bigram feature:

1. Use ground-truth to build a bigram lexicon which has bigram and its corresponding frequency
2. tokenize each word in tesseract into letter bigram
3. number of bigrams is (length-1)
4. find the frequency for bigrams each word in tesseract
5. calculate bigr

Most frequent symbol:

1. tokenize each word into letter unigram
2. find the most frequent one and the count

Non-alphabetical symbols.

1. tokenize each word into letter unigram
2. count number of alphabetical symbols l1
3. l2=length-l1
4. feature is l2/l1

Levenshtein distance

Refer *edit\_distance* function in NLTK