APPROACH TO OPTIMIZE HANGMAN GAME FOR ACCURACY BETTER THAN 50 %

- **Inputs given-**
  - Word length= len
  - Predict letters in less than 06 wrong attempts
  - Train Data -250000 words
  - Understood will be tested on entirely disjoint data

- **Preparation for Strategy**:
  - Initially I tried optimized the heuristics approach by selecting the letter with highest frequency, but after few predictions the new dictionary was getting empty.

  - Since the access of disjoint test data is not available, so after certain practice runs I found out that most of the more than 50% words in the test data set were either derived from the training data set or were the combinations of the subsets of the words in training dictionary.

  - Hence I explored the machine learning approached and found N-gram, RNN (LTSM and GRU) and GPT2 and other Pre-trained model can be used. As all these ML models were not working straight in this model hence it was planned to use N-gram and define similar functions for enhancing accuracy.

- **Strategy Followed**:

  - I have used count and probability based strategy.
  - Initially heuristics function was used for prediction of letters from coming more than once in training dictionary. The letters are predicted with highest number of occurrence but not predicted before.
  - Possibly of Vowels in each word is worked out on the basis of average and prediction of number of vowels kept limited to that.
  - Further in case no prediction
  - Futher, N-gram function using NLTK has been deployed. The training dataset is prepared for N-gram as substrings from full set of words in Train data.
  - Come back to heuristics approach max occurance from the full dictionary for prediction in case of no prediction of letter not possible from the above substrings of train dictionary words.

Further Scope of Improvement:

  - There is a plan to further enhance accuracy by deploying/integrate the function RNN (LSTM and GRU)/GPT2 or the similar function to enhance further accuracy
  - In the heuristic approach the letter with second highest occurrence may also be chosen once the dictionary goes empty.
  - These approached will be decided to be deployed after successful result.