Code Documentation

(All are self explanatory from code , code is well commented with doc strings)

XGboost.py:

Variables:

max_num_features : tokens are padded/trancated to this length (default : 10)
pad_size : no of next/previous tokens in feature vector (def : 1)
boundary_letter : seperator(def : -1)
space_letter : padding character (def 0)
max_data_size : Trainingdata size

Methods:

context_window_transform :

Converts token sequence to vector of feature vectors

Feature vector of each preprocessed token is current token's encoding appended with encodings of previous and next tokens

No of prev/next tokens to be included in feature vector for context in given by pad_size

train_model:

Trains a model on date of size max_data_size and saves the model
No of training rounds is 10 by default can be changed in line 117

loadModel:

Loads previously trained Model xgb_model_3200000 by default
Can be changed by changing next lines

Normalise:

Normalises the sequence of tokes provided as string with '###' as delimiter and prints normalised sentence