# CS771A - Assignment 2

## 1 Question 1

Give detailed calculations explaining the various design decisions you took to develop your decision tree algorithm. This includes the criterion to choose the splitting criterion at each internal node (which essentially decides the query word that Melbo asks when that node is reached), criterion to decide when to stop expanding the decision tree and make the node a leaf, any pruning strategies and hyperparameters etc.

Answer:
The splitting criterion at each internal node is decided based on the lowest entropy attainable by splitting the remaining set of words. We start with an infinite entropy and calculate each possible lower entropy, using the proportion of the words in each group to find the lowest possible entropy split.

The splitting takes place by comparing the differences in each bit(0 or 1) after each pair of words is converted to their binary form, which gets stored as the mask. The decision tree splitting is then chosen based on the lowest entropy.

Entropy is chosen as the frequency of each possible outcome in the binary decision obtained through calculating the mask. The entropy is calculated using the formula

$$E = \Sigma(p_i log(n_i))$$

where $p_i$ represents the proportion of the words in the split class to the total number of words and $n_i$ represents the number of words in the split class.

A loop is used to generate the mask for each pair of $(a,idx)$ to generate a mask for every word depending on how close it is to a. The split obtained is then used to calculate the entropy and check if the generated entropy is lower than the minimum.

The pruning strategy used is to limit the maximum depth of the decision tree to 12. Since the dictionary used will be 5000 to 6000 words, the maximum depth of 12 is a reasonable pruning strategy to optimize the training time of the model.

## 2 Question 2

The python program 'submit.py' has the my_fit() function which trains the model according to the specifications.

Preprint. Under review.