in alpha influenced the number of topics a single document represented, with higher alpha values promoting greater topic diversity.

### 2.3 Synthetic corpus

Blei conducted testing of the model on synthetic corpus. His approach first fits a real corpus into a HLDA model. Then, he will randomly pick out a path along the tree. He will also take a Dirichlet prior with hyperparameter alpha to partition the number of words along the path. After that, it is randomly sampling the words from the topics.

**Outcomes:**

* The synthetic corpus was successfully generated
* Label Switching Issue: During the comparison of recovered hierarchies with the known ground truth, a label switching problem was encountered. This issue arises when the model assigns different labels to the same underlying topics across different runs or within the hierarchy, complicating direct comparisons.

Attempts to automate the process of comparing hierarchical structures and assessing topic similarity were unsuccessful due to the complexities. Hence, To ensure accuracy, the decision was made to manually verify the recovered topic hierarchies, despite the time-consuming nature of this method.