MAKEFILE

ASSIGNMENT-1

CS689A- Computational Linguistics for indian Languages

Question (1)

a) Unicode :rule 1)halant is add to the corresponding consonants
2)after consonants add vowel अ

Question (2)

a) syllables :rule 1) Breaking at consonants consonants
2)Breaking at vowels as i defined in notebook
3) consider halant as vowel
b) Bigram_frequencies :function use: find_ngrams
c) Used libraries-collections

Question (3)

a) BPE: used libraries collections,re
b) Remaining same as question 2

Question (4)

Precision is around100% for 1k BPE tokens

And recall is around 0% for 1k BPE tokens

a) Precision = TruePositives / (TruePositives + FalsePositives)
b) Recall = TruePositives / (TruePositives + FalseNegatives)
c) F_Measure = (2 * Precision * Recall) / (Precision + Recall)

Question (5)

a) Used libraries-pyconll for extractions of lemma

Question (6)

a) I have made the graph between frequency vs rank for zipfian distribution
b) Token follow zipfian
c) Bpe tokens not follow zipfian
d) Syllables follow zipfian
e) Characters follow zipfian

f) Lemma follow zipfian
g) Libraries used: matplotlib

Question (7)

a) First i match original word with lemma after that characters that are left in original word append to any of list that i call it suffix