TODOs:

1. Data Preprocessing:
   1. Cleaning (removing tags – non Arabic letters or diacritics – HTML tags).
   2. Tokenization.
   3. Remove sentence in correct.
   4. Correct diacritics error.
2. Word Features
   1. Bag of words.
   2. TF-IDF.
   3. CBOW (word embeddings).
   4. BERT.
3. Char Features
   1. CBOW (word embeddings).
   2. TF-IDF.
   3. BERT.
4. Classifier
   1. RNN.
   2. LSTM.
5. Rule based Corrections.
   1. It is a space, 0 or one of the following: .
   2. It is the first letter of the Arabic word.
   3. It has Sukoon as a predicted primary diacritic.
   4. If it is, set the current diacritic to Kasra, by setting the probability of its class to 1 and the others to 0.
   5. If it is or, set the diacritic of the previous character to Fatha.
   6. If it is and the last letter of the word, allow only Fatha, Fathatan, or no diacritic choices by zeroing the probabilities of the other classes.
   7. If it is and not the last letter of the word, set Fatha on the previous character.
   8. If it is the first letter in the word, forbid Sukoon.
   9. If it is not the last character of the word, prohibit any Tanween diacritic from appearing on it.
   10. If it is the last letter, prohibit Fathatan unless this character is or .
   11. If it is a space, 0 or any of the following characters: , set the choice to no-diacritic.

Notes to search:

* Relation between Word encoder and character encoder.