Submission by Hemant Yadav (hemantya@iiitd.ac.in)

- Trexquant Hangman Challenge Registration - Quantitative Researcher - Early Career - India - HEMANT YADAV.
- Accuracy achived 53%.

Description of the strategy.

The strategy is completely Data driven without any Human induced bias.

- Two models are trained using **ONLY** the provided words.
  - N gram.
  - CNN based character BERT.

While True:

- Step 1: If not a single character is guessed successfully return unigram scores.

- Step 2: Once one character is successfully guessed.

  - Using character BERT get the topk possible words and their likelihood scores.
  - Calcualte the prior scores using the N gram model on these possible words.
  - Using the likelihood and prior to calculate the posterior score on the possible words.
  - Return the unigram scores calculated on (1) the possible words and (2) scale them using the psoterior scores.

- Step 3: Use the unigram scores to guess the most probable character that is not guessed until now.

NOTE

- Scores are log probabilites for numerical stability.
- Unigram score means log probability distribution over the possible characters.

Hyperparameter choice

- I used a 3 gram model because the test set was disjoint.

Future work

- Introducing human bias to be used as a heuristic in either step 2 or 3 or both. This can help in reducing the search space of possible words (Step 2) or characters (Step 3).

- Introducing data bias. This means expanding the number of **WORDS** in addition to the provided words. This is a little bit tricky because again, we have to introduce human knowledge on doing data extension.
- The method as of now is very simple. The accuracy can be improved maybe with a detailed hyperparamter search.

Final thoughts.

- I am not in favor of adding human knowledge/bias when designing ML based systems. I am in favor of increasing data.