MAKEFILE ASSIGNMENT-1

Computational Linguistics for indian Languages

CS689A

22111021 (Drashtant Singh Rathod)

Question (1)

a) Corrected Unicode rule

1) After each consonant add अ but not for consonant with halant

2) Halant is added to consonant by adding chr(2381)

Question (2)

a) Finding syllables

:rule

1) Breaking at consonants

2)Breaking at vowels as i defined in code

3) considering halant as vowel

Syllable are extracted by vowel sound ending.

b) Bigram_frequencies :

library use: 2-gram finder

c) Used libraries-collections

Question(3)

a) BPE: used libraries collections,re

b) Remaining same as question 2

Question (4)

Precision is around 98% for 1k BPE tokens And recall is around 4% for 1k BPE tokens

Precision = TruePositives / (TruePositives + FalsePositives)

Recall = TruePositives / (TruePositives + FalseNegatives)

F_Measure = (2 * Precision * Recall) / (Precision + Recall)

Question (5)

a) Used libraries-pyconll for extractions of lemma and treeparser is used to extract tokentree

Question (6)

a) I have made the graph between frequency vs rank for zipfian distribution

Libraries used: matplotlib

Token follows zipfian

Bpe tokens not follows zipfian

Syllables follow zipfian

Characters follows zipfian

Lemma follow zipfian

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